AI Solution Architecture

RFP Intelligence & Opportunity Navigator

Technical Architecture Document | Version 1.0

Executive Overview

The RFP Intelligence & Opportunity Navigator is an AI-powered platform that enables government contractors to instantly understand, evaluate, and respond strategically to any RFP. The system automates requirement extraction, compliance assessment, pursuit recommendations, and risk analysis—reducing initial RFP review time from 6-12 hours to under 20 minutes.

The platform addresses a core challenge: RFPs are long (40-200+ pages), jargon-heavy, and structurally inconsistent across agencies. Manual analysis requires semantic understanding, extraction, classification, and synthesis—tasks uniquely suited to modern LLM and RAG architectures.

Architecture Overview

The system implements a **Hybrid RAG + Fine-Tuned Classifier + Multi-Agent Architecture**. A foundation LLM handles semantic understanding and summarization, fine tuned classification models handle structured extraction tasks, and four specialized agents coordinate through an orchestration layer to produce comprehensive RFP analysis.

Core Components:

1. **Document Processing Pipeline:** OCR, parsing, chunking, and embedding of RFP documents
2. **RAG Retrieval System:** Vector database with hybrid search for context-aware extraction
3. **Fine-Tuned Classifiers:** Specialized models for requirement categorization, eligibility detection, and risk scoring
4. **Multi-Agent Orchestration:** Four coordinated agents handling extraction, compliance, pursuit evaluation, and synthesis
5. **Hallucination Controls:** Forced citation, confidence thresholds, and RAG grounding for extraction accuracy

LLM Selection & Configuration

RFP analysis requires models with large context windows (RFPs routinely exceed 100 pages), strong instruction-following for structured extraction, and reliable grounding capabilities to prevent hallucination of requirements.

Use Case	Recommended Model	Rationale
Primary Extraction & Synthesis	Claude 3.5 Sonnet or GPT-4 Turbo	128K-200K context window handles full RFPs; strong structured output; excellent instruction following
High-Volume Queries	GPT-4o-mini or Claude 3 Haiku	Cost-efficient for repeated extraction queries and amendment comparisons

On-Premise Option	Llama 3.1 70B or Mixtral 8x22B	Full data sovereignty for sensitive government contractors; no external API calls

Model Requirements Specification:

- **Input Formats:** PDF, HTML, Word, OCR-processed scanned documents • **Context Window:** Minimum 128K tokens; prefer 200K for full-document analysis without chunking
- **Latency Requirement:** < 8 seconds per major extraction query
- **Temperature:** 0.1-0.2 for extraction tasks (deterministic); 0.4 for synthesis and summarization

RAG Pipeline Architecture

The RAG system grounds all extractions in source document content, enabling citation backed outputs and reducing hallucination risk.

Document Processing Pipeline

1. **Ingestion:** PDF parsing via PyMuPDF; OCR fallback via Tesseract for scanned documents; DOCX via python-docx
2. **Structure Extraction:** Section hierarchy detection (L/M/C sections); table extraction to structured JSON; list parsing
3. **Chunking:** Hierarchical semantic chunking preserving section boundaries (see chunking strategy below)
4. **Embedding:** Vector encoding via embedding model; metadata tagging (section, page, document type)
5. **Indexing:** Storage in vector database with hybrid search index

Embedding & Vector Storage

Component	Specification
Embedding Model	OpenAI text-embedding-3-large (cloud) or InstructorXL (on premise) for domain-specific nuance; 1536-3072 dimensions
Vector Database	Pinecone (managed) or Weaviate (self-hosted) with hybrid search (dense vectors + sparse BM25)
Retrieval Method	Hybrid search (semantic + keyword); top-15 retrieved, reranked to top-5 via cross-encoder
Reranker Model	ms-marco-MiniLM-L-12-v2 or BGE-reranker-large; < 500ms latency budget

Chunking Strategy for RFP Documents

Government RFPs present unique chunking challenges: structural inconsistency across agencies, nested section hierarchies, and mixed narrative/tabular content. The system employs hierarchical semantic chunking:

Content Type	Chunking Approach	Rationale
Section Headings (L, M, C)	Preserve as metadata; chunk content within sections	Enables section-level filtering; maintains hierarchy
Requirements Lists	Each requirement = individual chunk with parent section linkage	Atomic retrieval; full traceability
Narrative Prose	Semantic chunking (600-1000 tokens) with 15% overlap	Captures cross-paragraph context
Tables (CLIN, pricing)	Extract to structured JSON; embed as single chunk with table metadata	Preserves row/column relationships
Attachments (SOW, PWS)	Process separately with cross reference to main RFP	Targeted retrieval from specific attachments

Fine-Tuned Classification Models

Specialized classifiers handle structured extraction tasks where fine-tuning outperforms few shot prompting, providing higher accuracy and faster inference for high-volume operations.

Classifier	Base Model	Training Data	Target Performance
Requirement Type	DeBERTa-v3-base	15,000 labeled requirements	F1 > 0.90 (12 categories)
Mandatory vs. Desirable	RoBERTa-base	8,000 labeled examples	Recall > 0.95 (mandatory)
Eligibility Flags	DistilBERT	5,000 eligibility clauses	Precision > 0.92
Risk Indicator	BERT-base	3,000 risk-tagged clauses	AUC > 0.85

Requirement Taxonomy (12 Categories): Technical, Staffing, Compliance/Certifications, Security, Financial/Pricing, Past Performance, Small Business, Deliverables, Schedule, Quality Assurance, Data Rights, Reporting

Multi-Agent Orchestration

Four specialized agents coordinate through a LangGraph-orchestrated pipeline to produce comprehensive RFP analysis:

Agent Definitions

Agent	Responsibilities
Extraction Agent	Extracts requirements, deadlines, deliverables from RFP; produces structured JSON requirement lists with page citations; identifies mandatory vs. desirable items
Compliance Agent	Flags mandatory certifications (ISO, CMMI, FedRAMP), required forms, security clearances, small business requirements, insurance minimums; cross-references against company capability matrix
Pursuit Evaluation Agent	Scores opportunity fit based on capability alignment, past performance relevance, competitive landscape; generates bid/no-bid recommendation with confidence score and rationale
Synthesis Agent	Compiles outputs into structured JSON and human-readable summaries; generates compliance matrices, risk analysis reports, and executive briefing; produces exportable artifacts

Agent Workflow

1. **Document Ingestion:** RFP uploaded → OCR/parsing → chunks embedded → indexed in vector store
2. **Extraction Phase:** Extraction Agent queries vector store → LLM extracts structured requirements → classifiers tag each requirement

3. **Compliance Phase:** Compliance Agent receives requirement list → cross references company capabilities → flags gaps and mandatory certifications 4. **Evaluation Phase:** Pursuit Agent aggregates outputs → retrieves similar past bids → generates bid/no-bid recommendation

5. **Synthesis Phase:** Synthesis Agent compiles all outputs → generates final deliverables (compliance matrix, executive summary, risk report)

Hallucination Prevention & Output Validation

Accurate requirement extraction is critical—missing a mandatory requirement could disqualify a proposal. The system implements multiple safeguards:

- **Forced Citation:** Every extracted requirement must include source page number and section reference; claims without citations are rejected
- **Confidence Thresholds:** Extractions below 70% confidence flagged with "Needs Human Review" tag; low-confidence items surfaced prominently in UI
- **RAG Grounding:** All extraction queries grounded against retrieved document chunks; LLM cannot generate requirements not present in source material • **Structured Output Enforcement:** JSON schema validation ensures all outputs conform to expected structure; malformed outputs trigger regeneration
- **Recall Bias:** Model tuned to favor recall over precision—better to surface a potential requirement for human review than miss it entirely

- **OCR Quality Alerts:** Pages with OCR confidence < 85% trigger warning; user prompted to upload cleaner version or manually verify extracted content

System Outputs

- **JSON Requirement Lists:** Structured data for integration with proposal management tools
- **Compliance Matrices:** Exportable spreadsheets mapping requirements to company capabilities and compliance status
- **Bid/No-Bid Recommendation Brief:** Executive summary with recommendation, confidence score, key factors, and risk assessment
- **Competitor Landscape:** Analysis of likely competitors based on past awards and incumbent information
- **Risk Analysis Report:** Detailed breakdown of compliance risks, capability gaps, and mitigation recommendations

Infrastructure & Monitoring

Requirement	**Specification**
End-to-End Latency	< 3 minutes for full RFP analysis (50-100 pages); < 8 seconds per individual query
Deployment	AWS/Azure cloud (SOC 2 compliant) for SaaS; private cloud/on premise option for sensitive contractors
Token Optimization	Efficient chunking strategy; cached retrieval for repeated queries; prompt compression for cost control
Monitoring	Real-time alerts for extraction failures; drift monitoring on requirement vocabulary; token usage tracking; accuracy metrics vs. human review
Retraining Cadence	Monthly evaluation of classifier performance; quarterly re indexing of embeddings to capture agency language shifts; continuous prompt refinement based on user corrections

[End of AI Solution Architecture Document]