iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: https://doi.org/10.22214/ijraset.2026.85064

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue X Oct 2026- Available at www.ijraset.com

A Survey on Retrieval-Augmented and Transformer-Based Approaches for Legal AI: Grounding, Reasoning, and Multilingual Outcome Prediction

Neeraja R A[1], Sherin Shaji[2], Devende M R[3], Ajmalsha S R[4], Mrs. Viji C[5]
[1, 2, 3, 4]*Department of Computer Science (Data Science), St. Thomas Institute for Science and Technology, Trivandrum, India*
[5]*Assistant Professor, Department of Computer Science (Data Science) St. Thomas Institute for Science and Technology, Trivandrum, India*

Abstract: ***Legal artificial intelligence (AI) systems increasingly draw on retrieval-augmented generation (RAG), transformer-based embeddings, multi-stage structured reasoning, and ensemble learning to support tasks such as legal question answering, legal judgment prediction, and multilingual outcome prediction. Hallucination, limited domain grounding, and uneven performance across languages and legal systems remain open problems. This survey synthesizes four recent peer-reviewed studies used as core sources: a recursive-feedback retrieval-augmented generation framework for legal question answering (LQ-RAG); a four-stage large language model (LLM) framework for legal judgment prediction combining knowledge infusion, case-law retrieval, multi-hop reasoning, and generative judgment synthesis (LegalReasoner); a retrieval-augmented generation system evaluated across six embedding-LLM pairings on a small corpus of Indian statutory documents; and a stacked-ensemble system for multilingual legal outcome prediction from case summaries in English, Telugu, Tamil, and Kannada. Rather than treating these four studies as an exhaustive literature base, we use them as core studies and interpret their reported findings critically against their own stated methodology, sample sizes, evaluation metrics, and self-identified limitations, supplementing the discussion with the related work each study itself cites. Across the reviewed studies, we find scope-limited evidence that retrieval grounding, fine-tuned or domain-adapted embeddings, and stacked ensembles improved reported performance relative to the corresponding baselines evaluated within those studies; these improvements are dataset-, language-, and metric-specific and should not be read as a general ranking across architectures or studies. We identify recurring gaps, including small or proof-of-concept datasets, reliance on a proprietary LLM as an automated evaluator, absence of legal-domain-expert human evaluation, narrow language and jurisdiction coverage, and the broader challenge of temporal drift in legal knowledge, and we outline evidence-based directions for future work. This review is presented as a focused technical survey rather than a systematic literature review, and its conclusions are bounded accordingly.***

Keywords: ***retrieval-augmented generation; legal judgment prediction; legal question answering; large language models; multilingual natural language processing; ensemble learning; hallucination mitigation.***

I. INTRODUCTION

Legal decision-making depends on the accurate retrieval of statutes and precedent, correct application of legal reasoning, and transparent justification of outcomes. These requirements make legal text a demanding domain for natural language processing (NLP): legal documents combine domain-specific terminology, long and structurally nested arguments, and jurisdiction-specific conventions that differ across common-law systems, civil-law and statutory systems, and international human-rights bodies. When large language models (LLMs) are applied to this domain without safeguards, they are prone to hallucination - the generation of plausible but unverified or fabricated content, including citations to non-existent cases [3]. In legal practice this is not a cosmetic failure: several reported real-world incidents involve legal filings that cited fabricated case law produced by an LLM [1], and independent audits of commercial legal-AI research tools have reported non-trivial rates of mis-citation even in products that internally use retrieval, as noted within the related-work discussion of one of the core studies reviewed here [3].

Retrieval-augmented generation (RAG), structured multi-stage reasoning, domain-adapted transformer embeddings, and ensemble or stacked-model strategies have each been proposed as partial mitigations. Retrieval grounds a model's output in specific retrieved passages rather than relying solely on parametric memory [1], [3]. Multi-stage reasoning pipelines attempt to decompose a legal prediction task into interpretable sub-steps - for example, knowledge infusion, case-law retrieval, multi-hop reasoning, and judgment synthesis [2]. Domain-adapted embeddings (such as legal-domain or regional-language transformer models) aim to represent legal and multilingual text more faithfully than general-purpose encoders [4]. Ensemble or stacking approaches combine multiple base predictors to reduce the overfitting that individual classifiers can exhibit on limited legal datasets [4]. Each of these strategies is documented, and evaluated to differing degrees of rigor, in the four core studies synthesized in this survey.

A cross-cutting methodological point, corrected here relative to looser framings sometimes used in this space, concerns how legal systems are classified. It is inaccurate to describe the European Court of Human Rights (ECHR) as a "common-law" jurisdiction; the ECHR adjudicates under the European Convention on Human Rights, a treaty-based supranational instrument, and does not map cleanly onto the common-law/civil-law dichotomy used for national legal systems. Throughout this survey, where studies evaluate models across different legal contexts, we describe those contexts concretely - for example, cases before the European Court of Human Rights, criminal cases from the Supreme People's Court of China, and statutory or judicial documents from Indian courts - rather than assigning them to a single, potentially misleading, legal-family label.

This survey differs from a summary of four individual papers in that it organizes their reported methods and results around shared technical themes (retrieval grounding, reasoning, representation, multilinguality, ensembling, evaluation, dataset limitations, and computational cost), explicitly separates each study's own claims from claims that require independent verification, and states plainly where the four studies do not permit a comparison (for example, no study benchmarks its approach directly against another of the four). The survey is organized around the following review questions (RQs), each scoped to what the reviewed material can actually support:

- RQ1: How are retrieval-based approaches used to improve grounding and reduce hallucination in legal AI, and what do the reviewed studies report about the residual limitations of retrieval grounding?
- RQ2: How do structured multi-stage reasoning and knowledge-integration methods support legal judgment prediction, and how much of the reported improvement can be attributed to each stage?
- RQ3: How do transformer-based and domain- or language-specific representations affect legal text processing and outcome prediction across languages?
- RQ4: What is the distinct role of ensemble or stacked multi-model strategies, as opposed to multi-component pipelines or retrieval-generation architectures more loosely described as "ensembles"?
- RQ5: What limitations related to dataset scale, multilinguality, evaluation methodology, computational cost, fairness, and temporal change remain open in the reviewed literature?

These questions are addressed using the four core studies as primary evidence and the literature those studies themselves cite as secondary, contextual evidence. Section II describes the review methodology and its limitations. Section III summarizes related work as reported within the core studies. Sections IV through VIII address RAG, multi-stage reasoning, embeddings, multilingual prediction, and ensembling, respectively. Section IX compares evaluation metrics. Section X compares datasets and computational cost. Section XI presents a cross-study comparative table and synthesis. Section XII discusses limitations and threats to validity, Section XIII outlines research gaps and future directions, and Section XIV concludes.

II. LITERATURE SEARCH AND REVIEW METHODOLOGY

This review is designed as a focused technical survey rather than a systematic literature review. The distinction is important because the scope and evidence base determine how the findings should be interpreted. The review does not claim exhaustive coverage of the legal artificial intelligence (AI) literature, and no PRISMA-style screening procedure or quantitative count of identified, screened, or excluded publications is reported.

The analysis is centered on four peer-reviewed core studies that address complementary aspects of legal AI: retrieval-augmented question answering with a recursive evaluation and feedback mechanism [1]; multi-stage large language model (LLM)-based legal judgment prediction combining knowledge integration, retrieval, reasoning, and judgment synthesis [2]; retrieval-augmented generation evaluated across multiple embedding–LLM combinations for reducing hallucinations in legal guidance [3]; and multilingual legal outcome prediction using transformer-based embeddings and stacked classification models [4]. These four studies form the primary evidentiary basis of the survey and are compared according to their tasks, datasets, languages, architectures, evaluation metrics, reported results, and limitations.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Additional literature discussed in this review is drawn from the related-work and reference sections of the four core studies. These sources are used primarily to provide technical context for topics such as legal-domain language models, retrieval-augmented generation, legal judgment prediction, reasoning approaches, and domain-specific or multilingual embeddings. Because these secondary sources were identified through the core studies rather than through an independent database search, they are treated as contextual evidence rather than as independently screened members of the core evidence set. Their bibliographic details should therefore be independently verified before direct citation in a final submitted manuscript.

The four core studies were selected because they collectively cover the principal technical themes examined in this survey: retrieval grounding, multi-stage legal reasoning, domain-specific representation learning, multilingual processing, and ensemble or stacked prediction. They also provide empirical evaluations with quantitative metrics and sufficient methodological information to support a structured technical comparison.

Because the reviewed studies investigate different tasks and experimental settings, their results are not treated as directly interchangeable. Differences in datasets, legal jurisdictions, languages, prediction objectives, evaluation metrics, and experimental configurations are considered when interpreting reported performance. The comparison therefore focuses on the conditions under which each method was evaluated rather than attempting to establish a universal ranking of the approaches. The review process consists of four main stages: identification of the four core studies, extraction of their methodological and experimental characteristics, thematic comparison across the selected technical areas, and synthesis of common limitations and research gaps. The resulting analysis is intended to identify patterns across the reviewed approaches while keeping the conclusions bounded by the evidence available in the selected studies.

III. RELATED WORK

The related work is organized around the technical themes represented by the four core studies rather than as a paper-by-paper historical review. The discussion covers legal-domain language models, retrieval-augmented generation, legal judgment prediction, domain-specific and multilingual embeddings, and reasoning and evaluation. Where a work is discussed only because it appears in the related-work section of a core study, the discussion is treated as contextual rather than as independently verified evidence.

A. Legal-domain large language models

Recent legal AI research has explored the adaptation of large language models to specialized legal terminology, reasoning patterns, and domain-specific tasks. The related work discussed in [1] and [3] includes several legal-domain models developed through continued pretraining, fine-tuning, or instruction adaptation. Examples include HanFei, LawGPT_zh, LawGPT, LexiLaw, Lawyer LLaMA, SaulLM-7B, Aalap, and InternLM-Law, which have been developed for different legal and linguistic settings. The studies cited in [1] and [3] indicate that domain-specific adaptation can improve the handling of legal terminology and task-specific information. However, domain adaptation does not by itself eliminate reliability concerns. The reviewed literature continues to identify challenges related to hallucination, bias, knowledge cutoffs, and dependence on the quality and coverage of training data [1], [3]. Therefore, legal-domain language models should be considered as components of broader legal AI systems rather than as complete solutions to legal reliability.

B. Legal retrieval-augmented generation systems

Retrieval-augmented generation (RAG) combines information retrieval with language generation so that a model can use externally retrieved information when producing an answer. The related work discussed in [1] and [3] includes several legal RAG systems designed for question answering, case retrieval, legal drafting, and knowledge augmentation.

Examples include DISC-LawLLM, CBR-RAG, LexDrafter, Knowledge-Augmented BERT2BERT (KAB), LawRAG, and an Augmented Question-guided Retrieval (AQgR) approach for Indian case law [1], [3]. These systems differ in their retrieval strategies and in the extent to which retrieval is combined with reasoning, re-ranking, or other processing stages.

The related work also includes general RAG approaches such as SELF-RAG, which incorporates retrieval with self-reflection, and RAG-end2end [3]. Although these approaches are not specific to legal applications, they provide useful technical foundations for legal retrieval-grounded systems.

The reviewed literature suggests that retrieval can improve access to external legal information, but retrieval quality remains dependent on the completeness, relevance, and authority of the underlying knowledge source. Consequently, RAG should be evaluated not only according to the quality of generated text but also according to whether the retrieved evidence actually supports the resulting answer.

C. Legal judgment prediction and neural architectures

Legal judgment prediction (LJP) predates the recent LLM-based literature. Earlier approaches relied on manually designed features and conventional machine-learning methods for predicting outcomes in Chinese criminal cases. Subsequent neural approaches included hierarchical attention networks and TOPJUDGE, an iterative model that retrieves relevant paragraphs from similar cases [2].

Transformer-based approaches have subsequently been applied to legal judgment prediction. The related work in [2] includes hierarchical BERT-based approaches for encoding facts and applicable legal articles in ECHR cases, as well as Legal-BERT, which is pretrained on legal-domain corpora.

More recent architectures summarized in [4] include Bilinear Convolutional Neural Networks and Attention Networks (BCA), Topology-aware Multitask Learning, Modified Hierarchical-Attention Networks, KnowPrompt4LJP, LegalMind, Italian-Legal-BERT models, and POSTURE50K. These approaches address different aspects of legal prediction, including representation of legal facts, multitask learning, prompt-based prediction, event-chain generation, domain adaptation, and extreme multi-label classification [4].

The progression from conventional machine learning to neural and transformer-based approaches reflects the increasing use of richer representations for complex legal documents. However, performance remains dependent on dataset characteristics, task formulation, legal domain, and the amount and quality of available training data.

D. Domain-specific and multilingual embeddings

Embedding models are important in legal AI because they provide numerical representations of legal documents, passages, or sentences for retrieval and downstream prediction. The literature summarized in [4] distinguishes general-purpose multilingual models such as multilingual BERT (mBERT) and XLM-R from models adapted to legal domains or specific languages.

Legal-domain models such as Legal-BERT and its regional adaptations, including InLegalBERT and InCaseLawBERT, are designed to capture legal terminology and semantic relationships. IndicBERT is relevant to multilingual applications involving Indian languages, while FastText uses subword representations that can be useful when vocabulary variation or previously unseen terms are present [4].

The reviewed literature indicates that domain-specific and language-oriented representations can provide advantages for particular legal tasks. However, the magnitude of any improvement depends on the task, dataset, language, and evaluation metric. Therefore, embedding models should be compared within clearly defined experimental settings rather than treated as universally superior or inferior.

E. Reasoning, evaluation, and reliability

Legal reasoning often requires a model to connect facts, legal provisions, precedents, and multiple pieces of supporting evidence. The related work discussed in [3] includes chain-of-thought approaches and other techniques intended to improve structured reasoning in legal tasks.

The reviewed literature also highlights the importance of evaluation methodology. Automated evaluation can provide scalable measurements, but text similarity or language-model-based evaluation does not necessarily establish substantive legal correctness. This is particularly relevant when evaluating generated legal answers, where a response may appear coherent while containing unsupported or incorrect claims.

The studies therefore point toward the need for evaluation procedures that separately consider retrieval quality, factual correctness, legal reasoning, source grounding, and citation accuracy. Human evaluation by legal-domain experts would be particularly valuable for assessing whether generated responses are legally meaningful and adequately supported.

Overall, the related work shows a shift toward combining retrieval, domain adaptation, structured reasoning, multilingual representations, and multi-model strategies. At the same time, the literature remains affected by differences in datasets, jurisdictions, languages, evaluation procedures, and computational resources. These differences must be considered when interpreting and comparing the findings of the four core studies.

IV. RETRIEVAL-AUGMENTED GENERATION FOR LEGAL AI

Retrieval-augmented generation (RAG) combines a retrieval component with a language-generation model. Instead of relying only on information stored in the model parameters, the system first retrieves relevant passages from an external knowledge source and then uses those passages to generate a response.

This approach is particularly relevant to legal applications, where answers often need to be grounded in statutes, regulations, case law, or other authoritative legal documents. The four core studies include two different applications of retrieval-based methods. Gupta et al. [3] investigate a proof-of-concept RAG system for reducing hallucinations in legal guidance, while Wahidur et al. [1] propose Legal Query RAG (LQ-RAG), which extends retrieval with additional processing, evaluation, and feedback mechanisms.

A. A comparatively lightweight pipeline: statutory-document RAG evaluated across six embedding-LLM pairings [3]

Gupta et al. [3] evaluate a RAG-based system using a collection of 17 Indian ordinances and regulations comprising approximately 297 pages. The documents are processed into smaller passages, converted into vector representations, stored in a retrieval system, and subsequently provided to a language model during answer generation.

The study evaluates six combinations of embedding models and language models without additional fine-tuning. Several evaluation measures are reported, including ROUGE, BLEU, Mean Reciprocal Rank (MRR), embedding-based similarity, and BERTScore. Among the reported configurations, the combination of Llama 2 with the all-mpnet-base-v2 embedding model achieved a BERTScore F1 of 56.87% [3].

The study is explicitly presented as a proof-of-concept evaluation and uses a relatively small legal-document collection. Therefore, the reported results provide evidence for the feasibility of retrieval-grounded legal question answering within the evaluated setting, but they should not be interpreted as evidence of general performance across legal domains or larger legal corpora

B. A more elaborate architecture with fine-tuning and a recursive feedback loop [1]

Wahidur et al. [1] propose Legal Query RAG (LQ-RAG), a more elaborate retrieval-based framework for legal question answering. The system combines hybrid retrieval with dense and lexical retrieval, re-ranking, query processing, and an iterative evaluation-feedback mechanism.

The framework also incorporates a legal-domain embedding model and parameter-efficient adaptation of the underlying language model. The authors evaluate retrieval using measures such as Hit Rate and Mean Reciprocal Rank (MRR), and use an automated language-model-based evaluator as part of the feedback process [1].

Compared with a simpler retrieve-then-generate pipeline, the additional processing stages are intended to improve retrieval relevance and response quality. The study reports improvements in Hit Rate and MRR over the corresponding baseline configurations [1]. However, the additional components also increase computational and response-time requirements. The authors report higher response time for the complete LQ-RAG pipeline compared with a simpler RAG configuration.

An important limitation is that the evaluation framework relies on an automated language-model evaluator rather than independent human evaluation by legal experts [1]. Consequently, the reported improvements should be interpreted as improvements under the study's evaluation procedure rather than as direct evidence of improved legal correctness.

C. Synthesis

The two studies demonstrate different levels of complexity in retrieval-grounded legal AI. The statutory RAG study [3] focuses on a relatively small collection of Indian legal documents and evaluates multiple embedding–language-model combinations. LQ-RAG [1], in contrast, introduces additional retrieval, re-ranking, query-processing, and feedback stages.The results of the two studies cannot be directly combined because they use different datasets, tasks, evaluation measures, and experimental configurations. Nevertheless, both studies support the broader observation that retrieval can provide external grounding for legal language-generation systems. At the same time, retrieval alone does not guarantee correct legal reasoning or reliable answers; the quality and coverage of the retrieved sources, the generation model, and the evaluation procedure all influence the final result.

V. MULTI-STAGE REASONING AND LEGAL JUDGMENT PREDICTION

Wang et al. [2] propose LegalReasoner, a four-stage framework for legal judgment prediction (LJP) built on a LLaMA-2 (7B parameter) base model. The framework combines legal knowledge infusion, case-law retrieval, multi-hop reasoning, and judgment synthesis to address the different stages involved in legal outcome prediction.

The first stage performs legal knowledge infusion using contrastive learning to adapt the language model to legal- domain information. The second stage uses a Graph Attention Network (GAT) for case-law retrieval. The graph represents cases and statutes as nodes, with relationships such as citations, shared legal concepts, or similar fact patterns represented through edges. The third stage applies a six-layer transformer with hierarchical attention to perform multi-hop reasoning over the case information together with the retrieved legal materials.

The fourth stage performs judgment synthesis using a generative adversarial network (GAN) consisting of transformer-based generator and discriminator components [2].

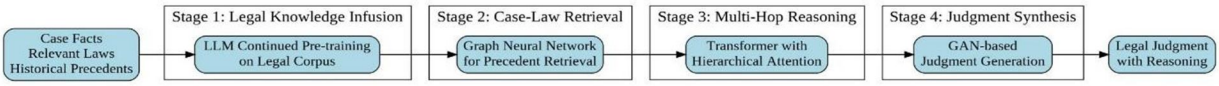


Fig. 1. Overview of the four-stage LegalReasoner pipeline (legal knowledge infusion, case-law retrieval, multi-hop reasoning, and judgment synthesis), reproduced from Wang et al. [2].

LegalReasoner is evaluated on two datasets representing different legal settings: European Court of Human Rights (ECHR) cases and CAIL2018, a large-scale Chinese criminal-case dataset [2]. The ECHR dataset contains 11,478 cases and uses a binary label indicating whether a specific Convention article was violated. The authors use a 9,000/1,000/1,478 training, validation, and testing split. CAIL2018 contains 2,676,075 criminal cases, from which the authors sample 200,000 training cases, 20,000 validation cases, and 20,000 testing cases because of computational-resource constraints [2]. On the ECHR test set, LegalReasoner reports an accuracy of 0.855, macro-F1 of 0.851, and AUC-ROC of 0.914. The corresponding values for the strongest baseline, LEGAL-PEGASUS, are 0.827 accuracy, 0.821 F1, and 0.889 AUC-ROC. On the sampled CAIL2018 dataset, LegalReasoner reports an accuracy of 0.893, F1 of 0.889, AUC-ROC of 0.951, and mean absolute error (MAE) of 5.43 months for prison-term prediction, compared with 0.871 accuracy, 0.867 F1, 0.935 AUC-ROC, and 6.31 months MAE for the reported baseline [2]. These results indicate improvements over the corresponding baselines evaluated in the study. However, the results from the two datasets should not be treated as directly comparable because they involve different legal systems, languages, prediction tasks, and dataset characteristics. The abstract of [2] also reports an average accuracy increase of 7.8%, but this value does not directly reconcile with the individual dataset improvements reported in the results section. Therefore, the 7.8% figure is not used as a standalone result in this survey. The study also reports an ablation analysis of the four stages. Removing individual components results in reductions in the reported performance, indicating that the different stages contribute to the overall framework. However, because the components interact with one another, the results do not establish that any single stage is independently responsible for the complete performance improvement. The model also has substantial computational requirements. The authors report approximately 36.5 hours of training time, around 120 ms of inference time per case, and a model size of approximately 14 GB [2]. Error analysis identifies misinterpretation of facts, incorrect application of law, insufficient context, and complex reasoning as major sources of prediction errors. These findings indicate that multi-stage reasoning can improve prediction performance while still leaving difficulties in interpreting complex legal information. Overall, LegalReasoner demonstrates how legal knowledge integration, retrieval, and structured reasoning can be combined within a single LJP framework. Its reported improvements are specific to the datasets and experimental settings evaluated in [2]. Broader evaluation across additional jurisdictions, languages, datasets, and legal tasks would be required to determine the generalizability of the approach.

VI. EMBEDDING SELECTION AND DOMAIN ADAPTATION

The three studies that use pretrained or fine-tuned sentence or document embeddings [1], [3], and [4] illustrate that embedding performance in legal AI depends on the dataset, language, task, classifier, and evaluation metric. The reviewed studies do not support the conclusion that a single embedding family is universally superior.

In [3], five distinct general-purpose sentence-transformer embedding models (with all-mpnet-base-v2 used in two pairings) are evaluated in six embedding–LLM pairings for retrieval-augmented legal question answering. The evaluated embeddings include all-mpnet-base-v2, paraphrase-mpnet-base-v2, all-MiniLM-L12-v2, paraphrase-MiniLM-L6-v2, and all-distilroberta-v1. The study reports that all-mpnet-base-v2 performs strongly across most evaluation measures, while all-distilroberta-v1 achieves the highest reported ROUGE-1 score. This indicates that the best embedding can vary according to the evaluation metric even within the same experimental setting [3].

In [1], the authors fine-tune a GIST Large Embedding v0 model on synthetic legal query-context pairs to obtain a legal-domain embedding model referred to as GIST-Law-Embed. The model is compared with several untuned embedding models. The authors report that the fine-tuned embedding achieves an average Hit Rate of 51% and an average Mean Reciprocal Rank (MRR) of 40% at top-K = 5, with reported improvements of 13% in average Hit Rate and 15% in average MRR relative to the untuned embedding baseline [1]. The corresponding figure in [1] shows average Hit Rate rising from roughly 45% to 51% and MRR from roughly 34% to 40%, i.e., about 6 percentage points in absolute terms, so the reported percentages are relative gains.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

These results are obtained on the study's in-house information-retrieval benchmark and therefore should be interpreted within that experimental setting.

In [4], four pretrained embedding models—RoBERTa, FastText, InLegalBERT, and IndicBERT—are used to generate representations for multilingual legal outcome prediction. These representations are subsequently evaluated with several classifiers, including Random Forest, Decision Tree, Support Vector Machine, K-Nearest Neighbors, Multilayer Perceptron, XGBoost, LightGBM, and Naive Bayes. The study also uses stacked ensemble models to combine predictions from multiple classifiers [4].

The embedding models represent different approaches to legal text representation. RoBERTa provides a general transformer-based representation, FastText captures subword information, InLegalBERT is adapted to Indian legal documents in English, and IndicBERT is designed for multilingual Indian-language applications [4]. The related work discussed in [4] also includes other legal-domain models such as Legal-BERT and regional adaptations including InCaseLawBERT.

Domain adaptation is particularly relevant to legal AI because legal documents contain specialized terminology, formal expressions, citations, and domain-specific patterns. A model trained primarily on general text may therefore represent legal language differently from a model exposed to legal-domain data. However, domain adaptation does not automatically guarantee better performance across every task or language.

The findings across the reviewed studies therefore support a task-dependent view of embedding selection. The embedding that performs well for retrieval in a legal question-answering system cannot automatically be assumed to perform best for multilingual legal outcome classification. Differences in datasets, languages, legal domains, model architectures, and evaluation metrics prevent direct numerical comparison between the reported embedding results.

Overall, embedding selection should be treated as an important experimental design decision in legal AI. General-purpose embeddings, legal-domain embeddings, and multilingual representations each have potential advantages depending on the application. Future evaluations would benefit from larger legal datasets, consistent benchmarking procedures, native multilingual legal text, and controlled comparisons between general-purpose and domain-adapted embedding models.

VII. MULTILINGUAL LEGAL OUTCOME PREDICTION

Prabhakar et al. [4] study legal outcome prediction from case summaries rather than full judgments. Their dataset (IJJD) contains 600 judgments of the Supreme Court of India and various High Courts (1950–2020) across 11 legal domains, labelled with four outcomes: Appeal Allowed, Appeal Dismissed, Petition Dismissed, and Petition Allowed, whose distribution is reported as approximately balanced. Extractive and abstractive summaries were generated for each judgment because no pre-existing summaries were available; the English summaries were then machine-translated into Telugu, Tamil, and Kannada and spot-checked by legal researchers. The summaries are embedded with RoBERTa, FastText, InLegalBERT, and IndicBERT, used without further fine-tuning [4].

The authors compare languages and summary types using two-sample z-tests over performance scores pooled across embedding–classifier configurations (64 observations per group). For English versus Telugu, mean scores of 50.77 and 44.41 ($z = -4.17$) are reported as a statistically significant difference favouring English, which the authors interpret as a potential bias toward English; for extractive versus abstractive summaries, mean scores of 48.55 and 46.63 ($z = -1.19$, two-tailed $p = 0.23$) are not significantly different [4]. Because these scores pool many configurations, including weak base classifiers, the tests describe the average behaviour of the whole configuration grid rather than the performance of any one deployable system. Among the highest-accuracy metamodel configurations reported, test accuracy is 86.66–91.47% for English and Telugu summaries (Table 5 of [4]) and 85.40–87.20% for Tamil and Kannada (Table 9 of [4]); the authors note that stacking narrows, but does not remove, the English–Telugu gap [4].

The four core studies provide limited but complementary coverage of languages and legal systems. The reviewed evidence includes English-language legal question answering, English and Chinese legal judgment prediction, and multilingual legal outcome prediction involving English, Telugu, Tamil, and Kannada. However, the language coverage remains narrow relative to the diversity of legal systems and languages in which legal AI may be deployed.

LegalReasoner [2] evaluates its framework on English-language European Court of Human Rights cases and Chinese criminal cases. This provides evaluation across two different legal settings and languages, but it does not establish that the framework generalizes across a wider range of legal systems or language families. Differences in legal terminology, case structure, legislation, and judicial practice may affect model performance when the system is transferred to another jurisdiction.

The multilingual study [4] extends legal outcome prediction to English, Telugu, Tamil, and Kannada. The study uses 600 summarized judgments covering 11 legal domains and evaluates several embedding and classification combinations.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

However, the regional-language versions were produced through machine translation of English summaries and were subsequently spot-checked by legal researchers [4]. Therefore, the results demonstrate multilingual evaluation within the study's specific dataset and translation process rather than full validation on independently authored legal documents in each target language.

The two retrieval-augmented generation studies [1], [3] are primarily evaluated using English-language legal material. This creates a gap between retrieval-grounded legal question answering and multilingual legal processing. In particular, the reviewed studies do not provide a comparable evaluation of retrieval quality and generated-answer reliability across multiple Indian languages.

Language diversity also interacts with legal-domain adaptation. A model trained or fine-tuned on legal text in one language may not transfer reliably to another language because legal terminology, syntax, citation conventions, and available training data differ. Multilingual representations such as IndicBERT [4] may provide broader language coverage, but the evidence in the reviewed studies is not sufficient to determine whether multilingual models consistently outperform language-specific or domain-specific representations.

Another limitation is jurisdictional coverage. The core studies focus on selected legal environments, including the European Court of Human Rights, Chinese criminal law, Indian statutory documents, and Indian court judgments [1]-[4]. These settings represent different legal contexts, but they do not cover the full range of common-law, civil-law, mixed, and other legal systems. Consequently, performance observed in one jurisdiction should not automatically be assumed to transfer to another.

Overall, the reviewed evidence indicates that multilingual and cross-jurisdictional legal AI remains insufficiently evaluated. Future studies should use larger datasets containing natively authored legal text in multiple languages, involve legal-domain experts in evaluation, and test the same or comparable systems across different jurisdictions. Such evaluations would help determine whether improvements observed in one language or legal system remain valid when the system is applied in a different linguistic and legal context.

VIII. ENSEMBLE AND MULTI-MODEL STRATEGIES

Prabhakar et al. [4] is the only one of the four core studies that uses ensembling in the strict statistical sense, i.e., combining the predictions of several independently trained classifiers. Eight base classifiers—Random Forest, Decision Tree, SVM, k-NN, MLP, XGBoost, LightGBM, and Naive Bayes—are trained on fixed pretrained embeddings, and their predictions are then used as input features to a meta-learner, for which the same eight algorithms were tried [4]. This differs from the multi-stage pipeline of LegalReasoner [2] and the retrieval–generation loop of LQ-RAG [1], in which different components perform different functions rather than several models making competing predictions on the same task.

The base classifiers overfit. The highest test accuracies reported for base models in the English and Telugu settings are 40.70–45.21% (Table 4 of [4]), and one base configuration reaches 100% training accuracy but only 42.55% test accuracy. To limit overfitting at the meta level, the authors report using out-of-fold predictions from five-fold cross-validation, regularization, early stopping for XGBoost and LightGBM, and shuffling of the data, with hyperparameters tuned by nested grid search [4]. The resulting metamodels reach test accuracies of 86.66–91.47% in the English and Telugu settings with a smaller train–test gap than the base models (for example, 94.10% train versus 91.45% test for the XGBoost metamodel on extractive English summaries) [4]. The authors also report pairwise error correlations of about 0.6 among the tree-based learners and about 0.3–0.4 for the remaining pairs, which they take as evidence that the base classifiers are complementary [4]. The study further compares its results with earlier deep-learning results on the same corpus from the same group (35.62–50.14% accuracy) [4], [6]; whether those baselines used identical inputs and splits is not fully specified.

Several details limit how far this result can be interpreted. The improvement is measured on a single dataset of 600 judgments with a 75:25 train–test split, which implies a test set of roughly

150 cases, yet the confusion matrices in [4] contain 480 (extractive English) and 451 (extractive Telugu) instances; the manuscript does not explain this difference, and its description of shuffling the train and test data does not state how the held-out partition was kept separate from meta-learner training. These points do not show that the reported metamodel accuracies are inflated, but they mean the figures should be treated as provisional until the evaluation protocol is clarified or the experiments are replicated on an independent test set. Within its own evaluation, then, the study supports the narrower claim that stacking reduces the overfitting seen in individual classifiers trained on a small legal dataset, which addresses RQ4 for this dataset only. It does not show that stacking is superior to fine-tuned end-to-end models, or that the gain transfers to larger corpora or to other legal tasks.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

IX. EVALUATION AND BENCHMARKING

A. *Metrics and comparability*

The four core studies use different evaluation metrics because they address different legal AI tasks, including legal judgment and outcome classification, retrieval, and open-ended answer generation [1]- [4]. Table I summarizes the principal evaluation metrics reported by each study.

TABLE I. EVALUATION METRICS REPORTED BY THE FOUR CORE STUDIES.

Study	Task type	Metrics reported
LegalReasoner [2]	Legal judgment / outcome classification and regression	Accuracy, Macro-F1, AUC-ROC, Mean Absolute Error (prison-term months)
Study	Task type	Metrics reported
LQ-RAG [1]	Legal question answering / retrieval / generation	Hit Rate, Mean Reciprocal Rank (MRR), cosine similarity, Answer Relevance, Context Relevance, Groundedness, Accuracy, Exact Match, BLEU, ROUGE
Statutory RAG (PoC) [3]	Legal question answering / retrieval / generation	ROUGE-1/2/L, BLEU, MPNET cosine similarity, BERTScore F1, Mean Reciprocal Rank
Multilingual LOP [4]	Four-class legal outcome classification	Accuracy, F1 score; post-hoc explainability via LIME

The metrics reported across the four studies measure different aspects of system performance and therefore should not be treated as directly comparable. Accuracy and macro-F1 in [2] and [4] measure classification performance, but the underlying tasks differ in label structure, language, and dataset characteristics. For example, LegalReasoner [2] evaluates binary violation prediction on the ECHR dataset and charge classification and prison-term prediction on CAIL2018, whereas [4] evaluates four-class legal outcome prediction. Consequently, an accuracy value from one study cannot be interpreted as directly equivalent to an accuracy value from another study. The retrieval and generation metrics used in [1] and [3] measure different properties from classification metrics. Hit Rate and Mean Reciprocal Rank (MRR) primarily evaluate retrieval effectiveness, while ROUGE, BLEU, and BERTScore assess characteristics of generated responses or their similarity to reference text. These measures therefore cannot be directly compared with classification accuracy or F1-score.

In [1], Answer Relevance, Context Relevance, and Groundedness are evaluated using an LLM-based evaluator. The authors identify the absence of human evaluation by legal-domain experts as a limitation of their evaluation methodology [1]. Therefore, these results should be interpreted as performance under the study's automated evaluation procedure rather than as independent evidence of legal correctness.

The reported numerical results should consequently be interpreted within the experimental setting of each study. This survey does not combine values such as 91.45% classification accuracy in [4] and 56.87% BERTScore F1 in [3] into a single performance measure, because they represent different evaluation objectives. Differences in datasets, languages, legal systems, task definitions, and evaluation procedures further limit direct cross-study comparison.

B. *Reported results within each study*

For Legal Query RAG (LQ-RAG), Wahidur et al. [1] evaluate retrieval using Hit Rate and Mean Reciprocal Rank (MRR). The study also reports answer relevance, context relevance, and groundedness measures. The authors report a 13% improvement in Hit Rate and a 15% improvement in MRR for the fine-tuned legal embedding model compared with the untuned embedding baseline [1]. At the system level, the study reports a 23% relative improvement in average relevance score over a naive RAG configuration (65% to 80%) and a 14% relative improvement over a RAG configuration using the fine-tuned language model without the complete feedback framework [1]. The statutory RAG study [3] uses several automatic text-generation and semantic-similarity metrics, including ROUGE, BLEU, Mean Reciprocal Rank, embedding-based similarity, and BERTScore. Among the evaluated embedding–language-model combinations, the Llama 2 with all-mpnet-base-v2 configuration achieves a reported BERTScore F1 of 56.87% [3]. The study is explicitly presented as a proof-of-concept evaluation using a relatively small collection of 17 Indian ordinances and regulations comprising approximately 297 pages. Consequently, the reported metrics should be interpreted as

evidence from a limited experimental setting rather than as a general measure of legal question-answering performance.

LegalReasoner [2] uses accuracy, macro-F1, AUC-ROC, and mean absolute error (MAE) to evaluate legal judgment prediction. On the ECHR dataset, the framework reports 85.5% accuracy, 85.1% macro-F1, and 0.914 AUC-ROC. On the sampled CAIL2018 dataset, it reports 89.3% accuracy, 88.9% F1, 0.951 AUC-ROC, and an MAE of 5.43 months for prison-term prediction [2]. These results are reported relative to the corresponding baselines in the study.

The multilingual legal outcome prediction study [4] evaluates several embedding and classifier combinations using classification metrics, including accuracy and F1-score. The study reports a maximum test accuracy of 91.47% for the evaluated configurations. However, this result is specific to the study's dataset, language, classifier, and experimental setup. The dataset contains 600 summarized judgments covering 11 legal domains, with regional-language versions generated through translation. Therefore, the reported accuracy should not be interpreted as a general measure of multilingual legal AI performance.

Overall, the evaluation evidence demonstrates that the reviewed approaches are assessed using task-specific metrics rather than a common benchmarking framework. A more reliable comparison of legal AI systems would require shared datasets, standardized evaluation protocols, broader multilingual coverage, and human evaluation by legal-domain experts.

X. DATASET CHARACTERISTICS AND COMPUTATIONAL CONSIDERATIONS

The datasets used across the four core studies differ substantially in scale, language, legal domain, and label structure, which limits direct comparison of their reported results. Table II summarizes the primary datasets and evaluation tasks used in the four studies.

TABLE II. PRIMARY DATASETS USED IN THE FOUR CORE STUDIES.

Study	Dataset	Scale	Language(s)	Task / label
LegalReasoner [2]	ECHR case corpus	11,478 cases (9,000/1,000/1,478 split)	English	Binary: violation of a specific Convention article
LegalReasoner [2]	CAIL2018 (sampled)	200,000/20,000/20,000 sampled from 2,676,075 total cases	Chinese	Charge classification and prison-term regression
LQ-RAG [1]	In-house IR benchmark	1,803 training docs; 6 evaluation sets (123-910 docs each)	English (book-corpus derived)	Passage retrieval (Hit Rate, MRR)
LQ-RAG [1]	QA / instruction fine-tuning corpora	Legal_QA 97,500; Alpaca_cleaned 52,800; plus SQuAD_v2, TruthfulQA, and several small legal-domain sets (Table 2 of [1])	English	Question answering / instruction following
Statutory RAG PoC [3]	17 Indian ordinances/regulations	297 pages	English	Open-ended legal question answering
Multilingual LOP [4]	IJJD case summaries	600 judgments, 11 domains, 1950-2020	English, Telugu, Tamil, Kannada (translated)	Four-class outcome classification

The datasets differ not only in size but also in task definition, language, and legal-domain coverage. LegalReasoner [2] is evaluated on substantially larger case corpora, whereas the statutory RAG study [3] uses a relatively small collection of 17 Indian ordinances and regulations. The multilingual legal outcome prediction study [4] uses 600 summarized judgments covering 11 legal domains, with regional-language versions produced through translation. These differences are important when interpreting the reported performance results.

Computational requirements also vary across the reviewed approaches. LegalReasoner [2] reports approximately 36.5 hours of training, an inference time of about 120 ms per case, and a model footprint of approximately 14 GB. LQ-RAG [1] uses parameter-efficient LoRA adaptation and 4-bit quantization to reduce fine-tuning requirements, but the complete retrieval and feedback pipeline is reported to require approximately 14.6 seconds compared with 7.2 seconds for the corresponding naive RAG configuration in the authors' five-question evaluation. The statutory RAG study [3] uses a small document collection and identifies computational constraints as one reason for its proof-of-concept scope.

These differences illustrate a trade-off between model complexity, retrieval and reasoning stages, computational requirements, and evaluation scope. More elaborate systems may introduce additional retrieval, reasoning, or feedback components, while simpler pipelines may require fewer computational resources. However, the studies use different hardware, datasets, and experimental configurations, so their reported computational costs should not be interpreted as directly comparable measurements.

Overall, the reviewed datasets and computational settings demonstrate that legal AI performance depends not only on the model architecture but also on dataset scale, language, task formulation, and available computational resources. Future evaluations would benefit from larger and more diverse legal datasets together with standardized reporting of training resources, inference latency, model size, and evaluation procedures.

XI. CROSS-STUDY COMPARATIVE SYNTHESIS

Table III summarizes the four core studies side by side. Cells are populated only with information verified from the source papers; "Not reported" indicates the source paper does not provide that information. The table is intended to support technical comparison of methods and scope, not to declare an overall winner across studies that differ in task, language, dataset, and metric.

TABLE III. CROSS-STUDY COMPARISON OF THE FOUR CORE STUDIES.

Study	Task	Legal context	Language	Main method	Retrieval / reasoning component	Embedding	Key metric(s)	Headline reported result	Key limitation (self-reported)
LQ-RAG [1]	Legal QA	General legal corpora (Library Genesis-sourced); not restricted to one jurisdiction	English	Fine-tuned embedding + LoRA-merged generative LLM (HFM) inside a multi-agent RAG loop	Hybrid BM25 + DPR retrieval, re-ranking, ReAct agent, GPT-4 evaluation agent, feedback loop	GIST Large Embedding v0 (fine-tuned)	Hit Rate, MRR, Answer/Context Relevance, Groundedness	+13% Hit Rate, +15% MRR; +23% relative relevance over naive RAG (system)	Relies on proprietary GPT-4 evaluator; no domain-expert human evaluation; closed-domain context relevance/groundedness below 50%
LegalReasoner [2]	Legal judgment prediction	European Court of Human Rights cases; Chinese criminal	English; Chinese	Four-stage LLM pipeline: contrastive knowledge infusion, GNN retrieval,	Graph Attention Network case-law retrieval; multi-hop	LLaMA-2 (7B) with continued pretraining	Accuracy, Macro-F1, AUC-ROC, MAE	Accuracy 0.855 (ECHR) / 0.893 (CAIL2018 subsample); abstract cites	High computational cost (36.5 h train, ~14 GB model); GAN generation head not benchmarked against standard decoding;
Study	Task	Legal context	Language	Main method	Retrieval / reasoning component	Embedding	Key metric(s)	Headline reported result	Key limitation (self-reported)
		cases (Supreme People's Court)		hierarchical-attention multi-hop reasoning, GAN judgment synthesis	transformer reasoning			7.8% avg. accuracy gain (see Note, Section V)	abstract figure requires verification
Statutory RAG PoC [3]	Legal QA	Indian ordinances and regulations	English	Untuned RAG across six embedding-LLM pairings	Single-pass dense retrieval (no re-ranking or agentic loop)	all-mpnet-base-v2 (best pairing, untuned)	ROUGE-1/2/L, BLEU, MPNET similarity, BERTScore F1, MRR	BERTScore F1 56.87% (best pairing)	17-document proof-of-concept corpus; explicitly not generalized by the authors; DOI/date metadata inconsistent in the source document (Section XII)
Multilingual LOP [4]	Four-class legal outcome prediction	Indian Supreme Court and High Court judgments (11 domains)	English, Telugu, Tamil, Kannada (machine-translated)	Stacked ensemble (8 base classifiers + metamodel) over pretrained embeddings	Not applicable (no retrieval component)	RoBERTa, FastText, InLegalBERT, IndicBERT (all untuned)	Accuracy, F1	Test accuracy 91.45-91.47% (best English configuration)	n = 600 dataset; machine-translated (not natively authored) regional-language text; statistically significant English-over-regional-language gap

Read together, the four studies converge on a small number of broad, but individually scope-limited, patterns. Retrieval grounding is reported to improve relevance or reduce unsupported generation relative to un-grounded or less-sophisticated baselines within each RAG study [1], [3], but retrieval reliability itself remains incomplete even in the most elaborate system reviewed, whose own closed-domain groundedness scores stayed below 50% [1]. Structured, multi-stage reasoning components are each reported to contribute positively to legal judgment prediction accuracy in an ablation setting, with domain-specific knowledge infusion contributing the most in the one study that measures this decomposition [2], though the study's headline aggregate accuracy figure has an internal inconsistency that requires resolution. The reviewed studies report performance differences associated with embedding choice and domain or language adaptation, although the magnitude and direction of these differences vary across tasks, datasets, languages, and evaluation metrics [1], [3], [4]. Stacked ensembling substantially mitigates the overfitting exhibited by individual classifiers on a small, real dataset, at the cost of interpretability of any single model's decision process [4]. Because none of the four studies benchmarks its approach against another of the four, and because they use different tasks, languages, datasets, and metrics, no cross-architecture ranking (for example, "RAG outperforms multi-stage reasoning" or "ensembling outperforms fine-tuning") is supported by this evidence base.

XII. LIMITATIONS AND THREATS TO VALIDITY

A. *Limitations of the reviewed studies*

1) Small or proof-of-concept datasets: the statutory RAG evaluation in [3] uses only 17 documents (297 pages), and the multilingual outcome-prediction dataset in [4] contains 600 judgments; both sets of authors explicitly acknowledge these as scale-limited and caution against over-generalizing their results.

2) Reliance on a proprietary evaluator: LQ-RAG's automated evaluation agent is GPT-4, a proprietary, closed-weight model whose scoring behavior is not independently auditable; the authors of [1] identify this as a limitation of their own study.

3) Absence of legal-domain-expert human evaluation: none of [1], [2], or [3] reports evaluation by practicing lawyers or judges; [1] explicitly names the lack of domain-expert feedback as a limitation.

4) Machine-translated, not natively authored, non-English legal text: the Telugu, Tamil, and Kannada data in [4] were produced by machine-translating English summaries and only spot-checked by legal researchers, rather than being independently authored or comprehensively validated in each target language.

5) Narrow jurisdictional and linguistic coverage: LegalReasoner [2] is evaluated on two jurisdictions (ECHR and Chinese criminal law); the multilingual study [4] is restricted to Indian courts and three related South Indian languages; neither generalization across jurisdictions nor across unrelated language families is tested.

6) An architecturally unvalidated design choice: the GAN-based judgment-synthesis stage in [2] is not compared against a standard autoregressive decoding head under otherwise matched conditions, limiting attribution of its measured contribution.

7) An internal numerical inconsistency: the headline "7.8% average accuracy increase" in the abstract of [2] does not evidently reconcile with the per-dataset gains reported in the same paper's results section and requires clarification before being cited as a standalone figure.

8) Inconsistent bibliographic metadata: the PDF of [3] carries two different DOIs (one in the header, 10.1109/ACCESS.2024.0429000, and one in the page footer, 10.1109/ACCESS.2026.3675624), an unfilled publication-date placeholder alongside a "Volume 11, 2023" footer despite an apparent 2025/2026 publication timeline, and should not be cited from this manuscript without checking the authoritative IEEE Xplore record.

9) LLM-based automated text metrics (ROUGE, BLEU, BERTScore) are used in [1] and [3] without accompanying human judgment of answer correctness, and are known in the broader NLP literature to correlate imperfectly with factual accuracy; neither study discusses this limitation of its own metric choice.

B. *Limitations of this survey*

1) The four core studies are not a complete representation of the legal AI literature; no database search or systematic screening was performed, and the survey should not be read as exhaustive.

2) The studies use different datasets, languages, tasks, and evaluation metrics, so no meta-analytic statistical aggregation across them is possible or attempted; the comparative table in Section XI is descriptive, not statistical.

3) Conclusions in this survey are bounded strictly by what the four core studies report and by the evidence they themselves cite; no additional independent experiments, literature search, or expert consultation was performed for this revision.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

XIII. RESEARCH GAPS AND FUTURE DIRECTIONS

The following research directions arise from limitations explicitly reported in the reviewed studies and from gaps identified when the four core studies are considered together.

1) Multilingual, retrieval-grounded legal question answering. Neither RAG study reviewed here ([1], [3]) evaluates retrieval or generation in a language other than English, while the multilingual study [4] evaluates classification, not open-ended retrieval-augmented question answering, in regional languages. Combining these two lines of work

2) — grounded generation in Telugu, Tamil, Kannada, or other under-resourced legal languages — is an open gap directly implied by the current division of labor across the reviewed studies.

3) Human, legal-domain-expert evaluation. The most methodologically elaborate RAG system reviewed [1] relies on a proprietary LLM as its sole evaluator and explicitly identifies the absence of domain-expert feedback as a limitation; incorporating structured evaluation by practicing lawyers or judges would address a gap the authors themselves name.

4) Temporal updating of legal knowledge. Statutes are amended and precedents are overturned over time, and none of the four core studies evaluates model behavior under a temporal shift in the underlying law (for example, retraining or retrieval-index staleness as legal rules change). This remains an important open issue for legal AI systems because legal information can change over time; we note this as a general research challenge rather than attributing an explicit temporal-drift finding to any of the four studies, none of which specifically tested this.

5) Fairness and bias evaluation. LegalReasoner's own future-work discussion explicitly names "ethical and bias considerations" as unaddressed; none of the four core studies reports a bias or fairness audit of its predictions across protected characteristics or across litigant demographics, an omission consistent across all four studies reviewed here.

6) Robustness across legal domains and jurisdictions. LegalReasoner is tested on two jurisdictions and the multilingual study on one country's courts across three related regional languages; cross-jurisdictional and cross-legal-family generalization (for example, applying a model trained on Indian statutory text to a different national legal system) is untested in any of the four studies.

7) Shared, comparable evaluation benchmarks. Section IX shows that the four studies use almost entirely non-overlapping metrics for related but distinct tasks; the field lacks a benchmark that would allow retrieval-augmented generation, multi-stage reasoning, and outcome-classification approaches to be compared on a common basis without conflating incompatible quantities.

8) Retrieval reliability, specifically. Even the most advanced retrieval architecture reviewed here [1] reports groundedness and context-relevance scores below 50% in closed-domain testing; retrieval reliability, not merely retrieval-augmented generation as a category, requires further targeted study.

9) Explainability and citation verification. Post-hoc explainability is applied only in [4] (via LIME, applied to classification predictions, not generation); neither RAG study reports a mechanism for verifying that a generated legal citation actually corresponds to the retrieved source passage, which is precisely the type of hallucination (fabricated case citations) that motivates this entire body of work.

10) Computational efficiency at deployment scale. LegalReasoner's roughly 14 GB model footprint and 36.5-hour training time, and LQ-RAG's near-doubling of response latency relative to a naive baseline, both indicate an unresolved accuracy/cost trade-off; neither study reports deployment-oriented optimization (for example, distillation, caching, or reduced-agent configurations) as an evaluated alternative rather than a stated future intention.

11) Larger, more rigorously validated, and more representative multilingual and low-resource-jurisdiction datasets. Both [3] (17 documents) and [4] (600 judgments, machine-translated regional-language text) are explicitly scale-limited proof-of-concept studies by their own authors' description; expanding dataset scale and using natively authored (rather than machine-translated) regional-language legal text would directly address a limitation both sets of authors state themselves.

XIV. CONCLUSION

This survey has examined four recent, peer-reviewed studies spanning retrieval-augmented legal question answering [1], [3], multi-stage legal judgment prediction [2], and multilingual, ensemble-based legal outcome prediction [4], using them as core evidence rather than as an exhaustive account of legal AI research. Addressing the review questions posed in Section I: retrieval-based approaches are reported to improve grounding and relevance relative to un-grounded or less elaborate baselines, but residual failures in context relevance and groundedness persist even in the most sophisticated system reviewed, indicating that retrieval helps provide external grounding without resolving reliability outright (RQ1).

Structured, multi-stage reasoning components are associated with improved legal judgment prediction performance in the study that evaluates their contributions through ablation. However, the reported results do not establish that any individual component is universally responsible for the overall improvement, and one aggregate accuracy figure stated in the study's abstract requires clarification (RQ2). Transformer-based and domain- or language-adapted representations are associated with improved performance relative to generic or non-adapted alternatives across the reviewed studies, but the size and consistency of this advantage depends on the specific dataset, language, task, and metric, and does not license a claim that any single embedding is universally superior (RQ3). Ensembling in the strict statistical sense, distinct from multi-stage pipelines and retrieval-generation architectures, is evaluated in one study using stacked classifiers and is reported to improve performance relative to individual classifiers on the study's legal dataset (RQ4). Finally, the reviewed literature leaves open a consistent set of limitations: dataset scale, reliance on proprietary and unaudited evaluators, absent legal-domain-expert human evaluation, narrow language and jurisdiction coverage, unaddressed temporal drift in legal knowledge, and an unresolved accuracy/computational-cost trade-off (RQ5).

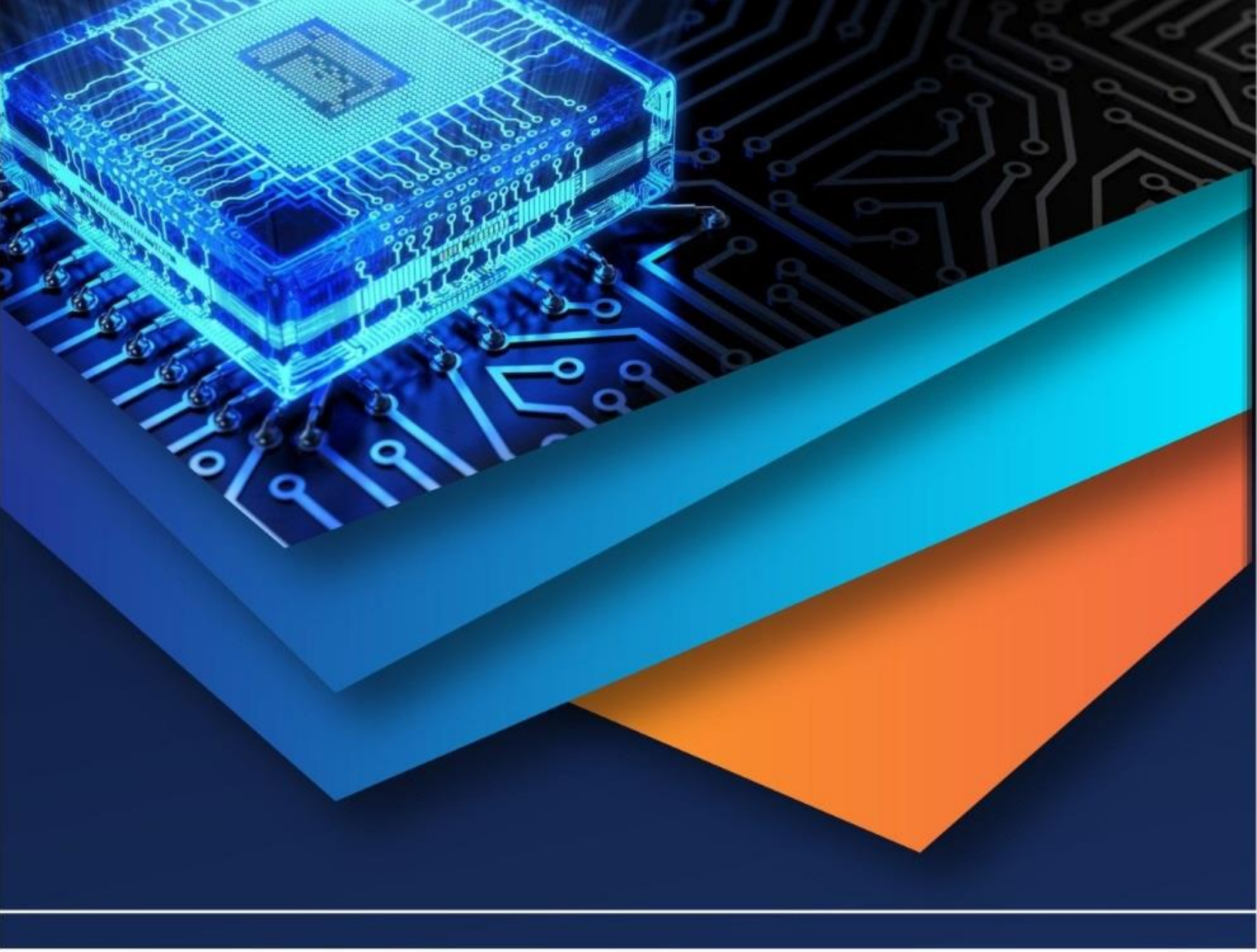
These observations are synthesized from four specific studies and the literature they themselves cite. This pattern is observed across the tasks and languages represented in the reviewed studies, although broader evaluation is required to determine whether it generalizes across legal domains, jurisdictions, and deployment settings. The four approaches should not be interpreted as directly ranked against one another because they address different tasks, languages, datasets, and experimental settings. Future work that closes the gaps identified in Section XIII-particularly multilingual retrieval-grounded question answering, human-expert evaluation, cross-jurisdictional robustness testing, and shared evaluation benchmarks - would strengthen the empirical basis for any future claim about which architectural strategies are most effective for legal AI, and under what conditions.

XV. ACKNOWLEDGMENT

The authors acknowledge the use of Claude (Anthropic) for assistance with drafting, synthesis, language refinement, and organization of portions of this manuscript. The authors reviewed the generated material and are responsible for the accuracy and final content of the manuscript.

REFERENCES

[1] R. S. M. Wahidur, S. Kim, H. Choi, D. S. Bhatti, and H.-N. Lee, "Legal Query RAG," IEEE Access, vol. 13, pp. 36978-36994, 2025, doi: 10.1109/ACCESS.2025.3542125.
[2] X. Wang, X. Zhang, V. Hoo, Z. Shao, and X. Zhang, "LegalReasoner: A multi-stage framework for legal judgment prediction via large language models and knowledge integration," IEEE Access, vol. 12, pp. 166843-166854, 2024, doi: 10.1109/ACCESS.2024.3496666.
[3] S. Gupta, V. Agrawal, Y. Negi, S. Karunanithi, and A. Balakrishnan, "Reducing hallucinations in legal AI: A retrieval augmented generation-based model for accurate legal guidance," IEEE Access, early access, 2026, doi: 10.1109/ACCESS.2026.3675624.
[4] P. Prabhakar, T. Gaddam, D. Chennupalle, and P. B. Pati, "Transformer-based embeddings and stacking classifiers for judgment outcome prediction in multilingual legal texts," Scientific Reports, vol. 16, art. 18305, 2026, doi: 10.1038/s41598-026-47652-x.
[5] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020, pp. 9459-9474.
[6] S. H. Varshini, G. Sarayu Varma, P. Prabhakar, and P. B. Pati, "AI in the courtroom: Enhancing legal decision-making through predictive modelling," Journal of Information & Knowledge Management, 2024.
[7] V. Malik et al., "ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation," arXiv:2105.13562, 2021.
[8] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in English," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 4317-4323.
[9] C. Xiao et al., "CAIL2018: A large-scale legal dataset for judgment prediction," arXiv:1807.02478, 2018.
[10] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," in Findings Assoc. Comput. Linguistics: EMNLP, 2020, pp. 2898-2904.
[11] J. E. Hu et al., "LoRA: Low-rank adaptation of large language models," in Proc. Int. Conf. Learn. Represent., 2021, pp. 1-53.
[12] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv:2307.09288, 2023.
[13] AI@Meta, "Llama 3 Model Card," 2024. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
[14] A. Vaswani et al., "Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 1-11.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)