Journal of Artificial Intelligence and Legal Technology
e-ISSN 3123-786X
Vol. 2, No. 2, August 2026

Toward an Intelligent Legal Education Simulator for Medical Negligence Litigation in India: A Hybrid AI/ML Framework and System Design

Niraja Jain[1,*], Rajeev Kumar[2], Golnoosh Manteghi[3]

[1] MITADT University, Pune; Kuala Lumpur University of Science & Technology (KLUST), Malaysia
[2,3] Moradabad Institute of Technology, Moradabad, Uttar Pradesh, India; Kuala Lumpur University of Science & Technology (KLUST), Malaysia
**Corresponding author. Email addresses: neeraja.jain2002@gmail.com.*

Abstract. Medical negligence litigation in India has grown steadily in volume and complexity, yet no dedicated statute governs it, and legal education rarely offers students practice-oriented training in medico-legal reasoning, pleading drafting, or precedent analysis. This paper addresses that gap by proposing the Intelligent Medical-Negligence Litigation Simulator (IMN-LS), a hybrid AI/ML architecture for legal education that combines retrieval-augmented generation (RAG), legal and clinical named-entity recognition, knowledge-graph representation, and probabilistic evidence reasoning to support explainable, pedagogically scaffolded simulation of medical-negligence cases. The paper makes three contributions: (i) a thematically organised, critical synthesis of legal judgment prediction, explainable-AI, and case-retrieval research that identifies why none of this work addresses medico-legal pedagogy in the Indian context; (ii) a modular system architecture together with an evaluation framework — spanning automated metrics (F1, ROUGE [1], BLEU [2], retrieval precision) and structured human-expert evaluation — designed to support empirical validation in subsequent work; and (iii) a research agenda covering the technical, ethical, and institutional barriers to deployment. This is a conceptual design study: no system has yet been implemented or empirically tested, and the paper reports no performance results. Its contribution lies in the architecture, evaluation framework, and research roadmap themselves, intended to guide the system's future implementation and validation. The work contributes to Intelligent Systems and e-governance research by offering a blueprint for AI-supported legal education in India.

Keywords: Legal-AI; Medical Negligence Litigation; Legal Education Simulator; Retrieval-Augmented Generation; Explainable AI; India.

Received Aug 2026 / Revised Sep 2026 / Accepted Sep 2026

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

INTRODUCTION

Artificial intelligence (AI) — particularly natural language processing (NLP), retrieval-augmented generation (RAG), and hybrid symbolic–statistical architectures — is transforming professional domains, and law is no exception. Legal judgment prediction (LJP), automated case retrieval, and explainable-AI (XAI) decision support have matured rapidly, most visibly for well-digitised, high-resource jurisdictions [3], [4].

Within this broader trend, India presents a distinctive and under-served case. Medical-negligence litigation is rising steadily, driven by growing public awareness, expanding private healthcare utilisation, and evolving jurisprudence on duty of care, standard of care, and patient rights [5]. Yet India has no dedicated statute or section under Bharatiya Nyay Sanhita (earlier IPC) provision for medical negligence: claims are adjudicated through a patchwork of consumer-protection, tort, and occasionally criminal law, applying judge-made standards such as the Bolam test [6] as domesticated in Jacob Mathew v. State of Punjab [7]. Against this fragmented and doctrinally demanding backdrop,

Indian law schools continue to rely mostly on doctrinal, casebook-based instruction, with little practice-oriented training in drafting pleadings, organising clinical evidence, or applying medico-legal reasoning to a live fact pattern.

This paper asks: can modern AI/ML methods support an interactive educational simulator that lets law students practise medical-negligence pleadings against realistic case scenarios — with AI-assisted issue-spotting, precedent retrieval, and feedback on drafting quality — in a way that is explainable and pedagogically appropriate rather than a black-box verdict predictor? We answer in the design: this paper proposes and specifies such a system, the Intelligent Medical-Negligence Litigation Simulator (IMN-LS).

1.1 Research gap

Existing Legal-AI research clusters around three strands: judgment/outcome prediction, case retrieval, and (more recently) explainability. None of these strands, individually or combined, targets medico-legal pedagogy for Indian medical-negligence litigation. Specifically:

- Judgment-prediction systems (e.g., on ILDC;[5]) are built to predict outcomes, not to scaffold a student's construction of an argument.
- Retrieval and case-similarity tools index judgments but rarely fuse legal text with structured clinical facts — diagnosis, treatment timeline, standard-of-care benchmarks — that medical-negligence reasoning depends on.
- XAI research in law [8] addresses transparency and fairness in prediction, but has not been applied to a teaching context where the goal is formative feedback rather than a verdict.
- Recent evaluations of RAG-based legal tools show that grounding reduces, but does not eliminate, citation hallucination [9], [10] — a risk that is amplified, not diminished, in an educational setting where students may not yet be equipped to detect a plausible-looking fabrication.

This gap — the absence of an explainable, clinically-literate, pedagogically-oriented simulator for Indian medical-negligence litigation — is the problem this paper addresses.

1.2 Objectives and research questions

The paper pursues three objectives:

1. To critically synthesise Legal-AI research relevant to medical-negligence litigation and identify, with evidence, why it does not meet the pedagogical need described above (Objective 1).
2. To specify a modular system architecture for IMN-LS — covering document ingestion, knowledge representation, pleading-generation, decision-support, and explainability — appropriate to Indian medico-legal data constraints (Objective 2).
3. To propose an evaluation framework and research agenda that would allow future implementation to be validated technically, pedagogically, and ethically (Objective 3).

These translate into three research questions: (RQ1) What do existing legal judgment-prediction, retrieval, and XAI systems offer — and fail to offer — for medico-legal pedagogy in India? (RQ2) What system architecture can combine retrieval-augmented generation, medical NER, knowledge-graph reasoning, and explainability in a way suited to India's data and doctrinal environment? (RQ3) What technical, ethical, and institutional conditions must be met before such a system could be built and evaluated responsibly?

1.3 Contributions

In summary, this paper contributes: (i) a thematically organised, critical literature review connecting Legal-AI research directly to the identified pedagogical gap; (ii) a proposed modular architecture (IMN-LS) with defined data sources, processing pipeline, and system modules; (iii) a proposed evaluation framework combining automated metrics with structured human-expert assessment; and (iv) a consolidated research agenda addressing the technical, legal-ethical, and operational barriers to deployment. The paper does not report an implementation or empirical results; it is a design and scoping study intended to ground subsequent development and evaluation work.

LITERATURE REVIEW

This section synthesises Legal-AI research along three themes directly relevant to IMN-LS — outcome prediction, explainability, and retrieval — and closes with a critical gap analysis that connects each theme back to the objectives in Section 1.2.

2.1 Legal judgment prediction and outcome modelling

Early legal judgment prediction (LJP) work used classical machine learning or shallow neural networks over case facts; transformer-based models such as LEGAL-BERT [3] and long-context architectures subsequently improved performance on lengthy judgments with multiple structural sections (facts, reasoning, holdings). For India specifically, the Indian Legal Documents Corpus (ILDC) supplies roughly 35,000 Supreme Court judgments with decision and, for a subset, expert-explanation annotations [5], and the recent NyayaAnumana/InLegalLlama resource extends this with the largest Indian LJP dataset and a specialised language model to date [4].

More recent work increasingly pairs retrieval of relevant precedent with classification or generation — retrieval-augmented designs intended to ground predictions in actual jurisprudence rather than allow unconstrained generation [11]. Robustness studies further probe how sensitive such models are to irrelevant facts or spurious correlations [12].

Critical assessment: LJP research demonstrates that Indian judgments can be modelled computationally, and that Indian-specific datasets and models are emerging [4]. However, LJP systems are built and evaluated as outcome predictors, not as pedagogical tools; they offer no scaffolding for a student's own drafting or reasoning, and none of the reviewed systems incorporates structured clinical-fact reasoning of the kind medical-negligence cases require.

2.2 Explainable AI and argument mining

A parallel body of work extracts arguments and reasoning structure from judgments rather than producing bare predictions: argument-mining pipelines parse claims, supporting evidence, and conclusions, while hybrid systems that combine symbolic or probabilistic reasoning (e.g., Bayesian networks) with statistical NLP aggregate evidence and assign interpretable weights to legal elements such as duty, breach, causation, and damage. Systematic reviews of explainable AI in the legal domain report persistent concerns around transparency, fairness, and accountability across the field [8].

A closely related and increasingly urgent strand concerns reliability rather than interpretability: empirical audits of commercial RAG-based legal-research tools find that grounding in retrieved documents substantially reduces, but does not eliminate, hallucinated or unsupported citations [9], [10].

Critical assessment: explainability and argument-mining research is directly relevant to IMN-LS's explainability layer (Section 3.3.5), but it has been developed for legal-research and decision-support use cases, not for formative teaching. The hallucination-reliability literature is a direct design constraint for IMN-LS: it implies that RAG grounding must be treated as risk-reduction, not risk-elimination, and that pedagogical safeguards (Section 5.2.1) are not optional.

2.3 Case retrieval, citation networks, and legal knowledge graphs

A third strand builds retrieval infrastructure and knowledge representations: systems indexing large corpora of judgments and statutes for keyword or semantic search, knowledge graphs of legal concepts and precedent relations, and citation-network analysis that surfaces landmark or heavily-relied-upon cases.

Critical assessment: these tools are necessary building blocks for the case-similarity and knowledge-graph components of IMN-LS (Sections 3.3.2 and 3.3.4), but existing retrieval and knowledge-graph work targets legal text alone. None of the reviewed systems links legal reasoning to structured clinical facts — diagnosis, treatment timeline, standard-of-care indicators — which is precisely the cross-domain reasoning medical-negligence adjudication requires.

2.4 Synthesis: how the reviewed work relates to this study

Read together, Sections 2.1–2.3 show a field that is technically mature for legal text alone — prediction, explanation, and retrieval are all active, improving research areas, increasingly supported by India-specific resources [4] — but that has not been assembled into a tool for teaching medical-negligence litigation in India. Four gaps recur across all three themes and map directly onto the design of IMN-LS: (a) jurisdictional and domain specificity, since most datasets and pretrained models remain skewed toward U.S., European, or Chinese law; (b) absence of cross-domain (legal–clinical)

reasoning; (c) explainability that is not yet oriented toward instructional feedback; and (d) persistent hallucination risk in generative components, even under RAG. Section 3 specifies an architecture designed to close these four gaps for the Indian medical-negligence context specifically, rather than to advance general-purpose LJP or retrieval performance.

PROPOSED SYSTEM: DESIGN AND METHODOLOGY

This section specifies the proposed IMN-LS architecture: its data sources, processing workflow, core modules, and the model-adaptation strategy that would be used to build it. Because no implementation yet exists, the emphasis is on making the design specific and replicable enough that it — and the evaluation framework in Section 4 — could be executed and reported on in a follow-on empirical study.

3.1 Design approach and rationale

A hybrid architecture — combining retrieval-augmented generation, domain-adapted NER, a structured knowledge graph, and rule-informed probabilistic reasoning — is proposed rather than a single end-to-end generative model, for three reasons directly tied to the gaps identified in Section 2.4. First, purely generative approaches carry a documented hallucination risk that is unacceptable in a teaching tool meant to model correct legal reasoning [9], [10]; retrieval grounding and a symbolic knowledge layer constrain outputs to verifiable source material. Second, medical-negligence reasoning is inherently cross-domain (legal and clinical), which a knowledge graph can represent explicitly, whereas a text-only model cannot. Third, explainability — essential for formative feedback — is easier to guarantee when reasoning steps are traceable through a graph and citation trail than when they are implicit in a generative model's weights.

3.2 Data sources

The simulator is designed to rely exclusively on authentic, verifiable, open-access, or institutionally licensed legal and medical data:

3.2.1 Legal data sources

- Supreme Court of India judgments (supremecourt.gov.in) [13]
- High Court judgments (individual High Court portals)
- District Courts / e-Courts platform (ecourts.gov.in) [14]
- National Consumer Disputes Redressal Commission (NCDRC) judgments (ncdrc.nic.in) and State Consumer Commission repositories [15]
- Indian Kanoon (indiankanoon.org) for structured access [16]
- SCC Online and Manupatra (licensed databases) for comprehensive legal research

3.2.2 Medical standards and guidelines

- National Medical Commission (NMC) regulations [17]
- World Health Organization (WHO) clinical guidelines [18]
- National Health Mission (NHM) protocols [19]
- Public-sector and AIIMS clinical guidelines
- MedlinePlus / PubMed Central for medical-procedure reference [20]

3.2.3 Annotation and benchmark datasets

India lacks a ready-made medical-negligence training corpus, so baseline model development would draw on: AILA (AI for Legal Assistance) datasets where available; published Indian legal-NER datasets; and global legal-NLP benchmarks such as COLIEE and CaseHOLD, used for architecture validation rather than domain fine-tuning.

Sampling and access: Given the absence of a purpose-built corpus, the proposed sampling strategy is purposive rather than random — drawing NCDRC and relevant High Court medical-negligence judgments identified by keyword and subject-classification search, cross-checked against Indian Kanoon indexing, until thematic saturation across common negligence scenarios (diagnostic error, delayed treatment, informed-consent failure, surgical error) is reached. This sampling logic, and the resulting corpus, would need to be reported in full in any implementation paper to support replication.

3.3 System workflow and core modules

The simulator follows a sequential, seven-stage workflow: (1) input acquisition of court judgments and related legal-medical documents; (2) document ingestion and pre-processing; (3) knowledge representation in a structured graph; (4) analytical processing — issue identification, pleading generation, case-similarity assessment; (5) decision-support and explainability, backed by transparent reasoning paths; (6) the simulation interface, where students draft pleadings and receive structured feedback; and (7) learning-output analytics comparing learner reasoning against judicial decisions. The overall technical architecture of the simulator is presented in Figure 1.

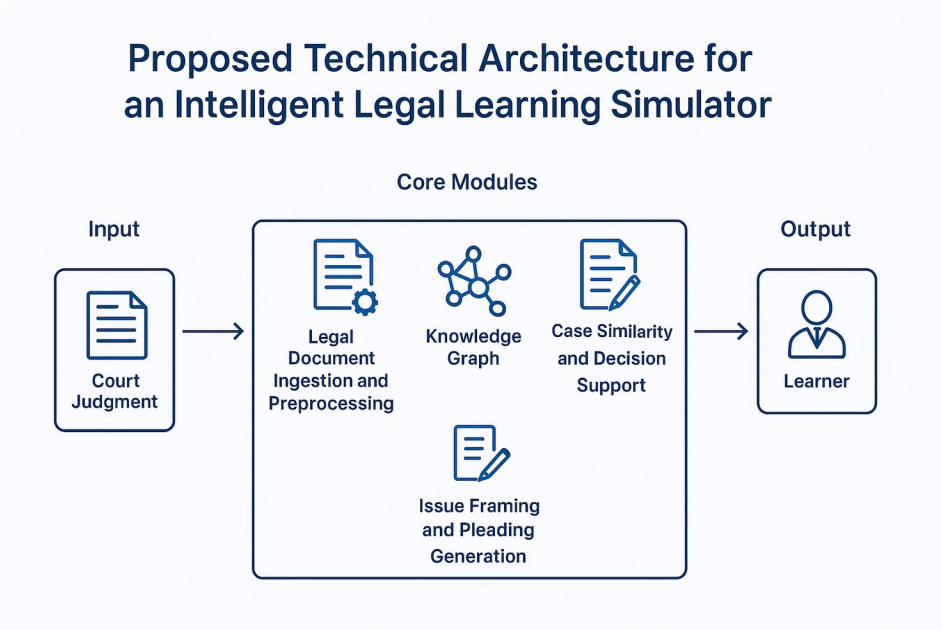


Figure 1. Proposed technical architecture for the Intelligent Medical-Negligence Litigation Simulator (IMN-LS).

3.3.1 Legal document ingestion and pre-processing module

Performs document retrieval, OCR correction for scanned judgments, segmentation into facts/issues/arguments/evidence/findings, Legal Named Entity Recognition (parties, statutes, medical procedures, expert testimony), medical entity extraction via domain-adapted models such as BioBERT [21] or ClinicalBERT [22], and terminology normalisation (e.g., standardising "breach of duty" / "lack of skill" / "deviation from standard of care" to a single canonical label).

3.3.2 Knowledge graph construction module

A Medical-Negligence Knowledge Graph (MNKG) represents patient condition, diagnostic steps, treatment timeline, alleged negligence points, roles of doctors and hospitals, the legal standards of duty/breach/causation/damages, judicial outcomes, and cited precedents. The graph supports case-similarity retrieval, interpretability, knowledge-based reasoning, and explicit linkage between legal facts and medical procedure — directly addressing the cross-domain-reasoning gap identified in Section 2.4.

3.3.3 Issue framing and pleading-generation module

Guides the learner through issue identification from fact summaries, AI-structured notes on argument framing, generation of complaint drafts and written statements following court-style templates, and case-based reinforcement in which suggestions are grounded in retrieved precedent to constrain hallucination. This module is explicitly scoped to provide educational scaffolding, not legal advice.

3.3.4 Case-similarity and decision-support module

Retrieves comparator cases via semantic similarity over transformer embeddings (e.g., LegalBERT, Longformer-legal), MNKG graph similarity, medical-scenario matching (e.g., wrong diagnosis, delayed surgery, lack of informed consent), and identified judicial-reasoning patterns (reliance on expert versus circumstantial evidence), so learners can see how comparable fact patterns were argued and decided.

3.3.5 Explainability and transparency module

Provides highlighted textual evidence, reasoning chains, citation maps linking arguments to precedent, confidence scores, and visual explanations (flow charts, argument trees). This module is the primary mitigation for the hallucination and over-reliance risks discussed in Sections 2.2 and 5.2.1.

3.3.6 Learner simulation interface

Comprises interactive fact sheets, an AI-assisted drafting workspace, a comparative view against real case pleadings, scenario-based quizzes, and an evaluation dashboard tracking improvement over time.

3.4 Model adaptation strategy

Given the scarcity of Indian legal-NLP training data, the proposed adaptation strategy is: (i) domain-adapting general transformer models on available Indian legal and medical corpora; (ii) progressive fine-tuning across layered datasets (general legal text, then medico-legal text, then negligence-specific annotated data); and (iii) retrieval-augmented generation to constrain the pleading-generation and decision-support modules to verifiable source material, informed by the hallucination-mitigation literature reviewed in Section 2.2 [9], [10].

3.5 Validity, reliability, and replicability

Because IMN-LS is not yet built, validity and reliability are addressed here as design commitments to be discharged during implementation, rather than as reported results. Content validity would be established through domain-expert review (law faculty and practising advocates for legal content; clinicians for medical content) of the annotation scheme and knowledge-graph schema before any model training. Inter-annotator reliability would be reported as a Cohen's/Fleiss' kappa on a held-out annotation sample prior to full-corpus labelling. Construct validity for the pleading-generation module would be assessed by comparing AI-assisted student drafts against both AI-unassisted drafts and expert-authored reference pleadings, as specified in Section 4. Replicability requires that the final corpus composition, annotation guidelines, model checkpoints, and evaluation protocol all be published alongside any implementation study — a commitment made explicit here so it can be checked against the eventual results paper.

PROPOSED EVALUATION FRAMEWORK

This section specifies how IMN-LS would be evaluated once implemented. The framework below is the proposed instrument for a subsequent empirical study, provided in enough detail to be executed and audited against.

4.1 Evaluation metrics

Table 1 defines the metrics an implementation study would report, separating automated technical metrics from human-judged pedagogical and legal-validity metrics, consistent with the recommendation in Section 2.2 that automated grounding be verified rather than assumed.

Metric	Definition	Role in evaluating IMN-LS
F1-score	Harmonic mean of precision and recall	Accuracy of issue identification, NER, and clinical-fact extraction

ROUGE	N-gram overlap between generated and reference text	Quality of generated pleading drafts against expert-written references
BLEU	Precision-based text-generation metric	Structural correctness of generated legal summaries
Retrieval precision	Proportion of retrieved cases that are truly relevant	Reliability of precedent suggestions shown to learners
Human expert rating	Structured scoring by law faculty, advocates, and clinicians on usefulness, correctness, and coherence	Pedagogical and legal validity, which automated metrics cannot capture alone
Learning-gain measures	Pre/post assessment of student pleading quality and issue-spotting accuracy	Whether the simulator improves learning outcomes, not only output quality

Table 1. Proposed evaluation metrics for a future empirical study of IMN-LS.

4.2 Evaluation design

A future implementation study would be expected to report: (i) module-level automated metrics (F1, ROUGE, BLEU, retrieval precision) on a held-out annotated test set, with confidence intervals or significance tests (e.g., paired bootstrap resampling) against a defined baseline such as unconstrained LLM generation without retrieval grounding; (ii) blinded human-expert evaluation, with at least two independent raters per item and inter-rater reliability reported; and (iii) a controlled or quasi-experimental comparison of learning outcomes between simulator-assisted and traditional moot-court-style instruction, as proposed in Section 6.4. Reporting statistical significance is only meaningful once (i)–(iii) are executed with real data; it is listed here as a requirement for that later study, not asserted for the present one.

DISCUSSION

This discussion interprets the proposed design in relation to the objectives set out in Section 1.2 and the literature reviewed in Section 2.

5.1 Relation to research objectives and prior work

Objective 1 (critical synthesis) is addressed in Section 2, which shows that LJP, XAI, and retrieval research are individually mature but collectively silent on medico-legal pedagogy — consistent with, and extending, observations in prior systematic reviews of explainable legal AI [8] into the specific context of Indian medical-negligence teaching. Objective 2 (architecture) responds directly to the four gaps identified in Section 2.4 by pairing a legal-text pipeline with a clinical-fact pipeline inside a shared knowledge graph, which no reviewed system does. Objective 3 (evaluation framework) responds to the hallucination-reliability findings of Dahl et al. [9] and Magesh et al. [10] by requiring that any future claim of "grounded" or "safe" generation be backed by the retrieval-precision and human-expert metrics in Table 1, rather than asserted on the strength of RAG architecture alone.

5.2 Implications

5.2.1 Theoretical implications

The design suggests that explainability in a pedagogical Legal-AI system is not a single property but a composite of traceable retrieval, symbolic knowledge-graph reasoning, and confidence signalling — a more granular decomposition than "explainable AI" is typically given in prediction-focused legal-AI work, and one that may be useful beyond this specific application.

5.2.2 Practical implications

For Indian legal education specifically, the design implies that a usable simulator cannot be built from legal text alone: clinical-fact pipelines, domain-expert annotation, and institutional partnership with law schools are prerequisites, not

optional extensions. This raises the resourcing bar for implementation considerably above a typical text-only legal-NLP research prototype.

5.3 Limitations

The central limitation of this study is definitional: it is a design and scoping paper, not an implementation or evaluation. No claims are made, or should be inferred, about IMN-LS's actual accuracy, usefulness, or learning impact — those can only be established once the system in Section 3 is built and evaluated against the framework in Section 4. A second limitation is that the literature review, while thematically organised and critical, is necessarily selective given the volume of Legal-AI publications; it prioritises work most directly relevant to the four gaps in Section 2.4 rather than offering an exhaustive field survey. Third, the data-source list in Section 3.2 identifies where data could come from but does not yet report an assembled, annotated corpus, which remains a significant undertaking given the fragmentation problems described in Section 6.1.

IMPLEMENTATION CHALLENGES AND RESEARCH AGENDA

Building and validating IMN-LS involves technical, legal-ethical, and operational challenges. Each is paired below with a concrete research opportunity, consolidating what would otherwise be a purely descriptive challenges list into an actionable agenda.

6.1 Technical challenges

6.1.1 Limited structured Indian case-law data

Most Indian judgments are published as unstructured PDF or text, without standardised headings or metadata; medical-negligence cases additionally suffer from small sample sizes, dispersed jurisdictional sources, and inconsistent lower-court digitisation.

- Research opportunity: automated case-parsing pipelines for segmentation, metadata extraction, and precedent tagging; annotation schemes aligned with Indian legal-reasoning patterns; open, anonymised medical-negligence datasets to support reproducibility.

6.1.2 Complexity of legal reasoning and domain ambiguity

Medical-negligence reasoning requires interpreting medical standards, expert evidence, burden of proof, and doctrinal tests (Bolam [6]; Jacob Mathew [7]) — multi-step reasoning that exceeds simple NLP classification.

- Research opportunity: hybrid architectures combining LLMs with rule-based and causal reasoning; RAG with domain-constrained retrieval to reduce hallucination [9]; argument-mining techniques to reconstruct issue-wise judicial reasoning.

6.1.3 Evaluating legal-AI output

Standard NLP metrics cannot alone assess legal correctness; evaluation must incorporate legal validity, factual consistency, and citation accuracy.

- Research opportunity: India-specific legal-AI evaluation frameworks; human expert scoring protocols involving law faculty, advocates, and clinicians (formalised in Section 4).

6.2 Legal and ethical challenges

6.2.1 Risk of over-reliance

Students may treat AI output as definitive legal advice; incorrect model behaviour in a high-stakes domain risks misguiding future practitioners.

- Research opportunity: mandatory disclaimers and guided-critique modules; confidence explanations flagging uncertain reasoning steps; safe-design frameworks for law-teaching simulations.

6.2.2 Privacy, confidentiality, and anonymisation

Medical-negligence cases routinely contain patients' personal health information requiring careful anonymisation before use in training.

- Research opportunity: anonymisation tools tailored to Indian medico-legal documents; synthetic-data generation where real patient detail is not essential.

6.2.3 Absence of a dedicated statutory framework

India's lack of a specific medical-negligence statute produces variable judicial interpretation across courts and jurisdictions.

- Research opportunity: models explicitly designed to handle cross-jurisdictional variance; capturing judicial trends as explanatory scaffolds for students.

6.3 Operational challenges

6.3.1 Curricular integration

Adoption requires alignment with existing pedagogical structures — moot courts, clinical courses, blended learning.

- Research opportunity: co-design of training modules with Indian law schools; controlled academic studies of learning outcomes (Section 4.2).

6.3.2 Interdisciplinary coordination

Development requires sustained collaboration among AI researchers, law faculty, practising advocates, medical experts, and data engineers, with attendant terminology and coordination costs.

- Research opportunity: structured interdisciplinary research clusters; shared glossaries and domain ontologies.

6.4 Consolidated research agenda

4. Build a benchmark, anonymised corpus of Indian medical-negligence judgments with structured metadata (issues, statutes cited, medical specialities, precedents applied).
5. Develop Indian-context argument-mining pipelines combining symbolic reasoning with transformer models.
6. Develop explainable-AI techniques specific to legal argumentation and formative feedback, building on general XAI-for-law research.
7. Run a controlled comparison of simulator-based learning against traditional moot-court training, using the evaluation framework in Section 4.
8. Define an institutional ethics framework — evaluation committees, disclosure norms — for deploying AI in legal education.

CONCLUSION

Medical-negligence litigation in India sits at the intersection of a fragmented statutory landscape, growing case volume, and legal training that remains largely doctrinal. This paper has proposed IMN-LS, a hybrid AI/ML architecture intended to give law students explainable, precedent-grounded practice in medico-legal pleading and reasoning, and has shown through a thematic, critical review that existing legal judgment-prediction, retrieval, and explainable-AI research — while individually mature — does not address this pedagogical need. The proposed modular architecture, evaluation framework, and consolidated research agenda together constitute the paper's contribution: a specific, checkable blueprint rather than a general call for AI in legal education. The work reports no implementation or empirical results; its value lies in defining, with enough precision to be executed and audited, what building and validating such a system would require. Realising that blueprint — assembling and annotating a benchmark corpus, implementing the modules specified in Section 3, and running the evaluation in Section 4 — is the

immediate next step, and one with implications extending beyond this single application into legal education, judicial studies, and AI research more broadly.

REFERENCES

[1] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. Text Summarization Branches Out, 74–81.

[2] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318.

[3] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., & Aletras, N. (2020). LEGAL-BERT: The muppets straight out of law school. Findings of EMNLP 2020.

[4] Nigam, S. K., Balaramamahanthi, D. P., Mishra, S., Shallum, N., Ghosh, K., & Bhattacharya, A. (2025). NyayaAnumana and InLegalLlama: The largest Indian legal judgment prediction dataset and specialized language model for enhanced decision analysis. Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025), 11135–11160.

[5] Malik, V., Sanjay, R., Nigam, S. K., Ghosh, K., Guha, S. K., & Modi, A. (2021). ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. arXiv:2105.13562.

[6] Bolam v. Friern Hospital Management Committee, [1957] 1 WLR 582 (UK).

[7] Jacob Mathew v. State of Punjab, (2005) 6 SCC 1.

[8] Joshi, G., Mali, S., & Fegade, P. G. (2024). A systematic review on explainable AI in legal domain. International Journal for Research in Applied Science and Engineering Technology (IJRASET).

[9] Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. arXiv:2401.01301.

[10] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. Journal of Empirical Legal Studies.

[11] Gao, Y., et al. (2024). Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997.

[12] Shui, R., Cao, Y., Wang, X., & Chua, T.-S. (2023). A comprehensive evaluation of large language models on legal judgment prediction. arXiv:2310.11761.

[13] Supreme Court of India. (n.d.). Judgments and Orders. https://supremecourt.gov.in

[14] National Judicial Data Grid. (n.d.). District and Taluka Court Judgments. https://ecourts.gov.in

[15] National Consumer Disputes Redressal Commission. (n.d.). Judgments Archive. https://ncdrc.nic.in

[16] Indian Kanoon. (n.d.). Legal document repository. https://indiankanoon.org

[17] National Medical Commission. (n.d.). Ethics and Medical Practice Regulations. https://www.nmc.org.in

[18] World Health Organization. (n.d.). Clinical Guidelines Repository. https://www.who.int

[19] National Health Mission. (n.d.). Clinical Protocols and Standard Treatment Guidelines. https://nhm.gov.in

[20] PubMed Central. (n.d.). Biomedical and Clinical Research Repository. https://www.ncbi.nlm.nih.gov/pmc

[21] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.

[22] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv:1904.05342.