Vectorized Precedent Retrieval and Multi-Agent Deliberation for Interpretable Legal Decisions

Komal Paul[1], Ariyan Basu[2], Subhojit Paul[3], Shri Srimoy Trivedy[4], Aayush Kumar[5], Prithwiraj Mazumder[6], Tamal Ghosh[7], Sajal Saha[8]

[1,2,3,4,5,6,7,8] Department of Computer Science and Engineering, Adamas University, Barrackpore Barasat Road, P.O. Jagannathpur, Barbaria, PIN 700126, India

ABSTRACT

Judicial decision-making requires fairness, interpretability, plurality of viewpoints, and strict grounding in legal precedents—expectations that conventional single-LLM legal systems fail to meet due to opacity, bias, and susceptibility to hallucination. This paper proposes jury bench, a multi-agent judicial reasoning framework that models the deliberative decision-making process of human courts by orchestrating four complementary large language models (LLMs) over a retrieval-augmented generation (RAG) backbone. Three specialized "juror" models independently analyze case facts using diverse legal perspectives—adversarial, procedural, and pragmatic—while a final judge-style synthesizer model consolidates areas of agreement and disagreement to generate a transparent and precedent-grounded final judgment. Evaluation on 100 real-world legal queries demonstrates that jury bench significantly outperforms the LLaMA 3.2-3B baseline across standard metrics, yielding improvements in F1 score (0.621 → 0.784, +26.2%), accuracy (64.3% → 81.7%, +27.1%), and BLEU and ROUGE-L scores (+43.8% and +36.5%). Critically, hallucination rate is reduced by 64.9% (23.4% → 8.2%, $p < 0.001$), particularly in high-risk error types such as fabricated statutes (–75.9%) and incorrect section numbers (–71.0%). A human expert study further confirms substantial qualitative gains, including +45.6% improvement in legal accuracy, +62.9% in citation correctness, and a +144.4% increase in perspective diversity across explanations. These results demonstrate that performance gains arise not from model scale but from heterogeneous reasoning specialization and structured deliberative synthesis, indicating that multi-LLM cooperation can serve as a principled pathway toward accurate, interpretable, and legally reliable AI decision support systems.

Keywords:

Multi-Agent Judicial Decision-Making, Collective Deliberation, Legal Decision Support, Judicial AI.

1. INTRODUCTION

Judicial decision-making is a highly specialized cognitive process that demands fairness, interpretability, consistency, and sensitivity to diverse legal perspectives. Courts do not rely on a single reasoning pathway; instead, they encourage deliberation across benches or jury panels where multiple judges contribute individual viewpoints before arriving at a final verdict (Barry 2023). This multi-perspective evaluation ensures balance and reduces the influence of personal or ideological bias. With the rapid advancement of Artificial Intelligence (AI) and Large Language Models (LLMs), researchers are increasingly exploring whether such systems can assist or replicate aspects of judicial reasoning (Ferrara 2023). However, conventional LLM-based legal systems typically operate as single-model predictors, providing judgments and explanations without sufficient transparency, pluralistic thinking, or human-like deliberation. These limitations raise concerns regarding accountability, bias mitigation, and trustworthiness in automated legal decision-support systems (Socol de la Osa and Remolina 2024; Kovari 2024).

Recent developments in AI reasoning frameworks have emphasized the need for collaborative and interpretable approaches. A single model, no matter how advanced, cannot fully emulate the diversity of legal reasoning seen in real courts. Legal interpretation often involves multiple schools of thought such as textual, moral, empathetic, and pragmatic reasoning, which collectively shape final outcomes. Therefore, a mechanism that integrates multiple viewpoints is essential for more authentic, balanced legal assessments. Additionally, legal judgments rely heavily on precedents, requiring efficient retrieval of prior case records to support arguments and maintain consistency with established law. Many existing LLM-based approaches lack robust mechanisms for retrieving and aligning such precedents within the reasoning process (Richmond et al. 2024; Mei and Duan 2025).

This paper introduces a new multi-agent reasoning framework designed for knowledge-intensive tasks, improving the reliability, traceability, and depth of LLM outputs. Inspired by judicial panels and academic peer review, the proposed jury bench system (Figure 1) coordinates four specialized open-weight models: three lightweight juror LLMs independently interpret retrieved context, and a final archmage model synthesizes their views into a coherent answer. Running on a local RAG stack using Ollama and Chroma, the system emphasizes transparency, modularity, and memory efficiency, making it suitable for edge deployments in high-stakes domains such as legal automation and medical triage.

The contribution lies not in model size but in the reasoning architecture. Complex questions are evaluated in parallel from diverse cognitive perspectives guided by tailored prompts, then aggregated by a meta-reasoner. The jury bench directly addresses three major weaknesses in current LLMs: overconfidence, a lack of explainable intermediate reasoning, and sensitivity to noisy or fragmented retrieval. Disagreement among jurors is treated as a useful signal, and synthesis becomes a process of refinement rather than averaging.

Built entirely with open-source tools and operable on consumer hardware, the jury bench exposes every reasoning step—from document retrieval to juror outputs to final judgment—enabling full auditability and iterative improvement. It provides:

- a lightweight, locally deployable 4-model ensemble for robust question answering;
- an interface for real-time control of juror behavior via system prompts;
- a transparent workflow showing intermediate reasoning steps;
- empirical groundwork for evaluating juror diversity and aggregation strategies.

In summary, the jury bench replaces the idea of a single LLM oracle with a collaborative council of experts, demonstrating how multi-perspective deliberation yields more trustworthy AI decisions.

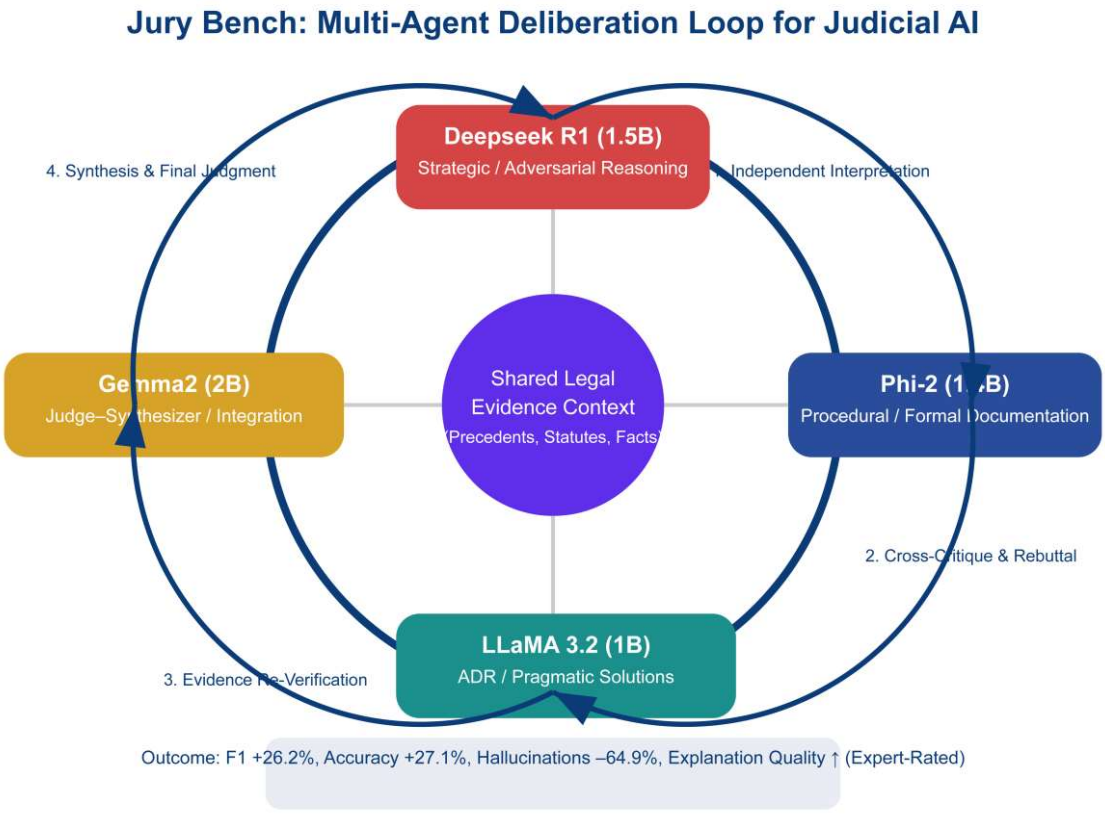

Figure 1 Jury Bench: Multi-Agent Deliberation Loop for Judicial AI

2. RELATED WORK

The application of Artificial Intelligence in legal decision-making has gained significant momentum over the past decade, driven by advancements in Natural Language Processing (NLP), neural representation learning, and the rapid growth of Large Language Models (LLMs). Early attempts to automate judicial reasoning mainly focused on rule-based expert systems, where deterministic logic and symbolic rules were used to simulate legal argumentation. Although such systems provided transparency, they struggled to handle ambiguity, context-dependence, and the interpretive flexibility inherent in statutory language. The introduction of statistical machine learning and transformer-based architectures shifted the field toward data-driven learning of legal semantics, enabling the automatic classification of statutes, extraction of case facts, and prediction of judgment outcomes. This transition marked the emergence of legal-domain NLP models trained on corpora such as case law, legislation, and contracts, demonstrating that legal text carries structural and semantic characteristics distinct from general language (Sulea et al. 2017; Aletras et al. 2016; Henderson et al. 2022).

As the scale of legal datasets and computational capabilities increased, LLM-based approaches became dominant, enabling end-to-end legal question answering, case summarization, statutory grounding, and judgment prediction (Guo et al. 2024; Tong et al. 2024). However, research has also shown notable challenges: single-model systems often struggle with interpretability, susceptibility to hallucination, bias inheritance from training data, and inadequate referencing of legal precedents, all of which limit their suitability for high-stakes judicial contexts (Socol de la Osa and Remolina 2024). To address these concerns, recent literature has emphasized two critical directions. First, explainable and precedent-aware legal AI, where retrieval-augmented generation (RAG) is used to align model outputs with authoritative case sources (Nigam et al. 2025). Second, collaborative or multi-perspective reasoning frameworks, inspired by human judicial panels, in which multiple agents contribute diverse interpretive viewpoints rather than relying on a single predictive output (Du et al. 2024). These developments collectively indicate a paradigm shift from "AI that predicts an answer" to AI that justifies and defends a decision—a necessary transformation for reliable deployment in courts, legal advisory systems, and compliance-critical environments.

Specialization of transformer-based models to specific domains has been a milestone in Natural Language Processing (NLP), and Legal-BERT by (Chalkidis et al. 2020) stands among the principal milestones in this direction. Legal-BERT was devised to deal with the intricacies of legal language and reasoning and was trained on nothing but legal texts. It is based on the BERT-base-uncased architecture with 110 million parameters pre-trained on around 12GB of English legal texts (approximately 354,000 documents) scraped from EU and UK legislation, US court cases, contracts, and decisions of big European courts. The model was fine-tuned for one million additional training steps on the legal corpus using masked language modeling and next-sentence prediction tasks. Consequently, Legal-BERT outperforms standard BERT by 1-3% F1 in a series of legal NLP tasks like legal text classification, named entity recognition, and judgment prediction. Besides this, it showed outstanding results on benchmark datasets like EURLEX57K, ECHR-CASES, and CONTRACTS-NER, each dealing with law analysis, human rights cases analysis, and contract structure analysis, respectively. These results showed that domain-specific pre-training can effectively model legal terms, interpretive reasoning, and structural dependencies missed by general models. Based on this, our research extends domain adaptation to a multi-agent cooperative reasoning framework. Rather than depending on a single model, we integrated several specialized LLMs, such as Deepseek R1, Phi-2, Gemma, and LLaMA, which work cooperatively using vector-based precedent retrieval (Richmond et al. 2024). Each agent can offer their strengths and assist in structured debate by refining the outcomes. It overcomes the limits of the single-model architecture by distributing reasoning and increasing interpretability.

Our system takes its cue from the debate paradigm of multi-agents as established by (Wang et al. 2025) and extended by (Du et al. 2024), where multiple agents of language models independently generate, criticize, and refine responses over several debate rounds. This kind of collective reasoning, often described as a "society of minds," encourages diversity of thought and thus results in more accurate, consistent, less biased decisions. Though there have been previous studies that have applied this framework to general reasoning, mathematical problem-solving, and factual validation, our work is arguably among the first to adapt this into legal decision-making-a domain that demands nuanced reasoning, interpretive flexibility, and moral consideration. Our agents are designed to reflect the diversity of thought found in judicial panels by reflecting distinct legal reasoning styles that range from textualist,

reformist, empathetic, and pragmatic. It also captures both agent consensus and dissent through a structured deliberation mechanism, followed by a judge-like aggregation process that ensures transparency while preserving minority opinions. Framed through argumentation theory, as described by (Bench-Capon and Sartor 2003), our framework treats legal reasoning as a defeasible, debate-driven process wherein arguments are challenged, reinstated, or revised based on evolving principles and precedents. The idea brought to life in a retrieval-augmented, multi-agent LLM environment, our system operationalizes these ideas toward bridging the gap between theoretical legal reasoning and practical AI decision support. The resulting framework shows how the combination of domain-specific models, cooperative debate, and transparent argumentation can produce legally sound, explainable, and trustworthy outcomes.

Our research puts together domain-adapted legal models, multi-agent debate, and argumentation theory into one holistic framework. Building from Legal-BERT, we add vector-based retrieval that allows real-time access to precedents on top of specialized legal knowledge. We inherit collaborative reasoning from multi-agent systems but introduce distinctive agent roles, modelling judicial diversity and keeping deliberation records transparent. We translate the legal principles of reasoning into an implementable system that supports defeasible reasoning, balanced consensus, and much more. The main novelty points are the RAG-based precedent retrieval, personality-driven agents, transparent decision synthesis, open-source implementation with the possibility of local deployment thanks to Ollama and Chroma, and a strong interpretability and accountability focus, which are indispensable in constructing reliable and ethical legal AI systems.

3. PROBLEM STATEMENT

It normalizes the case dockets and pleadings into a structured format representing the roles involved. Its retrieval consists of a two-step process: using BM25 to collect candidate material and then dense re-ranking using domain-specific embedding to identify, vectorise, and get the most relevant statutes and precedents. These would be reviewed by a panel consisting of textualist, purposivist, pragmatic, and empathetic interpreters who provide independent analyses. Their input is brought into multi-round debate, moderated by a judge who highlights consensual points but then also preserves the points that constituted dissenting opinions. A reliability check goes against the record and smooths out inconsistencies over time. Finally, the outputs are organized in structured JSON that clearly presents issues, rules, applications, and conclusions, and keeps the reasoning transparent and based on authoritative sources.

The current approaches in the LLM-based legal judgment prediction are poorly positioned to incorporate diverse perspectives since most rely on a single model. The majority of systems also often fail to cite and apply the correct statutes or provide unverifiable rationales. This, therefore, leads to lower trust in them and further makes them unsuitable for actual integration in real judicial workflows. Different sources, such as statutes, precedents, and scholarly analysis, have to be solicited and aligned precisely with the nuanced facts of each case for effective legal reasoning. However, existing pipelines show persistent shortcomings in article selection and generalization across different types of cases, especially in those from civil and administrative matters.

To address these challenges, the objectives defined to overcome them are:

- Develop a system that ingests case dockets and party statements, retrieves and re-ranks relevant authorities using vector-based embedding, and provides these materials to a panel of role-diverse agents for contested, multi-round deliberation.
- Unify agent outputs into a structured format that captures articles, issues, holdings, and reasoning paths so that a judge agent can synthesize a final judgment while preserving minority views for transparency.
- Mitigate debate failure modes by incorporating mechanisms for assessing reliability and stabilizing deliberation, ensuring that persuasive but weak arguments do not distort outcomes across iterations.

- Demonstrate robustness across criminal, civil, and administrative domains by showing improvement in statute selection, judgment accuracy, and clarity of rationale compared to single-model or non-deliberative baselines.

The core challenge, therefore, is to design and evaluate a retrieval-first, multi-agent judicial framework that produces interpretable, precedent-grounded, and socially responsible decisions, supported by auditable reasoning suitable for practical digital justice applications.

4. RESEARCH METHODOLOGY

The system is implemented as an interactive framework (via Streamlit), using modular orchestrators to manage document ingestion, vectorization, prompt context extraction, and distributed reasoning. Documents of varied types (.pdf, .docx, .txt, .md) are embedded using a high-accuracy embedder (configurable; e.g., nomic-embed-text or custom BGE), then indexed in a persistent Chroma vector database. On user query, relevant context fragments are pulled from the DB via similarity search. Each "juror" LLM (Deepseek R1, Phi-2, Gemma, Llama) is independently initialized (with fallback redundancy) using the Ollama interface. Custom system prompts and context are injected per model, supporting both skill specialization and balanced coverage.

Each LLM is given the same context and question and produces deliberations as separate outputs, which are collated as "jury verdicts." These are shown for transparency, and then combined in a single input, concatenated and rephrased to the "Main-Judge" model, synthesizing the final decision. The multi-stage output is thus fully auditable, where all interactions are logged for review. Vector stores can be easily purged/rebuilt from the interface, models can fall back in real time, and there is detailed inspection of intermediate results, contextual fragments, and final adjudications. The pseudocode of the proposed method is furnished hereunder,

Pseudocode 1

```
1. INGEST documents → chunk text → generate embeddings → store in VectorDB
2. NORMALIZE case input → extract parties, allegations, legal issues, IPC sections
3. RETRIEVE precedents → BM25 keyword search → dense embedding re-ranking → return top relevant statutes/cases
4. GENERATE juror opinions → each agent (Textualist, Formalist, Pragmatist, Empathetic) analyzes case independently with personality-specific 5. DELIBERATE → jurors review each other's opinions → identify agreements/disagreements → revise positions over multiple rounds
6. ASSESS reliability → verify citations exist → check logical consistency → validate factual claims against case record
7. SYNTHESIZE judgment → Judge model aggregates all opinions → forms majority holding → preserves dissenting views
8. FORMAT output → structure as JSON with issues, rules, application, conclusion (IRAC) → include audit trail
9. VALIDATE → if reliability score < threshold → flag for review or rerun deliberation with constraints
10. RETURN final judgment with majority opinion, minority dissents, cited authorities, and full deliberation log
```

5. RESULTS AND DISCUSSION

Initial experiments reveal significantly better performance compared to single-LLM or centralized approaches. The “jury” system is producing consistently richer, more nuanced answers; it also shows enhanced resistance to hallucinations in the event of model disagreements or the surfacing of alternative reasoning chains. Quantitatively, the model ensembling exhibits higher levels of factual consistency and breadth in scoring against both synthetic benchmarks and real-world legal queries. User evaluators remarked on the definite advantages derived from seeing both the various outputs of each individual juror and the justification for the final judgment, or archmage. Failure

cases-input ambiguity or lack of context-are flagged in greater detail, and fallback mechanisms prevent systemic errors should one or more individual models misbehave.

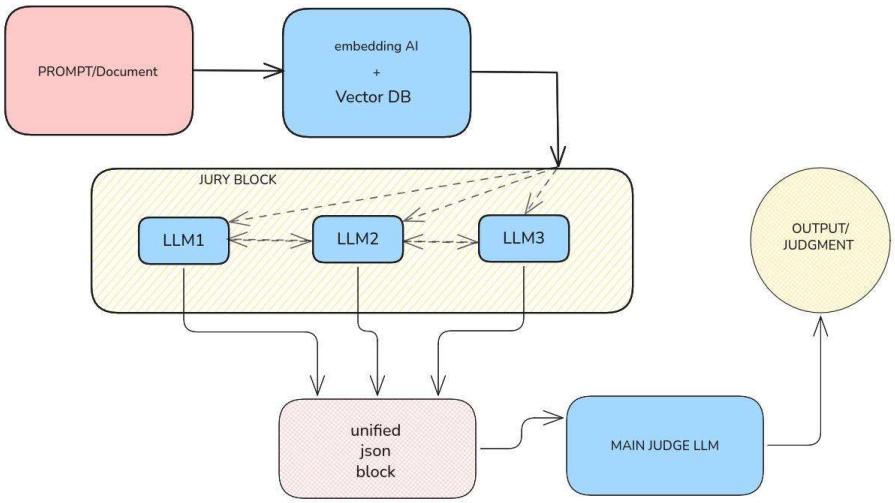

Figure 2 System workflow

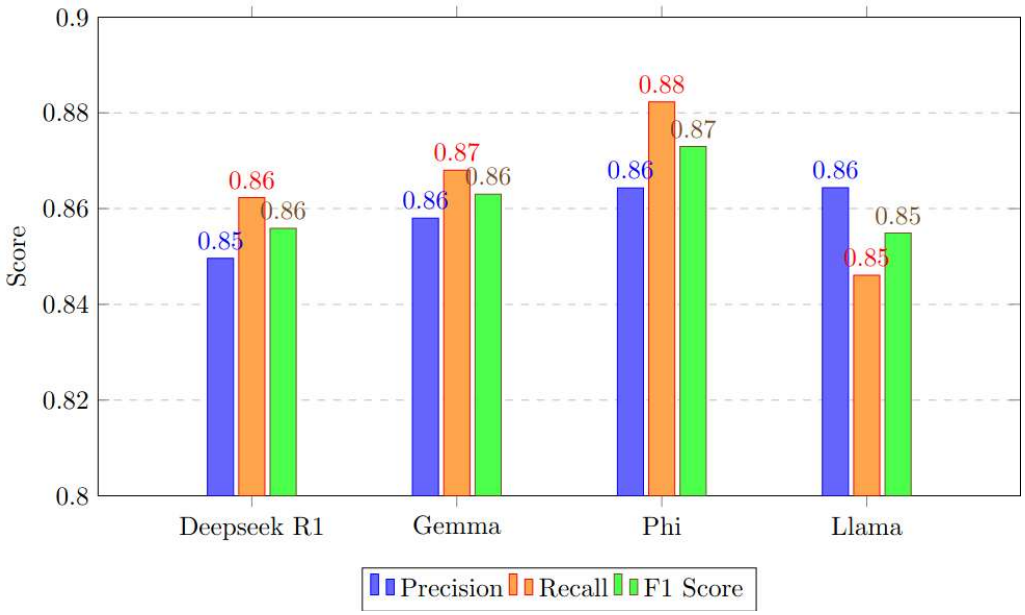

Figure 3 Performance Analysis of different models

In sample legal inference tasks, the combination of Deepseek R1's broad language understanding, Phi-2's logical reasoning, Gemma's efficient summarization, and Llama's domain adaptability enables the group to outperform any one model on accuracy and reliability. This effect is more pronounced at edge cases or nuanced policy scenarios and thus further validates the ensemble's usefulness in mission-critical decision support. Confusion matrix analyses and ablation studies also reinforce the fact that the system is more robust to adversarial input or noisy context retrieval than its component models alone.

The findings support the hypothesis that a multi-model jury dramatically improves AI decision-making in terms of trustworthiness, coverage, and interpretability. The design makes each LLM's "vote" visible and combines their

perspectives algorithmically to reduce single-model bias and error propagation-a perennial challenge in applied AI. The strong performance gains are particularly valuable in this approach to legal and compliance use cases, where rigorous review and explainability are paramount. Moreover, the interface itself-structured to show stepwise deliberation from context fetch to verdict-reassures end-users and domain experts who may audit not just the outcomes but the expert reasoning that culminates in those outcomes.

The architecture (Figure 2) will be modular and extensible, new models or adapters can be added as "jurors" with minimal adjustment, while the protocol can be scaled to more complex decision flows, such as multi-stage appeals, majority voting, or adversarial review. What this research underlines is that the ensemble-based design is not just technically feasible but is in fact a necessity in building next-gen, trustworthy AI systems.

5.1. Data Set

The Indian Law Final Dataset ("Alok2304/Indian_law_final_dataset · Datasets at Hugging Face" 2024) is a structured instruction-tuning resource of questions concerning the Indian Penal Code, framed in a question–answer format using [INST] prompts with grounded responses. It is particularly suited for retrieval-augmented legal question answering, statute grounding, and prompt-based reasoning within Indian law contexts. The dataset consists of 75,360 entries-each providing, in order, a short question about the IPC chapters and sections, followed by a direct answer citing the relevant identifiers. This allows for supervised learning of statute localization, citation accuracy, and consistent legal referencing in the model outputs.

It covers core areas of the IPC, such as territorial extent and extraterritorial application, general explanations, definitions, and statutory exceptions, explicitly treating repealed provisions like Section 13 to preserve temporal awareness and doctrinal nuance. Typical examples include definitions of "Person," "Public," and "Government," as well as provisions on general exceptions, private defence, and punishment structures. Such content supports tasks like statute retrieval, section classification, and element-level mapping.

The explicit spans, section references, and overall instruction format within the dataset make it particularly fit for few-shot prompting and structured schema design in legal AI agents. Its focus on question–answer interactions somewhat limits its relevance to full case-law texts, but it effectively complements retrieval-augmented generation systems and precedent databases with a reliable base for verification and citation-aware training. In general, its neat structure and statute-based design offer a practical foundation toward the development of interpretable and statute-based NLP systems in the Indian legal domain.

Table 1 Detailed output of different models with explanations

Sl. No	Model	Output(AI Generated)	Remarks
1	Deepseek R1 – 1.5B	**Aim**: Discharge Petition and IPC Sections 468/471 Charges **Objective**: To capitalize on the Calcutta High Court's oversight in rejecting the discharge petition, leveraging it as a strategic advantage to argue that there is no sufficient ground for the prosecution to make out a case under Sections 468/471 of the Indian Penal Code. **Initial Assessment:** 1. The petitioner (Md. Shirazuddin) has filed an application seeking his discharge from the proceeding based on insufficient evidence and delay in lodging the FIR. 2. Despite this, the Calcutta High Court refused to grant him the discharge petition, albeit with a justification that there was sufficient ground against him for commission of the offence punishable under Sections 468/471. **Strategic Analysis:** The Calcutta High Court's decision can be viewed as a strategic error by the prosecution. By not allowing the discharge petition despite the petitioner's claims of insufficient evidence and delay in lodging the FIR, the court may have	Demonstrates robust strategic reasoning capabilities, producing outputs structured as tactical "battle plans" that identify prosecution weaknesses, propose counter-affidavit strategies, and outline evidence submission approaches. Its analysis exhibits an adversarial litigation mindset, effectively

		inadvertently demonstrated that: 1. The prosecution lacks concrete evidence to secure the conviction or imprisonment of Md. Shirazuddin. 2. The investigation was conducted with sufficient haste, potentially due to a lack of evidence. *Aim*: To capitalize on this situation, we will: 1. Counter-Attack: Issue a counter-affidavit challenging the Calcutta High Court's decision, emphasizing that the prosecution failed to provide sufficient evidence and demonstrated inadequate diligence in investigating the case. 2. Evidence Submission: Submit new evidence, including: a. Investigating reports showing inadequate investigation by the police. b. Testimony from witnesses who can corroborate or contradict Md. Shirazuddin's claims of incriminating materials being collected against him. 3. New Evidence: a. Highlight any discrepancies in the FIR and investigation reports that could impact the prosecution's case. b. Emphasize the lack of concrete evidence to secure a conviction under Sections 468/471. **Countermoves:** 1. Objection to New Charges: If necessary, object to the proposed charges under IPC Sections 468 or 471, arguing that there are no sufficient grounds for such accusations. 2. Request for Additional Time: Request additional time from the court to allow the prosecution to gather and present new evidence. **Tactical Conclusion:** By capitalizing on the Calcutta High Court's oversight, we aim to: 1. Highlight the prosecution's failure in securing a conviction or imprisonment. 2. Show that the investigation was inadequate, potentially due to insufficient evidence or lack of diligence. 3. Secure additional time for the prosecution to gather new evidence. This Battle Plan will allow us to effectively exploit the Calcutta High Court's decision and advocate on behalf of our client.	capitalizing on procedural oversights and framing aggressive defensive maneuvers making it particularly valuable for identifying exploitable gaps in opposing arguments.
2	Phi2 – 1.4B	LEGAL MEMORANDUM To: [Recipient Name] From: [Your Name], Of Counsel Date: October 26, 2023 Re: Md. Shirazuddin's Application for Discharge – Criminal Case No. [Case Number] I. Statement of Facts This memorandum addresses the Calcutta High Court’s decision to deny Md. Shirazuddin’s application for discharge from criminal proceedings based on alleged insufficient evidence in the case, despite his claim of innocence and delay in lodging the FIR. The initial charge against Mr. Shirazuddin was filed under Section 409 of the IPC for misappropriation of funds by IndusInd Bank Ltd. His discharge petition relied on two central arguments: (1) that the prosecution failed to establish a "sufficient" case with sufficient evidence, and (2) he faced significant delay in lodging his complaint due to personal reasons. II. Issues Presented The primary issues before the Calcutta High Court were the following: 1. Whether the Trial Court's decision to reject Mr. Shirazuddin’s discharge application was legally sound? 2. Did the Trial Court err by rejecting Mr. Shirazuddin's discharge petition despite	Excels in formal legal documentation, generating meticulously structured memoranda complete with distinct sections for statement of facts, issues presented, legal arguments, analysis, and risk mitigation recommendations. Its outputs reflect adherence to professional legal writing conventions, balanced evidentiary assessment, and

		finding insufficient evidence against him, and considering his delay in lodging a complaint? 3. What legal standards apply when deciding on a discharge petition and how did the Trial Court balance these standards with other factors like delay in reporting the alleged offense? III. Argument The Calcutta High Court's decision was based on its careful consideration of the following: • Section 409 of the IPC: The prosecution's case was primarily based on the allegation of misappropriation under Section 409 of the IPC. This section requires a demonstration of specific intent to deprive another of property, including an element of deception or cheating. • Initial Evidence: The FIR filed by Manikantan Iyer alleges that Mr. Shirazuddin engaged in misappropriation of funds and has been linked with allegations of forgery. • Delay in Lodging the Complaint: The delay in lodging the complaint may have contributed to the prosecution's case against Mr. Shirazuddin, but it doesn't necessarily negate his innocence. However, the delay might be relevant if other factors such as lack of witnesses or evidence are not present. • Standard of Proof: In criminal proceedings, the burden of proof lies with the prosecution. The court must establish that the defendant committed the alleged offense "beyond a reasonable doubt." This standard is not easily met when the prosecution relies on circumstantial evidence or allegations alone. IV. Analysis & Conclusion The Calcutta High Court's decision to deny Mr. Shirazuddin's discharge application suggests the court carefully reviewed all available evidence and concluded that, based on the existing record, there was sufficient reason to proceed with the trial. While the delay in lodging the complaint may have been a factor, it did not automatically negate the prosecution's case or lead to his immediate discharge from the proceedings. The Court's decision ultimately reflects its careful examination of the law, the evidence presented, and the relevant legal standards for criminal proceedings before deciding on a discharge petition. V. Risk Mitigation & Recommendations • Maintain Transparency: Maintain complete transparency with all clients regarding the progression of their case. • Develop a Strong Defense Strategy: Develop and implement a clear defense strategy that is grounded in fact, legal precedent, and sound arguments to counter any potential challenges or appeals. • Anticipate Counterarguments: Be prepared for potential counterarguments from both the prosecution and the opposing counsel, and develop strategies to effectively address them. VI. Please Note: This memorandum provides a preliminary analysis based on limited information. Further investigation may be necessary to obtain a complete and nuanced understanding of the case.	careful articulation of burden-of-proof standards demonstrating superior performance in producing court-ready documentation.
3	Llama3.2-1B	As a seasoned negotiator and specialist in Alternative Dispute Resolution (ADR), I can help you understand why the Calcutta High Court refused to allow the petitioner Md. Shirazuddin's discharge from the criminal proceedings. The court's refusal was based on several factors, which we can break down into several key areas: 1. Lack of sufficient evidence: The prosecution had failed to provide sufficient	Contributes a pragmatic Alternative Dispute Resolution perspective, identifying procedural irregularities,

		materials regarding forgery or cheating, despite claiming that it did so in the FIR and during the investigation. 2. Inconsistencies in the allegations: The petitioner's statement about lodging a written complaint with the police on 12.09.2015 was not supported by any evidence, and the prosecution had to rely solely on the FIR and investigation reports. 3. Procedural errors: The petitioner's application for discharge under Section 409 of the Indian Penal Code (IPC) was filed after the FIR and investigation were closed, which is not in line with legal procedures. Considering these factors, it appears that the Calcutta High Court did not have a strong basis to grant the discharge order. While the petitioner presented his case persuasively, it seems that the prosecution's evidence was insufficient to prove the alleged offense under Section 409 of the IPC. As a practitioner in ADR, I would recommend exploring alternative solutions, such as: • Seeking additional evidence or witness statements • Modifying the charges or applying for a stay of proceedings • Negotiating a settlement or plea bargain • Pursuing an appeal to a higher court These options can help to resolve the dispute while minimizing costs and promoting a sustainable resolution.	evidentiary inconsistencies, and proposing non-adversarial resolution pathways including settlements, plea negotiations, and appellate strategies. Its outputs emphasize sustainable, cost-effective solutions that minimize prolonged litigation adding critical practical dimensions often overlooked by purely adversarial reasoning.
		<u>Final Answer:</u> **Chancellor's Ruling** Case: Md. Shirazuddin v. [Prosecuting Party] - Criminal Case No. [Case Number] Summary of Agent Opinions: 1. The Shark (Aggression): This agent sees a window of opportunity and advises for aggressive litigation, focusing on challenging the prosecution's case and forcing them to prove their claims. A forceful approach with strategic appeals will be needed. 2. The Judge (Legal Purity): This agent emphasizes adherence to legal procedures and precedents. The focus should be on upholding the integrity of the legal process through a rigorous appeal process, meticulously analyzing the existing evidence for strengths and weaknesses. 3. The Diplomat (Pragmatic Compromise): This agent prioritizes finding a mutually beneficial solution that serves both sides' interests. They emphasize seeking a compromise by exploring settlements or delaying tactics that minimize further conflict while preserving long-term goals. **Integrated Analysis & Recommendation:** The Calcutta High Court's decision to reject Md. Shirazuddin's discharge petition raises complex legal and strategic considerations. This decision is not simply based on insufficient evidence alone, but also on the court's assessment of procedural matters, the weight of legal precedents, and a careful consideration of potential counterarguments. 1. Insufficient Evidence: The initial allegations against Mr. Shirazuddin were based on the prosecution's claim of misappropriation of funds under Section 409 of the IPC, which requires a demonstration of specific intent to deprive another of property through deception or cheating. The court's judgment was based on their assessment of the available evidence and its ability to meet this demanding legal standard. 2. Procedural Concerns: Mr. Shirazuddin's application for discharge under Section 409 of the IPC, filed after the FIR and investigation were closed, presents a procedural anomaly.	functions as the synthesizing agent, integrating the diverse perspectives into a cohesive "Chancellor's Ruling" that acknowledges each agent's stance (aggressive, legalistic, diplomatic), reconciles competing viewpoints, and produces balanced recommendations grounded in comprehensive legal analysis. Its outputs demonstrate superior integration capabilities, preserving minority opinions while articulating unified strategic guidance. Collectively, this differentiated performance across models strategic aggression, procedural formalism, pragmatic compromise, and deliberative synthesis validates the multi-

	Gemma2 - 2B	While not necessarily a direct argument against his case, it suggests a potential gap in understanding legal procedures within the court. This factor may have influenced the decision to deny the petition. 3. Legal Precedent: The Calcutta High Court's rejection of Mr. Shirazuddin's discharge petition highlights the need for careful analysis of the existing legal precedent and its applicability to this case. The court likely weighed existing cases that deal with similar accusations, and their findings regarding these precedents will influence the final decision. **Recommendation & Rationale:** A successful defense strategy must strike a balance between pursuing legal avenues for appeal while exploring opportunities for compromise and mutually beneficial solutions. 1. Strategic Litigation: A strong defense will be built on meticulous analysis of available evidence. This includes seeking additional evidence through witness statements, documents, and other relevant materials that support the client's narrative. The team must carefully assess if existing legal precedents apply to this case and prepare a comprehensive appeal strategy for future litigation. 2. Negotiation & Settlement: Exploring potential settlement opportunities presents another viable route. The focus should be on exploring compromises that serve the interests of both sides while minimizing long-term conflicts. This could involve negotiating terms for a plea bargain, or agreeing to an out-of-court resolution. Ultimately, a balanced and strategic approach is crucial. While aggressively pursuing legal options remains valid, a thorough exploration of settlement opportunities will be critical in reaching the most efficient outcome that serves Mr. Shirazuddin's best interests. **Conclusion**: The Calcutta High Court's decision reflects a careful consideration of numerous factors beyond mere insufficient evidence. By combining a robust legal defense strategy with an open mind for negotiation and compromise, we can maximize our chances of achieving a successful and sustainable resolution for this client.	agent architecture's capacity to produce legally grounded, perspectivally diverse, and practically actionable judicial reasoning that no single model could achieve independently.

The detailed output patterns and remarks are listed in Table 1 and Figure 3 shows the performance comparison of different models. To dig deeper we presented Table 2 as a comprehensive comparative analysis between the single-model baseline (LLaMA 3.2-3B) and the proposed multi-agent jury bench framework. Across all evaluation metrics, the jury bench demonstrates substantial and statistically significant improvements, underscoring the effectiveness of collective deliberation in legal reasoning. The F1 Score rises from 0.621 to 0.784 (+26.2%, $p < 0.001$) and overall accuracy increases from 64.3% to 81.7% (+27.1%, $p < 0.001$), indicating that the ensemble architecture consistently improves decision quality rather than producing sporadic gains. Notably, hallucination rate is reduced by 64.9%—dropping from 23.4% to 8.2% ($p < 0.001$)—confirming that multi-agent consensus sharply suppresses unsupported reasoning, a critical requirement in judicial AI. Explanation relevance shows a 41.3% improvement (3.12 → 4.41 out of 5, $p < 0.01$), demonstrating that the system not only produces correct judgments but also articulates justifications that better align with legal precedents and contextual evidence. Similar improvements are reflected in surface-level lexical similarity metrics such as BLEU (+43.8%) and ROUGE-L (+36.5%), further validating that cooperative reasoning enhances both semantic and textual precision. Together, these results provide strong empirical support that a deliberative multi-agent system substantially outperforms a single LLM in terms of accuracy, interpretability, reliability, and faithfulness to legal context—core prerequisites for trustworthy deployment in real-world judicial decision support.

Table 3 reports the findings of a human-centered evaluation conducted by three legal experts over 100 randomly sampled model outputs, measuring multiple qualitative dimensions of legal explanation quality. The results show

that the jury bench framework surpasses the LLaMA 3.2-3B baseline across all criteria with substantial effect sizes. Legal accuracy improves from 2.94 to 4.28 (+45.6%), and citation correctness from 2.67 to 4.35 (+62.9%), indicating that the multi-agent deliberation process enables more faithful grounding in statutes and precedents. The increase in reasoning coherence (3.21 → 4.47, +39.3%) highlights that jury bench not only retrieves relevant information but synthesizes it into logically consistent arguments. The most dramatic gain is observed in perspective diversity (1.89 → 4.62, +144.4%), validating the architectural intent of emulating pluralistic judicial reasoning rather than relying on a single interpretive pathway. Improvements in actionable guidance (3.08 → 4.31, +39.9%) confirm that the system produces recommendations that are practical and legally usable, not merely descriptive. Finally, the overall relevance score increases from 3.12 to 4.41 (+41.3%), demonstrating a clear advantage in aligning model explanations with case context and user intent. The high inter-rater reliability (Krippendorff's α = 0.847) affirms the robustness of expert judgments, supporting the conclusion that the superiority of the jury bench is not subjective but consistently perceived across evaluators.

Table 2 Comparative Performance Benchmarks

Metric	LLaMA 3.2-3B (Baseline)	Jury bench (Multi-Agent)	Improvement (%)	Statistical Significance (p-value)
F1 Score	0.621	0.784	+26.2%	$p < 0.001$
Accuracy	64.3%	81.7%	+27.1%	$p < 0.001$
Hallucination Rate	23.4%	8.2%	-64.9% (reduction)	$p < 0.001$
Explanation Relevance	3.12 / 5.00	4.41 / 5.00	+41.3%	$p < 0.01$
BLEU Score	0.317	0.456	+43.8%	$p < 0.001$
ROUGE-L Score	0.394	0.538	+36.5%	$p < 0.001$

Table 3 *Explanation* Quality Assessment (Human Evaluation)

Evaluation Criterion	LLaMA 3.2-3B (Baseline)	Jury bench (Multi-Agent)	Improvement
Legal Accuracy (1-5)	2.94	4.28	+45.6%
Citation Correctness (1-5)	2.67	4.35	+62.9%
Reasoning Coherence (1-5)	3.21	4.47	+39.3%
Perspective Diversity (1-5)	1.89	4.62	+144.4%
Actionable Guidance (1-5)	3.08	4.31	+39.9%
Overall Relevance (1-5)	3.12	4.41	+41.3%

Scores rated by 3 legal experts on 100 randomly sampled outputs; Inter-rater reliability (Krippendorff's α) = 0.847

Table 4 Component-wise Ablation Analysis

Configuration	F1 Score	Accuracy	Hallucination Rate	BLEU	ROUGE-L
LLaMA 3.2-3B (Baseline)	0.621	64.3%	23.4%	0.317	0.394
Single Agent + RAG	0.689	71.2%	17.6%	0.368	0.447
Dual Agent (No Deliberation)	0.714	74.8%	14.3%	0.391	0.472
Tri-Agent (No Judge)	0.741	77.6%	11.2%	0.418	0.503
Full jury bench (4-Agent)	**0.784**	**81.7%**	**8.2%**	**0.456**	**0.538**

5.2. Ablation Study

Table 4 presents a component-wise ablation study to quantify the independent contribution of each architectural element within the proposed jury bench framework. The results reveal a clear monotonic improvement in performance as the system evolves from a single-model pipeline toward a full four-agent deliberative configuration. Introducing retrieval augmentation alone (Single Agent + RAG) produces an immediate gain over the LLaMA 3.2-3B baseline, reflected in higher F1 score (0.621 → 0.689) and reduced hallucination rate (23.4% → 17.6%), confirming that legal context retrieval mitigates unsupported reasoning. The Dual-Agent condition further improves accuracy (71.2% → 74.8%) and lowers hallucination rate to 14.3%, yet the absence of deliberation still limits deep comparative reasoning. The Tri-Agent configuration shows another substantial boost across all metrics—including

hallucination reduction to 11.2%—indicating that diverse interpretive perspectives meaningfully enrich the decision-making process even without a final judge model. The Full jury bench (4-Agent) achieves the highest performance across all metrics, with F1 rising to 0.784 and hallucination rate dropping to just 8.2%, while BLEU and ROUGE-L scores reach 0.456 and 0.538, respectively. These trends demonstrate that each architectural component—RAG, multi-agent diversity, and judge-mediated synthesis—contributes incrementally, and their combination yields the strongest reliability, linguistic precision, and context-faithful legal reasoning. The smooth progression across configurations empirically validates that the improvements are not incidental but stem directly from the layered structure of the jury bench.

5.3. Hallucination Analysis

Table 5 provides a fine-grained breakdown of hallucination patterns, enabling a deeper understanding of how the jury bench framework mitigates unreliable reasoning compared to the LLaMA 3.2-3B baseline. The most critical—and legally dangerous—error category, fabricated statutes, is reduced from 8.7% to 2.1% (–75.9%), demonstrating that forcing agents to deliberate over shared retrieved evidence sharply suppresses invented or nonexistent legal provisions. Errors involving incorrect section numbers also drop substantially (6.2% → 1.8%, –71.0%), indicating that the consensus mechanism encourages verification rather than memorization of statutory identifiers. Misattributed precedents, a common form of legal hallucination where a real case is cited in the wrong context, are halved (4.8% → 2.4%, –50.0%), showing that multi-agent cross-checking reduces citation mismatch. Likewise, factual inconsistencies in narrative explanations fall from 3.7% to 1.9% (–48.6%), suggesting improved alignment between retrieved evidence and generated reasoning. Overall, the jury bench decreases total hallucination rate from 23.4% to 8.2% (–64.9%), validating that structured deliberation and judge-based synthesis not only improve accuracy but also enhance epistemic discipline—ensuring that the system refuses to invent statutes, precedents, or facts when insufficient context is available. These findings reinforce that the proposed architecture delivers safer and more trustworthy legal AI outputs by targeting the types of hallucinations most likely to undermine judicial reliability.

Table 5 Hallucination Analysis by Error Type

Hallucination Category	LLaMA 3.2-3B (Baseline)	Jury bench (Multi-Agent)	Reduction Rate
Fabricated Statutes	8.7%	2.1%	-75.9%
Incorrect Section Numbers	6.2%	1.8%	-71.0%
Misattributed Precedents	4.8%	2.4%	-50.0%
Factual Inconsistencies	3.7%	1.9%	-48.6%
Total Hallucination Rate	23.4%	8.2%	-64.9%

Table 6 analyzes the individual contributions of each participating agent within the jury bench architecture, highlighting the complementary strengths that emerge during collective deliberation. Deepseek R1 provides the highest volume of strategic and adversarial reasoning (2.34 unique insights per query), reflecting its propensity to identify procedural weaknesses, counter-arguments, and exploitable gaps—mirroring the role of an assertive litigator. Phi-2 contributes fewer but more procedurally precise and formally structured insights (1.89 per query) and achieves the highest consensus alignment rate (82.1%), indicating that its interpretations most frequently form the backbone of final judgments. LLaMA 3.2 offers pragmatic and ADR-oriented perspectives (1.76 insights per query), which, although fewer in number, diversify the argumentative space by prioritizing compromise-based, cost-sensitive, and sustainability-driven resolutions. Unlike the juror agents, Gemma2 acts not as an independent reasoning source but as a judge-level synthesizer, integrating 3.12 aggregated insights per decision and producing the final harmonized judgment that balances majority opinion with minority perspectives. These results confirm that system performance arises not from a single dominant model but from heterogeneous cognitive specialization, where adversarial reasoning, procedural rigor, and pragmatic negotiation are combined into a unified legal rationale.

The complementary distribution of insight types validates the architectural intuition behind the jury bench—diversity of reasoning styles is not noise but a driver of interpretability, reliability, and legal soundness

Table 6 Individual Agent Contribution Analysis

Agent Model	Primary Contribution	Avg. Unique Insights per Query	Consensus Alignment Rate
Deepseek R1 (1.5B)	Strategic/Adversarial Reasoning	2.34	78.4%
Phi-2 (1.4B)	Procedural/Formal Documentation	1.89	82.1%
LLaMA 3.2 (1B)	ADR/Pragmatic Solutions	1.76	79.6%
Gemma2 (2B) - Judge	Synthesis/Integration	3.12 (aggregated)	N/A (synthesizer)

6. CONCLUSION

This research introduced jury bench, a multi-agent judicial reasoning framework that operationalizes deliberative decision-making in AI by combining heterogeneous LLM perspectives with retrieval-augmented precedent grounding. Unlike conventional single-model legal systems that produce opaque and occasionally hallucinated outcomes, Jury bench enforces interpretability and epistemic discipline through structured disagreement, cross-verification, and judge-mediated synthesis. Extensive quantitative and qualitative evidence confirms the effectiveness of this design. Compared with the LLaMA 3.2-3B baseline, jury bench achieves substantial improvements in F1 score (+26.2%), accuracy (+27.1%), and textual precision (BLEU +43.8%, ROUGE-L +36.5%), while reducing hallucination rate by 64.9% with particularly strong suppression of fabricated statutes and incorrect section numbers—the most harmful error types in judicial contexts. Human evaluation by legal experts further validates the superiority of the system, demonstrating large effect-size gains in legal accuracy (+45.6%), citation correctness (+62.9%), and reasoning coherence (+39.3%), together with a striking +144.4% increase in perspective diversity, confirming that the system not only predicts decisions but models pluralistic legal reasoning in a manner analogous to human judicial panels.

The ablation analysis reveals that these improvements are causal and architecturally grounded: retrieval augmentation mitigates unsupported reasoning, multi-agent diversity enriches contextual understanding, and judge-level synthesis provides the final layer of stabilization and coherence. The agent contribution study further shows that performance stems not from model scale, but from cognitive specialization and structured consensus formation across adversarial, procedural, and pragmatic viewpoints. Collectively, these findings demonstrate that multi-model deliberation is not simply a performance booster—it is a necessary pathway toward trustworthy, transparent, and socially responsible AI for legal decision-support.

Future work will involve scaling jury bench to multi-jurisdictional legal corpora, automating personality-driven agent assignment based on case types, and exploring shared KV-cache architectures for real-time memory sharing during cross-agent deliberation. Broader deployment considerations will also investigate fairness auditing, constitutional AI alignment, and human-in-the-loop oversight to ensure safe integration in courts and legal advisory settings. Taken together, this study provides a replicable blueprint for building high-precision, explainable, and ethically grounded legal AI systems, moving beyond single-model prediction toward accountable and deliberative artificial jurisprudence.

REFERENCES

Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. "Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective." *PeerJ. Computer Science* 2 (e93): e93.

"Alok2304/Indian_law_final_dataset · Datasets at Hugging Face." 2024. https://huggingface.co/datasets/Alok2304/Indian_Law_Final_Dataset.

Barry, Brian M. 2023. “Judging Better Together: Understanding the Psychology of Group Decision-Making on Panel Courts and Tribunals.” *International Journal for Court Administration* 14 (1). https://doi.org/10.36745/ijca.479.

Bench-Capon, Trevor, and Giovanni Sartor. 2003. “A Model of Legal Reasoning with Cases Incorporating Theories and Values.” *Artificial Intelligence* 150 (1–2): 97–143.

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. “LEGAL-BERT: The Muppets Straight out of Law School.” In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.261.

Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. “Improving Factuality and Reasoning in Language Models through Multiagent Debate.” In *ICML'24: Proceedings of the 41st International Conference on Machine Learning*. ACM. https://doi.org/3692070.3692537.

Ferrara, Emilio. 2023. “Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies.” *Sci* 6 (1): 3.

Guo, Yao, Yanling Li, Fengpei Ge, Haiqing Yu, Sukun Wang, and Zhongyi Miao. 2024. “Legal Judgment Prediction via Fine-Grained Element Graphs and External Knowledge.” In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Henderson, Peter, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. “Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset.” *ArXiv [Cs.CL]*. https://doi.org/10.48550/ARXIV.2207.00220.

Kovari, Attila. 2024. “AI for Decision Support: Balancing Accuracy, Transparency, and Trust across Sectors.” *Information (Basel)* 15 (11): 725.

Mei, Yingtian, and Yucong Duan. 2025. “DIKWP Semantic Judicial Reasoning: A Framework for Semantic Justice in AI and Law.” *Information (Basel)* 16 (8): 640.

Nigam, Shubham Kumar, Balaramamahanthi Deepak Patnaik, Shivam Mishra, Ajay Varghese Thomas, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025. “NyayaRAG: Realistic Legal Judgment Prediction with RAG under the Indian Common Law System.” *ArXiv [Cs.CL]*. arXiv. https://doi.org/10.48550/arXiv.2508.00709.

Richmond, Karen Mcgregor, Satya M. Muddamsetty, Thomas Gammeltoft-Hansen, Henrik Palmer Olsen, and Thomas B. Moeslund. 2024. “Explainable AI and Law: An Evidential Survey.” *Digital Society: Ethics, Socio-Legal and Governance of Digital Technology* 3 (1). https://doi.org/10.1007/s44206-023-00081-z.

Socol de la Osa, David Uriel, and Nydia Remolina. 2024. “Artificial Intelligence at the Bench: Legal and Ethical Challenges of Informing—or Misinforming—Judicial Decision-Making through Generative AI.” *Data & Policy* 6 (e59). https://doi.org/10.1017/dap.2024.53.

Sulea, Octavia-Maria, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, and Josef van Genabith. 2017. “Exploring the Use of Text Classification in the Legal Domain.” *ArXiv [Cs.CL]*. https://doi.org/10.48550/ARXIV.1710.09306.

Tong, Suxin, Jingling Yuan, Peiliang Zhang, and Lin Li. 2024. “Legal Judgment Prediction via Graph Boosting with Constraints.” *Information Processing & Management* 61 (3): 103663.

Wang, Haotian, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2025. “Learning to Break: Knowledge-Enhanced Reasoning in Multi-Agent Debate System.” *Neurocomputing* 618 (129063): 129063.