**Final PRD (Rev 2) – Cursor AI Agent for LA County Fire Recovery**

🧱 **Section 1: Technical Architecture & Stack**

**1.1 Layered System Architecture**

|  |  |  |
| --- | --- | --- |
| **Layer** | **Description** | **Stack & Tools** |
| **Frontend** | User-facing chat interface, DOM parser | TypeScript, React, DOMParser API, WebSockets |
| **Inference & Routing** | Query understanding, context merging, fallback | OpenAI GPT-4.5 / o4 + Jurisdictional Router (custom logic) |
| **Embedding & Retrieval** | Semantic search for page/docs | Weaviate + OpenAI/Ada-002 or Cohere embeddings |
| **Memory Layer** | Session & recovery journey memory | Redis (short-term), Firestore/PostgreSQL (long-term) |
| **Admin & Alert Panel** | CMS for updates, alerts, and audits | Next.js (Admin UI), Firebase Auth, Cloud Functions |
| **Data Governance & Bias Auditing** | Model audits, fairness scoring | LangChain evals, custom prompt probes, Apache Superset for dashboards |
| **Infrastructure** | Hosting, CI/CD, observability | GCP (preferred), Docker, Cloud Run, Terraform, Datadog |
| **Security & API Layer** | Escalation hooks, user profiles | Node.js backend API, JWT auth, gRPC/REST bridges to external systems |

♻️ **Section 2: Content Freshness & Live Diff Engine**

**2.1 Re-indexing SLA**

* **Goal**: 90% of trusted documents re-ingested & embedded within **6 hours** of change.

**2.2 Implementation**

|  |  |  |
| --- | --- | --- |
| **Component** | **Function** | **Tools & Logic** |
| **Change Detection Agent** | Scrapes trusted URLs periodically | Puppeteer + Last-modified + DOM fingerprint hash |
| **Semantic Diff Engine** | Compares prior & current embeddings | Sentence Transformers + Cosine similarity threshold (e.g. < 0.85 = reprocess) |
| **Partial Re-Embedding** | Only chunks with meaningfully changed content are re-embedded | Smart chunking pipeline using LangChain splitters |
| **Alert Flagging** | Admin panel flags stale/inconsistent docs | Admin queue + confidence decay score shown |
| **Audit Trail** | Track doc change metadata | Timestamped logs with version hashes in Firestore |

🧩 **Section 3: Integration with Existing Systems**

**3.1 Real-Time Human Escalation Pathways**

|  |  |
| --- | --- |
| **Integration Area** | **Details** |
| **CRM Hook** | Integrate with existing LA County CRM (e.g. Salesforce Service Cloud, Zendesk, or Freshdesk) via REST/gRPC |
| **Ticket Sync** | Auto-create ticket with user query, session history, page URL, and AI summary |
| **Live Chat Integration** | Twilio Flex / Intercom / Freshchat plugin via iframe or webhook |
| **Context Transfer** | JSON payload with extracted user profile, history, location → CRM Notes field |
| **Two-Way Feedback Loop** | CRM operator can flag AI mistakes via sidebar tool (stored for model retraining) |

🧠 **Section 4: Bias, Hallucination & Transparency (Recap with Implementational Enhancements)**

**Bias Mitigation Enhancements**

* Automated fairness probes using test case libraries (e.g., Urban vs Rural LA, non-native English queries)
* Model response tagging by socio-linguistic impact risk
* Diversity index on trusted sources displayed in Admin dashboard

**Hallucination Prevention**

* Low-confidence fallback logic (score < 0.6 triggers visual uncertainty warning + fallback message)
* Prominent source citations:
  + **Low confidence** → **Source link bolded and prefixed: “Please double-check this here:”**
* Training samples where model is *penalized* for bluffing vs admitting gaps

🔐 **Section 5: Security & Compliance**

**5.1 Security Architecture**

|  |  |
| --- | --- |
| **Measure** | **Details** |
| **Input Sanitization** | All user inputs escaped, HTML tags stripped |
| **API Rate Limits** | JWT-based throttling (e.g., 100 req/hour per IP, burst protection) |
| **Role-Based Access Control** | Admin panel locked with Firebase Auth + role checks |
| **Escalation Access Logs** | All escalation payloads are logged, hashed, and audit-tracked |
| **Zero Retention on PII** | No long-term storage of personally identifiable information unless user explicitly opts in |
| **SOC 2 / FedRAMP alignment** | Infrastructure (GCP/AWS) inherits compliance; project configured to comply with CA state data handling mandates |
| **Penetration Testing Plan** | Scheduled prior to Phase 1 public launch; report to be submitted to LA County IT Security Office |

🔔 **Section 6: Proactive Alerting (Expanded)**

|  |  |
| --- | --- |
| **Feature** | **Detail** |
| **Alert Trigger Types** | Manual (admin sends), Location-based (ZIP match), Topic-based (based on recovery stage) |
| **User Opt-In UX** | Banner during chat: “Would you like updates on rebuilding in your area?” |
| **Delivery Mechanisms** | In-chat notification + Email/SMS (Phase 7) |
| **Personalization Fields** | ZIP code, damage type, interest flags (debris removal, rebuilding, reimbursement) |

📅 **Updated Final Agile Timeline**

|  |  |  |
| --- | --- | --- |
| **Sprint** | **Focus Area** | **Key Deliverables** |
| 1–2 | Base System Infra | Chat widget, embedding search, GCP setup |
| 3–4 | DOM/Context Fusion | DOM + jurisdiction-aware routing |
| 5–6 | Personalization + Memory | Redis + Firestore integration |
| 7–8 | Hallucination + Bias Mitigation | Uncertainty warnings, fairness test suite |
| 9–10 | Real-Time Escalation + CRM Integration | API bridge, ticket sync, context transfer |
| 11–12 | Notifications + Admin Panel | Alert builder, source tracker, fairness dashboard |
| 13–14 | Voice Input/Output PoC | Whisper-based input, basic TTS replies |
| 15–16 | QA, Red Team, Security Audit | Pen test, compliance checks, final launch gate |