Case Study

Internal Production Deployment — the Patrol Sight reference stack

HumanityAI'd — July 2026

CONFIDENTIAL

Summary

Before offering Patrol Sight to security buyers, HumanityAI'd built it as a real production system on its own infrastructure and ran it: a **34-service stack on a single RTX 4060 Ti**, a **93,728-person face index**, and RayNeo X3 Pro glasses streaming live to it. Recognition arrives in the officer's lens in **about 68 milliseconds**, measured stage by stage rather than estimated. This case study documents what was built, what was measured, and — as importantly — the production problems the deployment surfaced and how they were fixed.

The setup

- **Recognition host**: one NVIDIA RTX 4060 Ti (8 GB), sharing the GPU with other live AI workloads by CUDA time-slicing — a deliberately constrained, realistic environment rather than a dedicated benchmark rig.
- **The stack**: 34 services across five tiers — PostgreSQL 15, object storage and four Redis instances (stateful); connection pooling, authentication, business API, recognition orchestration and the command console (stateless); the GPU recognition engine, index worker and face-quality engine (GPU); and nginx, message bus and TURN relay (infrastructure). All 22 data volumes declared external so a teardown cannot destroy data.
- **The index**: 93,728 persons / 95,667 vectors in a single GPU FAISS shard using IVF-PQ compression.
- **The edge**: RayNeo X3 Pro AR glasses streaming hardware-encoded H.264 at 2560×1440 / 20 Mbps over 5 GHz WiFi 6, with results returned as JSON datagrams and rendered as colour-coded label pills in the lens.

What was measured

Metric	Result
Glasses-to-label, end to end	**~68 ms typical** (49 ms min, 103 ms max), measured stage by stage
Alternative WebRTC path, clock-synchronised	**p50 26–32 ms, p95 63–85 ms**
Effective recognition rate per stream	**14–20 fps**
Enrolled population	**93,728 persons** / 95,667 vectors, one GPU shard
Watchlist search latency	avg 5.4 ms, **P95 8.5 ms**, P99 11.7 ms — over 1,931 real searches
Vector searches actually performed	**2.4% of frames** (97.6% answered from the identity cache)
Network	**0% packet loss**, ~4 ms round trip
Bulk enrolment throughput	**40–60 images/second**

Metric	Result
Recognition range	Reliable ~5 m; 6–7 m in good light; 8–10 m in long-range mode
Biometric data sent off-premises	**Zero** — fully on-premises

Scale headroom, measured rather than extrapolated: recognition latency by index size runs 40–60 ms at 10k persons, 50–70 ms at 100k, 60–90 ms at 1M, and **80–130 ms at 13M**. IVF-PQ compression means 13 million faces occupy 962 MB instead of the 26 GB an uncompressed index would need — a 27× reduction. Index size is not the limiting factor; decode and detection are.

The engineering that produced those numbers

Three decisions did most of the work:

The two-layer tracker. Spatial bounding-box matching assigns a stable tracking identity; an embedding-and-identity cache then answers subsequent frames for the same person without touching the vector index. In practice the search is skipped on **97.6% of frames**, which cut average per-frame latency from 65 ms to 30 ms and GPU utilisation from 85% to 40%. This is what makes continuous patrol recognition viable on a single mid-range card.

Backpressure by design. The video ingest uses a newest-frame-wins reader with a frame-rate-capped worker, so the GPU never accumulates a queue. It is the difference between a system that stays at 68 ms and one whose latency climbs all shift.

Model choice by benchmark, not assumption. A 0–90° yaw study across 24 images found ArcFace R50 beat the lightweight MobileFaceNet in 19 of 21 comparisons (+0.054 mean similarity), with roughly **60% fewer false negatives at extreme pose**. Off-angle faces are the operational norm on patrol, so the heavier model was chosen deliberately and its GPU cost accepted.

Production problems the deployment surfaced — and the fixes

Running this on real, shared hardware surfaced the kind of failures only a live deployment reveals:

- **The wireless network was the bottleneck, not the AI.** On a 2.4 GHz hotspot with WiFi power-save enabled, round-trip latency was around 100 ms — larger than the entire recognition pipeline. Moving to 5 GHz WiFi 6 with power-save disabled brought it to ~4 ms and allowed 2560×1440 at 20 Mbps. This is now a specified deployment requirement rather than a discovery each customer makes.
- **A cache layer was creating false positives.** The deepest fallback (a PostgreSQL cosine search on cache miss) was removed from the glasses path because it produced false positives and cost 3–12 ms per miss. Removing a feature to improve accuracy is the correct trade in a security product.
- **The detector was hallucinating faces at high resolution.** Raising the detection confidence threshold from 0.35 to 0.60 eliminated 2–11 spurious detections per frame at the largest detector size.
- **Labels flickered and doubled in the lens.** Fixed with exponential smoothing plus spatial proximity matching, stale-label eviction, and duplicate-track suppression — and by fixing a bug where UUID tracking identifiers were parsed as integers, returning array indices instead of identities.
- **Login requests failed under load.** Traced to connection-pool exhaustion in session mode, shared between the authentication and business services. The pool was raised from 10 to 40 and the fix documented in our own notes as "*headroom, not a cure*" — the durable fix is transaction-mode pooling, and it is on the roadmap rather than quietly declared done.
- **Cold starts and boot races.** A GPU warm-up query removed ~170 ms from the first request, and a startup wrapper now waits for the driver and retries specifically on the FAISS CUDA initialisation race — so the stack recovers from a reboot with no human involvement.
- **An optimisation that was measured and rejected.** Network QoS marking was trialled, found to make the p95 latency tail roughly twice as bad, and reverted. Not every idea survives measurement, and we would rather report the ones that did not.

Every issue was diagnosed, fixed, redeployed and re-verified on the live system.

What this deployment does *not* prove

Stated plainly, because a security buyer will ask:

- It is **one site with a small number of devices**. The concurrent-device ceiling per GPU has not yet been established, and we will not estimate it — it is measured during the customer pilot and contracted on the measured figure.
- It runs on a **trusted internal network**, which is why the plaintext wearable transport has been acceptable so far. Encrypted transport is a scheduled hardening item, and until it ships the isolated-VLAN requirement is a control rather than a suggestion.
- It has **not been through an external security audit**, and it carries deployment-grade credential debt that is fully disclosed in the Technical Due-Diligence Report.
- It holds **no independent accuracy benchmark**. We have not submitted to NIST FRTE and do not imply a ranking; the substitute we offer is measured accuracy testing in the customer's own environment during the pilot.

Why it matters to a security operation

This was not a demonstration built for a meeting. It is the **same software, same deployment tooling, same operator console** a customer receives — proven to:

1. Identify a person in the officer's line of sight fast enough to act on during the encounter.
2. Search a nearly 94,000-person watchlist in single-digit milliseconds, with headroom measured to 13 million.
3. Run entirely on one on-premises server, with zero biometric data leaving the network.
4. Recover from reboots, survive shared-GPU contention, and absorb a full production hardening cycle with the incidents written down rather than buried.

From here to your operation

The POC Playbook turns this into a **6-week pilot at one of your sites**: your patrol area, your watchlist, your wireless network, your success metrics — with an information-security and data-protection gate before any live watchlist is loaded, a shadow-running week to establish the false-positive rate before any officer acts on a match, and the pilot's watchlist, thresholds and audit history carrying straight into rollout.