**Product Requirements Document (PRD)**

**Cursor AI Agent - LA County Fire Recovery Assistant**

Version: 4.0 (Combined)

Owner: Maharaj Ji

Objective: Build a context-aware, ethically-governed AI chatbot that integrates into your site and dynamically answers user queries based on current page context and continuously updated trusted documents, with advanced query understanding and proactive assistance capabilities. The agent will provide real-time guidance to residents affected by LA County fires, maintaining strict ethical AI principles and bias mitigation, while ensuring efficient development and robust security. 1

**Executive Summary**

The Cursor AI Agent will serve as a contextually-aware, ethically-governed chatbot integrated into fire recovery assistance websites, providing real-time guidance to residents affected by LA County fires. 1 The agent will dynamically understand user location context, scrape current page content in real-time, provide relevant information with proper citations, and proactively assist users while maintaining strict ethical AI principles and bias mitigation. 1 This comprehensive PRD combines a deep commitment to ethical AI and user experience with granular technical specificity and a pragmatic, well-structured agile timeline, ensuring a final product that is not only efficient and secure but also trustworthy, equitable, and continuously improving.

**Section 1: Overall Objectives & Phased Development Approach**

**1.1 Overall Objectives**

The overarching objective is to build a context-aware, ethically-governed AI chatbot that integrates into your site and dynamically answers user queries based on current page context and continuously updated trusted documents, with advanced query understanding and proactive assistance capabilities. 1

**1.2 Phased Development Approach (7 Phases)**

The project is structured into seven distinct phases, each with clear objectives and success criteria, allowing for breaking down a complex AI project into manageable, measurable chunks. 1

* **Phase 1: Initialization & Frontend Widget**: Create embedded chatbot widget with real-time page context awareness and ethical AI foundation. 1
* **Phase 2: Document & Link Integration with Bias Detection**: Build semantic knowledge base from provided documents and URLs with continuous updates and bias monitoring. 1
* **Phase 3: Advanced AI Logic & Ethical Query Understanding**: Implement sophisticated NLP for intent recognition, bias mitigation, and ethical response generation. 1
* **Phase 4: Proactive Features & Advanced User Experience**: Implement proactive assistance capabilities and advanced user experience features. 1
* **Phase 5: Human-in-the-Loop Integration & Advanced Testing**: Implement comprehensive human oversight system and advanced testing framework. 1
* **Phase 6: Administration & Ethical Content Management**: Comprehensive admin system for ethical content management and AI governance. 1
* **Phase 7: Advanced Analytics & Multimodal Accessibility Roadmap**: Implement sophisticated analytics and plan for multimodal accessibility features. 1

**Section 2: Technical Architecture & Stack**

**2.1 Layered System Architecture**

The system's architecture is designed with distinct layers, each utilizing specific technologies and tools to ensure robust functionality and scalability. 1

|  |  |  |
| --- | --- | --- |
| Layer | Description | Stack & Tools |
| Frontend | User-facing chat interface, DOM parser | TypeScript, React, DOMParser API, WebSockets 1 |
| Inference & Routing | Query understanding, context merging, fallback | OpenAI GPT-4.5 / o4 + Jurisdictional Router (custom logic) 1 |
| Embedding & Retrieval | Semantic search for page/docs | Weaviate + OpenAI/Ada-002 or Cohere embeddings 1 |
| Memory Layer | Session & recovery journey memory | Redis (short-term), Firestore/PostgreSQL (long-term) 1 |
| Admin & Alert Panel | CMS for updates, alerts, and audits | Next.js (Admin UI), Firebase Auth, Cloud Functions 1 |
| Data Governance & Bias Auditing | Model audits, fairness scoring | LangChain evals, custom prompt probes, Apache Superset for dashboards 1 |
| Infrastructure | Hosting, CI/CD, observability | GCP (preferred), Docker, Cloud Run, Terraform, Datadog 1 |
| Security & API Layer | Escalation hooks, user profiles | Node.js backend API, JWT auth, gRPC/REST bridges to external systems 1 |

**2.2 Ethical AI Layer**

This foundational layer integrates ethical considerations directly into the system's core. 1

* **Bias Detection Engine**: Real-time bias identification and mitigation. 1
* **Hallucination Prevention**: Confidence-based response filtering. 1
* **Transparency Module**: Uncertainty disclosure and source attribution. 1
* **Fairness Monitor**: Demographic representation and equity analysis. 1

**2.3 Advanced Frontend Integration**

The frontend is designed for advanced user interaction and accessibility. 1

* **Multimodal Interface**: Voice, text, and visual input capabilities. 1
* **Accessibility Engine**: WCAG 2.1 AA+ compliance with future multimodal support. 1
* **Proactive Notification System**: Location and interest-based suggestions. 1
* **Ethical Transparency Interface**: Confidence scores, bias warnings, uncertainty disclosure. 1

**2.4 Enhanced Backend Infrastructure**

The backend provides robust support for AI governance and advanced analytics. 1

* **Ethical AI Governance API**: Policy enforcement and compliance monitoring. 1
* **Advanced Analytics Engine**: Sophisticated metrics and predictive capabilities. 1
* **Human-AI Collaboration Platform**: Seamless handoff and learning integration. 1
* **Multimodal Processing Pipeline**: Future-ready architecture for voice and visual processing. 1

**Section 3: Content Freshness & Live Diff Engine**

**3.1 Re-indexing SLA**

**Goal**: 90% of trusted documents re-ingested & embedded within 6 hours of change. 1

**3.2 Implementation**

To ensure content accuracy and timeliness, a sophisticated re-indexing system is implemented. 1

|  |  |  |
| --- | --- | --- |
| Component | Function | Tools & Logic |
| Change Detection Agent | Scrapes trusted URLs periodically | Puppeteer + Last-modified + DOM fingerprint hash 1 |
| Semantic Diff Engine | Compares prior & current embeddings | Sentence Transformers + Cosine similarity threshold (e.g. < 0.85 = reprocess) 1 |
| Partial Re-Embedding | Only chunks with meaningfully changed content are re-embedded | Smart chunking pipeline using LangChain splitters 1 |
| Alert Flagging | Admin panel flags stale/inconsistent docs | Admin queue + confidence decay score shown 1 |
| Audit Trail | Track doc change metadata | Timestamped logs with version hashes in Firestore 1 |

**Section 4: Ethical AI Governance Framework**

This framework integrates ethical considerations into every layer and phase of development, ensuring the AI agent is not just functional but also trustworthy and equitable. 1

**4.1 Bias Mitigation Strategies**

* **Real-time Bias Detection**: Automated fairness probes using test case libraries (e.g., Urban vs Rural LA, non-native English queries) and model response tagging by socio-linguistic impact risk. 1
* **Demographic Representation Analysis**: Monitoring and ensuring fair representation in content and responses. 1
* **Multi-perspective Content Validation**: Validating content against multiple viewpoints. 1
* **Historical Bias Pattern Recognition**: Identifying and addressing historical biases in data. 1
* **Diversity Index**: Displaying a diversity index on trusted sources in the Admin dashboard. 1

**4.2 Hallucination Prevention**

* **Confidence Thresholds**: Low-confidence fallback logic (score < 0.6 triggers visual uncertainty warning + fallback message). 1
* **Source Verification Requirements**: Prominent source citations: Low confidence → Source link bolded and prefixed: “Please double-check this here:”. 1
* **Uncertainty Disclosure Protocols**: Clear communication about limitations. 1
* **Training Penalties**: Training samples where the model is penalized for bluffing vs admitting gaps. 1
* **Fact-checking**: Against authoritative sources. 1

**4.3 Transparency Principles**

* **Clear Confidence Score Display**: Showing confidence scores (0-100%) to users. 1
* **Source Attribution**: For all claims, with reliability scores. 1
* **Uncertainty Acknowledgment**: "I'm not sure about this" indicators and "How confident am I?" toggle. 1
* **Bias Warning Alerts**: Displayed when detected. 1
* **Ethical Citation Format**: "According to (Reliability: 95%), [information]... [Confidence: 87%]" 1

**4.4 Fairness**

* **Fairness Monitor**: Dedicated monitoring for demographic representation and equity analysis. 1
* **Equitable Support**: Ensuring equal quality assistance across all demographic groups. 1
* **Bias-Free Information Access**: Eliminating discriminatory patterns in information delivery. 1

**Section 5: User Experience (UX) Features & Accessibility**

**5.1 Frontend Widget Integration**

* **Widget UI**: Embedded chat interface with branding and user-friendly layout. 1
* **Trigger Conditions**: Automatically appear on relevant pages (e.g., after 10s or on scroll). 1
* **Context Tagging**: Script scrapes <title>, meta tags, and key headings from current page. 1
* **Ethical AI Indicators**: Confidence scores, uncertainty disclosure, and bias warnings displayed. 1

**5.2 Basic Chat Interface with Ethical AI Features**

* **UI Components**: Minimizable chat bubble with accessibility indicators, message history with timestamps and confidence scores, typing indicators with processing transparency, quick action buttons with bias-aware suggestions, citation display for source references, uncertainty disclosure badges, and a "How confident am I?" toggle for transparency. 1
* **Structured Responses**: Step-by-step procedures with confidence indicators. 1

**5.3 Mobile Optimization**

* Responsive design for all screen sizes. 1

**5.4 Accessibility Foundation**

* WCAG 2.1 AA compliance with roadmap for multimodal interactions. 1

**5.5 Personalization Without Privacy Invasion**

* **Session-Based Personalization**: Adapted responses based on user's current session, relevant information prioritization, contextual interface adjustments, and progressive disclosure optimization. 1
* **Preference Learning**: Response format preferences, information depth preferences, communication style adaptation, and accessibility needs recognition. 1

**5.6 Multimodal Accessibility Roadmap**

* **Phase 1 - Voice Integration**: Voice query input capabilities (Whisper-based input), audio response generation (basic TTS replies), voice-based navigation, and accessibility compliance for visually impaired users. 1
* **Phase 2 - Visual Processing**: Image intake and analysis, document photo processing, visual accessibility features, and screen reader optimization. 1
* **Phase 3 - Advanced Multimodal**: Video snippet processing, real-time transcription, multi-language audio support, and advanced accessibility features. 1

**Section 6: Advanced AI Logic**

**6.1 Enhanced AI Logic Layer**

* **Advanced NLP Engine**: Classifies intent, detects location/context, handles ambiguity. 1
* **Bias Mitigation System**: Identifies and corrects biased responses in real-time. 1
* **Hallucination Detection**: Identifies and flags potentially fabricated information. 1
* **Response Synthesizer**: Combines extracted data into clear, ethical responses. 1
* **Ambiguity Clarification**: Detects unclear queries and asks clarifying questions. 1
* **Fallback Router**: Detects irrelevant queries and triggers appropriate fallback. 1

**6.2 Advanced Intent Classification System**

* **Primary Intent Categories**: Information Seeking, Status Checking, Process Guidance, Emergency/Urgent, Comparative, Location-Specific, Ambiguous/Unclear. 1
* **Multi-level Context Integration**: Immediate query context, page/site context, user session history, jurisdictional context, temporal context. 1

**6.3 Ambiguity Clarification System**

* **Ambiguity Detection Triggers**: Multiple possible interpretations, missing critical context (location, timeframe, specific program), conflicting information requirements, unclear pronouns or references. 1
* **Clarification Strategies**: Structured follow-up questions, multiple choice options, context-aware suggestions, progressive disclosure of information. 1

**6.4 Enhanced Response Generation Framework**

* **Structured Responses**: Step-by-step procedures with confidence indicators. 1
* **Ethical Citation Format**: "According to (Reliability: 95%), [information]... [Confidence: 87%]" 1
* **Bias-Aware Guidance**: Include multiple perspectives when appropriate. 1
* **Confidence Scoring**: Transparent scoring to trigger fallback when needed. 1
* **Alternative Perspective Integration**: Proactively suggest different viewpoints. 1

**6.5 Advanced Fail-Safe Fallback System**

* **Graceful Degradation**: "I don't have specific information about that (Confidence: 23%), but here's who can help..." 1
* **Ethical Uncertainty Disclosure**: Clear communication about limitations. 1
* **Contact Routing**: Provide relevant contact information based on detected location/topic. 1
* **Escalation Triggers**: Flag complex queries for human review. 1
* **Learning Integration**: Log fallback cases for knowledge base improvement. 1

**6.6 Advanced Query Understanding**

* **Multi-turn Conversation Management**: Context retention across conversation turns, reference resolution for pronouns and implicit references, progressive information gathering, conversation state management. 1
* **Complex Query Decomposition**: Break down multi-part questions, prioritize query components, sequential response delivery, comprehensive answer synthesis. 1

**Section 7: Human-in-the-Loop Integration**

This section details the comprehensive human oversight system and seamless integration with existing LA County systems. 1

**7.1 Live Handoff Flow**

* Seamless transition from AI to human support. 1
* Complete context transfer including conversation history. 1
* User preference preservation. 1
* Continuity maintenance. 1

**7.2 Escalation Triggers**

* Confidence threshold breaches (<70%). 1
* Bias detection alerts. 1
* Potential hallucination flags. 1
* User frustration indicators. 1
* Emergency/urgent query detection. 1
* Complex legal/medical questions. 1

**7.3 Information Transfer Protocol**

* Complete conversation history. 1
* User context and preferences. 1
* Attempted solutions and outcomes. 1
* Confidence scores and uncertainty areas. 1
* Detected bias or hallucination concerns. 1

**7.4 Learning Loop Integration**

* **Human Feedback Integration**: Expert review of AI responses, correction and improvement suggestions, bias identification and mitigation, quality assurance validation. 1
* **Continuous Improvement Pipeline**: Regular model fine-tuning based on human feedback, bias mitigation strategy updates, knowledge base refinement, response quality enhancement. 1

**7.5 Integration with Existing Systems**

* **CRM Hook**: Integrate with existing LA County CRM (e.g. Salesforce Service Cloud, Zendesk, or Freshdesk) via REST/gRPC. 1
* **Ticket Sync**: Auto-create ticket with user query, session history, page URL, and AI summary. 1
* **Live Chat Integration**: Twilio Flex / Intercom / Freshchat plugin via iframe or webhook. 1
* **Context Transfer**: JSON payload with extracted user profile, history, location → CRM Notes field. 1
* **Two-Way Feedback Loop**: CRM operator can flag AI mistakes via sidebar tool (stored for model retraining). 1

**Section 8: Proactive Features & Alerting**

**8.1 Proactive Assistance System**

* **Location-Based Proactive Notifications**: Relevant updates for user's detected location, new resource availability alerts, deadline reminders for time-sensitive processes, weather-related safety updates. 1
* **Interest-Based Suggestions**: Proactive information based on user's browsing patterns, relevant resource recommendations, process completion assistance, follow-up guidance for multi-step procedures. 1
* **Contextual Recommendations**: Page-specific relevant information, next-step suggestions based on current context, related resource discovery, preventive guidance. 1

**8.2 Alert Trigger Types**

* Manual (admin sends), Location-based (ZIP match), Topic-based (based on recovery stage). 1

**8.3 User Opt-In UX**

* Banner during chat: “Would you like updates on rebuilding in your area?” 1

**8.4 Delivery Mechanisms**

* In-chat notification + Email/SMS (Phase 7). 1

**8.5 Personalization Fields**

* ZIP code, damage type, interest flags (debris removal, rebuilding, reimbursement). 1

**Section 9: Security & Compliance**

**9.1 Security Architecture**

A robust security architecture is detailed to protect user data and system integrity. 1

|  |  |
| --- | --- |
| Measure | Details |
| Input Sanitization | All user inputs escaped, HTML tags stripped 1 |
| API Rate Limits | JWT-based throttling (e.g., 100 req/hour per IP, burst protection) 1 |
| Role-Based Access Control | Admin panel locked with Firebase Auth + role checks 1 |
| Escalation Access Logs | All escalation payloads are logged, hashed, and audit-tracked 1 |
| Zero Retention on PII | No long-term storage of personally identifiable information unless user explicitly opts in 1 |

**9.2 SOC 2 / FedRAMP Alignment & CA State Data Handling Mandates**

* Infrastructure (GCP/AWS) inherits compliance; project configured to comply with CA state data handling mandates. 1

**9.3 Ethical AI Governance Framework (Policy Enforcement, Audit and Accountability)**

* **Policy Enforcement**: Automated policy compliance checking, ethical guideline adherence monitoring, bias mitigation requirement enforcement, transparency standard maintenance. 1
* **Audit and Accountability**: Comprehensive audit trails, decision explanation capabilities, bias incident reporting, corrective action tracking. 1

**Section 10: Administration & Ethical Content Management**

**10.1 Enhanced Administration Tools**

|  |  |  |
| --- | --- | --- |
| Tool | Purpose | Features |
| Ethical AI Dashboard | Monitor bias, hallucination, and confidence metrics | Real-time bias alerts, hallucination detection, confidence distribution 1 |
| Advanced Admin Panel | Manage content with ethical oversight | Link management, bias scoring, ethical review workflows 1 |
| AI Governance Center | Oversee AI behavior and compliance | Policy enforcement, audit trails, compliance reporting 1 |
| Training Set Builder | Create ethically-balanced training data | Bias detection, demographic representation, quality assurance 1 |

**10.2 Ethical Content Management System**

* **Bias-Aware Content Management**: Multi-perspective content validation, demographic representation analysis, alternative viewpoint integration, historical bias pattern recognition. 1
* **Quality Assurance Workflows**: Automated ethical screening, human review requirements, bias mitigation strategies, transparency requirement compliance. 1

**10.3 AI Governance Framework**

* **Policy Enforcement**: Automated policy compliance checking, ethical guideline adherence monitoring, bias mitigation requirement enforcement, transparency standard maintenance. 1
* **Audit and Accountability**: Comprehensive audit trails, decision explanation capabilities, bias incident reporting, corrective action tracking. 1

**10.4 Audit Trail**

* Track document change metadata with timestamped logs and version hashes in Firestore. 1

**Section 11: Testing Frameworks**

**11.1 Advanced Testing Framework**

* Comprehensive testing for ethical AI compliance and system performance. 1

**11.2 Bias Testing Suite**

* Demographic representation testing, language pattern bias detection, systematic bias scenario testing, intersectional bias analysis. 1

**11.3 Hallucination Detection Testing**

* Fabricated information identification, source verification accuracy, confidence calibration testing, uncertainty disclosure validation. 1

**11.4 Ethical AI Compliance Testing**

* Transparency requirement verification, fairness metric validation, accountability mechanism testing, explainability standard compliance. 1

**11.5 Penetration Testing Plan**

* Scheduled prior to Phase 1 public launch; report to be submitted to LA County IT Security Office. 1

**11.6 QA, Red Team, Security Audit**

* Pen test, compliance checks, final launch gate. 1

**Section 12: Success Criteria & KPIs**

**12.1 Ethical AI Metrics**

|  |  |  |
| --- | --- | --- |
| Metric | Target | Measurement Method |
| Bias Detection Accuracy | ≥ 95% | Systematic bias testing 1 |
| Hallucination Detection | ≥ 90% | Fabrication identification testing 1 |
| Confidence Calibration | ≥ 85% | Confidence vs. accuracy correlation 1 |
| Transparency Compliance | 100% | Uncertainty disclosure verification 1 |
| Fairness Score | ≥ 90% | Demographic equity analysis 1 |

**12.2 Advanced Performance Indicators**

|  |  |  |
| --- | --- | --- |
| Metric | Target | Measurement Method |
| Response Accuracy | ≥ 92% | Manual verification with bias consideration 1 |
| Proactive Suggestion Relevance | ≥ 80% | User engagement with suggestions 1 |
| Multi-turn Conversation Success | ≥ 90% | Context retention accuracy 1 |
| Human Handoff Success | ≥ 95% | Smooth transition verification 1 |
| Accessibility Compliance | WCAG 2.1 AA+ | Automated and manual accessibility testing 1 |

**12.3 Business Impact with Ethical Considerations**

* **Equitable Support**: Ensure equal quality assistance across all demographic groups. 1
* **Bias-Free Information Access**: Eliminate discriminatory patterns in information delivery. 1
* **Transparent Decision Making**: Clear explanation of AI reasoning and limitations. 1
* **Inclusive User Experience**: Accessible design for users with varying abilities. 1
* **Ethical AI Leadership**: Demonstrate responsible AI implementation in public service. 1

**Section 13: Risk Mitigation**

**13.1 Ethical AI Risks**

* **Bias Amplification**: Multi-layer bias detection and mitigation strategies. 1
* **Hallucination Risks**: Confidence thresholds and fact-checking requirements. 1
* **Transparency Failures**: Mandatory uncertainty disclosure and source attribution. 1
* **Fairness Violations**: Continuous demographic equity monitoring. 1
* **Accountability Gaps**: Comprehensive audit trails and explainable AI. 1

**13.2 Advanced Technical Risks**

* **Multimodal Complexity**: Phased rollout with thorough testing. 1
* **Accessibility Failures**: Continuous compliance monitoring and user feedback. 1
* **Proactive Feature Overreach**: User control and preference management. 1
* **Human-AI Handoff Issues**: Seamless transition protocols and fallback systems. 1

**Section 14: Project Timeline**

**14.1 Phased Development Approach (Summary)**

The project will follow a 7-phase development approach, ensuring a structured and measurable path to delivery. 1

**14.2 Detailed Agile Timeline (Sprints)**

A granular 16-sprint timeline provides a clear roadmap for project execution. 1

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