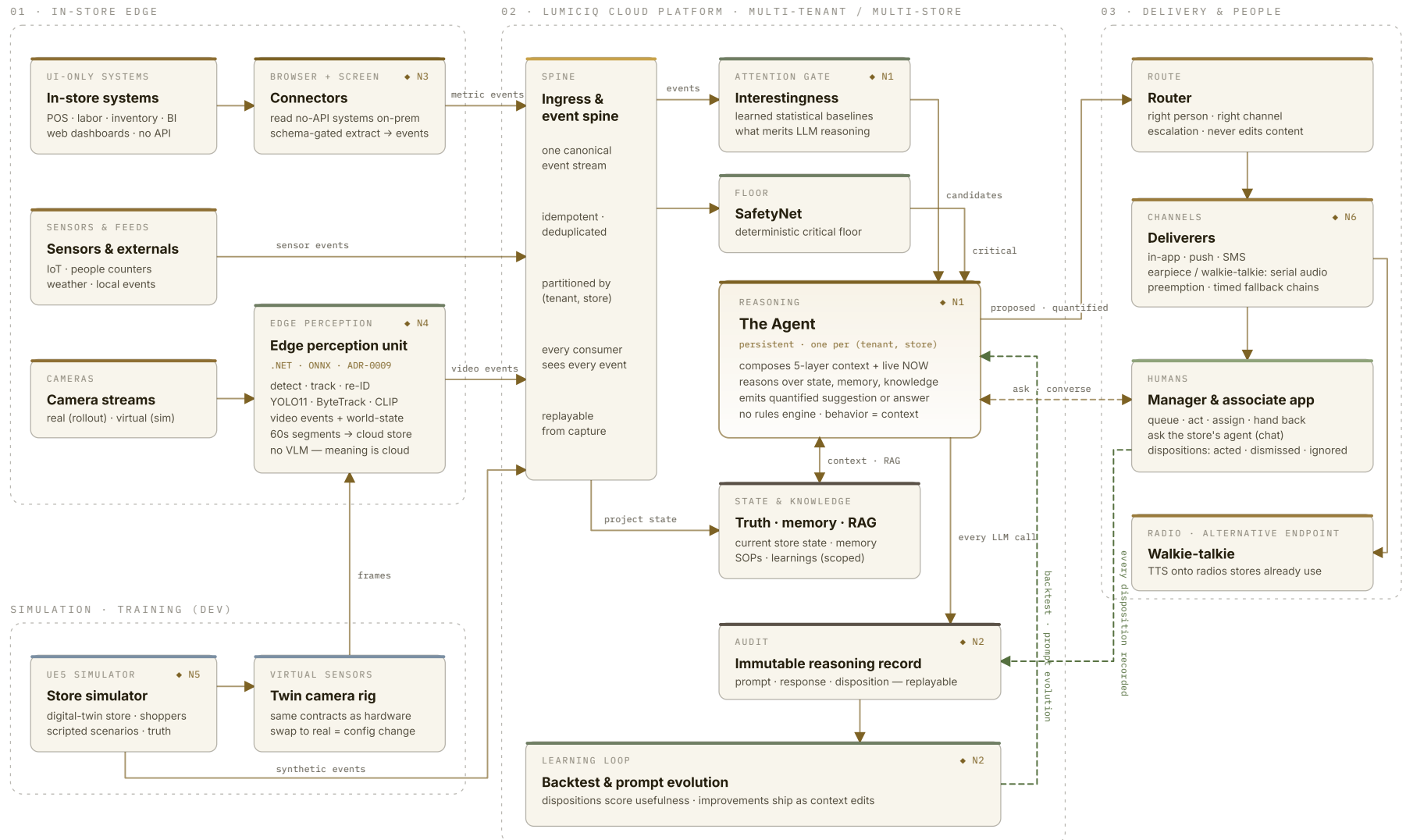


LumicIQ — what we're building, end to end

■ STORE SIGNAL SOURCE ■ AGENT / LLM ■ SPINE ■ DETECT / LEARN ■ DELIVER ■ STORAGE ■ SIMULATION ■ DATA / CONTROL FLOW
--- LEARNING FEEDBACK --- CONVERSATION ◆ N# CANDIDATE NOVELTY AREA



§1 The components, in brief

Each row expands to a short plain-language description. Follow the diagram left to right: signals are captured in the store, reasoned about in the cloud, and delivered to people.

01 · IN-STORE EDGE — TURNING WHATEVER THE STORE ALREADY HAS INTO SIGNALS

CONNECTORS **Connectors — dashboard & screen** ♦ N3

On-premise connectors, running on a store computer, that turn systems the platform doesn't own — point-of-sale, labor, inventory, BI — into sensor streams **without vendor integration work**. They attach to the operator's already-authenticated browser session (no stored credentials), navigate deterministically, and use an LLM for exactly one step: extracting a declared set of metrics into a typed schema. Readings that fail schema validation are never emitted.

Two capture modes ship today: a **dashboard connector** that walks the URL-addressable report views of a web back-office (site adapters make each new vendor a small config; Loyverse is the first), and a **screen connector** — a human-in-the-loop companion where the operator shares a tab and the connector reads the numbers on it. Both emit structured, idempotent events onto the same spine as every physical sensor; all *reasoning* stays with the platform Agent.

EDGE PERCEPTION **Edge perception unit** ♦ N4

A per-store perception unit (ADR-0009) that runs the vision **at the edge** — a store box, a customer datacenter, or a cloud-hosted instance — and sends only derived data across the trust boundary. Deterministic code does the perception: object detection (ONNX / YOLO11), multi-object tracking (ByteTrack), and appearance re-identification (CLIP galleries) that carries a person's identity across track loss and camera hand-off. From the tracked world it emits **typed video events** (a shopper lingered, an aisle is crowded, someone is waiting unassisted) plus a periodic **world-state aggregate** — occupancy and dwell per zone — onto the spine.

Crucially, **no VLM runs at the edge**. Expensive visual "meaning" is a **cloud cascade**: the edge ships trigger evidence and rolling 60-second video segments to a cloud world-store, and the platform re-inspects the actual pixels only when a question or a shelf-level judgment makes it worth the spend. Perception at the edge, meaning in the cloud — the **Overwatch deck** is the full-pipeline reference implementation this port is derived from.

SENSORS & FEEDS **Existing sensors & external context**

Whatever the store already operates — IoT sensors, people counters, inventory feeds — plus external context such as weather and local events, pulled by per-source adapters with deduplication, rate limiting, and health monitoring. The platform is deliberately **sensor-agnostic**: it standardizes signals, it does not prescribe hardware.

SIMULATION · TRAINING — A SYNTHETIC STORE WITH GROUND TRUTH

UE5 SIMULATOR **Store simulator + twin camera rig** ♦ N5

A digital-twin store built in Unreal Engine 5 — inventory, shoppers, and scripted scenarios with known ground truth. Its virtual camera rig and event generators speak the **exact same contracts as production sensors**, so the entire loop — video → events → attention gate → agent → human — can be trained, exercised, and regression-tested synthetically. Moving to physical cameras is a configuration swap, not a rebuild.

02 · CLOUD PLATFORM — FROM EVENT STREAM TO REASONED SUGGESTION

SPINE **Ingress & event spine**

All signals become one canonical event stream — idempotent, deduplicated, partitioned by (tenant, store), with every consumer seeing every event. Everything downstream is **replayable from this stream**, which is what makes the audit and learning loop possible.

ATTENTION GATE **Interestingness detector + SafetyNet** ♦ N1

A universal, **learned statistical attention gate** — per store / event type / time-of-week baselines, with no per-customer rules — that decides which events merit expensive LLM reasoning. This is the cost-control mechanism that makes continuous LLM supervision of a high-volume physical event stream economically viable.

Beneath it sits the **SafetyNet**: a small deterministic floor of critical conditions the platform must never miss, independent of both the learned gate and the LLM.

REASONING **The Agent** ♦ N1

A persistent LLM reasoner per (tenant, store). For each gated event it composes a five-layer context — platform → vertical → tenant → store → situation — plus the live "now", reasons over current state, its own memory, and scoped knowledge, and emits either a **quantified suggestion** or a conversational answer.

There is **no rules engine anywhere**: customer configuration describes the business, it never encodes rules. Behavior lives in composed context, so it can evolve without code changes.

AUDIT + LEARNING **Immutable reasoning record → learning loop** ♦ N2

Every LLM call — the exact composed prompt, layer versions, retrieved context, and response — is stored as an immutable, reconstructable record. Every suggestion's disposition (acted on, assigned, dismissed, ignored, expired) joins it, scoring **how useful each alert actually was**, so candidate prompt changes are **backtested against recorded history** and shipped through experiments. The product improves weekly as measured context edits, not deployments.

03 · DELIVERY & PEOPLE — THE RIGHT EAR AT THE RIGHT MOMENT

ROUTE + CHANNELS Router & deliverers ◆ N6

The router resolves who should hear about a suggestion and on which channel, with escalation policies — and **never edits agent-authored content**. Specified channels include in-app, push, and SMS; two audio endpoints are planned — a wearable **earpiece** and a **walkie-talkie bridge** (text-to-speech onto the radios many stores already carry) — both treated as priority-preemptive serial audio media: critical interrupts lower-priority speech, queues are depth-limited, and timed fallback chains hop channels (audio → push → SMS) until someone is reached.

HUMANS Manager & associate app

A mobile-first floor app: a live queue of suggestions with quantified impact and the agent's reasoning on tap — act, assign to an associate, dismiss, or hand back — plus a direct conversation with the store's agent.

Every suggestion's **disposition is recorded** — acted on, assigned, dismissed, ignored, expired — alongside the reasoning record that produced it, so the usefulness of every alert is measurable per store, per event type, and per prompt version, and feeds directly into adjustment and learning.

§2 Candidate areas of novelty — for counsel review

Six mechanisms we believe may be distinctive, each expanding into *what it is* and *why it may be distinctive*. These are engineering-identified candidates, not legal conclusions.

◆ N1 Attention-gated, agent-primary store operations

WHAT IT IS

A three-stage decision funnel — deterministic safety floor → universal learned statistical attention gate → persistent per-store LLM agent — that replaces a rules engine entirely. The gate learns per store / event type / time-of-week baselines with no per-customer configuration and decides which real-world events are worth LLM reasoning.

WHY IT MAY BE DISTINCTIVE

Prior systems either hard-code rules per deployment or invoke models indiscriminately. Here the gate is the mechanism that makes **continuous LLM supervision of a high-volume physical event stream economically viable**, and tenant configuration describes the business rather than encoding behavior.

◆ N2 Replayable reasoning substrate & prompt evolution

WHAT IT IS

Every LLM invocation is recorded immutably with its exact composed prompt layers, versions, and retrieved context, so any historical decision is reconstructable. Every suggestion's disposition — acted on, dismissed, ignored, expired — is joined to the record that produced it, yielding a **per-suggestion usefulness signal**. Candidate behavior changes are backtested by replaying recorded history and scored against those dispositions, then shipped through controlled experiments.

WHY IT MAY BE DISTINCTIVE

A closed-loop method for **versioning, auditing, and evolving the behavior of an operational AI** where improvement is a measured context edit — no code deploy, with replay-verified fidelity against real past events.

◆ N3

On-prem connectors as universal retail sensors

WHAT IT IS

On-premise connectors that ride the operator's authenticated session on customer hardware and turn UI-only systems into events — a **dashboard connector** that navigates legacy web back-offices deterministically, and a **screen connector** where the operator shares a tab. Each uses an LLM solely for schema-bound metric extraction, gates every reading through validation, and emits idempotent events onto the same spine as physical sensors — with cost controls (change-detection gating, action-replay caching, cheap-model-first cascade).

WHY IT MAY BE DISTINCTIVE

A method for converting **UI-only systems with no API into first-class, auditable sensor streams** — the integration bottleneck of retail — with the LLM confined to a single validated step.

◆ N4

Edge perception with question-time visual re-analysis

WHAT IT IS

Deterministic perception runs **at the edge** — detection, tracking, and re-identification — and decides **when raw frames are worth keeping** (densely around events, sparsely otherwise), indexed by time, not embeddings, and shipped as rolling segments to a cloud world-store. **No VLM runs at the edge.** When a question or a shelf-level judgment arrives, the cloud answering model itself elects to re-inspect the retained pixels **in the context of the new question**, and vision re-runs only then; claims are separately verified against the tracked geometry before being committed.

WHY IT MAY BE DISTINCTIVE

Splitting **deterministic perception (cheap, local, always-on)** from **probabilistic meaning (expensive, cloud, on-demand)** lets the system answer about details that were never extracted at capture time, at a fraction of always-on vision-LLM cost — while keeping raw video inside the store's trust boundary until a specific question justifies moving a clip.

◆ N5

Synthetic-to-real sensor equivalence for training

WHAT IT IS

A UE5 digital-twin store whose virtual cameras and event generators implement the exact production sensor contracts, with scripted scenarios and known ground truth. The full pipeline — video → events → gate → agent → human delivery — is trained, exercised, and regression-tested synthetically; physical cameras replace virtual ones as a configuration change.

WHY IT MAY BE DISTINCTIVE

Combined with N2, the simulator doubles as the **benchmark substrate for behavior evolution**: prompt and model changes are scored against simulated scenarios with ground truth before touching a real store.

◆ N6

Priority-preemptive delivery to floor staff

WHAT IT IS

A routing policy resolves recipient and channel per suggestion with escalation; the audio endpoints — a wearable earpiece and a walkie-talkie bridge that speaks onto the radio hardware stores already use — treat worn or handheld audio as a serial medium with preemption rules (critical interrupts lower-priority speech, per-user queue depth limits) and timed cross-channel fallback chains (audio → push → SMS) until delivery is confirmed. Each delivery and its eventual disposition is recorded.

WHY IT MAY BE DISTINCTIVE

A delivery method matched to a hands-busy retail floor: **AI-generated, quantified guidance arbitrated across interruption-aware channels**, closing the loop from sensor to spoken word.

The combination may be the strongest story. Beyond the individual mechanisms, the end-to-end system — heterogeneous "whatever-the-store-has" ingestion (including the on-prem connectors, edge perception, and simulated sensors) into one replayable event stream, attention-gated LLM reasoning with an immutable audit substrate and measured behavior evolution, and preemptive human-in-the-loop delivery — may support system-level claims. This document is an engineering description prepared for a preliminary patentability discussion; it is not a legal analysis, and no prior-art search has been performed.

Sources: the platform requirements suite and ADRs in the product repo (architecture, agent & LLM, suggestion & routing, video-frame contract, ADR-0009 edge perception & world-store), the connectors track (dashboard & screen), and the Overwatch reference design. Diamond markers are engineering-identified candidates only.

Confidential — internal and counsel use.

LumicIQ

FEATURE ARCHITECTURE
· JUL 2026