# Product Requirements Document (PRD)

## Cursor AI Agent - LA County Fire Recovery Assistant

\*\*Version\*\*: 3.0

\*\*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.

## 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. 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.

---

## Phase 1: Initialization & Frontend Widget

\*\*Objective\*\*: Create embedded chatbot widget with real-time page context awareness and ethical AI foundation

### 1.1 Frontend Widget Integration

| \*\*Task\*\* | \*\*Description\*\* | \*\*Checkpoint\*\* |

|----------|----------------|----------------|

| Widget UI | Embedded chat interface with branding and user-friendly layout | ✅ UI renders and opens/closes properly |

| Trigger Conditions | Automatically appear on relevant pages (e.g. after 10s or on scroll) | ✅ Behavior verified via user test |

| Context Tagging | Script scrapes `<title>`, meta tags, and key headings from current page | ✅ Output shows correct tags in dev console |

| Ethical AI Indicators | Confidence scores, uncertainty disclosure, and bias warnings | ✅ Transparency indicators display correctly |

### 1.2 Page Context Extraction System

\*\*Real-time Scraping\*\*: JavaScript scraper extracts semantic info from active webpage

\*\*Context Elements\*\*:

- Page title and meta descriptions

- H1-H3 headings for topic identification

- URL patterns for location detection

- Key content blocks and form elements

- User interaction patterns and session context

\*\*Location Detection\*\*: Automatic city/jurisdiction identification from URL and content

\*\*Topic Classification\*\*: Categorize page content (debris removal, permits, insurance, etc.)

\*\*Multi-level Context Merging\*\*: Combine immediate page context with user session history and broader site navigation patterns

### 1.3 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

- "How confident am I?" toggle for transparency

\*\*Ethical AI Interface Elements\*\*:

- Confidence score display (0-100%)

- "I'm not sure about this" indicators

- Bias warning alerts when detected

- Alternative perspective suggestions

- Clear source attribution with reliability scores

\*\*Mobile Optimization\*\*: Responsive design for all screen sizes

\*\*Accessibility Foundation\*\*: WCAG 2.1 AA compliance with roadmap for multimodal interactions

### 1.4 Success Criteria for Phase 1

- Widget embeds successfully on all target sites

- Real-time context extraction works on 5+ test pages

- UI opens/closes and handles basic interactions

- Context data appears correctly in dev console

- Mobile responsiveness verified across devices

- Ethical AI indicators display properly in all scenarios

\*\*Deliverable\*\*: Working widget with static responses, real-time context extraction, and ethical AI transparency features

---

## Phase 2: Document & Link Integration with Bias Detection

\*\*Objective\*\*: Build semantic knowledge base from provided documents and URLs with continuous updates and bias monitoring

### 2.1 Backend & Data Layer Architecture

| \*\*Module\*\* | \*\*Description\*\* | \*\*Checkpoint\*\* |

|------------|----------------|----------------|

| Context Parser | Extracts semantic info from active webpage (location, topic, etc.) | ✅ Parses correctly on 3+ test pages |

| Document Indexer | Caches and semantically indexes contents of provided links/docs | ✅ All URLs/docs indexed without error |

| Bias Detection Engine | Monitors content for potential bias patterns and flags concerns | ✅ Detects 95%+ of known bias test cases |

| Retriever Engine | Matches user queries to the most relevant content chunks | ✅ Query returns top 3 semantically relevant results |

| Knowledge Updater | Monitors doc/URL changes & re-ingests periodically or via webhook | ✅ Triggered update adds new info to index |

### 2.2 Document Processing Pipeline with Bias Mitigation

\*\*Input Sources\*\*:

- HTML pages from government websites

- Google Docs (public links)

- PDF documents

- Dynamic content from recovery sites

\*\*Processing Steps\*\*:

- Content extraction and cleaning

- Bias detection and flagging

- Text chunking with overlap

- Semantic embedding generation

- Metadata tagging (source, date, topic, jurisdiction, bias risk score)

- Fact-checking against authoritative sources

\*\*Bias Detection Framework\*\*:

- Language pattern analysis for discriminatory content

- Demographic representation analysis

- Authoritative source verification

- Multi-perspective content validation

- Historical bias pattern recognition

### 2.3 Semantic Indexing System with Quality Scoring

\*\*Embedding Model\*\*: OpenAI text-embedding-ada-002 or equivalent

\*\*Chunk Strategy\*\*:

- 500-word chunks with 50-word overlap

- Preserve section headers and context

- Maintain source attribution

- Include bias risk assessment

\*\*Enhanced Metadata Schema\*\*:

```json

{

"source\_url": "string",

"jurisdiction": "string",

"topic\_category": "string",

"last\_updated": "timestamp",

"content\_type": "string",

"reliability\_score": "number",

"bias\_risk\_score": "number",

"fact\_check\_status": "string",

"demographic\_representation": "object",

"alternative\_perspectives": "array"

}

```

### 2.4 Real-time Content Monitoring with Quality Assurance

\*\*Change Detection\*\*: Monitor source URLs for content updates

\*\*Update Triggers\*\*:

- Scheduled crawling (daily for critical pages)

- Webhook notifications when available

- Manual refresh via admin panel

- Bias pattern alerts

\*\*Quality Assurance Pipeline\*\*:

- Automated fact-checking against authoritative sources

- Bias detection screening

- Multi-perspective validation

- Human review triggers for sensitive content

### 2.5 Success Criteria for Phase 2

- All provided URLs successfully indexed with bias scoring

- Semantic search returns relevant results for test queries

- Update system detects and ingests content changes

- Retrieval accuracy >85% for domain-specific queries

- Bias detection accuracy >95% on test dataset

- Processing time <5 minutes for new document ingestion

\*\*Deliverable\*\*: Bot can respond using reference material from real documents/URLs with proper citations, bias warnings, and confidence scores

---

## Phase 3: Advanced AI Logic & Ethical Query Understanding

\*\*Objective\*\*: Implement sophisticated NLP for intent recognition, bias mitigation, and ethical response generation

### 3.1 Enhanced AI Logic Layer

| \*\*Module\*\* | \*\*Description\*\* | \*\*Checkpoint\*\* |

|------------|----------------|----------------|

| Advanced NLP Engine | Classifies intent, detects location/context, handles ambiguity | ✅ Queries correctly classified with 97%+ accuracy |

| Bias Mitigation System | Identifies and corrects biased responses in real-time | ✅ Bias detection >95% accuracy |

| Hallucination Detection | Identifies and flags potentially fabricated information | ✅ Hallucination detection >90% accuracy |

| Response Synthesizer | Combines extracted data into clear, ethical responses | ✅ 80%+ user satisfaction in A/B test |

| Ambiguity Clarification | Detects unclear queries and asks clarifying questions | ✅ Ambiguity detection >85% accuracy |

| Fallback Router | Detects irrelevant queries and triggers appropriate fallback | ✅ 100% fallback on gibberish/edge cases |

### 3.2 Advanced Intent Classification System

\*\*Primary Intent Categories\*\*:

- \*\*Information Seeking\*\*: "How do I apply for debris removal?"

- \*\*Status Checking\*\*: "When can I return to my property?"

- \*\*Process Guidance\*\*: "What's the next step in rebuilding?"

- \*\*Emergency/Urgent\*\*: "I need immediate help"

- \*\*Comparative\*\*: "What's the difference between Phase 1 and Phase 2?"

- \*\*Location-Specific\*\*: "Who do I contact in Pasadena?"

- \*\*Ambiguous/Unclear\*\*: Requires clarification before proceeding

\*\*Multi-level Context Integration\*\*:

- Immediate query context

- Page/site context

- User session history

- Jurisdictional context

- Temporal context (time-sensitive information)

### 3.3 Ethical AI Governance Framework

\*\*Bias Mitigation Strategies\*\*:

- Real-time bias detection in responses

- Demographic representation analysis

- Alternative perspective suggestions

- Language pattern analysis for discriminatory content

- Historical bias pattern recognition

\*\*Hallucination Prevention\*\*:

- Confidence thresholds for response generation

- Source verification requirements

- Uncertainty disclosure protocols

- Fact-checking against authoritative sources

- Response reliability scoring

\*\*Transparency Principles\*\*:

- Clear confidence score display

- Source attribution for all claims

- Uncertainty acknowledgment

- Alternative perspective suggestions

- Bias warning alerts

### 3.4 Ambiguity Clarification System

\*\*Ambiguity Detection Triggers\*\*:

- Multiple possible interpretations

- Missing critical context (location, timeframe, specific program)

- Conflicting information requirements

- Unclear pronouns or references

\*\*Clarification Strategies\*\*:

- Structured follow-up questions

- Multiple choice options

- Context-aware suggestions

- Progressive disclosure of information

### 3.5 Enhanced Response Generation Framework

\*\*Structured Responses\*\*: Step-by-step procedures with confidence indicators

\*\*Ethical Citation Format\*\*: "According to [Source Name] (Reliability: 95%), [information]... [Confidence: 87%]"

\*\*Bias-Aware Guidance\*\*: Include multiple perspectives when appropriate

\*\*Confidence Scoring\*\*: Transparent scoring to trigger fallback when needed

\*\*Alternative Perspective Integration\*\*: Proactively suggest different viewpoints

### 3.6 Advanced Fail-Safe Fallback System

\*\*Graceful Degradation\*\*: "I don't have specific information about that (Confidence: 23%), but here's who can help..."

\*\*Ethical Uncertainty Disclosure\*\*: Clear communication about limitations

\*\*Contact Routing\*\*: Provide relevant contact information based on detected location/topic

\*\*Escalation Triggers\*\*: Flag complex queries for human review

\*\*Learning Integration\*\*: Log fallback cases for knowledge base improvement

### 3.7 Success Criteria for Phase 3

- Intent classification accuracy >97% on test dataset

- Bias detection accuracy >95% in real-time responses

- Hallucination detection >90% accuracy

- Response relevance rating >90% in user testing

- Fallback system catches 100% of off-topic queries

- Citation accuracy rate >98%

- Ambiguity detection >85% accuracy

- Average response generation time <3 seconds

\*\*Deliverable\*\*: Ethically-governed bot returns accurate, bias-aware, citation-based answers with confidence scores and uncertainty disclosure

---

## Phase 4: Proactive Features & Advanced User Experience

\*\*Objective\*\*: Implement proactive assistance capabilities and advanced user experience features

### 4.1 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

\*\*Interest-Based Suggestions\*\*:

- Proactive information based on user's browsing patterns

- Relevant resource recommendations

- Process completion assistance

- Follow-up guidance for multi-step procedures

\*\*Contextual Recommendations\*\*:

- Page-specific relevant information

- Next-step suggestions based on current context

- Related resource discovery

- Preventive guidance

### 4.2 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

\*\*Complex Query Decomposition\*\*:

- Break down multi-part questions

- Prioritize query components

- Sequential response delivery

- Comprehensive answer synthesis

### 4.3 Personalization Without Privacy Invasion

\*\*Session-Based Personalization\*\*:

- Adapted responses based on user's current session

- Relevant information prioritization

- Contextual interface adjustments

- Progressive disclosure optimization

\*\*Preference Learning\*\*:

- Response format preferences

- Information depth preferences

- Communication style adaptation

- Accessibility needs recognition

### 4.4 Success Criteria for Phase 4

- Proactive suggestions relevance >80%

- User engagement increase >40% with proactive features

- Multi-turn conversation success >90%

- Complex query resolution >85%

- Personalization accuracy >80%

\*\*Deliverable\*\*: Proactive, personalized AI assistant with advanced query understanding capabilities

---

## Phase 5: Human-in-the-Loop Integration & Advanced Testing

\*\*Objective\*\*: Implement comprehensive human oversight system and advanced testing framework

### 5.1 Human-in-the-Loop Architecture

\*\*Live Handoff Flow\*\*:

- Seamless transition from AI to human support

- Complete context transfer including conversation history

- User preference preservation

- Continuity maintenance

\*\*Escalation Triggers\*\*:

- Confidence threshold breaches (<70%)

- Bias detection alerts

- Potential hallucination flags

- User frustration indicators

- Emergency/urgent query detection

- Complex legal/medical questions

\*\*Information Transfer Protocol\*\*:

- Complete conversation history

- User context and preferences

- Attempted solutions and outcomes

- Confidence scores and uncertainty areas

- Detected bias or hallucination concerns

### 5.2 Learning Loop Integration

\*\*Human Feedback Integration\*\*:

- Expert review of AI responses

- Correction and improvement suggestions

- Bias identification and mitigation

- Quality assurance validation

\*\*Continuous Improvement Pipeline\*\*:

- Regular model fine-tuning based on human feedback

- Bias mitigation strategy updates

- Knowledge base refinement

- Response quality enhancement

### 5.3 Advanced Testing Framework

\*\*Bias Testing Suite\*\*:

- Demographic representation testing

- Language pattern bias detection

- Systematic bias scenario testing

- Intersectional bias analysis

\*\*Hallucination Detection Testing\*\*:

- Fabricated information identification

- Source verification accuracy

- Confidence calibration testing

- Uncertainty disclosure validation

\*\*Ethical AI Compliance Testing\*\*:

- Transparency requirement verification

- Fairness metric validation

- Accountability mechanism testing

- Explainability standard compliance

### 5.4 Success Criteria for Phase 5

- Human handoff success rate >95%

- Context transfer accuracy >98%

- Bias detection in testing >95%

- Hallucination detection >90%

- Human feedback integration >90% implementation rate

\*\*Deliverable\*\*: Comprehensive human-AI collaboration system with advanced ethical AI testing

---

## Phase 6: Administration & Ethical Content Management

\*\*Objective\*\*: Comprehensive admin system for ethical content management and AI governance

### 6.1 Enhanced Administration Tools

| \*\*Tool\*\* | \*\*Purpose\*\* | \*\*Features\*\* |

|----------|-------------|--------------|

| Ethical AI Dashboard | Monitor bias, hallucination, and confidence metrics | Real-time bias alerts, hallucination detection, confidence distribution |

| Advanced Admin Panel | Manage content with ethical oversight | Link management, bias scoring, ethical review workflows |

| AI Governance Center | Oversee AI behavior and compliance | Policy enforcement, audit trails, compliance reporting |

| Training Set Builder | Create ethically-balanced training data | Bias detection, demographic representation, quality assurance |

### 6.2 Ethical Content Management System

\*\*Bias-Aware Content Management\*\*:

- Multi-perspective content validation

- Demographic representation analysis

- Alternative viewpoint integration

- Historical bias pattern recognition

\*\*Quality Assurance Workflows\*\*:

- Automated ethical screening

- Human review requirements

- Bias mitigation strategies

- Transparency requirement compliance

### 6.3 AI Governance Framework

\*\*Policy Enforcement\*\*:

- Automated policy compliance checking

- Ethical guideline adherence monitoring

- Bias mitigation requirement enforcement

- Transparency standard maintenance

\*\*Audit and Accountability\*\*:

- Comprehensive audit trails

- Decision explanation capabilities

- Bias incident reporting

- Corrective action tracking

### 6.4 Success Criteria for Phase 6

- Ethical AI compliance >98%

- Bias detection accuracy >95%

- Policy enforcement coverage 100%

- Audit trail completeness >99%

\*\*Deliverable\*\*: Comprehensive ethical AI governance system with advanced content management

---

## Phase 7: Advanced Analytics & Multimodal Accessibility Roadmap

\*\*Objective\*\*: Implement sophisticated analytics and plan for multimodal accessibility features

### 7.1 Advanced Analytics Dashboard

\*\*Ethical AI Metrics\*\*:

- Bias detection rates and patterns

- Hallucination frequency and types

- Confidence score distributions

- Uncertainty disclosure rates

- Human escalation patterns

\*\*Granular KPIs\*\*:

- Response accuracy by demographic group

- Bias incident rates by content type

- Confidence calibration accuracy

- Transparency metric compliance

- Fairness score distributions

### 7.2 Multimodal Accessibility Roadmap

\*\*Phase 1 - Voice Integration\*\*:

- Voice query input capabilities

- Audio response generation

- Voice-based navigation

- Accessibility compliance for visually impaired users

\*\*Phase 2 - Visual Processing\*\*:

- Image intake and analysis

- Document photo processing

- Visual accessibility features

- Screen reader optimization

\*\*Phase 3 - Advanced Multimodal\*\*:

- Video snippet processing

- Real-time transcription

- Multi-language audio support

- Advanced accessibility features

### 7.3 Predictive Analytics

\*\*Trend Analysis\*\*:

- Bias pattern evolution

- User behavior predictions

- Content gap identification

- Proactive resource allocation

\*\*Risk Assessment\*\*:

- Potential bias emergence

- Hallucination risk factors

- User satisfaction predictions

- System performance forecasting

### 7.4 Success Criteria for Phase 7

- Advanced analytics operational with ethical AI metrics

- Multimodal roadmap defined with clear milestones

- Predictive accuracy >80% for user behavior

- Accessibility compliance >WCAG 2.1 AA standard

\*\*Deliverable\*\*: Advanced analytics system with ethical AI monitoring and multimodal accessibility roadmap

---

## Enhanced System Architecture

### Ethical AI Layer

- \*\*Bias Detection Engine\*\*: Real-time bias identification and mitigation

- \*\*Hallucination Prevention\*\*: Confidence-based response filtering

- \*\*Transparency Module\*\*: Uncertainty disclosure and source attribution

- \*\*Fairness Monitor\*\*: Demographic representation and equity analysis

### Advanced Frontend Integration

- \*\*Multimodal Interface\*\*: Voice, text, and visual input capabilities

- \*\*Accessibility Engine\*\*: WCAG 2.1 AA+ compliance with future multimodal support

- \*\*Proactive Notification System\*\*: Location and interest-based suggestions

- \*\*Ethical Transparency Interface\*\*: Confidence scores, bias warnings, uncertainty disclosure

### Enhanced Backend Infrastructure

- \*\*Ethical AI Governance API\*\*: Policy enforcement and compliance monitoring

- \*\*Advanced Analytics Engine\*\*: Sophisticated metrics and predictive capabilities

- \*\*Human-AI Collaboration Platform\*\*: Seamless handoff and learning integration

- \*\*Multimodal Processing Pipeline\*\*: Future-ready architecture for voice and visual processing

---

## Enhanced Success Metrics & KPIs

### Ethical AI Metrics

| \*\*Metric\*\* | \*\*Target\*\* | \*\*Measurement Method\*\* |

|------------|------------|------------------------|

| Bias Detection Accuracy | ≥ 95% | Systematic bias testing |

| Hallucination Detection | ≥ 90% | Fabrication identification testing |

| Confidence Calibration | ≥ 85% | Confidence vs. accuracy correlation |

| Transparency Compliance | 100% | Uncertainty disclosure verification |

| Fairness Score | ≥ 90% | Demographic equity analysis |

### Advanced Performance Indicators

| \*\*Metric\*\* | \*\*Target\*\* | \*\*Measurement Method\*\* |

|------------|------------|------------------------|

| Response Accuracy | ≥ 92% | Manual verification with bias consideration |

| Proactive Suggestion Relevance | ≥ 80% | User engagement with suggestions |

| Multi-turn Conversation Success | ≥ 90% | Context retention accuracy |

| Human Handoff Success | ≥ 95% | Smooth transition verification |

| Accessibility Compliance | WCAG 2.1 AA+ | Automated and manual accessibility testing |

### Business Impact with Ethical Considerations

\*\*Equitable Support\*\*: Ensure equal quality assistance across all demographic groups

\*\*Bias-Free Information Access\*\*: Eliminate discriminatory patterns in information delivery

\*\*Transparent Decision Making\*\*: Clear explanation of AI reasoning and limitations

\*\*Inclusive User Experience\*\*: Accessible design for users with varying abilities

\*\*Ethical AI Leadership\*\*: Demonstrate responsible AI implementation in public service

---

## Enhanced Risk Mitigation

### Ethical AI Risks

\*\*Bias Amplification\*\*: Multi-layer bias detection and mitigation strategies

\*\*Hallucination Risks\*\*: Confidence thresholds and fact-checking requirements

\*\*Transparency Failures\*\*: Mandatory uncertainty disclosure and source attribution

\*\*Fairness Violations\*\*: Continuous demographic equity monitoring

\*\*Accountability Gaps\*\*: Comprehensive audit trails and explainable AI

### Advanced Technical Risks

\*\*Multimodal Complexity\*\*: Phased rollout with thorough testing

\*\*Accessibility Failures\*\*: Continuous compliance monitoring and user feedback

\*\*Proactive Feature Overreach\*\*: User control and preference management

\*\*Human-AI Handoff Issues\*\*: Seamless transition protocols and fallback systems

This enhanced PRD addresses all the critical gaps you identified, including comprehensive ethical AI governance, advanced query understanding, proactive features, detailed human-in-the-loop integration, sophisticated KPIs, and a clear multimodal accessibility roadmap. The system is designed to be not just functional, but ethically responsible and inclusive.