IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155143

A Multi-Agent Retrieval-Augmented Generation Framework for Context-Aware Legal Document Analysis

Dr. C N Shariff[1], Aaftab Zohra[2], K Sowmya[3], K Rakshitha[4], Aishwarya G[5]

Professor, Department of Artificial Intelligence and Machine Learning,

Ballari Institute of Technology and Management, Ballari, India[1]

Department of Artificial Intelligence and Machine Learning,

Ballari Institute of Technology and Management, Ballari, India[2-5]

Abstract: Legal document analysis requires high accuracy, traceability, and semantic understanding. While large language models (LLMs) provide strong generative capabilities, they suffer from hallucinations and lack of grounding in authoritative sources. This paper presents a Multi-Agent Retrieval-Augmented Generation (RAG) framework for legal document analysis. The system integrates semantic retrieval, vector embeddings, and collaborative agent-based reasoning to produce context-aware legal responses. A modular architecture consisting of retrieval, summarization, precedent discovery, and fact-checking agents aims to improve reliability and explainability. The framework is designed for scalable enterprise deployment and evaluated using grounding-based and qualitative evaluation metrics.

Index Terms: Retrieval-Augmented Generation, Legal NLP, Multi-Agent Systems, Vector Databases

I. INTRODUCTION

Recent advances in large language models have significantly improved natural language processing tasks. However, deploying these models in legal environments remains challenging due to hallucinations, lack of interpretability, and absence of verifiable grounding.

Legal professionals rely on accurate interpretation of lengthy documents including case laws and contracts. Traditional keyword search systems are insufficient for semantic reasoning. Retrieval-Augmented Generation (RAG) combines retrieval with generation to ground responses in external knowledge sources.

This paper proposes a multi-agent RAG framework that de-composes legal reasoning into specialized agents. The system is designed to improve traceability and reduce hallucination.
The main contributions are:

- A collaborative multi-agent RAG architecture
- Integration of semantic retrieval and precedent discovery
- A scalable legal document processing pipeline
- An evaluation framework for legal AI assistants

II. BACKGROUND AND RELATED WORK

Retrieval-Augmented Generation (RAG) was introduced by Lewis et al. [1], combining dense document retrieval with sequence-to-sequence generation for knowledge-intensive NLP tasks. Subsequent advances in large language models such as GPT-3 [2] and transformer architectures [3] significantly improved generative reasoning capabilities.

Transformer-based representation models including BERT [4] and Sentence-BERT [5] enabled efficient semantic embedding and similarity search. These embedding models form the foundation of modern vector retrieval systems.
Recent research explores multi-agent reasoning architectures where specialized agents collaborate to solve complex tasks [6]. Such systems improve interpretability and modular reasoning through task decomposition.

Vector databases such as ChromaDB support scalable embedding storage and fast similarity retrieval [7]. Evaluation frameworks for embedding models, such as the Massive Text Embedding Benchmark (MTEB) [8], provide

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155143

standardized evaluation metrics.

In the legal domain, AI-assisted document analysis has gained attention. LegalBench [9] evaluates reasoning capabilities of language models in legal tasks, while recent work emphasizes grounding and retrieval accuracy in legal NLP systems [10]. Despite these advances, few systems integrate multi-agent reasoning with retrieval-augmented pipelines for legal applications.

III. PROBLEM DEFINITION

Key challenges include:
- Complexity of legal document interpretation
- Hallucination in generative models
- High computational cost
- Lack of explainability

The goal is to design a grounded, explainable legal assistant.

IV. PROPOSED MULTI-AGENT RAG FRAMEWORK

The system combines document ingestion, semantic indexing, retrieval, and agent collaboration.

A. System Architecture

Figure 1 shows the system architecture.

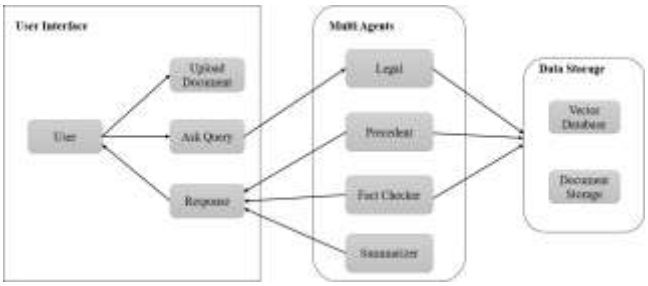

Fig. 1. Overview of the proposed multi-agent RAG architecture for legal document analysis

The architecture includes user interface, multi-agent reason-ing, and storage layers.

B. Document Processing Pipeline

Legal documents are extracted using PyMuPDF and seg-mented into overlapping chunks.

C. Embedding and Retrieval

Chunks are embedded using MiniLM and stored in a vector database. Queries retrieve top-*k* relevant segments that are subsequently validated through grounding evaluation.

D. Multi-Agent Collaboration

Agents include:
- Retrieval Agent
- Summarization Agent
- Precedent Agent
- Fact-Checking Agent

V. METHODOLOGY

The proposed framework follows a structured Retrieval-Augmented Generation (RAG) pipeline enhanced by multi-agent collaboration. The methodology consists of document preprocessing, semantic embedding, vector retrieval, and coordinated agent reasoning.

A. Document Preprocessing and Chunking

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155143

Uploaded legal documents are processed using PyMuPDF for text extraction. The extracted text is normalized and segmented into overlapping chunks of approximately 400–600 tokens with 15–20% overlap. Overlapping segmentation pre-serves contextual continuity and reduces semantic fragmenta-tion during retrieval.

B. Embedding Generation

Each text chunk is converted into a dense semantic embed-ding using MiniLM-L6-v2 [11], producing vector represen-tations used for semantic retrieval. The model maps textual content into a 384-dimensional vector space optimized for semantic similarity. Cosine similarity is used as the distance metric for retrieval.

C. Vector Indexing and Retrieval

Embeddings are stored in ChromaDB, a persistent vector database supporting efficient similarity search. During query processing, the user query is embedded and compared against stored vectors. The system retrieves the top-k relevant chunks (typically $k = 5$) as contextual evidence for reasoning.

D. Agent Roles

The retrieved evidence is processed by four specialized agents:

- **Retrieval Agent:** Filters and ranks retrieved chunks based on semantic relevance.
- **Summarization Agent:** Generates structured summaries highlighting key legal clauses.
- **Precedent Agent:** Identifies related case laws using external legal databases.
- **Fact-Checking Agent:** Verifies that generated responses are grounded in retrieved evidence.

Algorithm 1 Multi-Agent RAG Workflow

```
Extract and preprocess document text
Generate embeddings and store in vector database
Embed user query and retrieve top-k chunks
Apply multi-agent reasoning pipeline
Generate grounded final response
```

This workflow ensures that generation is continuously constrained by retrieved evidence throughout the reasoning process.

E. Implementation Environment

The system was implemented using Python with Flask for the web interface, LangChain for agent orchestration, and ChromaDB for vector storage. Experiments were conducted on a workstation with an Intel Core i5 processor and 8GB RAM.

TABLE I
GROUNDING CONFIDENCE SCORES

Query Category	Average Confidence Score
Clause Interpretation	0.82
Case Summary	0.78
Precedent Retrieval	0.85

VI. EXPERIMENTAL DESIGN

The experimental setup evaluates the effectiveness of the proposed multi-agent RAG framework in comparison to base-line systems.

A. Dataset

The evaluation dataset consists of approximately 160 cu-rated legal documents, including court judgments and contracts spanning multiple legal domains. A benchmark query set of 50 legal questions was constructed to evaluate system performance.

B. Baseline Systems

The proposed system is compared against:

- **LLM-only baseline:** Direct generation without retrieval
- **Single-agent RAG:** Standard retrieval-augmented pipeline

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155143

- **Multi-agent RAG:** Proposed collaborative architecture

C. *Evaluation Metrics*

Performance is measured using:

- Grounding confidence score from the fact-checking agent
- Evidence coverage ratio for response traceability
- Latency and token consumption for efficiency
- Human evaluation of correctness and usefulness

All experiments are conducted under consistent computa-tional settings to ensure fair comparison.

VII. RESULTS AND ANALYSIS

Instead of relying solely on traditional information retrieval metrics, the evaluation focuses on grounding quality and response reliability produced by the multi-agent framework. This evaluation framework is designed to measure how effec-tively the system generates responses that are supported by retrievable legal evidence.

A. *Grounding Confidence Evaluation*

The fact-checking agent assigns a confidence score to each generated response based on alignment with retrieved evidence. The score ranges from 0 to 1, where higher values indicate stronger grounding in source documents. The score is computed as the ratio of grounded statements to total generated statements, determined by alignment with retrieved evidence:

$$\text{Confidence} = \frac{\text{Grounded Statements}}{\text{Total Statements}}$$

Confidence scores represent the proportion of generated statements supported by retrieved evidence. Higher scores indicate stronger grounding and reduced hallucination. These results indicate that the multi-agent architecture maintains strong grounding across diverse legal tasks, particularly in precedent retrieval where citation consistency is highest.

B. *Evidence Coverage Analysis*

Evidence coverage measures the proportion of generated responses directly supported by retrieved document chunks. Higher coverage indicates reduced hallucination and stronger traceability.

Qualitative inspection shows that responses generated by the multi-agent pipeline consistently reference retrieved le-gal clauses and precedents. This behavior indicates that the retrieval component successfully supplies relevant contextual evidence to downstream agents.

C. *Human Evaluation*

A group of five evaluators rated system responses on clarity, relevance, and factual grounding using a 5-point Likert scale. Preliminary feedback indicates improved interpretability compared to single-agent systems.

VIII. DISCUSSION

Overall, the evaluation highlights the practical benefits of agent-based decomposition in legal reasoning workflows, and the results suggest that multi-agent collaboration enhances modularity and interpretability in legal reasoning tasks. Specialized agents allow task decomposition, improving both retrieval accuracy and response grounding.

While the multi-agent architecture introduces additional computational overhead, the benefits in explainability and reliability outweigh the cost in enterprise scenarios. The modular design also supports scalability and future extension.

IX. LIMITATIONS AND FUTURE WORK

The current framework depends on the quality of embed-dings and the availability of legal datasets. External precedent retrieval may be affected by API limitations. Future work includes expanding datasets, optimizing agent communication, and integrating domain-specific fine-tuned models.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155143

X. CONCLUSION

This paper presents a multi-agent RAG framework tailored for legal document analysis. By integrating semantic retrieval and collaborative reasoning, the system improves grounding, interpretability, and enterprise readiness. Future work will focus on large-scale evaluation and optimization for real-world deployment.

REFERENCES

[1]. P. Lewis et al., "Retrieval-Augmented eneration for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.
[2]. T. Brown et al., "Language Models are Few-Shot Learners," NeurIPS, 2020.
[3]. A. Vaswani et al., "Attention is All You Need," NeurIPS, 2017.
[4]. J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transform-ers," NAACL, 2019.
[5]. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT Networks," EMNLP, 2019.
[6]. Y. Wang et al., "Multi-Agent Retrieval-Augmented Generation," ACL, 2024.
[7]. Chroma Research Team, "Chroma: The AI-native Open-source Embed-ding Database," 2023.
[8]. N. Muennighoff et al., "MTEB: Massive Text Embedding Benchmark," arXiv:2210.07316, 2023.
[9]. L. Guha et al., "LegalBench: A Benchmark for Measuring Legal Reasoning," NeurIPS Datasets and Benchmarks, 2023.
[10]. S. Pipitone and G. Houir Alami, "LegalBench-RAG: Bench-marking Retrieval-Augmented Generation for Legal Reasoning," arXiv:2408.10343, 2024.
[11]. W. Wang et al., "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression," NeurIPS, 2020.
[12]. C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," JMLR, 2020.
[13]. A. Radford et al., "Language Models are Unsupervised Multitask Learners," OpenAI Technical Report, 2019.
[14]. V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Ques-tion Answering," EMNLP, 2020.
[15]. R. Thoppilan et al., "LaMDA: Language Models for Dialog Applica-tions," arXiv:2201.08239, 2022.

