# AI Solution Architecture

This section details the technical AI architecture for the Document Drafting & Review Copilot, including LLM selection, retrieval-augmented generation for firm knowledge bases, clause extraction methodology, and security-first design principles.

## Architecture Overview

The Copilot implements a **RAG-Augmented LLM Architecture** with specialized fine-tuned classifiers for clause extraction and risk detection. The system combines a foundation model for drafting and summarization with retrieval over firm-specific templates, clause libraries, and legal precedent databases.

**Core Components:**

1. **Foundation LLM:** Primary generation model for drafting, summarization, and conversational Q&A
2. **RAG Pipeline:** Retrieval over firm templates, clause library, and legal research databases
3. **Clause Classifier:** Fine-tuned model for extracting and categorizing contract clauses
4. **Risk Scoring Engine:** Classification model identifying problematic language and missing protections
5. **Document Processing Pipeline:** OCR, parsing, and chunking for ingested legal documents

## LLM Selection & Deployment Options

Legal document handling requires models with strong instruction-following, long-context capability, and deployment options that satisfy attorney-client privilege requirements.

| Deployment | Recommended Model | Rationale |
|---|---|---|
| **Cloud (Enterprise)** | GPT-4 Turbo via Azure OpenAI (data processing opt-out enabled) | 128K context for full contract analysis; Azure data residency controls; SOC 2 Type II; no training on customer data |
| **Private Cloud** | Claude 3.5 Sonnet via AWS Bedrock (VPC deployment) | 200K context; no data retention; VPC keeps data within firm's AWS environment |
| **On-Premise** | Llama 3.1 70B (quantized) or Mixtral 8x22B | Full data sovereignty; no external API calls; runs on firm infrastructure (requires 2x A100 GPUs or equivalent) |

**Temperature Settings:** Drafting tasks use temperature 0.3-0.5 for controlled creativity with firm-style adherence; extraction and classification tasks use temperature 0.1-0.2 for deterministic outputs; research summarization uses 0.4 for balanced synthesis.

## RAG Architecture for Legal Knowledge

### Document Corpus Structure

The RAG system maintains separate vector collections for different knowledge types, enabling targeted retrieval based on task context:

| Collection | Contents | Use Case |
|---|---|---|
| Firm Templates | Approved motion templates, contract boilerplate, letter formats by practice area | Drafting with firm-specific style and structure |
| Clause Library | Extracted clauses tagged as preferred, acceptable, prohibited; organized by type | Clause suggestion, risk comparison, alternative generation |
| Prior Work Product | Historical motions, briefs, memos (redacted/anonymized where required) | Style matching, argument patterns, precedent identification |
| Legal Research | Case law, statutes, regulations (via Westlaw/LexisNexis API or firm library) | Research summarization, citation support, applicability analysis |

### Embedding & Retrieval Configuration

- **Embedding Model:** text-embedding-3-large (OpenAI) or Legal-BERT embeddings for on-premise; 1536+ dimensions
- **Vector Database:** Pinecone (cloud) or Qdrant (self-hosted) with metadata filtering by practice area, document type, and date
- **Chunking Strategy:** Legal-aware chunking preserving clause boundaries (not mid-sentence splits); 800-1200 tokens per chunk; hierarchical chunking for contracts (section → subsection → clause)
- **Retrieval:** Hybrid search (semantic + keyword BM25) with cross-encoder reranking; top-k=15 retrieved, reranked to top-5 for context injection

## Clause Extraction & Classification

A dedicated fine-tuned model handles clause identification and categorization, operating independently from the generation LLM for accuracy and auditability.

### Model Architecture

- **Base Model:** DeBERTa-v3-large fine-tuned on CUAD dataset (Contract Understanding Atticus Dataset) plus firm-specific labeled examples
- **Task:** Multi-label sequence classification identifying clause boundaries and types (41 standard categories + firm-custom categories)
- **Training Data:** CUAD (510 contracts, 13,000+ annotations) augmented with 500+ firm-labeled documents
- **Performance Target:** >92% F1 score on clause boundary detection; >88% accuracy on clause type classification

### Clause Categories

Standard categories include: Indemnification, Limitation of Liability, Termination, Confidentiality, IP Assignment, Non-Compete, Governing Law, Dispute Resolution, Force Majeure, Change of Control, Warranty, Representations, Payment Terms, Insurance Requirements, and Data Protection.

## Risk Detection & Flagging

The risk engine combines rule-based checks with LLM-powered analysis to identify problematic language, missing protections, and deviation from firm standards.

1. **Rule-Based Checks:** Regex patterns for known red-flag phrases ("unlimited liability," "sole discretion," "waives all claims"); missing clause detection against checklist

2. **Semantic Risk Scoring:** LLM analyzes extracted clauses against firm-preferred alternatives; scores deviation severity (Low/Medium/High)
3. **Comparative Analysis:** Vector similarity between document clauses and clause library; flags clauses with low similarity to any "preferred" examples
4. **Output:** Risk summary with clause-level annotations, suggested alternatives from library, and confidence scores

## Document Ingestion Pipeline

- **Format Support:** DOCX (native parsing via python-docx), PDF (PyMuPDF + OCR fallback via Tesseract for scanned documents)
- **Structure Extraction:** Heading hierarchy detection, table parsing, list identification; preserves document outline for navigation
- **Metadata Tagging:** Auto-classification of document type (motion, contract, letter, memo); party extraction; date identification
- **Version Handling:** Track Changes preserved; redline comparison uses diff algorithm with semantic grouping of changes

## Security & Confidentiality Architecture

| Control | Implementation |
|---|---|
| Data Encryption | AES-256 at rest; TLS 1.3 in transit; client-side encryption option for highly sensitive matters |
| Access Control | Role-based (Partner, Associate, Paralegal); matter-level permissions; SSO integration (Okta, Azure AD) |
| No Training on Data | Zero data retention agreements with LLM providers; opt-out of model improvement; on-premise option eliminates external transmission |
| Audit Logging | Complete interaction history: queries, retrieved context, generated outputs, user edits; immutable logs for ethics compliance |
| Ethical Walls | Matter-based data isolation; conflict-flagged matters excluded from cross-matter retrieval |

*[End of AI Solution Architecture Section]*