Product Requirements Document (PRD)

AI-Powered Research Agent with RAG Capabilities

Document Information

- **Version**: 1.0
- **Date**: June 1, 2025
- **Author**: [Your Name]
- **Project Code**: RA-2025-001

---

1. Executive Summary

1.1 Product Vision

Develop an intelligent research agent system that combines web-based research capabilities with Retrieval-Augmented Generation (RAG) to produce comprehensive, contextual research reports by leveraging both external web sources and internal document repositories.

1.2 Business Objectives

- **Primary**: Create a portfolio-grade AI agent showcasing multi-agent orchestration, RAG implementation, and enterprise-ready architecture
- **Secondary**: Develop reusable components for future AI consulting engagements
- **Tertiary**: Demonstrate hybrid local/cloud deployment capabilities for diverse client needs

1.3 Success Metrics

- Generate comprehensive research reports in <10 minutes
- Achieve 90%+ source accuracy and relevance
- Support multiple document formats (PDF, DOCX, TXT)
- Demonstrate scalable architecture patterns

---

2. Product Overview

2.1 Product Description

The Research Agent is a multi-agent AI system that automatically conducts comprehensive research on user-specified topics by:

- Searching and analyzing web-based sources
- Processing and querying uploaded document repositories
- Synthesizing findings into structured, professional reports
- Providing source attribution and confidence scoring

2.2 Target Users

- **Primary**: Business analysts requiring market research
- **Secondary**: Content creators needing research-backed articles
- **Tertiary**: Consultants preparing client reports

2.3 Key Differentiators

- Multi-agent collaborative research approach
- Hybrid external/internal knowledge synthesis
- Enterprise-grade vector database integration
- Flexible deployment options (local/cloud)

---

3. Functional Requirements

3.1 Core Features

3.1.1 Research Orchestration

- **FR-001**: System shall accept research topics via REST API or web interface
- **FR-002**: System shall decompose complex topics into focused research tasks
- **FR-003**: System shall coordinate multiple specialized agents for parallel processing
- **FR-004**: System shall provide real-time progress updates during research execution

3.1.2 Web Research Capabilities

- **FR-005**: System shall search multiple web sources using configurable search APIs
- **FR-006**: System shall extract and clean content from web pages
- **FR-007**: System shall validate source credibility and recency
- **FR-008**: System shall handle rate limiting and error recovery

3.1.3 Document Repository Management

- **FR-009**: System shall ingest PDF, DOCX, and TXT documents

- **FR-010**: System shall implement intelligent chunking strategies
- **FR-011**: System shall generate and store document embeddings in vector database
- **FR-012**: System shall support semantic search across document repository

3.1.4 Report Generation

- **FR-013**: System shall synthesize findings into structured reports
- **FR-014**: System shall generate executive summaries with key insights
- **FR-015**: System shall provide source citations and confidence scores
- **FR-016**: System shall export reports in PDF and markdown formats

3.2 Agent Specifications

3.2.1 Document Processing Agent

- Ingests and processes uploaded documents
- Implements chunking and embedding generation
- Manages vector database operations
- Maintains document metadata and relationships

3.2.2 Web Research Agent

- Executes web searches using multiple APIs
- Extracts and cleans web content
- Validates source quality and relevance
- Implements content deduplication

3.2.3 RAG Query Agent

- Performs semantic searches against document repository
- Retrieves relevant context for research topics
- Implements hybrid search (semantic + keyword)
- Manages retrieval result ranking

3.2.4 Analysis Agent

- Synthesizes information from multiple sources
- Identifies patterns and key insights
- Performs cross-source validation

- Generates structured findings

3.2.5 Report Writer Agent

- Creates professional report formatting
- Generates executive summaries
- Implements citation management
- Ensures consistent tone and style

3.2.6 Quality Assurance Agent

- Validates report completeness and accuracy
- Checks source attribution
- Performs final quality review
- Generates confidence scores

---

4. Non-Functional Requirements

4.1 Performance Requirements

- **NFR-001**: Research completion within 10 minutes for standard topics
- **NFR-002**: Support concurrent processing of up to 5 research requests
- **NFR-003**: Document ingestion rate of 100 pages per minute
- **NFR-004**: Vector search response time <2 seconds

4.2 Scalability Requirements

- **NFR-005**: Support document repositories up to 10,000 documents
- **NFR-006**: Handle embedding databases up to 1 million vectors
- **NFR-007**: Horizontal scaling capabilities for agent workers

4.3 Reliability Requirements

- **NFR-008**: System availability of 99% during operation
- **NFR-009**: Graceful degradation when external services are unavailable
- **NFR-010**: Automatic retry mechanisms for failed operations

4.4 Security Requirements

- **NFR-011**: Secure API key management for external services

- **NFR-012**: Document encryption at rest and in transit
- **NFR-013**: User authentication and authorization (Version 2)

---

5. Technical Architecture Overview

5.1 System Architecture Patterns

- **Microservices Architecture**: Loosely coupled agent services
- **Event-Driven Architecture**: Asynchronous communication between agents
- **Layered Architecture**: Clear separation of concerns

5.2 Data Flow Architecture

1. **Ingestion Layer**: Document processing and web content extraction
2. **Processing Layer**: Multi-agent research orchestration
3. **Storage Layer**: Vector database and metadata management
4. **Presentation Layer**: Report generation and user interface

5.3 Integration Architecture

- **API Gateway**: Centralized request routing and management
- **Message Queue**: Asynchronous task processing
- **Service Mesh**: Inter-service communication and monitoring

---

6. Version Roadmap

6.1 Version 1.0 - Local Deployment

- **Scope**: Single-machine deployment for development and demonstration
- **Target**: Local Windows/Mac environments
- **Architecture**: Monolithic with embedded components

6.2 Version 2.0 - Cloud Deployment

- **Scope**: Distributed deployment for production use
- **Target**: VM or Kubernetes cluster environments
- **Architecture**: Microservices with external dependencies

---

7. Constraints and Assumptions

7.1 Technical Constraints

- **TC-001**: Internet connectivity required for web research
- **TC-002**: Minimum 16GB RAM for local vector database operations
- **TC-003**: GPU acceleration optional but recommended

7.2 Business Constraints

- **BC-001**: Initial development budget focused on open-source tools
- **BC-002**: LLM API costs should remain under $50/month for development

7.3 Assumptions

- **AS-001**: Users have basic technical knowledge for local setup
- **AS-002**: Research topics are in English language
- **AS-003**: Document formats are standard business documents

---

8. Dependencies and Risks

8.1 External Dependencies

- OpenRouter API availability and pricing
- Milvus database stability and performance
- Web search API rate limits and costs

8.2 Technical Risks

- **Risk**: Vector database performance degradation with large datasets
- **Mitigation**: Implement database optimization and monitoring

8.3 Business Risks

- **Risk**: LLM API cost escalation
- **Mitigation**: Implement usage monitoring and fallback to local models

---

9. Acceptance Criteria

9.1 Version 1.0 Acceptance

- [ ] Successfully processes research requests end-to-end
- [ ] Generates professional reports with proper citations

- [ ] Demonstrates RAG capabilities with uploaded documents
- [ ] Runs reliably on local development environment

9.2 Version 2.0 Acceptance

- [ ] Deploys successfully to cloud infrastructure
- [ ] Handles multiple concurrent users
- [ ] Demonstrates scalability and monitoring capabilities
- [ ] Includes proper security and authentication measures