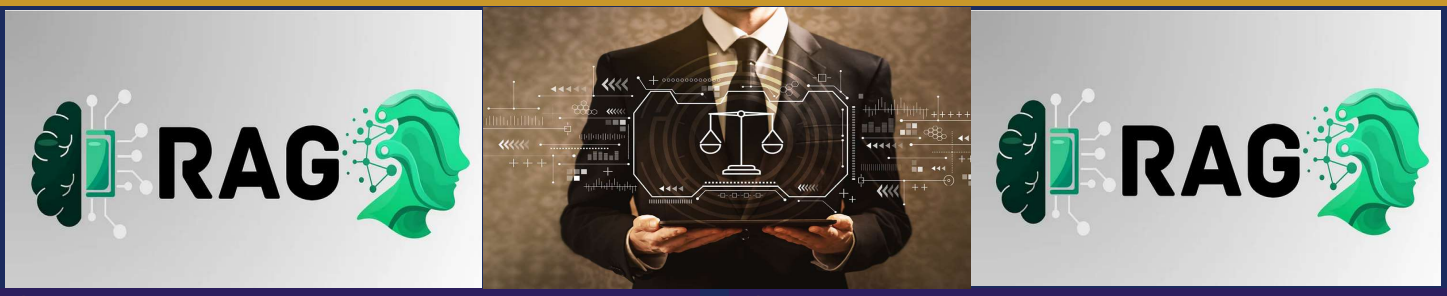
# LegalMind AI

## Intelligent Legal Document Processing and Retrieval Using Retrieval-Augmented Generation (RAG)

Manoj Kumar Konda , Agnes Maciocha | Department of Computing , South East Technological University Carlow, Ireland




## 1 Introduction

- Legal professionals spend up to 60% of their working time on document review and information retrieval (McKinsey, 2023). Keyword-based search — used by 85% of litigators (ABA, 2024) — fails to find semantically equivalent legal terms.
- Commercial platforms (Westlaw, LexisNexis) are expensive closed ecosystems that cannot process a firm's own private documents.
- General-purpose LLMs hallucinate at 58–88% rates on legal queries (Dahl et al., 2024) — extremely dangerous in high-stakes legal work.
- LegalMind AI solves this using RAG: every answer is grounded exclusively in the firm's own uploaded documents with verifiable source citations, eliminating hallucination risk.

**Figure 1. LegalMind AI — System Architecture: RAG Pipeline Block Diagram**

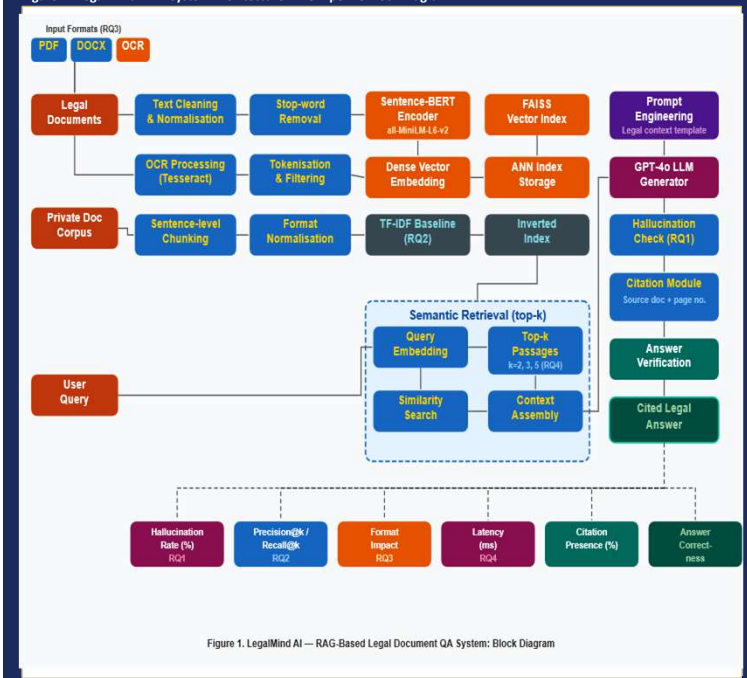



Figure 1. LegalMind AI — RAG-Based Legal Document QA System: Block Diagram

## 2 Literature Review

- Lewis et al. (2020): Foundational RAG — open-domain Wikipedia only; no legal domain, no private doc ingestion.
- Pipitone & Alami (2024) — LegalBench-RAG: retrieval stage only on public datasets; no end-to-end generation or citation quality.
- Wiratunga et al. (2024) — CBR-RAG: pre-built Australian case-law; no user-uploaded or multi-format document support.
- Dahl et al. (2024): Documents 58–88% LLM hallucination in law — identifies the problem; provides no working solution.
- Hindi et al. (2025) — IEEE Access RAG survey: public corpora only; no empirical keyword vs semantic comparison.

## 3 Research Objectives

This research evaluates Retrieval-Augmented Generation (RAG) for answering queries over privately held, unstructured legal documents — comparing it against standalone LLMs and keyword-based retrieval baselines. The system ingests PDF, DOCX, and OCR-scanned documents, embeds them using Sentence-BERT, retrieves relevant passages via FAISS, and generates grounded, cited answers using GPT-4o — across four empirical research questions (RQ1–RQ4).

## 4 The Data

Data consists of privately held legal documents from a corporate legal department — contracts, regulatory filings, and compliance reports. A 20-query annotated test set with human-verified ground-truth answers is used for evaluation.

| | | | | | |
|---|---|---|---|---|---|
| **Data Type:** | Legal Text Docs | **No. Documents:** | **10** | **Domain:** | Corporate Legal |
| **File Formats:** | PDF · DOCX · OCR | **Test Queries:** | 20 annotated | **Top-k Values:** | k = 1, 3, 5 |
| **Evaluation Metrics:** | Precision@k, Recall@k | **Hallucination:** | Measured (%) | **Latency:** | Response (ms) |

## 4 Research Questions

| | |
|---|---|
| **RQ1** | Does RAG significantly reduce hallucination rate vs standalone LLM on private legal documents? |
| **RQ2** | Does FAISS semantic retrieval outperform TF-IDF keyword baseline in Precision@k and Recall@k? |
| **RQ3** | Does document format (PDF vs DOCX vs OCR-scanned) affect retrieval completeness? |
| **RQ4** | How does varying top-k (k=1,3,5) affect answer quality vs response latency trade-off? |

## 5 Methodology

- Mixed-methods: empirical system evaluation + controlled quantitative comparison across RQ1–RQ4.
- Corpus: 10 privately held legal documents (native PDF, DOCX, OCR-scanned via Tesseract).
- Embeddings: Sentence-BERT (all-MiniLM-L6-v2) in FAISS flat-index for semantic retrieval; TF-IDF baseline for RQ2.
- Test set: 20 annotated legal queries, ground-truth answers, scored via RAGAS + human annotation.
- Metrics: Precision@k, Recall@k, Hallucination Rate (%), Citation Presence (%), Latency (ms).

## 5 Early Indicators

- Literature review completed — 5 core + 5 ACM papers reviewed, research gap identified.
- RAG architecture designed; technology stack selected; document ingestion module in progress.

## 6 Next Steps

- Complete document ingestion pipeline — native PDF, DOCX, and OCR via Tesseract.
- Implement FAISS semantic retrieval + TF-IDF keyword baseline for RQ2 comparison.
- Run all 4 experiments (RQ1–RQ4), collect all metrics and analyse results.
- Submit paper to IEEE Access (rolling) or NLLP Workshop @ EMNLP 2026.

LangChain · Hugging Face · ChatGPT-4o · scikit learn · ragas · jupyter · Streamlit · GitHub

**Contact: Manoj Kumar Konda**
Tel:0873763952
Email: C00317351@setu.ie

**References:** 1. Lewis et al. (2020). RAG for NLP. NeurIPS. 2. Dahl et al. (2024). Large Legal Fictions. J. Legal Analysis. 3. Pipitone & Alami (2024). LegalBench-RAG. arXiv. 4. Wiratunga et al. (2024). CBR-RAG. ICCBR 2024. 5. Hindi et al. (2025). RAG in Legal Tech. IEEE Access.