Volume 10, Issue 11, November – 2025 International Journal of Innovative Science and Research Technology
ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25nov1316

A Conceptual Framework for Precedent-Aware Retrieval-Augmented Generation in Case Law Analysis

Shatrunjay Kumar Singh[1]

Publication Date: 2025/12/01

Abstract: The external knowledge-based architecture of Retrieval-Augmented Generation (RAG) systems demonstrates strong potential for legal informatics through their ability to connect Large Language Models (LLMs) to external information. The typical design of RAG systems fails to match the common law system because they focus on semantic matching instead of following the legal principles of *stare decisis* and court organization and authority strength. The paper develops a conceptual framework for Precedent-Aware RAG (PA-RAG) which aims to connect these two systems. The system includes two main components: a precedent-aware retriever that uses jurisdictional authority and temporal recency and citation network centrality to rank cases and a legal-reasoning generator that creates structured outputs which can be verified. The research establishes a complete system design and develops specific evaluation criteria for legal applications to direct future system development. The paper evaluates the moral concerns surrounding these systems before presenting a plan for their deployment and testing process to create dependable AI-based legal research tools.

Keywords: *Precedent-Aware RAG, Legal Information Retrieval, Case Law Analysis, Stare Decisis, Knowledge Graphs in Law, Legal Artificial Intelligence, Retrieval-Augmented Generation, Large Language Models, Conceptual Framework, AI for Legal Research*.

How to Cite: Shatrunjay Kumar Singh (2025). A Conceptual Framework for Precedent-Aware Retrieval-Augmented Generation in Case Law Analysis. *International Journal of Innovative Science and Research Technology*, 10(11), 2099-2106. https://doi.org/10.38124/ijisrt/25nov1316

I. INTRODUCTION

Artificial Intelligence integration into legal practice needs systems which fulfill the strict evidence and reasoning requirements of the field. The paper demonstrates that RAG systems need complete reconstruction to fulfill legal requirements because they need to handle common law structures and doctrines directly (Barnett, 2024) (Hitt, 2016). A legal system needs to understand precedents to achieve reliability because it must apply stare decisis principles and identify between binding and persuasive cases and maintain correct jurisdictional order. The research creates a conceptual base for this transformation through the development of the Precedent-Aware RAG (PA-RAG) framework (Abdul-Azeez, 2024). The main contribution of this research involves creating a new system which properly evaluates legal authority to enable the most influential precedents to direct the generation process. The system achieves this through three essential design elements which include a precedent-aware retriever that uses legal knowledge graphs and detailed metadata to perform authority-based case re-ranking instead of semantic similarity and a legal-reasoning generator that produces structured verifiable outputs and a legal-specific evaluation metric taxonomy for future empirical validation. The research establishes a complete framework to direct legal AI development toward producing reliable and court-admissible results (Barnett, 2024).

The paper follows this structure: Section 2 establishes the required knowledge about legal precedents and standard RAG restrictions. Section 3 explains the fundamental elements and design structure of PA-RAG framework. Section 4 examines essential implementation requirements and technical obstacles that need resolution. Section 5 examines both ethical aspects and performance boundaries of the system. The final section of Section 6 presents a summary and outlines potential research paths for upcoming studies.

II. BACKGROUND: THE ANATOMY OF LEGAL PRECEDENT

➢ *The Doctrine of Stare Decisis*

The common law tradition depends on stare decisis as its fundamental principle because it establishes judicial decision-making that produces consistent and predictable results with fair outcomes. The system operates through a fundamental difference between cases that courts must follow and those they can use for guidance. A lower court within the same jurisdiction must adhere to binding precedents which serve as mandatory authorities (Magesh, 2024) (Jacob, 2014). The legal system enforces these rules through vertical application. A persuasive precedent lacks binding power because it includes decisions from other jurisdictions and dissenting opinions and scholarly work (Landes, 1976). The doctrine exists in two main aspects which include vertical stare decisis that requires lower courts to follow decisions

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25nov1316

from higher courts in their judicial system and horizontal stare decisis that requires courts to follow their previous decisions but allows them to reverse those decisions (De Brabandere, 2016) (Hitt, 2016).

➢ *The Architecture of Court Hierarchy*

The binding force of a judicial decision is intrinsically linked to its position within a pyramidal court structure, as illustrated in Table 1.

Table 1 Hierarchy of Legal Authority in the U.S. Federal System.

Court Level	Example	Binding On	Persuasive On
Supreme Court	U.S. Supreme Court (SCOTUS)	All lower federal and state courts on federal matters	-
Circuit Court	U.S. Court of Appeals for the 2nd Circuit	District courts within the 2nd Circuit	Other Circuit Courts
District Court	U.S. District Court for the S.D.N.Y.	None (establishes law of the case)	Other District Courts

➢ *The Language of Citations: Metadata as Legal Signal*

Legal citations operate as a complex metadata framework which stores essential details about case origins and their authoritative status. A typical citation (Brown v. Board of Education, 347 U.S. 483 (1954)) provides instant access to the case name and its publication details including the United States Reports volume and page number and decision year (Dong, 2023).

- *Standard RAG and Its Jurisprudential Shortcomings*

The retrieval-Augmented Generation (RAG) system improves LLMs through its ability to fetch suitable text sections from outside databases. The basic RAG system design shown in Figure 1 provides two main benefits to users which include automatic knowledge base updates and reduced processing requirements. The exclusive use of semantic similarity for legal analysis produces essential barriers that restrict its analytical capabilities. The legal field converts typical RAG system weaknesses into major operational breakdowns because it fails to identify between binding and persuasive authority (Duxbury, 2008) (Guillaume, 2011).

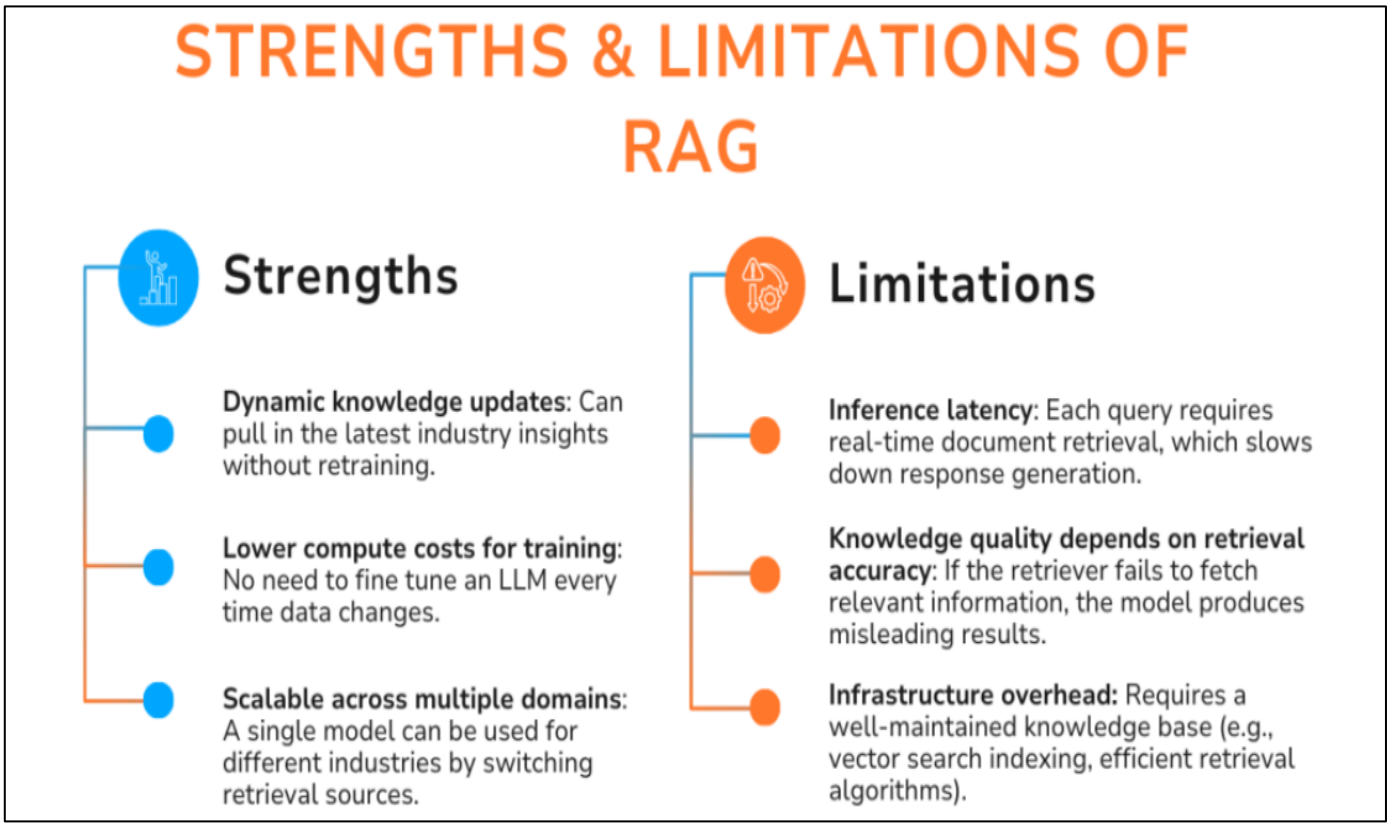

Fig 1 Strength and Limitations of RAG (Teneva, 2025).

The retrieval-Augmented Generation (RAG) system improves LLMs through its ability to retrieve suitable text sections from outside databases which guide its output production. The system uses dense vector embeddings to transform user queries and document chunks before executing a high-dimensional similarity search (e.g. through cosine similarity) (Guillaume, 2011). The LLM uses the most relevant *k* chunks from the search results to create its final answer. The current practice of using semantic similarity as the only method for legal analysis proves ineffective in most legal contexts. The legal system requires a different approach than the "bag-of-words" method because it operates through

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25nov1316

a "chain-of-authority." A standard RAG system would find a semantically appropriate passage from a dissenting opinion or an overturned lower court ruling while ignoring the essential binding precedent from a superior court that uses different wording (Duxbury, 2008). The system performs retrieval based on content without considering the author's identity or the time period or the obligation for others to follow their statements. The RAG paradigm requires fundamental changes to implement the jurisprudential structure which was previously described.

III. THE PRECEDENT-AWARE RAG (PA-RAG) FRAMEWORK

The Dynamic Legal RAG workflow in Figure 2 shows how legal-domain components including Legal Entity Recognition (LER) and specialized knowledge bases integrate into existing systems but these systems do not have a systematic approach to manage precedent doctrine. The framework fills this essential deficiency by incorporating fundamental legal reasoning principles into both retrieval operations and generation functions (Bench-Capon, 2005).

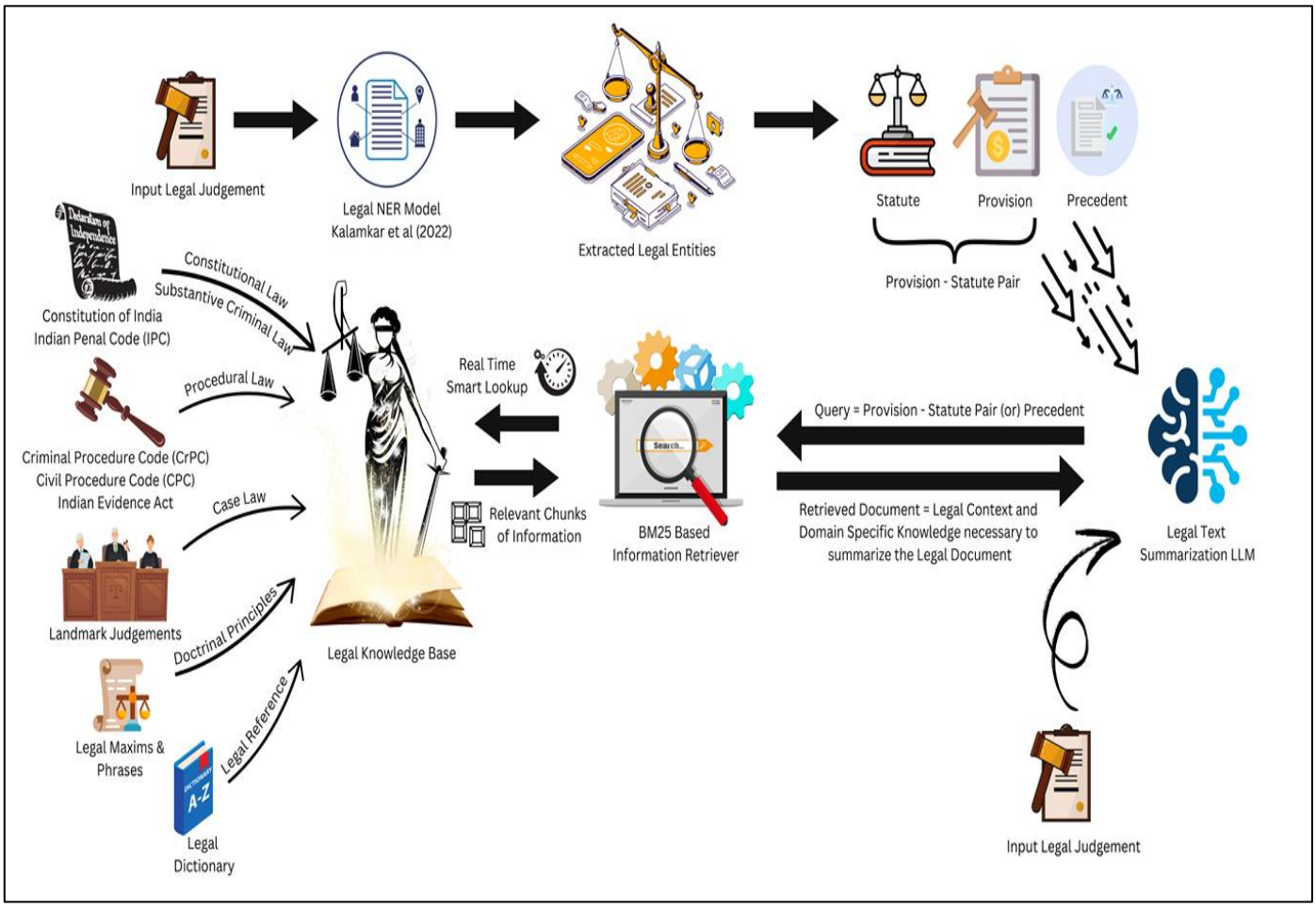

Fig 2 Dynamic Legal RAG Workflow (Ajay Mukund, 2025).

This architecture demonstrates core components for legal RAG which PA-RAG builds upon through its addition of precedent-awareness to statute-provision pairing and legal knowledge bases (Hinkle, 2015).

➢ *Core Principles*

The PA-RAG framework operates through four essential principles which form the basis of common law systems. The system uses these principles as built-in components which affect document representation and retrieval and synthesis operations.

- *Authority-Weighting:*

The legal system recognizes that judicial opinions carry different levels of legal significance. A PA-RAG system needs to move past semantic similarity evaluation because it should give greater importance to court decisions from higher courts that belong to a specific legal hierarchy. The system needs to include court metadata information to establish binding precedent priority over persuasive authority so it can generate responses based on the most legally significant sources (Hitt, 2016).

- *Temporal Reasoning:*

The law is not static. A precedent's validity depends on its non-overturn by a decision that occurred after it. A PA-RAG system requires a temporal model which detects the order of judicial decisions regarding particular legal matters. The system uses this capability to select current decisions that control the case while preventing the use of outdated or superseded legal precedents which standard RAG systems would otherwise depend on (Atkinson, 2005).

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25nov1316

- *Jurisdictional Scoping:*

The geographical area and subject-matter jurisdiction of the court that made the decision determines the relevance of each case. The PA-RAG framework needs to understand its environment so it can limit its data retrieval and analysis to the correct jurisdiction based on user input or the current legal matter. The system should select California state appellate decisions first when users search for California state law even though other states' decisions might be more semantically relevant (Hafner, 2002).

- *Citation Centrality:*

The significance of legal decisions becomes evident through their subsequent use by other courts. A legal decision becomes more important to the legal system when other jurists frequently use it in their decisions. PA-RAG uses citation network analysis to achieve this functionality. The system uses PageRank and custom legal centrality scores to detect fundamental legal cases which establish core principles of jurisprudence even when their text does not match the query best semantically (Al-Abdulkarim, 2016).

➢ *System Architecture:*

The system architecture of the Precedent-Aware RAG (PA-RAG) framework implements authority-weighting and temporal reasoning and jurisdictional scoping through a unified processing pipeline. The system architecture advances beyond basic retriever-generator functionality to implement an advanced re-ranking system which duplicates legal decision-making processes (Duxbury, 2008). The design foundation stems from current developments in preference-aware training which Figure 3 shows how generators learn to identify valuable information from unimportant data in a way that matches precedent identification from binding versus non-binding authorities (Hitt, 2016) (Atkinson, 2005). The main training goal becomes evident through this example because it shows how to move from a basic LLM that handles all documents equally to a complete generator system trained through PA-RAG which generates high-quality claims with proper citations by selecting "Golden Docs" instead of "Noisy Docs" (Bench-Capon, 2005).

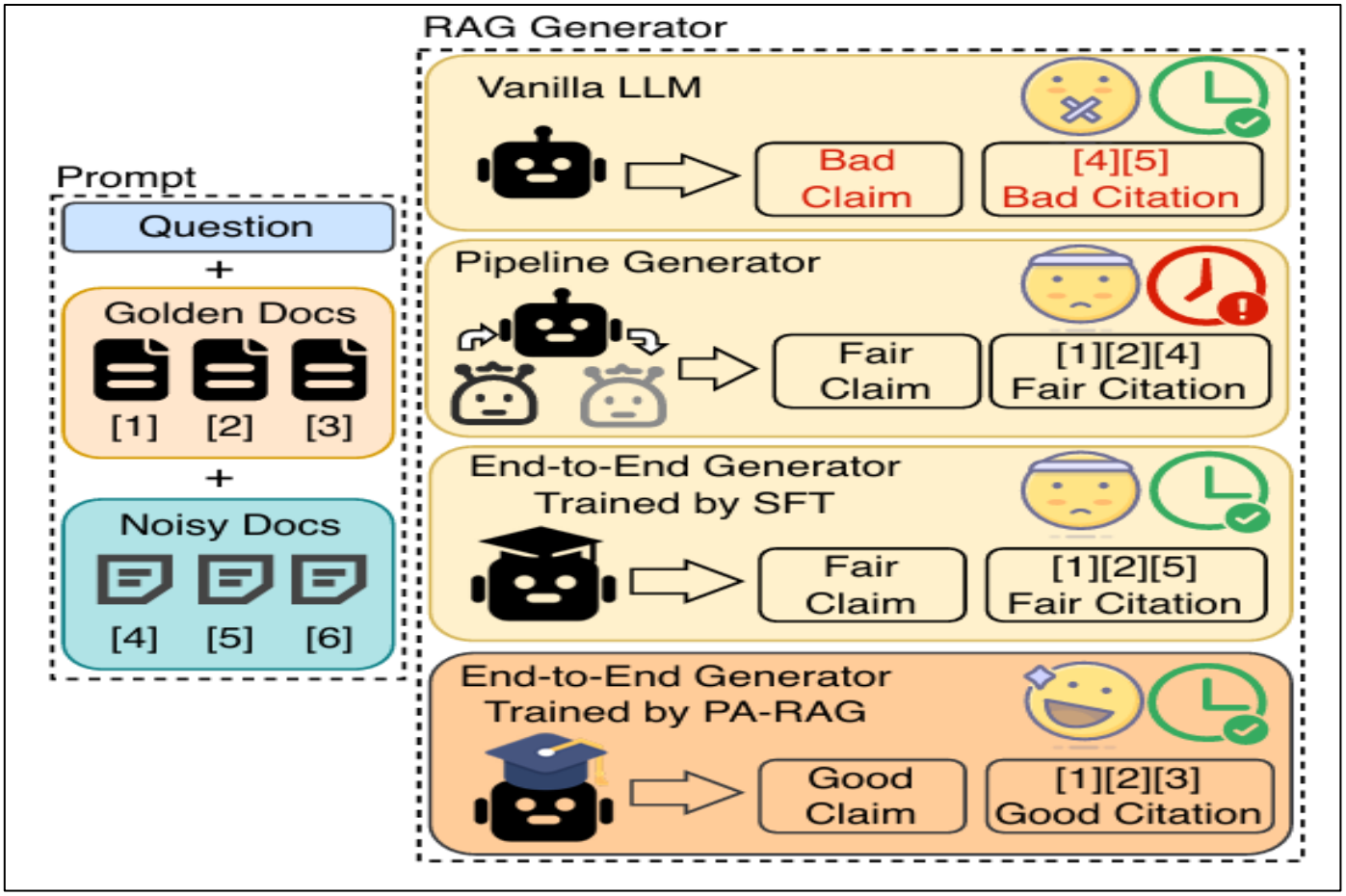

Fig 3 The PA-RAG Training Paradigm (Wu, 2024).

The PA-RAG architecture implements this preference-aware objective through a multi-stage process:

- *Multi-Modal Retrieval:*

The system starts with dense vector store semantic search to find an extensive list of candidate cases through textual similarity evaluation between legal queries and case content. The system achieves high retrieval of all possible relevant documents through this method (Zhou, 2024).

- *The Precedent-Aware Re-Ranker:*

The system uses this component to execute its fundamental principles. The re-ranker system evaluates the initial candidate set by generating a precedent-awareness

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25nov1316

score that combines multiple sub-scores. The system generates a composite precedent-awareness score through multiple sub-score evaluation.

✓ *Authority Score:*

The system generates Authority Score through court hierarchy metadata (Table 1) which gives supreme courts higher importance than lower courts (Ismail, 2025).

✓ *Temporal Score:*

The system generates Temporal Score through case date analysis which rewards current decisions while downgrading cases that have been reversed.

✓ *Jurisdictional Score:*

The system uses Jurisdictional Score to enhance cases from the specified jurisdiction which it determines based on the current query settings.

✓ *Citation Centrality Score:*

The system determines Citation Centrality Score through graph algorithm analysis of legal citation relationships to find essential cases in the field.

The re-ranker combines these signals to generate a weighted list which places legally superior cases first thus separating "Golden" from "Noisy" documents in legal settings (Ismail, 2025).

- *The Legal-Reasoning Generator:*

The generator receives the final document list which has been re-ranked by the system. The generator receives its guidance through advanced prompting techniques which enable it to understand the relative importance of its sources (Figure 3). The system produces a "Good Claim" which represents a legally valid and accurate response through "Good Citations" that start with the most influential precedents followed by other relevant authorities (Huang J. C., 2024).

The end-to-end system design merges retrieval operations with generation functions through legal precedent understanding to produce results that are both factually correct and jurisprudentially sound (Huang J. C., 2024).

IV. IMPLEMENTATION CONSIDERATIONS & CHALLENGES

The PA-RAG framework provides a theoretically correct method for legal AI to work with precedents yet its actual deployment encounters multiple substantial obstacles which need thorough evaluation. The development of a working system based on these principles needs to solve problems related to data quality and computational speed and legal knowledge representation and validation methods (Rane, 2024).

➢ *Data Acquisition and Curation Requirements*

The PA-RAG framework depends on having a complete legal database that has been carefully assembled. The system requires more than basic case text because it needs detailed structured metadata to perform precedent-based operations. The system requires the following essential data elements (Abdul-Azeez, 2024).

- *Court Hierarchy Mappings:*

A complete graph of jurisdictional relationships shows which courts must follow the decisions of other courts (Dong, 2023).

- *Temporal Metadata:*

Temporal Metadata contains exact decision dates together with an authentic historical database showing all instances of case modifications through overruling or reversal or substantive changes (Dong, 2023).

- *Complete Citation Networks:*

The complete collection of inter-case citations forms the base for legal PageRank calculations through citation centrality metrics (Dong, 2023).

The process of creating this dataset proves to be challenging because historical legal documents exist in unorganized formats and require specialized legal expertise to link cases through their authoritative connections.

➢ *Technical and Computational Complexities*

The implementation of the precedent-aware re-ranker system creates multiple technical obstacles. The system faces performance issues because it needs to analyze large citation networks in real time to calculate centrality scores which could make it unsuitable for interactive legal research. The ensemble weighting process of multiple scoring mechanisms needs advanced optimization techniques to prevent new biases from emerging while maintaining retrieval performance. The development of efficient methods for updating dynamic citation graphs and calculating approximate centrality values stands as a essential task for future engineering development (Al-Shboul, 2014).

➢ *Legal Knowledge Representation*

The main obstacle emerges from creating precise mathematical descriptions that capture the various legal status categories. The framework needs to identify both positive and negative legal content and also recognize specific doctrinal approaches which include case distinction and application limitation and critical evaluation. The process of converting complex legal reasoning into machine-readable logic demands extensive teamwork between legal specialists and artificial intelligence developers. The system will fail to understand precedent applications when it lacks precise representations of legal statuses which contain essential details (Huang Y. &., 2024).

➢ *Evaluation Methodology*

PA-RAG system performance validation requires metrics that extend beyond typical NLP assessment tools. The evaluation process requires legal-specific criteria which include three essential metrics:

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25nov1316

- *Authority Score:*
The system evaluates its capacity to detect and arrange the most influential legal cases.

- *Citation Fidelity:*
The evaluation system checks both the precision and suitable context of all referenced sources.

- *Hallucination Rate:*
The system tracks how often it generates fake legal cases and principles.

A complete evaluation process should compare the performance of the proposed RAG system to established baseline systems through legal expert assessments of difficult legal questions. The framework requires human evaluation through a curated set of complex legal queries to determine its success in generating legally valid and authoritative responses (Yang, 2025).

V. ETHICAL IMPLICATIONS AND LIMITATIONS

➢ *Ethical Implications*
The deployment of PA-RAG systems creates multiple essential ethical problems which need thorough analysis.

- *Amplification of Historical Biases:*
The historical nature of legal documents creates biases from past times which PA-RAG systems learn to reproduce through their training process. The system learns to maintain existing biases from historical data because it uses this information for training purposes. The system's ability to maintain outdated legal principles becomes possible through its authority-weighting mechanism which gives historical precedents more visibility and perceived validity (Xu, 2024).

- *The Black Box Problem and Accountability:*
The three components of the retriever and re-ranker and generator system make it difficult to achieve transparency. The exact sequence of reasoning that PA-RAG systems use to generate legal conclusions remains impossible to determine. The system's lack of transparency creates an accountability problem because it becomes impossible to determine who should take responsibility for adverse legal outcomes that result from system errors. The field requires both decision outcomes and their supporting reasons to be equally important (Zhao, 2024).

- *The Illusion of Objectivity and the De-Skilling of the Legal Profession:*
Users may mistake PA-RAG system outputs for absolute truth which would create an unsafe belief in their accuracy. The use of advanced tools for legal work could result in students and practitioners losing their ability to perform essential legal skills including manual case synthesis and analogical reasoning and precedent challenge. The law exists as an active interpretive practice rather than a static collection of data which users should avoid treating as absolute truth (Dong, 2023).

- *Access to Justice and the Digital Divide:*
The digital divide between people could grow because PA-RAG provides legal information access, but it might create new social inequalities. The development of advanced AI tools by well-funded law firms would establish a dual legal system which grants substantial benefits to users who have access to these systems while creating an increased justice gap for others (Barnett, 2024).

➢ *Limitations:*
The PA-RAG framework faces multiple technical and conceptual restrictions which extend beyond ethical considerations.

- *Dependence on Data Quality and Completeness:*
The system depends on high-quality data that includes complete information to achieve its performance goals. The system generates flawed results because it depends on complete and accurate metadata and a wide range of up-to-date knowledge sources. The system fails to detect legal precedents which exist outside its stored database (Huang Y. &., 2024).

- *Handling of Legal Novelty and Overruling:*
The system operates with a conservative approach that focuses on finding established legal precedents. The system faces difficulties when dealing with new legal matters because there are no established precedents to follow. The system faces a major obstacle in detecting when precedent lines undergo subtle changes that human jurists also find difficult to identify (Yang, 2025).

- *Computational and Resource Overhead:*
The precedent-aware re-ranking process along with real-time citation graph analysis requires more computational power than typical semantic search operations. The high computational requirements of this system create substantial infrastructure expenses which restrict its deployment for real-time operations (Yang, 2025).

- *The Human-in-the-Loop Imperative:*
PA-RAG lacks the ability to substitute human legal judgment as its main restriction. The system functions as a decision-support tool which helps users find relevant information more efficiently while improving their recall abilities, but it does not operate as an independent legal reasoning system. Legal professionals who have received proper training need to verify and interpret system outputs before they can use them in practice while maintaining full accountability for their decisions. The system shows how the law has been defined but human professionals need to apply this law to new situations which requires their exclusive expertise (Huang Y. &., 2024).

VI. CONCLUSION AND FUTURE WORKS

The research establishes PA-RAG as a fundamental concept which requires RAG paradigm transformation to fulfill legal reasoning requirements. The paper shows that legal retrieval needs more than semantic similarity because common law depends on court hierarchy and temporal

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25nov1316

precedent and citation network influence. The PA-RAG framework solves this essential problem through its new design which combines a precedent-aware retriever that uses authority and jurisdiction and timeliness and centrality for case re-ranking with a legal-reasoning generator that produces structured and verifiable outputs. The paper establishes a method to develop AI legal research tools which achieve both statistical precision and jurisprudential validity and forensic evidence reliability.

The upcoming research and development work will focus on multiple promising directions. The proposed framework requires immediate implementation and empirical verification through the use of legal-specific evaluation metrics which this study established for performance assessment against current systems. The following research should concentrate on improving temporal understanding through dynamic legal doctrine modeling which includes specific methods to identify when legal precedents become restricted or receive different treatment. The development of cross-jurisdictional reasoning abilities stands as a vital research priority because it enables systems to handle legal conflicts and determine the strength of evidence between different legal systems. The system requires future development to include explainable AI (XAI) methods which will reveal its retrieval and weighting operations for legal professionals to verify. The research community will create AI systems that enhance legal intelligence through accountable and ethical operations by following these research paths.

REFERENCES

[1]. Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C., & Ho, D. (2024). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. https://doi.org/10.48550/arxiv.2405.20362.

[2]. Jacob, M. (2014). Precedents and Case-Based Reasoning in the European Court of Justice. Cambridge University. https://doi.org/10.1017/cbo9781107053762.

[3]. Landes, W. M., & Posner, R. A. (1976). Legal Precedent: A Theoretical and Empirical Analysis. The Journal of Law and Economics, 19(2), 249–307. https://doi.org/10.1086/466868.

[4]. De Brabandere, E. (2016). The Use of Precedent and External Case Law by the International Court of Justice and the International Tribunal for the Law of the Sea. The Law & Practice of International Courts and Tribunals, 15(1), 24–55. https://doi.org/10.1163/15718034-12341311.

[5]. Hitt, M. P. (2016). Measuring Precedent in a Judicial Hierarchy. Law & Society Review, 50(1), 57–81. https://doi.org/10.1111/lasr.12178.

[6]. Dong, C., Yuan, Y., Chen, K., Cheng, S., & Wen, C. (2023). How to Build an Adaptive AI Tutor for Any Course Using Knowledge Graph-Enhanced Retrieval-Augmented Generation (KG-RAG). https://doi.org/10.48550/arxiv.2311.17696.

[7]. Duxbury, N. (2008). The Nature and Authority of Precedent. Cambridge University. https://doi.org/10.1017/cbo9780511818684.

[8]. Guillaume, G. (2011). The Use of Precedent by International Judges and Arbitrators. Journal of International Dispute Settlement, 2(1), 5–23. https://doi.org/10.1093/jnlids/idq025.

[9]. Bench-Capon, T., Atkinson, K., & Chorley, A. (2005). Persuasion and Value in Legal Argument. Journal of Logic and Computation, 15(6), 1075–1097. https://doi.org/10.1093/logcom/exi058.

[10]. Hinkle, R. K. (2015). Legal Constraint in the US Courts of Appeals. The Journal of Politics, 77(3), 721–735. https://doi.org/10.1086/681059.

[11]. Atkinson, K., & Bench-Capon, T. (2005). Legal Case-based Reasoning as Practical Reasoning. Artificial Intelligence and Law, 13(1), 93–131. https://doi.org/10.1007/s10506-006-9003-3.

[12]. Hafner, C. D., & Berman, D. H. (2002). The role of context in case-based legal reasoning: teleological, temporal, and procedural. Artificial Intelligence and Law, 10(1–3), 19–64. https://doi.org/10.1023/a:1019516031847.

[13]. Al-Abdulkarim, L., Bench-Capon, T., & Atkinson, K. (2016). A methodology for designing systems to reason with legal cases using Abstract Dialectical Frameworks. Artificial Intelligence and Law, 24(1), 1–49. https://doi.org/10.1007/s10506-016-9178-1.

[14]. Stone Sweet, A. (2002). Path Dependence, Precedent, and Judicial Power (pp. 112–135). Oxford University. https://doi.org/10.1093/0199256489.003.0004.

[15]. Rane, N., Choudhary, S. P., & Rane, J. (2024). Acceptance of artificial intelligence: key factors, challenges, and implementation strategies. Journal of Applied Artificial Intelligence, 5(2), 50–70. https://doi.org/10.48185/jaai.v5i2.1017.

[16]. Abdul-Azeez, O., Idemudia, C., & Ihechere, A. (2024). Best practices in SAP implementations: Enhancing project management to overcome common challenges. International Journal of Management & Entrepreneurship Research, 6(7), 2048–2065. https://doi.org/10.51594/ijmer.v6i7.1256.

[17]. Al-Shboul, M., Al-Saqqa, S., Rababah, O., & Ghnemat, R. (2014). Challenges and Factors Affecting the Implementation of E-Government in Jordan. Journal of Software Engineering and Applications, 07(13), 1111–1127. https://doi.org/10.4236/jsea.2014.713098.

[18]. Zhou, R. (2024). Advanced Embedding Techniques in Multimodal Retrieval Augmented Generation Comprehensive Study on Cross Modal AI Applications. Journal of Computing and Electronic Information Management, 13(3), 16–22. https://doi.org/10.54097/h8wf8vah.

[19]. Ismail, I., Kurnia, R., Widyatama, F., Wibawa, I. M., Brata, Z. A., Ukasyah, U., Nelistiani, G. A., & Kim, H. (2025). Enhancing Security Operations Center: Wazuh Security Event Response with Retrieval-Augmented-Generation-Driven Copilot. Sensors (Basel, Switzerland), 25(3), 870. https://doi.org/10.3390/s25030870.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/25nov1316

[20]. Huang, J., Cui, Y., Chen, L., Liu, J., Wang, T., Li, H., Wang, M., & Wu, J. (2024). Layered Query Retrieval: An Adaptive Framework for Retrieval-Augmented Generation in Complex Question Answering for Large Language Models. Applied Sciences, 14(23), 11014. https://doi.org/10.3390/app142311014.

[21]. Xu, K., Li, J., Zhang, K., Wang, Y., & Huang, W. (2024). CRP-RAG: A Retrieval-Augmented Generation Framework for Supporting Complex Logical Reasoning and Knowledge Planning. Electronics, 14(1), 47. https://doi.org/10.3390/electronics14010047.

[22]. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., & Cui, B. (2024). Retrieval-Augmented Generation for AI-Generated Content: A Survey. https://doi.org/10.48550/arxiv.2402.19473.

[23]. Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven Failure Points When Engineering a Retrieval Augmented Generation System. https://doi.org/10.48550/arxiv.2401.05856.

[24]. Huang, Y., & Huang, J. (2024). A Survey on Retrieval-Augmented Text Generation for Large Language Models. https://doi.org/10.48550/arxiv.2404.10981.

[25]. Yang, R., Ning, Y., Keppo, E., Liu, M., Hong, C., Bitterman, D. S., Ong, J. C. L., Ting, D. S. W., & Liu, N. (2025). Retrieval-augmented generation for generative artificial intelligence in health care. Npj Health Systems, 2(1). https://doi.org/10.1038/s44401-024-00004-1.

[26]. Teneva, M. (2025, May 20). RAG vs Fine Tuning: The Hidden Trade-offs No One Talks About. B EYE. https://b-eye.com/blog/rag-vs-fine-tuning/.

[27]. Ajay Mukund, S., & Easwarakumar, K. S. (2025). Optimizing Legal Text Summarization Through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation. Symmetry, 17(5), 633.

[28]. Wu, J., Cai, H., Yan, L., Sun, H., Li, X., Wang, S., ... & Gao, M. (2024). Pa-rag: RAG alignment via multi-perspective preference optimization. arXiv preprint arXiv:2412.14510.