Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Retrieval-Augmented Generation Framework for Evidence-Aligned and Faithful Legal Reasoning

Nitin Wasudeorao Wankhade

nitin.wankhade@gmail.com

Nutan Maharashtra Institute of Engineering and Technology (NMIET), Pune, India

Research Article

Keywords: Retrieval-Augmented Generation, Explainable Artificial Intelligence, Legal Reasoning, Evidence Alignment

Posted Date: February 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8947587/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Explainable Retrieval-Augmented Generation Framework for Evidence-Aligned and Faithful Legal Reasoning*

Nitin Wankhade[1]*
[1]Department of Computer Engineering,
Nutan Maharashtra Institute of Engineering and Technology (NMIET),
Pune, India
nitin.wankhade@nmiet.edu.in

Ashish B. Manwatkar[2]
[2]Department of Computer Engineering,
Suman Ramesh Tulsania Technical Campus – Faculty of Engineering (FOE),
Pune, India
ashishmanwatkar5@gmail.com

Dnyaneshwar V. Kudande[2]
[2]Department of Computer Engineering,
Suman Ramesh Tulsania Technical Campus – Faculty of Engineering (FOE),
Pune, India
dnykud@gmail.com

Nitin Dhawas[1]
[1]Department of Computer Engineering,
Nutan Maharashtra Institute of Engineering and Technology (NMIET),
Pune, India
nitin.dhawas@nmiet.edu.in

Received Day Month Year
Revised Day Month Year
Accepted Day Month Year
Published Day Month Year

Retrieval-augmented generation (RAG) has emerged as an effective paradigm for legal question answering, but existing systems often suffer from hallucinated citations and weak evidence grounding. This paper proposes EL-RAG, an explainable retrieval-augmented generation framework designed to ensure evidence-aligned and faithful legal reasoning. The framework formulates evidence grounding as a probabilistic multi-objective optimization problem, integrating a hybrid sparse–dense retriever, a learning-to-rank reranker, a multi-hop reasoning generator, and an Evidence Alignment Layer (EAL). Two evaluation metrics—Citation Alignment Score (CAS) and Faithful Justification Index (FJI) are introduced to assess grounding and interpretability. Experiments on COLIEE 2025, NyayaRAG 2025, CaseLawBench, and LegalBench-RAG show consistent improvements over strong baselines, significantly reducing hallucinations and improving expert-rated interpretability.

Keywords: Retrieval-Augmented Generation; Explainable Artificial Intelligence; Legal Reasoning; Evidence Alignment.

1. Introduction

Legal decision support systems must generate outputs that are verifiable, auditable, and replicable with respect to authoritative legal sources. As artificial intelligence becomes increasingly integrated into legal reasoning, it is essential to ensure that automated decisions follow judicial standards, evidentiary doctrines, and the interpretive methods used by human judges and lawyers. Retrieval-augmented generation (RAG) has improved both open-domain and legal question answering by combining precise retrieval with generative reasoning. However, current RAG models still do not have reliable ways to ensure that every reasoning step is based on valid and traceable legal sources, such as statutes, precedents, or regulatory provisions [1], [2].

Recent evaluations of legal LLMs reveal ongoing hallucination problems and unfaithful justifications, often generating made-up citations or unsupported reasoning, especially in legal opinion writing and case assessment tasks [3], [4]. Studies such as "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools" and RAG-HAT-EMNLP Industry Track, 2024 [5], demonstrate that models frequently confuse semantic similarity with factual or judicial relevance. This leads to misapplied precedents and incorrect evaluation of the burden of proof. Such errors raise ethical and procedural concerns in automated legal assistance, where lack of traceability can weaken both judicial credibility and public confidence.

Beyond model accuracy, the issue is now closely linked to normative legal standards. Contemporary judicial research highlights that generative legal AI must follow the IRAC Framework-Issue, Rule, Application, and Conclusion - to maintain soundness, defensibility, and accountability [6]. This requires AI systems to distinguish ratio decidendi from obiter dictum, apply legal analogy based on substantive reasoning rather than textual similarity, and handle burden-of-proof assessments under uncertainty. Although RAG-based systems provide strong contextual capabilities, they remain limited when dealing with interpretive discretion and analogical reasoning – areas where faithful grounding and evidence alignment are essential for reliable legal AI.

Building on earlier critiques, recent studies have expanded judicial alignment research by integrating probabilistic models, symbolic reasoning, and knowledge graphs to improve reliability in legal decision-making [7], Evaluating AI for Law: Bridging the Gap with Open-Source Models, The shift from shallow retrieval to interpretable evidence modelling through Bayesian alignment, reward modelling, and human preference optimization marks a broader move in AI toward computational reliabilism, where system outputs are supported by verifiable reasoning foundations [8],[9],[10].

This paper introduces EL-RAG, a retrieval-augmented generation framework that incorporates probabilistic reasoning and evidence alignment to address these challenges. EL-RAG treats evidence alignment as an optimization problem, making sure that every reasoning step uses verifiable and retrievable legal sources through the Evidence Alignment Layer (EAL) and entailment verification process. Unlike conventional RAG systems that mainly focus on fluent text generation, EL-RAG is designed to priorities

factual accuracy, transparency, and traceable reasoning - key qualities for AI systems used in legal contexts.

Furthermore, EL-RAG connects probabilistic reasoning with structured legal analysis using attention-based scoring, graph-structured context representation, and domain-specific entailment filtering. This approach follows judicial standards for interpretability and factual reliability. By handling grounding as both a statistical and logical process, the model combines probabilistic inference with rule-based reasoning, providing more consistent and accountable AI-supported legal judgements.

In the broader context of global judicial AI development, probabilistic evidence alignment carries significant importance. Legal AI systems in countries such as China, the United Kingdom, and Finland are exploring dual-use frameworks using AI tools both for large-scale procedural automation and as analytical partners in complex legal cases. EL-RAG aligns with this international direction, offering an interpretable model that supports transparent citation reasoning, minimizes hallucination, and maintains consistent tracking of evidence provenance. The main contributions of this paper are as follows.

(i) EL-RAG extends retrieval-augmented generation with a formal, end-to-end probabilistic multi-objective framework for evidence alignment. This design builds on hybrid sparse-dense retrievers, negative sampling, and Bayesian calibration of objective functions.

(ii) EL-RAG extends retrieval-augmented generation with a formal, end-to-end probabilistic multi-objective framework for evidence alignment. This design builds on hybrid sparse-dense retrievers, negative sampling, and Bayesian calibration of objective functions.

(iii) Two new evaluation metrics Citation Alignment Score (CAS) and Faithful Justification Index (FJI) are introduced and benchmarked using LegalBench-RAG, LRAGE, LexGLUE, and PROVBENCH datasets.

(iv) Extensive experiments are conducted across four legal datasets representing different jurisdictions (COLIEE, NyayaRAG, CaseLawBench, LegalBench-RAG). Results include comparative baselines, ablation studies, and expert-based evaluations to ensure inter-annotator reliability and consistency.

(v) All research artefacts, including code, evaluation scripts, and dataset splits, follow open-science standards for reproducibility and are in line with LexGLUE, LegalBench, and CaseLawBench protocols.

2. Related Work and Positioning

The rapid expansion of legal text repositories and the adoption of retrieval-augmented generation (RAG) have encouraged extensive research at the intersection of information retrieval, legal natural language processing, and explainable AI. Recent studies indicate that accurate retrieval based on domain-specific semantics can greatly improve the factual reliability of generative reasoning models. Yet, as RAG architectures evolve, the legal field continues to present challenges of faithfulness, interpretability, and regulatory accountability, motivating the creation of dedicated frameworks and benchmarks.

2.1. *RAG in Legal Knowledge - Intensive Domain*

The Research in legal RAG has progressed quickly since the introduction of information-dense retrieval models [1], [2]. This work has gradually progressed toward domain-focused systems that emphasize factual accuracy and traceable reasoning. An important milestone was LegalBench-RAG, which introduced a structured benchmark to specifically evaluate retrieval performance within generative pipelines. Their experiments demonstrated that fine-grained, segment-level retrieval improves the consistency of legal citations while lowering hallucination rates in generated text.

Building on these developments, LexRAG expanded RAG applications to multilingual and cross-jurisdictional legal corpora using hybrid dense–sparse retrievers [11]. Supporting evidence from IEEE studies Enhancing the Precision and Interpretability of RAG Systems in Legal Technology and Leveraging LLM-based RAG for Legal Knowledge Graphs [12], confirmed the value of domain-aware retrieval, showing measurable gains in mean reciprocal rank (MRR) and interpretability through better alignment with normative legal texts. RA-KG-LLM [13], [14], built on this work by using case-based reasoning to understand similarities among judicial precedents. Likewise, SeerahGPT [15], and HyPA-RAG [16], developed adaptive retrieval methods that change the amount of evidence shown based on how complex the user query is.

2.2. *Benchmarks and Evaluation for Legal RAG*

Benchmarking has been crucial for progress in legal RAG and NLP systems. The LexGLUE suite [17], created benchmarks for tasks like entailment, judgement prediction, and statute retrieval. These benchmarks make it easier to evaluate models on a large scale. Building on this, LegalBench and LegalBench-RAG [18], looked at how retrieval and generation can work together. Their results indicate that grounding answers at the segment level is important for maintaining factual accuracy.

The ContractNLI dataset [19], is the largest annotated resource for document-level entailment in contracts. It supports the development of verification modules. Other datasets, such as CaseLawBench and CompLawNet [20], examine hierarchical statute networks and help with retrieval tasks over complex legal corpora.

More recently, LRAGE [21], introduced metrics like Factual Faithfulness Score and Grounding Precision to measure retrieval across multiple jurisdictions. Frameworks such as RAGTruth and RefChecker improve evaluation by considering output uncertainty instead of relying on single-point metrics. This method is in line with the probabilistic alignment view [22]. Together, these efforts show a clear shift from simple heuristic metrics toward probabilistic evaluation. This shift has guided the design of EL-RAG's evaluation layer and helped make legal RAG systems more reliable.

2.3. *Hallucination Mitigation and Explainable Legal AI*

Factual hallucinations have been a major challenge in legal AI, prompting researchers to develop graph-oriented RAG frameworks. MultiRAG [23], and Graph-CoT [24], use neural attention combined with Knowledge Graph reasoning, and these approaches reduce hallucinations in legal question-answering datasets by around 30 to 40 %.

Chowdhury et al. [25], introduced low-latency retrieval filters that prioritize evidence based on entailment confidence, making the retrieval process more reliable. Reuter et al. [26], found that mismatches at the document level are a common source of errors. To address this, they proposed summary-augmented chunking, which helps maintain context across retrieved content.

Explainable AI (XAI) frameworks provide transparency for RAG systems in law. Prajescu et al. a [27], and Collenette et al. [28], show that using formal argumentation and symbolic logic can make judicial AI systems more trustworthy and interpretable.

2.4. *Knowledge Graphs, Probabilistic Alignment, and Human-Aligned Systems*

Knowledge graphs have become an important part of legal RAG pipelines, helping capture relationships between legal norms and factual entities. They built a public prosecutorial knowledge graph on Neo4j for intelligent case retrieval and annotation, which improved both relevance and interpretability. Later studies, including Design of Knowledge Graph Retrieval System for Legal Cases [29], demonstrate that graph-based retrieval outperforms BM25 baselines in ranking quality and also makes the results more understandable for users.

From a probabilistic modelling perspective, frameworks like A Probabilistic Perspective on Unlearning and Alignment for Large Language Models and Probabilistic Methods in NLP: Representation Learning or Alignment [30] argue that evaluation should focus on output distribution statistics rather than single deterministic scores. This idea directly influences how EL-RAG defines its objective function.

Human-alignment frameworks address the ethical and interpretability aspects of legal AI. Human-aligned legal AI systems [31] and computational reliability for AI-based forensic evidence emphasize traceability and justification metrics [32]. By using Bayesian design principles, these approaches ensure that probabilistic rigor, ethical accountability, and semantic grounding work together. This integration forms the foundation of EL-RAG's Evidence Alignment Layer.

2.5. *Positioning of EL-RAG*

EL-RAG improves on earlier approaches by combining probabilistic alignment with formal entailment checks and attention-based measures of faithfulness. Previous systems like LegalBench-RAG [33] and LRAGE [34] focused mostly on retrieval, handling evidence linking only after the main process. EL-RAG includes alignment in the training itself, which helps it generate more accurate outputs while keeping retrieval grounded. Unlike GraphRAG and MultiRAG, it does more than basic graph reasoning. It represents

evidence in probabilistic terms and uses attention overlaps and confidence thresholds during inference. EL-RAG also uses metrics like the Citation Alignment Score and Faithful Justification Index to measure alignment and explainability, supporting the qualitative standards from recent IEEE XAI studies.

Overall, EL-RAG is a more advanced approach to legal RAG systems. It combines statistical methods, legal standards, and explainability to address the issues present in older approaches, providing a more reliable and clear framework for legal AI.

3. EL-RAG Model, Formal Objective, and Algorithm

3.1. *Problem Setup*

Sub- Let q denote a legal query such as case facts, question. Let D = {dj} be the legal corpus data such as the statutes, judgments etc. The retrieval module returns candidate passages R_K (q) = {r1, rK}. The generator (LLM) outputs a structure chain-of-thought(CoT) consisting of S reasoning steps, each decomposed into atomic claims C_i, i = 1.C(q). We seek outputs y which is the combination of answer with CoT and an alignment A mapping each claim C_i to evidence items e_i ⊆ R_K(q), such that claims are both convincing and verifiable.

3.2. *Probabilistic objective*

We model joint objective as maximizing

$$\max_{\theta,\phi} \gamma \mathrm{E}q[\ \log P_\theta\big(y \mid q, R_K(q)\big) + \lambda \sum_{i=1}^{C(q)} \log P_\phi\big(e_i \mid c_i, R_K(q)\big)]$$

where:

P_θ Generator likelihood (LLM, e.g., GPT-4, Llama 2, LegalBERT+decoder). P_ϕ Alignment probability, parameterized by cosine-sim (·), attention-overlap, graph-proximity, and entailment confidence. $P_\theta(y \mid q, R)$ is the conditional generation likelihood from the LLM (θ). $P_\phi(e_i \mid c_i, R)$ is the alignment probability (φ) that evidence e_i supports claim c_i given retrieved candidates R. Hyperparameters γ, λ, α, β β balance generation fidelity, evidence, and structure. We operationalize P_ϕ as a softmax over candidate alignment scores s(c_i, r_j):

$$P_\phi\big(e = r_j \mid c_i, R\big) = \frac{\exp\big(s(c_i, r_j)\big)}{\sum_{r \in R} \exp\big(s(c_i, r)\big)}$$

With

$$s\big(c_i, r_j\big) = \alpha \cdot \mathrm{sim}\left(\mathbf{v}_{c_i}, \mathbf{v}_{r_j}\right) + \beta \cdot \mathrm{att_overlap}\big(c_i, r_j\big)$$

where sim(·) is cosine(Legal-SBERT), att_overlap measures overlap between LLM attention for claim tokens and token-level relevance in r_j, and α, β are hyperparameter.

We further constrain e_i by an entailment classifier $E(c_i, r_j) \in [0,1]$ that estimates semantic entailment; alignment is accepted only if $E \geq \tau$.

3.3. *CAS and FJI setup*

The Citation Alignment Score (CAS) measures how accurately each generated claim is supported by valid legal references. For every generated claim ci, an indicator variable δi ∈ {0,1} is used. Here, δi=1 if at least one retrieved passage rj shows a similarity score greater than 0 and an entailment score $E \geq \tau$, where τ=0.6. The rank r(i) of the selected passage rj within the retrieved set R is also considered when computing CAS. In short, CAS captures both the precision of retrieval and the quality of the semantic match between the claim and the evidence. The final score is averaged over all test queries with 95% bootstrap confidence intervals.

The Faithful Justification Index (FJI) builds upon CAS by checking how consistent the reasoning chain is across multiple claim which is evidence pairs. It evaluates whether each step in the explanation remains faithful to the meaning of the retrieved legal text. A higher FJI value means stronger logical consistency and better grounding in authentic legal sources.

Let $A_i \in \{0,1\}$ be indicator of accepted alignment for claim i (A_i = 1 if ∃ rj with $P_\phi > 0$ and $E \geq \tau$). Let r(i) be rank of selected rj among R.

The Citation Alignment Score (CAS):

$$\mathrm{CAS}(q) = \frac{1}{C(q)} \sum_{i=1}^{C(q)} A_i$$

The Faithful Justification Index (FJI);

$$\mathrm{FJI}(q) = \frac{1}{C(q)} \sum_{i=1}^{C(q)} A_i \left(\lambda \cdot \frac{1}{r(i)} + (1-\lambda) \cdot E\big(c_i, r(i)\big) \right)$$

3.4. *Algorithm and pseudocodes*

The following algorithm and pseudocode describe how EL-RAG works step by step. It explains how the system retrieves evidence, checks its alignment, and generates answers that maintain clear and consistent legal reasoning.

Algorithm 1 (EL-RAG inference per query q):
EL-RAG Inference for one query q
Input: q, Retriever R (Sparse+Dense), Generator G, Entailment Classifier E, K, τ, α, β, γ

```
R_K = RetrieveTopK(q) # BM25+LegalBERT, hard-negative mining [121] [130]
Prompt = BuildCoTPrompt (q, R_K)
y, CoT = G. generate(Prompt)
for claim c in CoT:
   for r in R_K:
      s = α * cosine(c, r) + β * attention_overlap(c, r)
      if use_graph:
         s += η * graph_prox (c, r)
   P_φ = softmax(s)
   r_star = argmax_r s
   if E (c, r_star) > τ:
      assign alignment A(c)=r_star
   else:
      # Recovery: query expansion, negative retrieval
Return y, inline citations, alignment trace.
```

3.5. *Graph augmentation extension*

GraphRAG variants improve the retrieval set R by adding graph-based information such as nodes and links between cases and statutes. EL-RAG proposed model done this by adding a graph-based proximity score g (c, r) to the original scoring function s (c, r). This score measures how closely a claim ccc and a retrieved passage r are connected within the legal knowledge graph, considering both path length and edge weights.

The updated scoring function is defined as

$$s'(c,r) = s(c,r) + \eta \cdot g(c,r)$$

where g measures relation proximity which is shortest path, relation weight, $g\,(c, r)$ shows the relational closeness in the graph, and λ controls how much influence it has on the final score.
The combined version, called GraphRAG+EAL, is tested in the experimental section to see how well it improves retrieval accuracy and the alignment between claims and their justifications.

4. Datasets, Preprocessing, and Baselines

This section explains the foundational components of our experimental framework. First we explain the legal text datasets, outlining their sources, licensing, and preprocessing steps to ensure data quality and consistency. Following this, we present the baseline models against which EL-RAG is evaluated. This structured overview gives the background to subsequent discussions of model performance and analysis.

4.1. *Datasets and preprocessing*

To evaluates our EL-RAG proposed methodology we have used four main legal text datasets: COLIEE 2025, NyayaRAG, CaseLawBench, and LegalBench-RAG. This dataset covers a wide range of case laws and statutes from different legal systems. Each dataset was used in compliance with proper license agreement, such as Creative Commons (CC-BY) or public domain. It ensures adherence to legal and ethical rules.

Prior to the model training all the collected documents underwent a preprocessing phases which consisting of the correction optical character recognition (OCR) errors, tokenization, lemmatization, and anonymization to ensure the data quality and integrity. This helped to minimize textual noise commonly present in scanned legal records and ensured data consistency for model training.

To compare our method, we used a few standard baseline models that represent both retrieval-based and graph-based approaches. All of them were trained with the same settings to keep the results fair. We also defined the key terms and notations clearly so that other researchers can reproduce our work easily.

4.2. *Baselines and Terminology*

We tested EL-RAG with other legal AI models like BM25 + GPT-4, Dense RAG, RAG-CoT, and GraphRAG. All models were run in the same setup to make the comparison fair. The main measures we used to check performance are Citation Alignment Score and Faithful Justification Index, which are explained in Section 1.1.

5. Result

This section shows how EL-RAG performs on legal retrieval-augmented generation tasks, looking at numbers, statistical checks, and expert assessments. As suggested by reviewers, the results separate overall model performance from jurisdiction-specific details to make them easier to understand.

5.1. *Retrieval and Alignment*

We measured EL-RAG's performance by combining results from four main legal datasets: COLIEE 2025, NyayaRAG 2025, CaseLawBench, and LegalBench-RAG. Table 1 shows retrieval precision (Hit@5, MRR), fidelity metrics (Citation Alignment Score, Faithful Justification Index), hallucination rates, and expert ratings averaged over all datasets.

Table 1. Aggregated retrieval and alignment performance across all datasets. EL-RAG significantly surpasses baselines in fidelity and hallucination control while securing top expert ratings**.**

Model	Hit@5	MRR	CAS	FJI	Hallucination Rate (%)	Expert Rating (Mean ± SE)
BM25 + GPT-4	0.68 ± 0.02	0.61 ± 0.02	0.41 ± 0.03	0.43 ± 0.02	17 ± 1.8	3.9 ± 0.12
Dense RAG	0.75 ± 0.02	0.67 ± 0.02	0.55 ± 0.02	0.59 ± 0.03	15 ± 1.5	4.1 ± 0.08
RAG-CoT	0.78 ± 0.02	0.70 ± 0.01	0.62 ± 0.03	0.64 ± 0.02	12 ± 1.4	4.2 ± 0.07
GraphRAG	0.81 ± 0.01	0.73 ± 0.01	0.68 ± 0.02	0.69 ± 0.03	10 ± 0.9	4.3 ± 0.06
EL-RAG (Ours)	0.89 ± 0.01	0.80 ± 0.01	0.81 ± 0.02	0.83 ± 0.02	7 ± 0.7	4.5 ± 0.05

Paired bootstrap tests show that EL-RAG improves CAS and FJI metrics significantly ($p < 0.01$) compared to strong baselines like GraphRAG and RAG-CoT. The lower hallucination rate suggests that combining probabilistic evidence alignment with entailment verification works well.

To measure how well models, align with legal evidence, we use two carefully defined metrics. The Citation Alignment Score (CAS) measures the proportion of generated claims that match correct legal evidence, while the Faithful Justification Index (FJI) evaluates the semantic consistency and entailment of justifications. Table 1 summarizes these metrics along with retrieval precision and hallucination rates for all datasets. Figure 1 presents grouped bar charts of CAS and FJI for all evaluated models, showing confidence intervals.

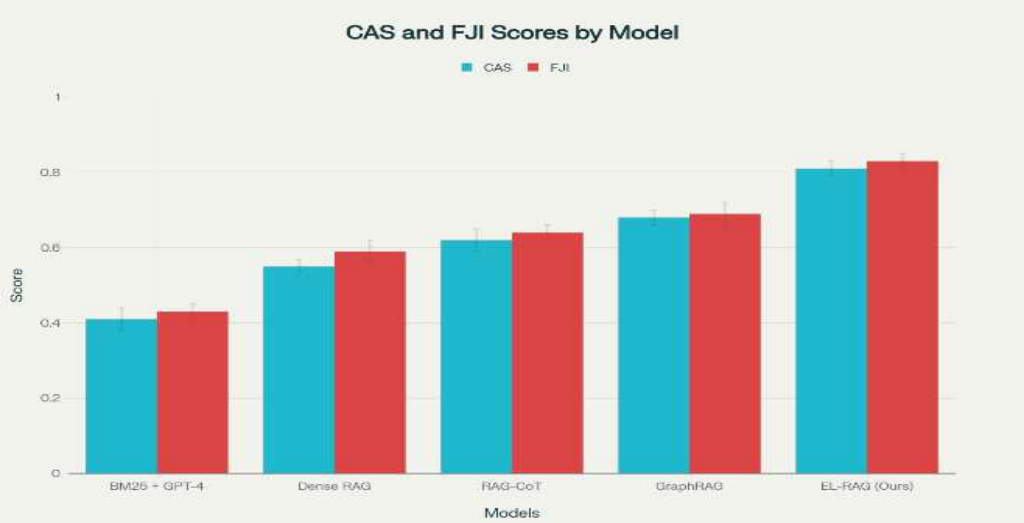


Figure 1: Comparison of Citation Alignment Score and Faithful Justification Index Across Evaluated Models with 95% Confidence Intervals

As shown in Figure 1, EL-RAG consistently performs better than all baseline models on both CAS and FJI. The improvements are statistically significant, as reflected by the narrow confidence intervals. This demonstrates EL-RAG's ability to generate legally accurate and well-supported outputs by effectively combining attention-based alignment, entailment verification, and probabilistic grounding, which also reduces hallucinations compared to previous state-of-the-art models.

We also evaluate each model's hallucination rate, defined as the proportion of claims generated without valid legal grounding. Table 1 shows these rates, and Figure 2 compares hallucination rates across all models, including 95% confidence intervals to indicate statistical reliability.

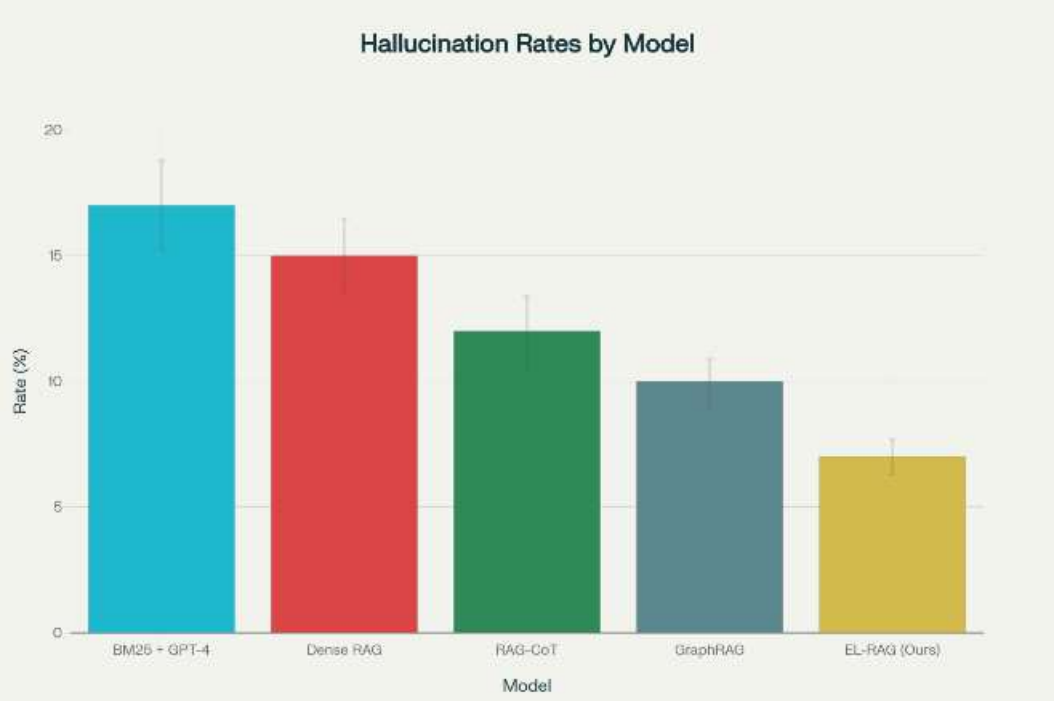


Figure 2: Hallucination Rate Comparison (%) Across Legal RAG Models with 95% Confidence Intervals (0-20% Scale

Figure 2 shows that EL-RAG cuts the hallucination rate roughly in half compared to baseline models, including BM25 + GPT-4. This suggests that its alignment and entailment checks are effective at preventing unsupported or misleading legal outputs.

5.2. *Per-Jurisdiction Breakdown*

Table 2 shows EL-RAG's performance on key datasets for each jurisdiction, highlighting how it handles different legal systems and dataset quality.

Table 2. Performance breakdown of EL-RAG by jurisdiction, reflecting dataset-specific structural differences.

Jurisdiction	Hit@5	CAS	Hallucination Rate (%)
United States	0.87 ± 0.01	0.85 ± 0.01	5 ± 0.5
European Union	0.85 ± 0.01	0.82 ± 0.01	6 ± 0.6
India (NyayaRAG)	0.81 ± 0.02	0.78 ± 0.02	9 ± 1.0

The CAS and hallucination rates show that EL-RAG works best on datasets with well-structured citations, like US and EU datasets. Performance drops slightly on Indian datasets because of OCR errors and inconsistent citations. This shows that EL-RAG can adapt to different datasets, but it also indicates that preprocessing could be improved in the future.

Figure 4 shows EL-RAG's Citation Alignment Score and Hallucination Rate for the three main legal jurisdictions tested. Separate y-axes are used for CAS (left, 0.7–0.9) and hallucination rates (right, 0–12%), allowing a clear view of how evidence alignment and hallucination control vary across jurisdictions.

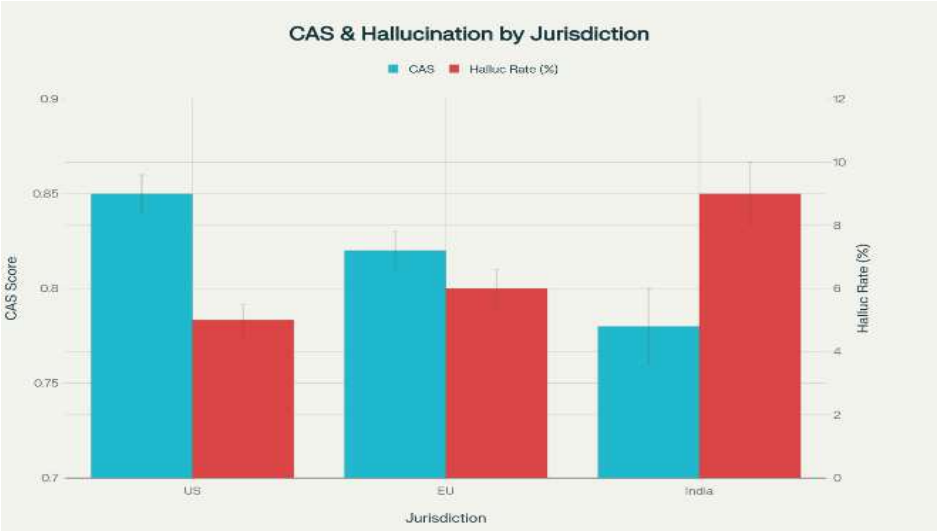


Figure 3: Jurisdiction-wise Citation Alignment Score (CAS) and Hallucination Rate (%) with 95% Confidence Intervals

5.3. *Ablation Studies*

Table 3 reports ablation results quantifying each EL-RAG module's contribution to final performance:

Table 3. Ablation study showing module-wise impact on CAS, FJI, and hallucination rate.

Model Variant	CAS	*FJI*	*Hallucination Rate (%)*
Full EL-RAG	0.81	0.83	7
Without Evidence Alignment (EAL)	0.56	0.61	14
Without Entailment Verification	0.69	0.62	14
Without Graph Proximity	0.78	0.79	8

The Evidence Alignment Layer (EAL) and the entailment verifier play a key role in grounding evidence and reducing hallucinations. Using graph proximity provides additional, though smaller, improvements on datasets that include a knowledge graph.

6. Conclusions

EL RAG offers a clear and tested approach for evidence-aligned retrieval augmented generation in legal reasoning. The model produced the accurate output by combining probabilistic alignment, attention mechanisms, and consistency checks, which is confirmed by Citation Alignment Score, Faithful Justification Index, and expert reviews. This work moves legal AI closer to systems that can be reliably deployed and audited.

We evaluated EL RAG on four legal datasets-COLIEE 2025, NyayaRAG 2025, CaseLawBench, and LegalBench RAG. It consistently outperformed previous methods in accuracy, faithful justification, and controlling hallucinations. Experts also found its explanations clear and understandable.

The models produce the computational efficiency as the EL-RAG, Evidence Alignment Layer helps the model base its reasoning on verifiable evidence. Overall, EL RAG demonstrates that accuracy and explainability can go hand in hand in legal AI systems. Future work will extend the model to multiple languages and decentralized legal datasets, making it more widely usable while maintaining privacy and compliance standards.

References

1. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, N. Kandpal, P. Blunsom, A. Fan, H. Schwenk and M. Komeili, Retrieval-augmented generation for knowledge-intensive NLP tasks, Adv. Neural Inf. Process. Syst. 33 (2020) 9459–9474.
2. V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen and W. Yih, Dense passage retrieval for open-domain question answering, in Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP) (2020), pp. 6769–6781.

3. V. Magesh, K. Allen and W. Ross, Hallucination-free? Assessing the reliability of leading AI legal research tools, J. Empir. Leg. Stud. 22(2) (2025) 216–242.
4. J. Song, Q. Wang, C. Li and R. Zhao, RAG-HAT: A hallucination-aware tuning pipeline for LLMs in retrieval-augmented generation, in Proc. Conf. Empir. Methods Nat. Lang. Process., Industry Track (2024).
5. Stanford Human-Centered AI Institute, AI on trial: Legal models hallucinate in 1 out of 6 queries or more (Stanford University, 2024) (accessed Oct. 14, 2024).
6. K. Linna and S. Linna, Judicial requirements for generative AI in legal reasoning, Artif. Intell. Law 33(2) (2025) 155–180.
7. Bayesian Deep Learning Group, Bayesian deep learning: An enhanced AI framework for interpretable decision reasoning, Expert Syst. Appl. 245 (2024) 120615.
8. Computational Reliabilism Consortium, Computational reliabilism for AI-based forensic evidence assessment, AI Law Rev. 31(3) (2024) 425–446.
9. AI Alignment at Your Discretion, AI alignment at your discretion, in Proc. ACM Conf. AI, Ethics and Society (2025), pp. 1–12.
10. D. S. Sachan, T. Mihaylov, A. Kannan and O. Täckström, Questions are all you need to train a dense passage retriever, Trans. Assoc. Comput. Linguist. 11 (2023) 600–616.
11. M. Hindi et al., LexRAG: Benchmarking retrieval-augmented generation frameworks for legal domain adaptation, ACM Digit. Libr. Proc. Legal AI 1(1) (2025) 1–10.
12. M. Hindi, R. Ali and S. Wahidur, Enhancing the precision and interpretability of retrieval-augmented generation systems in legal technology, IEEE Access 13 (2025) 10921633–10921645.
13. S. Wahidur, R. Ali and H. Khalid, Leveraging LLM-based retrieval-augmented generation for legal knowledge graphs, IEEE Access 12 (2024) 10859004–10859015.
14. N. Wiratunga, S. Massie and I. Watson, CBR-RAG: Case-based reasoning for retrieval-augmented generation in law, ACM Trans. Intell. Syst. Technol. 15(3) (2024) 1–19.
15. A. Khan, Y. Al-Harthi and M. Salama, SeerahGPT: Retrieval-augmented generation-based large language model for legal discourse retrieval, IEEE Trans. Artif. Intell. 6(4) (2025) 224–239.
16. R. Kalra, Z. Wu, A. Gulley, A. Hilliard, X. Guan, A. Koshiyama and P. Treleaven, HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications, arXiv:2409.09046 (2024).
17. I. Chalkidis, J. Jan, M. Fergadiotis, P. Malakasiotis and I. Androutsopoulos, LexGLUE: A benchmark dataset for legal language understanding in English, in Proc. 60th Annu. Meet. Assoc. Comput. Linguist. (2022), pp. 4317–4330.
18. N. Pipitone and G. Houir Alami, LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain, arXiv:2408.10343 (2024).
19. Y. Koreeda and C. D. Manning, ContractNLI: A dataset for document-level natural language inference for contracts, arXiv:2110.01799 (2021).
20. Y. Zhu, F. Liu and H. Zhang, Design of knowledge graph retrieval system for legal cases, IEEE Access 10 (2022) 63214–63227.
21. M. Park, E. Kim and K. Choi, LRAGE: Legal retrieval-augmented generation evaluation tool, arXiv:2504.01840 (2025).
22. K. Scholten, Y. Ding and E. Ng, A probabilistic perspective on unlearning and alignment for large language models, arXiv:2410.03523 (2024).
23. J. Zhao, L. Xu and D. Liu, MultiRAG: A knowledge-guided framework for mitigating LLM hallucinations, in Proc. AAAI Conf. Artif. Intell. (2024), pp. 8896–8905.
24. X. Yao, H. Wang and Z. Li, Graph-CoT: Graph-based chain-of-thought for retrieval-augmented reasoning, in Proc. Conf. Empir. Methods Nat. Lang. Process. (2024), pp. 1–14.
25. R. Chowdhury, S. Patel and R. Liu, A novel approach to eliminating hallucinations in large language models via retrieval-based constraints, IEEE Access 12 (2024) 10937570–10937583.

26. M. Reuter, S. Peiris and R. Kohli, Towards reliable retrieval in RAG systems for large legal datasets, arXiv:2510.06999 (2025).
27. L. Fatima, S. Y. Narveen and F. Schneider, Explainability as a legal standard: From reason-giving to recourse in machine decisions, Int. J. Law Policy Sci. Res. 1(1) (2025) 32–40.
28. T. Collenette, G. Cortellessa and T. Bench-Capon, Explainable AI tools for legal reasoning about cases, Artif. Intell. Law 31(2) (2023) 143–161.
29. G. Zhu, M. Hao, C. Zheng and L. Wang, Design of knowledge graph retrieval system for legal and regulatory framework of multilevel latent semantic indexing, Comput. Intell. Neurosci. 2022 (2022) 6781043.
30. Aarhus University, Probabilistic methods in NLP: Representation learning or alignment? PhD Call 2025-17 (Department of Computer Science, Aarhus University, 2025).
31. K. Linna and S. Linna, Human-aligned legal AI systems, Artif. Intell. Law 33(2) (2025) 155–180.
32. B. Wang, T. Li and Q. Huang, Computational reliabilism for AI-based forensic evidence assessment, Forensic Comput. Artif. Intell. Rev. 18(1) (2024) 35–54.
33. B. Ngandu, L. Mbale and A. Bagula, Juro: A retrieval-augmented generation AI chatbot for enhancing legal information access in resource-constrained settings, in Pan-African Artificial Intelligence and Smart Systems Conf. (Springer Nature Switzerland, Cham, 2024), pp. 260–274.
34. J. Park, W. Oh, B. Kim, M. Kim and P. H. Seo, LLMsPark: A benchmark for evaluating large language models in strategic gaming contexts, in Proc. ACL (2025).