**Final Enhanced PRD: Cursor AI Agent – LA County Fire Recovery Assistant**

🧠 **Strategic Enhancement Themes**

* **Bias-Resilient, Trustworthy Intelligence**
* **Multi-layered Context Fusion (DOM + Jurisdiction + User)**
* **Transparency in Uncertainty + Real-time Human Handoff**
* **Scalable Knowledge Ingestion + Personalization**
* **Proactive Alerting, Accessibility, and Ethical Governance**

✅ **Section 1: System Architecture**

**1.1 Modular Layered System**

* **Knowledge Sync Layer**: Crawling, chunking, tagging, versioning
* **Inference & Routing Layer**: Determines context path (jurisdiction, topic, current page)
* **Memory & Personalization Layer**: Stores session + long-term user journey context
* **Governance Layer** (NEW): Auditing, fairness scoring, hallucination detection triggers

🌐 **Section 2: Context Awareness & Location Intelligence**

* **DOM Parsing at Page Load**: Ingests headings, meta tags, and on-screen text
* **Breadcrumb Awareness**: Uses visible user path as context marker
* **Geo-Based Personalization**: County → City → Zip code fallback cascade

📚 **Section 3: Knowledge Base Management**

* **Trusted Document Registry**:
  + Sources are tagged with jurisdiction, recency, review status, and source diversity rating
* **Bias Mitigation Strategy**:
  + Quarterly audits of source diversity (ethnicity, geography, income coverage)
  + KPI: Bias/Fairness Score ≥ 90%
  + Test cases for socioeconomic/linguistic/cultural edge cases in LA County
* **Hallucination Guardrails**:
  + Embeds confidence scores in inference layer
  + Conditional prompting to explicitly warn users when uncertainty is high:  
    “I’m not fully confident in this answer. Here’s general guidance and a number to contact.”

🤖 **Section 4: Advanced Query Understanding**

* **Four-Level Context Merge**:
  1. User Input
  2. Current Page DOM
  3. Jurisdiction-specific policy
  4. LA County/FEMA universal policies
* **Intent & Ambiguity Classifier**:
  1. Determines: topic, intent (info-seeking vs next-action), jurisdiction, and urgency
  2. Clarifies ambiguities proactively: “Are you asking for Santa Monica or the County area?”
* **Thread-Aware Memory**:
  1. Thread depth: 3 past turns + user session flags (city, progress stage)

🗣️ **Section 5: Escalation and Human-in-the-Loop**

**5.1 Real-Time Escalation Pathway**

* **Trigger Conditions**:
  + Confidence score < 60%
  + Detected emergency keyword (e.g. “life-threatening,” “evacuate,” “injured”)
  + Bias detection triggered
* **Live Handoff Flow**:
  + Human support team receives:
    - Full query history
    - Page user was viewing
    - Inferred user state (“seeking permit help in unincorporated zone”)
  + Channels:
    - Live chat (phase 2)
    - SMS fallback
    - Direct phone dial-out link
* **Learning Loop**:
  + Human overrides or corrections directly inform model retraining for similar future cases

📥 **Section 6: Content Freshness & Admin Tools**

* **Live Diff Engine**:
  + Detects semantic (not just token-level) changes in updated pages
  + Triggers partial re-embedding + re-flagging of dependent intents
* **Admin Dashboard Features**:
  + New doc ingestion
  + Fairness scoring across cities
  + Flag/unflag hallucination reports
  + Alert trigger settings (for proactive messaging)

📈 **Section 7: Analytics, Auditing & KPIs**

|  |  |  |
| --- | --- | --- |
| **Metric** | **Goal** | **Notes** |
| Query Resolution Rate | >95% | Auto-classified |
| Confidence Disclosure Rate | 100% below 70% confidence | Shown with user alert |
| Hallucination Escalation Rate | <1% of queries | Auto-detected via hallucination prompts |
| Bias/Fairness Score | >90% | Measured via representative audit sets |
| Content Freshness SLA | 90% re-indexed in 6 hrs | Dashboard tracked |
| Task Completion Rate | 70% next-step success | e.g., download, form click |
| Human Escalation Latency | <30 sec to handoff | Measured weekly |

🧑**‍**🦽 **Section 8: Accessibility & Multimodal Roadmap**

**8.1 Current**

* Keyboard + screen reader support
* Mobile responsive chat widget

**8.2 Future (Phase 6+)**

* **Voice Input/Output**:
  + Dictation for queries
  + Audio summaries of responses
* **Image Intake**:
  + User uploads damage photo
  + Agent tags: "possible structural damage → rebuild permit likely"
* **Multimodal Output Options**:
  + Map overlays
  + Document highlighting
  + Video snippets (FEMA, county meetings)

📣 **Section 9: Proactive Notifications**

**9.1 Push Notification Framework (Phase 5+)**

* Admin-triggered alerts per:
  + User location (city or ZIP)
  + Expressed interest (e.g., rebuilding vs clean-up)
* **Sample Use Cases**:
  + “Debris pick-up window extended in Agoura Hills”
  + “New Cal OES reimbursement form live”
* User opt-in via session prompt

🧭 **Section 10: Ethical AI & Data Governance**

**10.1 Ethical AI Principles**

* **Fairness**: Equitable access and representation across all LA County communities
* **Transparency**: User-visible uncertainty and source citations
* **Accountability**: Escalation path + auditability of all agent responses
* **Privacy & Beneficence**: Only necessary data stored, anonymized before training

**10.2 Governance Plan**

* **Source-level tagging** for bias, recency, representation
* **Audit Logging**:
  + All queries stored with:
    - Confidence level
    - Source links used
    - Intent classification
* **Training Safeguards**:
  + Never train on raw user PII
  + Admin overrides required before any retraining cycle from human escalations

📅 **Updated Agile Timeline (With Enhancements)**

|  |  |  |
| --- | --- | --- |
| **Sprint** | **Focus Area** | **Key Deliverables** |
| 1–2 | Core agent + chat widget | Prompt routing, embedding search |
| 3–4 | DOM parsing + city-specific context | Page-context ingestion, jurisdiction fallback |
| 5–6 | Personalization layer | Recovery journey memory + assistant persona |
| 7–8 | Bias detection + hallucination guard | Audits, fairness KPI, uncertainty disclosure |
| 9–10 | Real-time escalation + human handoff | Live chat hook, context transfer pipeline |
| 11–12 | Notifications + admin dashboard | Alert builder, user segmentation |
| 13–14 | Accessibility + voice mode (R&D) | Speech-to-text PoC, voice reply stub |
| 15–16 | QA, stress tests, launch | Red-teaming, load tests, final sign-off |