# AI Agent Project V — Dashboard

**Date:** Oct 5, 2025  
**Owner:** Ernesto Olivo  
**Context:** Continuation of the AI Agent initiative (KB‑Legal‑Assistant: Immigration to Spain) with KB‑first routing + web fallback.

## Priority Goals (source of truth)

1. **Ship a ChatGPT‑style UI** at ai.olivogalarza.com with an input box and a transcript showing all user messages and LLM answers.
2. **RAG behavior:** The UI must answer from the **KB**; if confidence is low or KB is insufficient, it should **gracefully use open‑web sources** to answer.

**Doc location:** /mnt/data/Priority Goals of AI Agent Project.docx

## RoE v1.1 — Inmigración a España (Rules of Engagement)

* **Domain routing:** Focus on immigration to Spain. Use **KB first**; if low‑confidence, do **web fallback**. If outside domain but in KB, still answer + check web. If outside domain and not in KB → reply: *“Esta IA está especializada en temas de Inmigración a España.”*
* **Accuracy & grounding:** Support answers with KB or recent web sources. **Target ≥95% correctness**. **No fabricated citations.**
* **When unsure:** Reply exactly **“Respuesta incierta. Contáctenos.”**
* **Citations:** Hidden by default; reveal only on request.
* **Style:** Spanish, **3–6 sentences**, neutral legal tone; bullets for procedures; **bold key terms**. If >6 sentences → add **Resumen:** line.
* **Safety/guardrails:** One‑time disclaimer per session: *“Esta IA puede cometer errores; use la información de forma referencial y contáctenos ante cualquier duda.”*
* **Formatting:** Use bullets for procedures; highlight key terms.
* **Session memory:** Maintain rolling context per chat.

**Doc location:** /mnt/data/Rules of Engagement.docx

## Repos & Deployment

* **API (Next.js):** kb-legal-assistant → served at https://api.olivogalarza.com
  + Key files: app/api/rag/answer/route.ts (RoE routing & finalizer), app/api/chat/route.ts (chat endpoint)
* **UI (Next.js):** kb-legal-assistant-ui → served at https://ai.olivogalarza.com
  + Key file: RagChat.tsx (UI with collapsible sources & disclaimer)
* **Planned helpers:** domain.ts, format.ts, permissions.ts (authorize), prompt templates (planner/critic/finalizer)

## Known Issues / Open Threads

1. **Redis caching shows x‑cache: MISS** despite SSE writing two keys (request-specific cacheKey and stable cacheKeyFiles).
   * Hypotheses: key mismatch (naming/namespace), different **runtime** (Edge vs Node) or region, TTL/expire on write, response end before write flush, CORS/OPTIONS path, or JSON stringify differences.
2. **/lib/rag/rewrite.ts** (Spanish legal‑query normalizer) **not working as intended**; normalization/regEx coverage needs a robust rewrite + tests.
3. **DevTools tests not working** (Console/Network checks from Project IV); need deterministic test harness and clear pass/fail logs.
4. **RoE v1.1 enforcement**—verify in app/api/rag/answer/route.ts (domain routing, style, uncertainty phrase, disclaimer).

## Immediate Action Plan (concise)

* **A. Redis MISS debugging (priority)**
  + Log keys on **write & read**; add await client.set(...) with explicit **TTL/EX**, then await client.get(...) to assert persistence within the same request.
  + Normalize key schema: rag:chat:{hash(query)} and rag:kbfiles:v1 (stable). Centralize in a single cache.ts to avoid drift.
  + Ensure both API routes run on **Node runtime** (not Edge) when using Upstash SDK; align **region** and **URL/TOKEN** envs for read/write parity.
  + Add a /api/debug/cache?key=... read‑only endpoint to introspect values quickly.
* **B. Query Rewriter v2**
  + Replace ad‑hoc regex with a **deterministic pipeline**: lowercasing → accent fold → stopword trim → synonym map (ley/extranjería/visado), typo fix (DL distance ≤1 for frequent terms), **legal entities expansion** (art., artículo, arts.), plus **unit tests**.
  + Ship a **benchmark suite**: 30 real queries → expected rewritten forms; track pass‑rate.
* **C. DevTools Tests**
  + Provide a tiny **/api/health** JSON and **/api/rag/answer** deterministic mock when ?test=1 to stabilize Console/Network assertions.
  + Script reproducible checks: curl + jq snippets + screenshots; store under /docs/tests/.
* **D. RoE v1.1 Checks**
  + Add a **finalizer** step ensuring Spanish, 3–6 sentences, bold key terms, bullets for procedures; inject disclaimer once per session.

## References / Paths

* **Priority Goals:** /mnt/data/Priority Goals of AI Agent Project.docx
* **RoE v1.1:** /mnt/data/Rules of Engagement.docx
* **Embeddings:** data/kb/embeddings.json
* **Typical test prompt:** ¿qué dice el artículo 123 del código procesal?
* **Working dirs:** /Users/eos/kb-legal-assistant and /Users/eos/kb-legal-assistant-ui

## ## Scratchpad (for notes as we progress)