An Experimental Assessment of AI-Based Legal Decision-Making Systems in Contract Analysis and Risk Detection

Tatiana Suplicy Barbosa[1], Douglas de Castro[2], Anand Kumar Singh[3], Salvatore Vitale[4]

[1]Faculty of Law, Centro Universitário FAEL (UniFAEL), Lapa, Paraná, Brazil
[2]School of Law, Lanzhou University, Lanzhou, China
[3]National Law University Jodhpur, India
[4]Law School, Leonardo da Vinci University, Chieti, Italy

ABSTRACT: This comprehensive experimental study evaluates the performance, reliability, and practical applicability of AI-based legal decision-making systems in contract analysis and risk detection. Utilizing a corpus of 5,247 contracts with expert-validated annotations from 12 legal professionals, we benchmark four classes of AI systems—rule-based, supervised machine learning (XGBoost), fine-tuned transformer models (Legal-BERT), and large language models (GPT-4, Claude 3)—across multiple dimensions critical to legal practice. The Key Findings of this research : a) Performance Variability: Fine-tuned Legal-BERT achieved the highest overall clause classification F1-score (0.923, 95% CI [0.917, 0.929]), but exhibited significant degradation in cross-jurisdictional applications (28.4% performance drop from US to UK contracts). b) Risk Detection Gaps: All systems demonstrated decreasing recall with increasing risk severity. GPT-4 missed 18.2% of high-severity risks (severity ≥4), while Legal-BERT missed 12.3% of total risk severity weight (FNRP metric). c)Decision Inconsistency: LLMs showed substantial inconsistency, with GPT-4 achieving only 0.81 intra-model Jaccard similarity across identical inputs and 14.7% decision variation on identical clause phrasings. d)Domain-Specific Performance: Rule-based systems performed adequately on standardized agreements (NDA: F1=0.812) but failed catastrophically on complex contracts (M&A: F1=0.432). e) Cost-Effectiveness: Local fine-tuned models provided 92.3% of GPT-4's performance at 3.5% of the cost ($0.0087 vs $0.2478 per document).We introduce two novel legal-specific metrics—False-Negative Risk Penalty (FNRP) and Severity-Weighted F1 (SwF1)—that better capture the asymmetric cost structure of legal errors. Based on our empirical findings, we propose a three-tier human-in-the-loop deployment framework that reduces attorney review time by 64% while maintaining 99.7% risk coverage. The study establishes evidence-based performance thresholds for safe deployment, recommending against autonomous use of any system with FNRP > 0.15 or cross-jurisdiction performance degradation > 25%. Our findings challenge optimistic claims of AI autonomy in legal decision-making and provide a rigorous, reproducible framework for evaluating legal AI systems in practice.

Keywords: artificial intelligence in law, legal decision-making, contract analysis, risk detection, empirical evaluation, large language models, legal nlp, performance metrics, deployment framework.

I. INTRODUCTION

1. BACKGROUND AND MOTIVATION

The legal industry stands at a critical juncture. On one hand, corporate legal departments face unprecedented operational pressures: the average Fortune 500 company now reviews over 1,247 contracts

monthly with teams of just 3-5 specialized lawyers, creating a 72-hour average turnaround time that delays business operations, increases costs, and creates significant bottlenecks in procurement, mergers and acquisitions, and compliance processes [1]. Our survey of 42 Fortune 500 companies revealed that 92% of legal departments report being "understaffed relative to workload," with contract review constituting 34% of total lawyer hours.

Simultaneously, the market for AI-powered legal technology has exploded. The total global market for contract lifecycle management software reached $2.1 billion in 2023, with AI-powered solutions representing the fastest-growing segment at 34% year-over-year growth . Vendors increasingly promise "autonomous contract review" and "AI-driven risk detection" that can supposedly match or exceed human lawyer performance[2], [3].

However, beneath this promise lies a critical problem: there exists a profound disconnect between AI research benchmarks and real-world legal practice. Most AI systems are evaluated on narrow technical metrics (F1-scores, accuracy) without sufficient consideration of legal context, error asymmetry, or practical deployment constraints. The stakes for error in legal contract review are exceptionally high—missing a single critical clause can result in millions in liability, regulatory penalties, or loss of intellectual property rights [3]. A 2022 study by the Corporate Legal Operations Consortium found that 23% of companies experienced "significant contractual losses" due to inadequate review processes, with an average loss of $2.8 million per incident[4].

This study emerges from this critical tension between operational necessity and risk management. We aim to provide the comprehensive, rigorous evaluation that the legal profession urgently needs—one that measures AI performance against standards meaningful to legal practice, not just computational benchmarks.

2. RESEARCH GAP

Our analysis of 127 peer-reviewed studies on legal AI from 2018-2023 reveals significant and systematic gaps in current research. These gaps are not merely academic but have profound practical implications for deployment decisions:

- Limited Dataset Scope: 89% of studies use synthetic or limited-domain datasets (<1,000 contracts). For example, the widely cited CUAD dataset contains only 510 contracts focused primarily on software agreements [4]. This fails to represent the diversity of real-world contracts across industries, jurisdictions, and complexity levels. The result is that systems may perform well on standardized NDAs but fail catastrophically on complex M&A agreements or industry-specific contracts.
- Inappropriate Evaluation Metrics: 94% of studies report only standard NLP metrics (precision, recall, F1) without legal error weighting. This ignores the fundamental asymmetry of legal errors where false negatives (missing risks) are typically 10-100× more costly than false positives [5]. A system with 90% F1 but that misses the 10% most critical risks would be unacceptable in practice, yet this distinction is lost in standard metrics.
- Lack of Practical Validation: Only 2 studies systematically assess cross-jurisdictional robustness or decision consistency across identical inputs [6]. Legal practice is increasingly global—a contract drafted under English law differs substantially from one under Delaware law—yet most AI evaluations treat jurisdiction as irrelevant.
- Absence of Deployment Frameworks: No prior study provides evidence-based thresholds or practical deployment protocols validated against real legal workflows [5]. The transition from "benchmark performance" to "safe deployment" requires specific guidance that accounts for error types, costs, and human oversight requirements.

These gaps create a dangerous disconnect between AI research and legal practice. Systems may perform well on academic benchmarks but fail in critical real-world scenarios, potentially leading to catastrophic legal and financial consequences.

3. RESEARCH OBJECTIVES

This study aims to bridge these critical gaps through comprehensive experimental evaluation with four specific, measurable objectives:

3.1 Objective 1: Quantitative Performance Benchmarking

To measure and compare AI system performance across 8 contract domains (NDA, MSA, SOW, M&A, Employment, Licensing, Procurement, Real Estate) and 3 jurisdictions (US/Delaware, UK, Singapore) using both standard NLP metrics and novel legal-specific metrics.

3.2 Objective 2: Risk Detection Analysis

To quantify false-negative rates across 5 severity levels and analyze patterns in missed risks, with particular attention to high-severity (≥4) risks that pose the greatest financial and legal exposure.

3.3 Objective 3: Practical Applicability Assessment

To evaluate systems on dimensions critical to real-world deployment: explainability (quality and legal soundness of rationales), consistency (output stability across identical inputs), processing speed (throughput and latency), and cost-effectiveness (total cost of ownership).

3.4 Objective 4: Evidence-Based Framework Development

To establish minimum performance thresholds for safe deployment and develop a practical human-AI workflow validated through efficiency metrics and risk coverage analysis.

4. CONTRIBUTIONS

This paper makes four primary contributions that advance both academic research and legal practice:

4.1 A Comprehensive, Expert-Validated Benchmark Dataset

We provide a curated corpus of 5,247 contracts spanning 12 domains and 3 jurisdictions, with 1.2 million clause-level annotations and 38,471 risk labels validated by a panel of 12 legal experts. Unlike prior datasets, ours includes:

- Real-world procurement agreements from a multinational corporation (anonymized)
- Public filings from EDGAR with diverse agreement types
- Professionally drafted synthetic contracts for controlled cross-jurisdictional comparison
- Expert annotations with detailed severity scoring and rationale documentation

4.2 Legal-Specific Evaluation Metrics

We introduce and validate two novel metrics:

- False-Negative Risk Penalty (FNRP): Measures the proportion of total risk severity weight that was missed, weighted by severity (higher severity = higher penalty).
- Severity-Weighted F1 (SwF1): Modifies standard F1 to weight true positives by severity, penalizing misses of high-severity risks more heavily.

These metrics have been correlated with actual dispute outcomes using historical data from 1,842 contract disputes.

4.3 Empirical Performance Analysis Across System Generations

We provide detailed, comparative results across four AI system classes representing the evolution of legal AI: rule-based (symbolic), supervised ML (feature-based), fine-tuned transformers (Legal-BERT), and large language models (GPT-4, Claude 3). Our analysis goes beyond aggregate metrics to identify specific strengths, failure modes, and trade-offs that inform practical deployment decisions[6], [7].

4.4 Practical Deployment Framework with Evidence-Based Thresholds

Based on empirical findings, we propose a three-tier human-in-the-loop workflow that reduces attorney review time by 64% while maintaining 99.7% risk coverage. We establish evidence-based performance thresholds (e.g., FNRP < 0.15 for limited autonomous use) and provide a complete implementation guide including cost analysis, scalability considerations, and governance protocols.

4.5 Paper Organization

The remainder of this paper is organized as follows:

- Section 2 establishes the legal-theoretical foundation, discussing the unique characteristics of legal contracts, risk typology, and implications for AI modeling.

- Section 3 reviews related work across four system classes, identifying specific gaps that this study addresses.
- Section 4 details our research design and experimental framework, including dataset composition, annotation protocol, system configurations, and task definitions.
- Section 5 defines our evaluation metrics, explaining both standard NLP measures and our novel legal-specific metrics.
- Section 6 presents experimental results for clause analysis, including accuracy, domain-specific performance, error patterns, and alignment with expert judgments.
- Section 7 presents results for risk detection and decision-making, analyzing false-negative patterns, severity-based performance, cross-jurisdictional robustness, and decision consistency.
- Section 8 assesses explainability features and practitioner trust through expert evaluation and survey data.
- Section 9 synthesizes findings, discusses implications for legal practice, and addresses regulatory and ethical considerations.
- Section 10 proposes a practical deployment framework with detailed workflow specifications, governance protocols, and cost analysis.
- Section 11 addresses threats to validity and study limitations.
- Section 12 concludes with summary contributions, empirical impact, and future research directions.

Appendices provide complete documentation including annotation guidelines, prompt templates, statistical tables, and sample analyses.

II. LEGAL CONTRACT ANALYSIS AND RISK THEORY

1. THE STRUCTURAL COMPLEXITY OF LEGAL CONTRACTS

Legal contracts are not unstructured text but sophisticated semantic constructs with intricate architecture. Understanding this architecture is essential for designing effective AI systems. Contracts operate through several interlocking systems[8], [9]:

Hierarchical Structure: Contracts follow a document → section → clause → sub-clause hierarchy, but this is not merely organizational. Higher-level provisions (e.g., "Definitions" sections) establish terms that control interpretation throughout the document. A clause's meaning often depends on its position within this hierarchy—a "Limitation of Liability" clause in the general terms has different implications than one in a specific service schedule.

Defined Terms and Cross-Referential Systems: Contracts create their own internal dictionary through defined terms (terms beginning with capital letters, such as "Confidential Information," "Effective Date," "Services"). These terms may be referenced dozens of times throughout the document. More importantly, defined terms often have nested definitions: "Confidential Information" might be defined as "any information marked as confidential," which then depends on the definition of "marked" (does email subject line count?) and "confidential" (is orally disclosed information included?). AI systems must track these definition chains across potentially hundreds of pages[10], [11], [12].

Conditional Logic and Dependencies: Contractual obligations are typically conditional rather than absolute. Consider: "If Party A fails to deliver the Deliverables by the Delivery Date, and such failure is not cured within 30 days of written notice, then Party B may terminate this Agreement upon written notice." This single sentence contains multiple nested conditions, temporal dependencies, and procedural requirements. The "Delivery Date" might be defined elsewhere as "the later of June 1, 2024 or the date all Specifications are approved." AI systems must follow these dependency chains to assess risk.

Ambiguity Sources and Intentional Vagueness: Legal ambiguity arises not only from poor drafting but often from strategic choices. Common ambiguity types include[13], [14]:

- Vagueness: Terms like "reasonable efforts," "material adverse effect," or "industry standard" are intentionally vague to allow flexibility.
- Syntactic Ambiguity: "The Seller shall deliver the goods to the Buyer on the pallets with the documentation" (does "with the documentation" modify "deliver" or "pallets"?).
- Referential Ambiguity: Pronouns or incomplete references ("such obligation," "the foregoing," "herein").
- Temporal Ambiguity: "Within 30 days" (calendar days? business days? from when?).

In our corpus of 5,247 contracts, 37% of clauses contained at least one ambiguous term requiring legal interpretation beyond literal text analysis. This finding has profound implications for AI systems, which typically assume text means what it says.

2. LEGAL RISK TYPOLOGY AND COST STRUCTURE

Our risk taxonomy emerged from analyzing 1,842 actual contract disputes provided by a litigation analytics firm, supplemented by interviews with 47 practicing lawyers. The taxonomy captures not just risk types but their frequency, detectability, and financial impact:

Table 1. Legal risk taxonomy with empirical data.

Risk Category	Subtypes	Avg. Dispute Cost	Detection Difficulty	Frequency in Corpus
Financial Exposure	Unfavorable payment terms, uncapped liabilities, currency risk, price adjustment clauses, automatic renewal	$285,000	Medium	34.2%
Compliance & Regulatory	GDPR violations, SOX non-compliance, export control issues, industry-specific regulations (HIPAA, GLBA)	$1,200,000	High	18.7%
Liability & Indemnity	Asymmetric indemnification, exclusion of consequential damages, limitation of liability caps, warranty gaps	$3,400,000	Very High	22.4%
Termination & Dispute	Unilateral termination rights, ambiguous force majeure, unfavorable forum/choice of law, arbitration clauses	$890,000	Medium-High	24.7%

Risk Severity Scoring Framework: We developed a 5-point severity scale validated against actual negotiation outcomes:

- (Minor): Administrative issues, minor formatting errors, standard boilerplate variations. Negotiation impact: None or clerical fix.
- (Low): Non-material deviations from standard terms, minor drafting issues. Negotiation impact: Usually accepted after brief discussion.
- (Moderate): Material issues requiring negotiation but not deal-breaking. Examples: Payment terms 45 days vs 30 days, insurance requirements slightly above normal. Negotiation impact: Requires discussion and mutual concessions.
- (High): Significant issues that would require material concessions. Examples: Liability cap at 50% of contract value (vs standard 100%), indemnification slightly broader than standard. Negotiation impact: Likely to require escalation to senior management.

- (Critical): Deal-breaking issues requiring complete redrafting or likely termination. Examples: Unlimited liability for IP infringement, unilateral termination for convenience, jurisdiction in unfavorable location. Negotiation impact: Deal at serious risk without major revisions.

The critical insight from our dispute analysis is that severity levels correlate non-linearly with costs. A severity 5 risk is not 5× more expensive than a severity 1 risk—it's approximately 800× more expensive ($3.4M average vs $4,200). This non-linear relationship fundamentally changes how we should evaluate AI systems.

3. THE UNIQUE NATURE OF LEGAL DECISION-MAKING

Legal contract analysis differs fundamentally from standard classification tasks in ways that challenge current AI approaches[15], [16]:

Normative Interpretation vs. Descriptive Classification: Legal analysis asks not "What does this text say?" but "What should this text mean within the relevant legal framework?" For example, the phrase "best efforts" has been interpreted differently across jurisdictions. In Delaware courts, it means "efforts that are reasonable under the circumstances" (reasonable efforts standard). In English law, it historically meant "taking all steps a prudent and determined person would take" (higher standard). An AI system that simply identifies "best efforts" as a risk without jurisdictional context provides misleading guidance[17], [18].

Contextual Dependency at Multiple Levels: The risk level of a clause depends entirely on transaction context at multiple levels:

- Document Context: A limitation clause interacts with indemnification clauses elsewhere in the same document.
- Transaction Context: A $1 million liability cap might be reasonable in a $10 million transaction but catastrophic in a $100,000 transaction.
- Business Context: An exclusivity clause in a software license has different implications for a startup vs. an enterprise.
- Jurisdictional Context: Data processing clauses have different requirements under GDPR vs. CCPA vs. PIPEDA.

Current AI systems, even LLMs, struggle with these nested contextual dependencies because they typically analyze text in isolation from these broader contexts.

Precedent Sensitivity and Temporal Evolution: Legal interpretation evolves through case law. A clause that was standard last year might become risky this year due to a new court ruling. For example, the interpretation of "material adverse effect" clauses changed significantly after the 2021 AB Stable v. MAPS decision in Delaware. AI systems trained on historical contracts may not incorporate these evolving interpretations without explicit knowledge updates[19], [20].

Extreme Asymmetry of Error Costs: In legal practice, the cost ratio of false negatives (missing a risk) to false positives (over-flagging) typically ranges from 10:1 to 100:1, and in some cases exceeds 1,000:1. Consider:

- False Negative Cost: Missing an uncapped liability clause could result in $10M+ in damages.
- False Positive Cost: Flagging a standard clause requires a lawyer to spend 2 minutes confirming it's standard ($10-20 cost).

This asymmetry is fundamentally misaligned with standard ML optimization, which typically treats false positives and false negatives equally.

4. IMPLICATIONS FOR AI SYSTEM DESIGN

These theoretical considerations impose specific, often conflicting requirements on AI systems for legal contract analysis:

4.1 Requirement 1: Long-Range Dependency Modeling

Systems must track defined terms, cross-references, and conditional dependencies across entire documents, which can exceed 200 pages. This challenges the context window limits of even modern LLMs (typically 128K-200K tokens). More fundamentally, it requires understanding not just token proximity but semantic relationships across distant sections[21], [22].

4.2 Requirement 2: External Knowledge Integration

Models need access to several types of external knowledge:

- Legal databases (statutes, regulations by jurisdiction)
- Case law databases (precedent evolution)
- Industry standards and practices
- Company-specific policies and risk tolerances

Current systems typically lack systematic integration of this knowledge, relying instead on patterns learned from training data.

4.3 Requirement 3: Asymmetric Optimization Objectives

Loss functions must penalize false negatives far more heavily than false positives. This requires moving beyond standard cross-entropy loss to customized loss functions that incorporate severity weighting. More challengingly, it requires training data that captures not just "risk present/absent" but "cost if missed."

4.4 Requirement 4: Explainability Aligned with Legal Reasoning

Explanations should follow legal argument structure: rule (applicable legal principle), facts (what the clause says), application (how the rule applies to these facts), conclusion (risk assessment). Current explainability methods (attention maps, feature importance) do not produce this structure, making them less useful for legal practitioners.

Our experimental design in Section 4 addresses these requirements through careful task design, evaluation metrics, and analysis frameworks that highlight where current systems succeed and fail against these legal requirements.

III. RELATED WORK AND CRITICAL COMPARISON

1. RULE-BASED LEGAL EXPERT SYSTEMS: THE FIRST GENERATION

The earliest attempts at legal AI (1980s-2000s) employed rule-based or expert system approaches, encoding legal knowledge as if-then rules. These systems, including early commercial products like KIRA Systems and eBrevia (before their ML transitions), operated through pattern matching, keyword lists, and simple syntactic rules [8].

Technical Approach: These systems typically used:

- Regular expressions for clause patterns (e.g., "liability.shall not exceed.[amount]")
- Keyword proximity rules (e.g., "joint and several" within 50 characters of "liability")
- Template matching against known clause libraries

Strengths in Practice:

- Transparency: Every decision is traceable to specific rules.
- Determinism: Consistent outputs for identical inputs.
- No Training Data Requirement: Can be built from legal expertise alone.
- High Precision on Known Patterns: When rules match, they're usually correct.

Documented Limitations:

- Brittleness: Fail on novel phrasing. For example, a rule for "liability shall not exceed" misses "in no event shall liability exceed" or "the maximum liability hereunder is."
- Limited Coverage: Only handle predefined clause types; miss emerging or bespoke clauses.
- High Maintenance Costs: Rules must be manually updated for new phrasing or legal developments.
- No Semantic Understanding: Cannot infer risk from context or implied terms.

Our experiments confirm these limitations at scale. Rule-based systems achieved only 0.432 F1 on complex M&A agreements where bespoke drafting exceeded rule coverage, compared to 0.812 on standardized NDAs. More concerning, their recall dropped dramatically with increasing severity—they missed precisely the novel, high-stakes clauses that matter most.

2. MACHINE LEARNING FOR CONTRACT ANALYTICS: THE FEATURE ENGINEERING ERA

The 2010s saw the application of supervised machine learning to legal documents, moving beyond hand-crafted rules to learned patterns. Early approaches used extensive feature engineering with classifiers like SVM, Random Forests, and later XGBoost [9].

2.1 Technical Evolution

- Phase 1 (2010-2015): Simple features (n-grams, TF-IDF, basic syntax)
- Phase 2 (2015-2018): Rich feature engineering (syntactic parse features, legal-specific features like "contains defined term," "references other section")
- Phase 3 (2018-present): Integration with word embeddings (Word2Vec, GloVe) as features

2.2 Key Contributions

- The CUAD Dataset (2021): 510 contracts with 13,000+ labeled clauses enabled more robust evaluation [10]. However, CUAD's focus on software agreements limits generalizability.
- Feature-Based Approaches: Typical systems achieved F1-scores of 0.75-0.85 on clause classification but required extensive domain expertise for feature engineering.
- Early Risk Prediction: Some studies attempted to predict litigation risk from contract terms using features like "unbalanced indemnification," "unlimited liability," etc.

2.3 Persistent Challenges

- Feature Engineering Burden: Required legal and ML expertise to design effective features.
- Limited Generalization: Features tailored to one domain (e.g., software licensing) often failed in others (e.g., employment agreements).
- Context Limitations: Most approaches analyzed clauses in isolation, missing cross-document dependencies.
- Explainability Gaps: Feature importance scores (e.g., "the clause contains 'uncapped'") provided limited legal insight.

3. DEEP LEARNING AND TRANSFORMER MODELS: THE REPRESENTATION LEARNING BREAKTHROUGH

The advent of BERT (2018) and subsequent transformer models revolutionized legal NLP by enabling transfer learning from massive legal corpora. Domain-specific pre-training on legal text (Legal-BERT, CaseLaw-BERT, Lawformer) significantly improved performance on downstream legal tasks [11].

3.1 Technical Advances

- Domain-Specific Pre-training: Models pre-trained on legal corpora (court opinions, statutes, contracts) learned legal language patterns and concepts.
- Fine-tuning Paradigm: Pre-trained models could be fine-tuned on specific tasks (clause classification, risk detection) with relatively small labeled datasets.
- Contextual Understanding: Transformers captured longer-range dependencies than previous architectures, though still limited by context windows.

3.2 State-of-the-Art Performance

Recent studies report F1-scores up to 0.91 on clause classification using fine-tuned transformer models [12]. The best-performing models typically:

- Use legal-specific pre-training (Legal-BERT vs. generic BERT)
- Incorporate document structure features (section headings, defined terms)
- Use ensemble approaches combining multiple model types

3.3 Critical Gaps in Current Research

- Narrow Evaluation: Most studies evaluate on single-domain datasets (e.g., only software agreements) without cross-domain testing.
- Missing Risk Assessment: Focus on clause classification rather than holistic risk assessment.
- Ignoring Error Asymmetry: Use standard F1 without considering that missing a high-severity risk is much worse than missing a low-severity one.

- Limited Explainability: Attention maps and integrated gradients provide limited insight for legal practitioners.

4. LARGE LANGUAGE MODELS IN LEGAL DECISION-MAKING: THE NEW FRONTIER

The emergence of LLMs like GPT-4 and Claude 3 has prompted intense exploration of their capabilities in legal contexts. Initial studies showed remarkable performance on bar exam questions (GPT-4 scoring in the 90th percentile) and legal reasoning tasks [13].

4.1 Demonstrated Capabilities

- Zero-Shot Legal Analysis: Can analyze contracts without task-specific training.
- Complex Reasoning: Can follow multi-step legal reasoning chains.
- Natural Language Explanations: Generate human-readable rationales.
- Multi-Task Handling: Can perform classification, extraction, summarization, and Q&A in a single model.

4.2 Critical Issues Identified in Literature

- Hallucination: Tendency to generate plausible but incorrect legal citations or clauses [14]. One study found hallucination rates of 15-20% in legal analysis tasks.
- Prompt Sensitivity: Performance varies dramatically with prompt phrasing [15]. Small wording changes can change outputs significantly.
- Lack of Consistency: Different outputs for semantically identical inputs [16]. This is particularly problematic for legal applications requiring audit trails.
- Explainability Gaps: Rationales may be persuasive but legally incorrect [17]. The fluency of LLM outputs can mask substantive errors.
- Cost and Latency: API costs ($0.03-$0.12 per 1K tokens) and latency (seconds per query) limit practical deployment at scale.

4.3 Our Study's Position in This Landscape

While prior LLM studies in law have focused on capabilities (what can they do?), our study focuses on reliability (how consistently and accurately do they do it?) and practical applicability (what are the costs, risks, and deployment requirements?).

5. COMPARATIVE GAP ANALYSIS AND OUR CONTRIBUTION

Table 2. Systematic comparison with prior major studies.

Study	Year	Dataset Size	Max F1	Risk Assessment	Cross-Jurisdiction	Real-World Validation	Key Limitation
Hendrycks et al.	2021	1,024 docs	0.85	No	No	No	Narrow domain focus
Chalkidis et al.	2020	12,000 docs	0.89	Limited	No	No	No severity weighting
Zheng et al.	2022	3,500 docs	0.91	Yes	No	Pilot (3 firms)	Small validation
Bommasani et al.	2023	Various	0.87	No	No	No	No contract focus

Study	Year	Dataset Size	Max F1	Risk Assessment	Cross-Jurisdiction	Real-World Validation	Key Limitation
Our Study	2024	5,247 docs	0.923	Comprehensive	3 jurisdictions	42-company survey	N/A

Our study addresses four critical gaps in the literature:

5.1 Gap 1: From Single-Domain to Multi-Domain Evaluation

Prior studies typically evaluate on narrow domains (e.g., only software agreements). We test across 12 domains with systematic performance variation analysis.

5.2 Gap 2: From Technical Metrics to Legal-Specific Metrics

We move beyond F1 to introduce and validate FNRP and SwF1, which better capture the asymmetric cost structure of legal errors.

5.3 Gap 3: From Capability Demonstration to Reliability Assessment

While prior LLM studies show what's possible, we systematically measure consistency, hallucination rates, and failure modes across identical inputs.

5.4 Gap 4: From Academic Benchmark to Practical Deployment Framework

We provide evidence-based thresholds, cost analysis, and workflow specifications for real-world deployment—addressing the "now what?" question after evaluation.

Our contribution is not just another benchmark but a comprehensive framework for evaluating and deploying legal AI systems that accounts for the unique requirements and constraints of legal practice.

IV. RESEARCH DESIGN AND EXPERIMENTAL FRAMEWORK

1. STUDY DESIGN AND EXPERIMENTAL PROTOCOL

We designed a multi-phase, controlled experimental study spanning 9 months with rigorous methodology to ensure validity, reproducibility, and practical relevance:

1.1 Phase 1: Dataset Curation and Preparation (Months 1-3)

- Source Identification: Identified and secured access to multiple contract sources through partnerships with corporations, public databases, and legal service providers.
- Preprocessing Pipeline Development: Built and validated a comprehensive preprocessing pipeline including OCR, redaction, normalization, and segmentation.
- Ethics and Compliance: Obtained IRB approval, established data use agreements, and implemented rigorous PII protection protocols.

1.2 Phase 2: Expert Annotation and Ground Truth Establishment (Months 2-4)

- Expert Recruitment: Recruited 12 qualified lawyers through legal associations and professional networks, ensuring diversity in jurisdiction, practice area, and experience level.
- Annotation Protocol Development: Created detailed annotation guidelines through iterative refinement with pilot annotations.
- Annotation Process: Conducted supervised annotation with weekly calibration sessions, double annotation for 30% of documents, and adjudication for disagreements.

1.3 Phase 3: System Configuration and Experimental Execution (Months 4-7)

- System Selection: Selected representative systems from four AI classes based on market presence and technical capabilities.
- Configuration Optimization: For each system, optimized parameters through grid search on validation set.
- Prompt Engineering: For LLMs, developed and tested 200+ prompt variations to identify optimal formulations.
- Experimental Execution: Ran all systems on test set with randomization, blinding, and replication protocols.

1.4 Phase 4: Analysis and Framework Development (Months 7-9)

- Quantitative Analysis: Statistical analysis of performance metrics with confidence intervals and significance testing.
- Qualitative Analysis: Detailed error analysis through expert review of misclassifications.
- Framework Development: Created deployment framework based on empirical findings and validated through pilot implementation.

Key Methodological Innovations:

- Blinding Protocol: Legal experts were blinded to AI outputs during annotation; system outputs were evaluated against gold standards without knowledge of which system produced them.
- Cross-Jurisdictional Controlled Comparison: Using professionally drafted synthetic contracts that varied only by jurisdiction, we isolated jurisdictional effects from other variables.
- Consistency Testing: Created 100 semantically identical clauses with 5 syntactic variations each to measure decision consistency independent of content.
- Cost-Benefit Integration: Incorporated not just accuracy metrics but processing speed, latency, and cost per document into evaluation.

2. CONTRACT DATASETS: COMPOSITION AND CHARACTERISTICS

Our corpus represents one of the largest and most diverse collections of annotated contracts available for research:

Table 3. Dataset composition with detailed statistics.

Source	Documents	Avg Pages	Total Clauses	Domain Distribution	Jurisdictional Coverage
EDGAR 10-K/Q Filings	2,147	42.3	418,927	M&A (38%), Employment (22%), Licensing (19%), Other (21%)	Primarily US (91%), Some international (9%)
Corporate Procurement	1,892	15.1	567,120	NDA (31%), MSA (28%), SOW (18%), SLA (13%), Other (10%)	US (63%), UK (22%), EU (12%), Other (3%)
Synthetic Jurisdictional	1,208	28.4	241,600	8 template types × 3 jurisdictions	US/Delaware (33%), UK (33%), Singapore (34%)
Total	5,247	28.3	1,227,647	12 distinct domains	3 primary jurisdictions

Dataset Diversity Metrics:

- Document Length Distribution: Ranged from 2-page NDAs to 347-page M&A agreements
- Complexity Scores: We computed complexity scores based on: defined term density (avg 4.2 per page), cross-reference density (avg 8.7 per page), conditional clause percentage (avg 23.4%)
- Temporal Distribution: Contracts from 2015-2023, with concentration in 2019-2021 (78%)
- Industry Representation: Technology (34%), Financial Services (22%), Healthcare (18%), Manufacturing (14%), Other (12%)

Preprocessing Pipeline Details:

2.1 OCR and Text Extraction

- Tool: Abbyy FineReader 15 (enterprise edition)
- Accuracy: 99.8% verified on 500-page manual sample
- Handling of complex elements: Tables extracted with structure preservation, footnotes linked to reference points

2.2 Redaction and Anonymization

- PII Detection: Custom BERT model fine-tuned on legal PII patterns
- Redaction: Replaced with consistent placeholders ([COMPANY_A], [DATE_1], etc.)
- Validation: Manual review of 5% sample confirmed no PII leakage

2.3 Normalization and Cleaning

- Character encoding: Converted to UTF-8
- Whitespace: Normalized spaces, tabs, line breaks
- Section numbering: Standardized to consistent format (1.1, 1.1.1, etc.)
- Defined term identification: Rule-based + ML hybrid approach

2.4 Clause Segmentation

- Approach: Hybrid rule-based (section headings) + BERT-based classifier
- Training: 2,000 manually segmented clauses
- Accuracy: 96.4% F1 on held-out test
- Output: Clause boundaries with hierarchical relationships

2.5 Synthetic Jurisdictional Dataset Creation

To enable controlled cross-jurisdictional comparison, we commissioned the creation of professionally drafted contract variants. The process:

- Base Templates: Selected 8 common contract templates (NDA, MSA, Employment, License, etc.)
- Jurisdictional Adaptation: Qualified lawyers in each jurisdiction drafted jurisdictionally appropriate versions
- Controlled Variation: Only changed jurisdiction-specific elements (governing law, statutory references, etc.)
- Validation: Each variant reviewed by second lawyer in that jurisdiction

This approach allowed us to isolate jurisdictional effects from other variables—a methodological advance over prior studies.

3. GROUND-TRUTH ANNOTATION: PROTOCOL AND QUALITY ASSURANCE

3.1 Expert Panel Composition and Credentials

We assembled a panel of 12 lawyers with the following characteristics:

- Experience: Average 14.3 years (range: 8-22 years)
- Jurisdictional Representation: US (4: 2 Delaware, 2 NY), UK (4: all England & Wales), Singapore (4)
- Practice Areas: Corporate law (6), litigation/disputes (3), compliance/regulatory (3)
- Firm Types: Large law firms (7), in-house counsel (3), boutique firms (2)
- Diversity: 7 male, 5 female; average age 42.7

3.2 Annotation Protocol Development

We developed annotation guidelines through an iterative 4-phase process:

- Phase 1: Draft Guidelines based on legal literature and prior annotation schemes.
- Phase 2: Pilot Annotation with 3 lawyers annotating 50 contracts, identifying ambiguities and edge cases.
- Phase 3: Guideline Refinement incorporating pilot feedback, creating decision trees for common ambiguities.
- Phase 4: Training and Calibration 20-hour training for all annotators with weekly calibration sessions.

3.3 Final Annotation Tasks

3.3.1 Clause Segmentation and Typing

- Identify clause boundaries (start/end positions)
- Assign to one of 25 predefined types with hierarchical relationships
- Confidence rating (high/medium/low) for ambiguous cases

3.3.2 Risk Identification and Classification

- Binary decision: Does this clause contain legal/business risk?
- Risk type classification (4 main types, 12 subtypes)
- Flag for cross-referential risks (risk emerges from multiple clauses)

3.3.3 Severity Scoring

- 5-point scale (1-5) with detailed rubric
- Required justification for scores ≥3
- Consideration of: financial impact, likelihood, mitigation difficulty

3.3.4 Ambiguity and Interpretation Notes

- Flag clauses with multiple reasonable interpretations
- Document interpretive choices made
- Note jurisdictional dependencies

3.4 Quality Assurance Measures

- Double Annotation: 30% of documents (1,574) annotated independently by two experts
- Adjudication Process: Disagreements resolved by third expert or panel discussion
- Weekly Calibration: Regular meetings to discuss edge cases and maintain consistency
- Quality Metrics Tracking: Monitored individual annotator metrics for outliers
- Final Review: 10% sample reviewed by senior lawyer for quality control

3.5 Inter-Annotator Agreement Statistics

We calculated agreement metrics using multiple approaches:

3.5.1 Clause Segmentation

- Character-level overlap: 89.2%
- Cohen's κ for boundary agreement: 0.89 (almost perfect)

3.5.2 Clause Type Classification

- Exact match: 84.7%
- Cohen's κ: 0.82 (almost perfect)
- Hierarchical match (parent category): 91.3%

3.5.3 Risk Identification

- Binary agreement: 88.4%
- Cohen's κ: 0.76 (substantial)
- F1-score between annotators: 0.87

3.5.4 Severity Scoring

- Exact match: 72.1%
- Within 1 point: 94.3%
- Intraclass correlation coefficient: 0.81 (excellent)

3.5.5 Risk Type Classification

- Exact match: 79.8%
- Cohen's κ: 0.71 (substantial)

3.6 Annotation Cost and Time Investment

- Total Hours: 2,840 annotation hours
- Average per Document: 32.4 minutes (varied by complexity: 8 minutes for NDA, 68 minutes for complex M&A)
- Cost: $247,000 in professional fees (average $87/hour)
- Efficiency: Improved over time from 45 minutes/document in month 1 to 24 minutes/document in month 4

The resulting gold standard represents one of the most comprehensive and rigorously validated contract annotation sets available, providing a solid foundation for evaluating AI systems.

4. AI SYSTEMS EVALUATED: CONFIGURATIONS AND IMPLEMENTATION

We evaluated representative systems from four classes of AI approaches:

4.1 System 1: Rule-Based (RB) - KIRA Systems Configuration

- Implementation: Commercial rule-based contract analysis system
- Ruleset: 4,200 hand-crafted rules covering 35 clause types
- Rule Types: 68% pattern matching, 22% syntactic rules, 10% template matching
- Configuration: Optimized for high precision (default settings)
- Deployment: Local installation on Ubuntu server
- Processing: Single-threaded, rule-based engine

4.2 System 2: Supervised Machine Learning (S-ML) - Custom XGBoost Implementation

- Algorithm: XGBoost (version 1.7) with custom objective function
- Features: 150 engineered features including:
  Lexical: n-grams (1-3), character n-grams, TF-IDF weighted terms
  Syntactic: POS tag patterns, dependency parse features, clause length
  Legal-specific: Defined term presence, cross-reference count, section type
  Structural: Position in document, heading level, preceding/following clause types
- Training Data: 3,000 documents (70% of training set)
- Hyperparameters: Optimized via Bayesian optimization (100 iterations)

4.3 System 3: Transformer-Based (TB) - Legal-BERT Fine-Tuned

- Base Model: Legal-BERT-base-uncased (110M parameters)
- Pre-training: 12GB of legal text (cases, statutes, contracts)
- Fine-tuning: Two-stage approach:
  Task 1 (Clause Classification): Fine-tuned on 25-way classification
  Task 2 (Risk Detection): Multi-task learning with risk classification and severity regression
- Architecture Modifications:
  Added document structure embeddings (section hierarchy)
  Extended context window to 512 tokens with hierarchical processing for longer documents
  Custom loss function with severity-weighted false negative penalty
- Training Details: 10 epochs, batch size 16, learning rate 2e-5, AdamW optimizer

4.4 System 4: Large Language Models - GPT-4 and Claude 3

4.4.1 GPT-4 Configuration

- Model: gpt-4-0125-preview (128K context window)
- Temperature: 0.1 for consistency (lower than default 0.7)
- Prompt Strategy: Structured prompt chain with:

System prompt defining role and constraints
Task specification with examples
Output format requirements (JSON schema)
Reasoning chain prompting ("Think step by step...")

- API Parameters: max_tokens=4000, top_p=0.95, frequency_penalty=0.1

4.4.2 Claude 3 Configuration

- Model: claude-3-opus-20240229 (200K context window)
- Temperature: 0.1
- Prompt Strategy: Similar to GPT-4 but adapted to Claude's preferred format
- API Parameters: max_tokens=4000

4.4.3 Prompt Engineering Process

We developed prompts through an iterative optimization process:

- Base Prompt Creation: Initial prompt based on legal analysis frameworks
- A/B Testing: Tested 200+ variations on validation set
- Expert Feedback: Lawyers reviewed outputs and suggested improvements
- Final Optimization: Selected best-performing prompt through grid search

4.4.4 Example Prompt Structure (GPT-4)

text

System: You are a senior corporate lawyer with 15 years of experience reviewing contracts. Your task is to analyze contract clauses for legal and business risks.

4.4.5 Instructions

- 1. For each clause identified, determine if it contains material risk.
- 2. Classify risk type: Financial, Compliance, Liability, or Termination.
- 3. Assign severity 1-5 (5 = critical, deal-breaking).
- 4. Provide brief rationale following legal reasoning structure.
- 5. Output must be valid JSON with schema: {clauses: [{text, risk_present, risk_type, severity, rationale}]}

Clause: "[CLAUSE_TEXT]"

Think step by step: First identify key terms, then assess risk, then determine severity based on standard negotiation positions.

4.4.6 Cost Considerations and Optimization

We implemented several cost optimization strategies for LLMs:

- Caching: Cache identical clause analyses to avoid redundant API calls
- Batch Processing: Group clauses where possible within context limits
- Fallback Strategy: Use local model for low-risk, standard clauses; LLM only for complex/high-risk
- Token Optimization: Prune unnecessary text, use abbreviations where possible

Table 4. System configuration summary.

System	Parameters	Training Data	Inference Cost/1000 docs	Processing Speed	Hardware
Rule-Based	4,200 rules	Manual creation	$12.50	4.2 sec/page	4 vCPU, 16GB RAM
XGBoost	150 features	3,000 docs	$4.20	2.1 sec/page	2 vCPU, 8GB RAM

System	Parameters	Training Data	Inference Cost/1000 docs	Processing Speed	Hardware
Legal-BERT	110M params	12GB + fine-tuning	$8.70	7.8 sec/page	GPU (V100), 32GB RAM
GPT-4	~1.76T params	General web	$247.80	12.4 sec/page*	API
Claude 3	Unknown	General web	$198.40	10.8 sec/page*	API

*Includes API latency (avg 3.2s for GPT-4, 2.8s for Claude 3)

5. EXPERIMENTAL TASKS: DESIGN AND IMPLEMENTATION

We designed four experimental tasks that reflect real legal workflow stages while enabling controlled evaluation:

5.1 Clause Identification and Classification

Objective: Measure ability to identify clause boundaries and assign correct type labels.
Implementation:

- Input: Complete contract document in plain text with minimal formatting
- Processing: Each system processes document end-to-end
- Output Requirements:
  Clause boundaries (start/end character positions)
  Type label (1 of 25 categories)
  Confidence score (0-1)
- Evaluation Metrics:
  Boundary accuracy (IoU ≥ 0.8 for match)
  Type classification accuracy
  Combined F1 considering both boundary and type

5.2 Risk Classification and Severity Scoring

Objective: Assess risk detection accuracy and severity calibration.
Implementation:

- Input: Correctly identified clauses (from Task 1 or gold standard for ablation)
- Processing: Each clause analyzed independently
- Output Requirements:
  Risk present (binary)
  Risk type (4 categories)
  Severity score (1-5)
  Rationale (natural language)
- Evaluation Metrics:
  Binary classification metrics (precision, recall, F1)
  Severity-weighted metrics (SwF1, FNRP)
  Severity correlation with experts (Pearson's r)

5.3 Batch Processing and Efficiency Analysis

Objective: Measure practical throughput, latency, and cost in simulated real workload.
Implementation:

- Input: Batch of 100 contracts of varying types and complexities
- Processing: Systems process batch with parallelization where supported
- Metrics Collected:

Total processing time
Throughput (pages/hour)
Peak memory usage
API costs (for LLMs)
Latency distribution (P50, P90, P99)

- Constraints: Simulated real-world constraints (network latency for APIs, GPU memory limits)

5.4 Decision Consistency Analysis

Objective: Measure output stability across identical and near-identical inputs.
Implementation:

- Input Set 1: 100 identical clauses processed 10 times each
- Input Set 2: 100 semantically identical clauses with 5 syntactic variations each (500 total)
- Metrics:

Intra-model consistency (Jaccard similarity across runs)
Semantic consistency (agreement across syntactic variations)
Severity score standard deviation
Hallucination rate (generating content not in source)
Training/Validation/Test Split:
We used a stratified 70/15/15 split to ensure representative distribution:

- Training: 3,673 documents (all systems except rule-based)
- Validation: 787 documents (hyperparameter tuning, prompt optimization)
- Test: 787 documents (held-out final evaluation)

Stratification ensured proportional representation of:

- Document types (by domain)
- Jurisdictions
- Complexity levels (based on defined term density and length)
- Risk profiles (high/low risk concentration)

Statistical Analysis Plan:

- Primary Comparisons: Pairwise system comparisons on all metrics
- Confidence Intervals: Bootstrapped 95% CIs (10,000 samples)
- Significance Testing: Wilcoxon signed-rank test with Holm-Bonferroni correction
- Effect Sizes: Cohen's d for continuous metrics, Cramér's V for categorical
- Regression Analysis: To identify factors predicting error rates (document complexity, clause type, etc.)

This comprehensive experimental design enables us to answer not just "which system performs best?" but "under what conditions, at what cost, and with what reliability?"

V. EVALUATION METRICS AND LEGAL PERFORMANCE INDICATORS

1. STANDARD NLP METRICS: FOUNDATIONS AND LIMITATIONS

We report standard classification metrics but with careful interpretation considering legal context:

- Precision (Positive Predictive Value):

$$\text{Precision} = \frac{TP}{TP + FP}$$

Legal Interpretation: Proportion of flagged risks that are actually risks. High precision minimizes lawyer time wasted on false alarms.

- Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN}$$

Legal Interpretation: Proportion of actual risks that are detected. Critical for risk management but must be interpreted alongside severity.

- F1-Score (Harmonic Mean):

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Legal Limitation: Treats all errors equally. A system missing a severity 5 risk but catching all severity 1 risks could have high F1.

- Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Legal Limitation: In contract analysis, TN (correctly identifying no risk) dominates, making accuracy misleading. A system that never flags risks could have high accuracy but be useless.

Confidence Intervals: All metrics reported with 95% confidence intervals computed via bootstrapping (10,000 samples). For legal applications, the lower bound of the CI is often more important than the point estimate—we need confidence that performance won't fall below a threshold.

2. LEGAL RISK-AWARE METRICS: NOVEL CONTRIBUTIONS

Recognizing the limitations of standard metrics for legal applications, we developed and validated two novel metrics:

- False-Negative Risk Penalty (FNRP):

Motivation: In legal practice, not all false negatives are equal. Missing a severity 5 risk is much worse than missing a severity 1 risk. Standard recall treats them equally.

- Definition:

$$\text{FNRP} = \frac{\sum_{i=1}^{N} (S_i \times \mathbb{I}(\text{missed}_i))}{\sum_{i=1}^{N} S_i}$$

where:

- $S_i \in \{1,2,3,4,5\}$ is the severity score of risk i
- $\mathbb{I}(\text{missed}_i) = 1$ if risk i was missed by the system, 0 otherwise
- N = total number of actual risks in the document

Interpretation: FNRP represents the proportion of total risk severity weight that was missed. For example:

- FNRP = 0.10: System missed risks accounting for 10% of total severity weight
- FNRP = 0.25: System missed risks accounting for 25% of total severity weight

Example Calculation:

Consider a document with 4 risks:

- Severity 5 (critical) - MISSED
- Severity 3 (moderate) - DETECTED
- Severity 2 (low) - DETECTED
- Severity 1 (minor) - MISSED

Total severity weight = 5 + 3 + 2 + 1 = 11

Missed severity weight = 5 + 1 = 6

FNRP = 6/11 = 0.545

Despite detecting 2 of 4 risks (50% recall by count), the system missed 54.5% of severity weight due to missing the critical risk.

Validation Against Historical Outcomes: We correlated FNRP with actual dispute outcomes from 1,842 contract disputes. Systems with FNRP > 0.15 were associated with significantly higher dispute rates ($p < 0.001$, OR = 3.42).

Severity-Weighted F1 (SwF1):

Motivation: Standard F1 gives equal weight to all errors. We need a metric that penalizes missing high-severity risks more heavily.

Definition:

First, compute severity-weighted components:

$$\mathrm{TP}_{\mathrm{weighted}} = \sum_{i \in \mathrm{TP}} S_i$$

$$\mathrm{FP}_{\mathrm{weighted}} = \sum_{i \in \mathrm{FP}} 1 \text{ (constant penalty for false positives)}$$

$$\mathrm{FN}_{\mathrm{weighted}} = \sum_{i \in \mathrm{FN}} S_i$$

Then compute weighted precision and recall:

$$\mathrm{Precision}_w = \frac{\mathrm{TP}_{\mathrm{weighted}}}{\mathrm{TP}_{\mathrm{weighted}} + \mathrm{FP}_{\mathrm{weighted}}}$$

$$\mathrm{Recall}_w = \frac{\mathrm{TP}_{\mathrm{weighted}}}{\mathrm{TP}_{\mathrm{weighted}} + \mathrm{FN}_{\mathrm{weighted}}}$$

Finally:

$$\mathrm{SwF1} = 2 \times \frac{\mathrm{Precision}_w \times \mathrm{Recall}_w}{\mathrm{Precision}_w + \mathrm{Recall}_w}$$

Interpretation: SwF1 ranges 0-1 like standard F1, but a system that detects high-severity risks gets higher SwF1 than one that detects only low-severity risks, even with the same standard F1.

Example Comparison:

- System A: Detects 9 low-severity risks (severity 1-2) but misses 1 critical risk (severity 5)
- System B: Detects the 1 critical risk but misses 3 low-severity risks

Assuming 10 total risks (9 low, 1 critical):

- System A: Standard F1 = 0.90, SwF1 = 0.64
- System B: Standard F1 = 0.50, SwF1 = 0.71

SwF1 correctly identifies System B as better for legal risk management despite lower standard F1.

Contract-Level Risk Coverage (CLRC):

Definition:

$$\mathrm{CLRC} = \frac{\text{Number of contracts where all high-severity risks (≥4) were detected}}{\text{Total contracts}}$$

Interpretation: The percentage of contracts where the system caught all critical risks. This is particularly important for compliance and audit contexts where missing any critical risk is unacceptable.

Relationship to FNRP: CLRC is a stricter measure—a system could have moderate FNRP (0.10-0.15) but still miss at least one critical risk in many contracts, resulting in low CLRC.

3. DECISION CONSISTENCY METRICS

For legal applications, consistency is as important as accuracy. Inconsistent systems create audit trail problems and undermine trust.

- Intra-Model Consistency (Jaccard Similarity):

Definition: For stochastic models (LLMs, some ML models), we measure consistency across multiple runs on identical input:

$$J(A, B) = \frac{| A \cap B |}{| A \cup B |}$$

where A and B are sets of identified risks from two runs.

Interpretation: $J = 1.0$ means perfect consistency (identical outputs), $J = 0$ means no overlap. Legal applications typically require $J > 0.90$ for auditability.

- Cross-Jurisdiction Performance Drop (CJPD):

Definition:

$$\mathrm{CJPD} = \frac{\mathrm{F1}_{\mathrm{domestic}} - \mathrm{F1}_{\mathrm{foreign}}}{\mathrm{F1}_{\mathrm{domestic}}}$$

Interpretation: Percentage performance degradation when applying a system trained/optimized for one jurisdiction to another. For global deployment, CJPD should be < 20%.

Severity Score Standard Deviation:

Definition: Standard deviation of severity scores assigned to identical or semantically equivalent clauses.

Interpretation: High standard deviation indicates unreliable severity assessment, which is problematic for risk prioritization and workflow routing.

4. STATISTICAL VALIDATION AND SIGNIFICANCE TESTING

4.1 Bootstrapping for Confidence Intervals

For all metrics, we compute 95% confidence intervals using bootstrapping with 10,000 resamples. This is particularly important for:

- Asymmetric Distributions: Some metrics (especially FNRP) have skewed distributions
- Small Sample Effects: For rare risk types or high-severity risks
- System Comparisons: Overlap/non-overlap of CIs provides visual significance indication

4.2 Hypothesis Testing Protocol

- Primary Comparisons: Pairwise system comparisons on all metrics
- Test Selection: Wilcoxon signed-rank test for paired comparisons (same test set)
- Multiple Testing Correction: Holm-Bonferroni method to control family-wise error rate
- Significance Level: $\alpha = 0.01$ (conservative given multiple comparisons)
- Effect Size Reporting: Cohen's d for continuous variables, interpreted as: small (0.2), medium (0.5), large (0.8)

4.3 Regression Analysis for Error Prediction

We use multivariate regression to identify factors predicting error rates:

$$\text{Error Rate} = \beta_0 + \beta_1(\text{Complexity}) + \beta_2(\text{Length}) + \beta_3(\text{Jurisdiction}) + \epsilon$$

This helps identify conditions where systems are likely to fail.

4.4 Power Analysis:

Given our sample size (787 test documents, ~38,471 risk instances), we have:

- 99% power to detect effect sizes ≥ 0.15 (small-medium) for primary metrics
- 95% power to detect effect sizes ≥ 0.10 for subgroup analyses

This ensures we can detect practically significant differences, not just statistical significance.

5. PRACTICAL THRESHOLDS FOR DEPLOYMENT

Based on our analysis of dispute outcomes and expert consultations, we propose the following evidence-based thresholds for deployment decisions:

5.1 Red Flags (System Should Not Be Deployed)

- FNRP > 0.20 (misses >20% of severity weight)
- CJPD > 0.25 (>25% performance drop cross-jurisdiction)
- Intra-model consistency < 0.80
- Hallucination rate > 10%

5.2 Yellow Flags (Deploy with Enhanced Oversight)

- FNRP 0.15-0.20
- CJPD 0.15-0.25
- Intra-model consistency 0.80-0.90
- CLRC < 0.85

5.3 Green Flags (Suitable for Limited Autonomous Use)

- FNRP < 0.15

- CJPD < 0.15
- Intra-model consistency > 0.90
- CLRC > 0.90

These thresholds provide concrete guidance for legal departments evaluating AI systems, moving beyond vendor claims to evidence-based decision making.

6. EXPERIMENTAL RESULTS: CLAUSE ANALYSIS

6.1 Overall Clause Detection Accuracy

Table 5. Comprehensive clause classification performance.

System	Precision	95% CI	Recall	95% CI	F1-Score	95% CI	Processing Speed	Hardware Utilization
Rule-Based	0.823	[0.815, 0.831]	0.641	[0.632, 0.650]	0.721	[0.712, 0.730]	4.2 sec/page	CPU: 85%, RAM: 3.2GB
XGBoost	0.854	[0.847, 0.861]	0.781	[0.773, 0.789]	0.816	[0.808, 0.824]	2.1 sec/page	CPU: 92%, RAM: 4.8GB
Legal-BERT	0.931	[0.925, 0.937]	0.915	[0.909, 0.921]	0.923	[0.917, 0.929]	7.8 sec/page	GPU: 78%, RAM: 12.4GB
GPT-4	0.894	[0.887, 0.901]	0.868	[0.860, 0.876]	0.881	[0.874, 0.888]	12.4 sec/page*	API latency: 3.2s avg
Claude 3	0.876	[0.868, 0.884]	0.832	[0.824, 0.840]	0.854	[0.846, 0.862]	10.8 sec/page*	API latency: 2.8s avg

*Includes API latency (network + processing). Local processing time excludes latency.

6.1.1 Key Statistical Findings

- Legal-BERT significantly outperformed all other systems on F1-score ($p < 0.001$ for all pairwise comparisons). The effect size versus GPT-4 was large (Cohen's $d = 1.24$).
- The precision-recall trade-off varied systematically by system type:

Rule-based systems had highest precision (0.823) but lowest recall (0.641), reflecting their conservative design

Legal-BERT achieved the best balance (precision 0.931, recall 0.915)

LLMs showed intermediate performance with GPT-4 outperforming Claude 3 (difference statistically significant, $p = 0.003$)

- Processing speed showed inverse relationship with accuracy: The most accurate system (Legal-BERT) was 3.7× slower than the fastest (XGBoost), though still processed ~460 pages/hour.
- Confidence intervals revealed important patterns: While GPT-4's point estimate F1 (0.881) was close to Legal-BERT's (0.923), their 95% CIs did not overlap, indicating statistically significant difference.

6.1.2 Detailed Error Analysis by Clause Type

We analyzed performance across the 25 clause types to identify systematic patterns:

Table 6. Performance variation by clause type (top 10 by frequency).

Clause Type	Frequency	Rule-Based F1	Legal-BERT F1	GPT-4 F1	Difficulty Index*
Confidentiality	12.4%	0.892	0.971	0.954	Low (0.12)
Limitation of Liability	9.8%	0.843	0.952	0.928	Medium (0.34)
Termination	8.7%	0.812	0.941	0.917	Low (0.18)
Indemnification	7.9%	0.784	0.928	0.901	High (0.62)
Governing Law	7.2%	0.921	0.983	0.962	Very Low (0.08)
Warranty	6.8%	0.763	0.912	0.896	Medium (0.41)
Payment Terms	5.9%	0.801	0.936	0.919	Low (0.22)
Intellectual Property	5.4%	0.721	0.887	0.862	High (0.58)
Audit Rights	4.7%	0.892	0.971	0.953	Low (0.15)
Force Majeure	3.9%	0.941	0.987	0.974	Very Low (0.06)

*Difficulty Index: 0-1 scale based on expert disagreement rate and syntactic variation

6.1.3 Pattern Analysis

- High-Performance Clauses: Governing law, force majeure, and confidentiality clauses showed near-perfect performance (>0.95 F1) across all systems except rule-based. These clauses tend to use standardized language with limited variation.
- High-Difficulty Clauses: Indemnification and intellectual property clauses showed the largest performance gaps between systems. These clauses:
  Have high syntactic variation (37 distinct phrasings for indemnification in our corpus)
  Often contain complex cross-references
  Have significant jurisdictional variation
- Rule-Based Failure Modes: Rule-based systems performed particularly poorly on indemnification (F1=0.784 vs 0.928 for Legal-BERT) and IP clauses (0.721 vs 0.887). Manual inspection revealed these failures were due to:
  Novel phrasings not covered by rules
  Complex conditional structures (if-then-else in indemnification scope)
  Nested definitions (IP definitions referencing other documents)

6.2 Domain-Specific Performance Analysis

Table 7. F1 Scores by contract domain with statistical testing.

Domain	Documents	Rule-Based	Legal-BERT	GPT-4	Performance Gap*	p-value**
NDA	847	0.812	0.941	0.923	+12.9%	<0.001
MSA	921	0.784	0.928	0.901	+14.4%	<0.001
SOW	634	0.801	0.936	0.917	+13.5%	<0.001
Employment	512	0.721	0.912	0.896	+19.1%	<0.001
Licensing	489	0.694	0.887	0.862	+19.3%	<0.001
M&A	287	0.432	0.847	0.824	+41.5%	<0.001
Procurement	723	0.756	0.921	0.903	+16.5%	<0.001
Real Estate	342	0.712	0.898	0.881	+18.6%	<0.001
Weighted Average	5,247	0.721	0.907	0.886	+18.6%	<0.001

*Performance Gap = (Legal-BERT F1 - Rule-Based F1) / Rule-Based F1
**p-value for Legal-BERT vs Rule-Based comparison

Critical Findings by Domain:

M&A Agreements - The Worst Case for Rule-Based Systems:

Rule-based systems showed catastrophic failure on M&A agreements (F1=0.432). Detailed analysis revealed three primary failure modes:

- Bespoke Drafting: M&A agreements are highly customized with firm-specific language. Only 34% of clauses in M&A agreements matched known templates vs 78% in NDAs.
- Complex Cross-Referencing: M&A agreements average 42 cross-references per page vs 8 in NDAs. Rule-based systems struggled with chains like: "The limitations in Section 8.2 shall not apply to breaches of representations in Section 3.1(a)(iii) as they relate to Schedule 2.4(b)."
- Definition Dependencies: Defined terms often have multi-paragraph definitions with exceptions and qualifications that rule patterns couldn't capture.

NDA/MSA - Best Case for All Systems:

Standardized agreements showed strong performance across all systems, but with important nuances:

- Rule-based: Performed well on standard clauses but missed variations (e.g., "survival period" phrased as "shall survive termination" vs "shall continue after expiration")
- Legal-BERT/LLMs: Caught variations but sometimes over-flagged (higher false positives on creative but harmless drafting)

Employment Agreements - Jurisdictional Complexity:

Employment contracts showed significant jurisdictional variation in performance:

- US agreements: All systems >0.90 F1
- UK agreements: Rule-based dropped to 0.642, Legal-BERT to 0.831
- Singapore agreements: Further drops across all systems

The primary challenge was jurisdiction-specific statutory references (e.g., UK Equality Act 2010, Singapore Employment Act) that required specific legal knowledge.

Statistical Analysis of Domain Effects:

We performed a mixed-effects model to quantify domain impact:

Performance ~ System + Domain + System × Domain + (1 | Document)

Results showed:

- Domain had significant main effect ($F(7, 41976) = 142.3$, $p < 0.001$)
- System × Domain interaction was significant ($F(28, 41976) = 67.8$, $p < 0.001$), meaning performance degradation varied by system
- Rule-based systems showed strongest domain dependency ($\eta^2 = 0.38$), LLMs weakest ($\eta^2 = 0.12$)

This confirms that system choice should be domain-dependent—no single system performs best across all contract types.

6.3 Detailed Error Analysis

We conducted a comprehensive error analysis categorizing 12,847 misclassifications across all systems:

Table 8. Error type distribution with examples and impact.

Error Type	Rule-Based	Legal-BERT	GPT-4	Example	Impact Severity
Syntactic Variation	42%	12%	8%	"Liability shall be capped at" vs "Cap on liability is"	Medium
Semantic Ambiguity	28%	34%	41%	"Reasonable commercial efforts" (how defined?)	High
Cross-reference Miss	19%	38%	36%	Missed exception in referenced schedule	Very High
Hallucination	0%	0%	12%	Added non-existent jurisdiction clause	Critical
Jurisdictional	8%	12%	9%	UK vs US interpretation of "best efforts"	High
Other	3%	4%	3%	OCR errors, formatting issues	Low

Detailed Error Examples with Root Cause Analysis:

Example 1: Syntactic Variation (Rule-Based Failure)

- Clause: "Notwithstanding anything to the contrary herein, in no event shall Company's aggregate liability exceed..."
- Rule Trigger: "liability shall not exceed" (missed due to "in no event" phrasing)
- Root Cause: Rule looked for pattern "[liability] shall not exceed" but clause uses "in no event shall [liability] exceed"
- Impact: Missed liability cap detection. In this contract, the cap was unusually low (50% of fees), making this a severity 4 risk.

Example 2: Hallucination (GPT-4 Specific)

- Actual Text: "Governing law: This Agreement shall be governed by New York law."
- GPT-4 Output: "Governing law: This Agreement shall be governed by New York law. Note: Includes choice of forum clause specifying New York courts."

- Manual Review: No forum selection clause existed in document
- Root Cause: GPT-4 inferred common pairing (governing law often paired with forum selection) and hallucinated the pairing
- Impact: Created false risk flag, wasted lawyer time investigating non-existent clause

Example 3: Cross-Reference Chain Failure (All Systems)

- Clause 5.2: "Limitations set forth in Section 9 apply to all claims except as provided in Schedule B."
- Section 9.1: "Liability is capped at Fees paid in prior 12 months."
- Section 9.2: "The cap in 9.1 does not apply to IP infringement claims."
- Schedule B, Section 2: "For IP claims, liability is uncapped."
- System Output: All systems identified cap in Section 9.1 but only 23% followed the chain to Schedule B exception
- Root Cause: Need to maintain context across multiple sections and follow reference chains
- Impact: Severely underestimated liability exposure (actual: uncapped for IP, detected: capped)

Error Severity Distribution:

Not all errors are equal. We categorized errors by potential impact:

- Minor (42%): Wrong clause type but correct risk detection (e.g., misclassified as "Miscellaneous" instead of "Notices")
- Moderate (31%): Missed risk but low severity (1-2)
- Major (19%): Missed risk of severity 3-4
- Critical (8%): Missed risk of severity 5 or hallucination creating false critical risk

Critical Finding: Error severity distribution varied by system:

- Rule-based: Errors were mostly Major/Critical (64%) due to missing novel high-risk clauses
- Legal-BERT: Errors were more evenly distributed (Minor 38%, Critical 12%)
- GPT-4: Had highest Critical error rate (18%) due to hallucinations being treated as critical

6.4 Expert vs AI Agreement Analysis

Table 9. Alignment with expert judgments by multiple metrics.

System	Cohen's κ	Agreement Level	Major Disagreement Rate*	Severity Correlation (r)
Rule-Based	0.61	Substantial	14.3%	0.52
XGBoost	0.69	Substantial	11.8%	0.61
Legal-BERT	0.79	Almost Perfect	6.2%	0.78
GPT-4	0.71	Substantial	9.4%	0.67
Claude 3	0.68	Substantial	10.7%	0.63

*Major Disagreement = expert says risk ≥4, AI says no risk OR expert says no risk, AI says risk ≥4

6.4.1 Analysis of Disagreement Patterns

We analyzed 1,842 major disagreements (expert vs AI divergence on high-severity assessment) to identify patterns:

6.4.1.1 Pattern 1: Contextual Interpretation Gaps (47% of disagreements)

- Example: A limitation clause capped liability at "the greater of $1M or fees paid"
- Expert Assessment: Severity 2 (reasonable for $10M contract)
- AI Assessment: Severity 4 (flagged as unusually low cap)

- Root Cause: AI lacked transaction context (contract value)

6.4.1.2 Pattern 2: Jurisdictional Knowledge Gaps (28%)

- Example: Data processing clause referencing "adequate safeguards"
- Expert (UK): Severity 1 (references UK GDPR standard clauses)
- AI (US-trained): Severity 3 (flags as vague)
- Root Cause: AI lacked jurisdiction-specific knowledge of standard clauses

6.4.1.3 Pattern 3: Business Practice Understanding (15%)

- Example: Payment terms "net 60 days"
- Expert: Severity 1 (standard in this industry)
- AI: Severity 3 (flags as unusually long)
- Root Cause: AI lacked industry-specific norm knowledge

6.4.1.4 Pattern 4: Cross-Document Dependencies (10%)

- Example: Clause references "the Policy" defined in separate document
- Expert: Severity assessment depends on Policy content
- AI: Assessed based on clause text alone
- Root Cause: AI analyzed only the main document

6.4.2 Statistical Analysis of Agreement Factors

We ran a logistic regression predicting disagreement:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(\text{Complexity}) + \beta_2(\text{Jurisdiction}) + \beta_3(\text{System})$$

Results:

- Document complexity strongest predictor (OR = 2.34 per SD increase, $p < 0.001$)
- Non-US jurisdiction increased odds (OR = 1.87, $p < 0.001$)
- System type: Legal-BERT had lowest odds (reference), GPT-4 OR = 1.52, Rule-based OR = 2.41

6.4.3 Practical Implications

- AI performs best on standard US contracts where language and norms are predictable
- Performance degrades with complexity—AI should not be trusted on complex agreements without human review
- Jurisdictional adaptation is critical for global deployment
- Business context matters—AI needs industry and transaction-specific knowledge

6.4.4 Confidence Calibration Analysis

We examined whether system confidence scores correlated with accuracy:

Table 10. Confidence calibration by system.

Confidence Bin	Rule-Based Accuracy	Legal-BERT Accuracy	GPT-4 Accuracy	Expected*
0.0-0.5	42%	68%	61%	25%
0.5-0.7	67%	82%	74%	60%
0.7-0.9	84%	91%	83%	80%
0.9-1.0	96%	97%	89%	95%

*Expected = perfect calibration (accuracy = confidence)

6.4.5 *Findings*

- Legal-BERT showed best calibration—confidence scores meaningful predictors of accuracy
- GPT-4 was overconfident—high confidence (0.9+) predictions were wrong 11% of time vs 3% for Legal-BERT
- Rule-based showed reasonable calibration but with lower overall accuracy

This has important workflow implications: systems with good calibration can route low-confidence predictions for human review, improving efficiency.

VI. CONCLUSION

This study provides a rigorous, evidence-based assessment of AI-based legal decision-making systems for contract analysis and risk detection, directly addressing the persistent gap between algorithmic benchmark performance and the practical realities of legal practice. Through large-scale experimentation on 5,247 real-world contracts annotated by experienced legal professionals, we demonstrate that contemporary legal AI systems exhibit substantial heterogeneity not only in accuracy, but—more critically—in reliability, consistency, cross-jurisdictional robustness, and cost-effectiveness

An Experimental Assessment of AI.

Several core conclusions emerge. First, no existing AI system is suitable for fully autonomous legal decision-making in contract analysis. Even the best-performing model, a fine-tuned Legal-BERT, while achieving state-of-the-art clause classification performance (F1 = 0.923), exhibited non-trivial degradation under jurisdictional shift and remained vulnerable to cross-reference and contextual dependency failures. Large language models, despite their flexibility and expressive reasoning capabilities, demonstrated unacceptable levels of decision instability and hallucination for high-stakes legal environments, undermining auditability and practitioner trust. Rule-based systems, although transparent and deterministic, failed systematically on complex and bespoke agreements—precisely where legal risk is most consequential.

Second, the results confirm that standard NLP metrics are insufficient for evaluating legal AI systems. Metrics such as accuracy and F1-score obscure the asymmetric cost structure of legal errors. By introducing and empirically validating the False-Negative Risk Penalty (FNRP) and Severity-Weighted F1 (SwF1), this study shows that systems with superficially strong aggregate performance may nonetheless miss a disproportionate share of high-severity risks. Our findings establish that systems exceeding an FNRP threshold of 0.15 pose materially elevated legal exposure and should not be deployed without stringent human oversight.

Third, we demonstrate that deployment decisions must be domain- and context-sensitive. Performance varied markedly across contract types, with standardized agreements (e.g., NDAs, MSAs) representing a best-case scenario and complex instruments (e.g., M&A agreements) revealing fundamental system limitations. Jurisdictional transfer further exacerbated error rates, underscoring that legal AI cannot be treated as jurisdiction-agnostic language technology. Legal meaning remains embedded in statutory frameworks, precedent, and professional norms that are only partially encoded in data-driven models.

Fourth, and most importantly for practice, the study shows that well-designed human–AI collaboration outperforms any single-system approach. The proposed three-tier human-in-the-loop framework—routing low-risk, high-confidence outputs for automated handling while escalating ambiguous or high-severity cases to legal experts—achieved a 64% reduction in attorney review time while maintaining 99.7% coverage of critical risks. This finding reframes the role of AI in law: not as a substitute for legal judgment, but as a force multiplier when embedded within carefully governed workflows.

VII.FUTURE RESEARCH DIRECTIONS

The findings of this study point to several high-impact avenues for future research. First, there is a clear need for context-aware legal AI architectures capable of modeling long-range dependencies across entire contract ecosystems, including referenced schedules, external policies, and transaction-specific metadata.

Second, jurisdiction-sensitive modeling, potentially through modular legal knowledge integration or continual learning mechanisms, is essential for global deployment. Third, future work should explore cost-sensitive and severity-aware training objectives, aligning model optimization more closely with real legal risk profiles rather than symmetric classification loss functions. Finally, advancing legally grounded explainability, where system rationales mirror doctrinal legal reasoning rather than surface-level textual cues, remains a critical prerequisite for trust and regulatory acceptance.

In sum, this research establishes a reproducible, empirically grounded framework for evaluating and deploying AI systems in legal contract analysis. It challenges prevailing narratives of imminent legal AI autonomy and instead provides a pragmatic, evidence-driven roadmap for responsible adoption. By aligning technical evaluation with legal risk, jurisdictional reality, and professional practice, this study contributes both to the scientific literature on legal AI and to the informed, safe integration of AI into the legal domain.

REFERENCES

[1] D. A. Pashentsev and Y. G. Babaeva, "Artificial intelligence in law-making and law enforcement: Risks and new opportunities," *Vestnik Sankt-Peterburgskogo Universiteta. Pravo*, vol. 15, no. 2, 2024, doi: 10.21638/spbu14.2024.214.

[2] Dr. S. Borade, "DETEK-AI: A Web-based Deepfake Detection System," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 08, no. 05, 2024, doi: 10.55041/ijsrem31285.

[3] "A comprehensive study of Cybercrime and Digital Forensics through Machine Learning and AI," *Al Rafidain Journal of Engineering Sciences*, vol. 3, no. 1, 2025, doi: 10.61268/hff1pp49.

[4] P. Haley and D. N. Burrell, "Using Artificial Intelligence in Law Enforcement and Policing to Improve Public Health and Safety," *Law, Economics and Society*, vol. 1, no. 1, 2025, doi: 10.30560/les.v1n1p46.

[5] K. A. Talukder and T. F. Shompa, "ARTIFICIAL INTELLIGENCE IN CRIMINAL JUSTICE MANAGEMENT: A SYSTEMATIC LITERATURE REVIEW," *Non human journal*, vol. 1, no. 01, 2024, doi: 10.70008/jmldeds.v1i01.42.

[6] D. M. Makhmudova and S. M. Almufti, "Hybrid Metaheuristic Frameworks for Multi-Objective Engineering Optimization Problems," *Qubahan Techno Journal*, vol. 3, no. 1, pp. 1–14, Feb. 2024, doi: 10.48161/qtj.v3n1a23.

[7] S. M. Almufti and S. R. M. Zeebaree, "Leveraging Distributed Systems for Fault-Tolerant Cloud Computing: A Review of Strategies and Frameworks," *Academic Journal of Nawroz University*, vol. 13, no. 2, pp. 9–29, May 2024, doi: 10.25007/ajnu.v13n2a2012.

[8] L. S. F. Lin, "Organisational Challenges in US Law Enforcement's Response to AI-Driven Cybercrime and Deepfake Fraud," *Laws*, vol. 14, no. 4, 2025, doi: 10.3390/laws14040046.

[9] L. Tang and C. Shen, "Multimodal AI-driven object detection with uncertainty quantification for cardiovascular risk assessment in autistic patients," *Front Cardiovasc Med*, vol. 12, 2025, doi: 10.3389/fcvm.2025.1606159.

[10] M. Sarfraz, I. A. Sumra, B. Khalid, and E. Fatima, "AI-Driven Predictive Threat Detection and Cyber Risk Mitigation: A Survey," *Journal of Computing & Biomedical Informatics*, vol. 8, no. 2, 2025.

[11] S. M. Almufti, B. Wasfi Salim, and R. Rajab Asaad, "Automatic Verification for Handwritten Based on GLCM Properties and Seven Moments," *Academic Journal of Nawroz University*, vol. 12, no. 1, pp. 130–136, Feb. 2023, doi: 10.25007/ajnu.v12n1a1651.

[12] M. C. Dela Cruz, S. M. Almufti, and J. Bošković, "Portable Few-Shot Learning for Early Warning Systems in Small Private Online Courses: A CNN-Based Predictive Framework for Student Performance," *Qubahan Techno Journal*, vol. 3, no. 4, pp. 1–13, Dec. 2024, doi: 10.48161/qtj.v3n4a42.

[13] T. Miller, I. Durlik, E. Kostecka, S. Sokołowska, P. Kozlovska, and R. Zwolak, "Artificial Intelligence in Maritime Cybersecurity: A Systematic Review of AI-Driven Threat Detection and Risk Mitigation Strategies," 2025. doi: 10.3390/electronics14091844.

[14] Ç. Sıcakyüz, R. Rajab Asaad, S. Almufti, and N. R. Rustamova, "Adaptive Deep Learning Architectures for Real-Time Data Streams in Edge Computing Environments," *Qubahan Techno Journal*, vol. 3, no. 2, pp. 1–14, Jun. 2024, doi: 10.48161/qtj.v3n2a25.

[15] H. A. Hakim, C. B. E. Praja, and S. Ming-Hsi, "AI in Law: Urgency of the Implementation of Artificial Intelligence on Law Enforcement in Indonesia," *Jurnal Hukum Novelty*, vol. 14, no. 1, 2023, doi: 10.26555/novelty.v14i1.a25943.

[16] I. A. Olubiyi, Rahamat Oyedeji-Oduyale, and Damilola M.Adeniyi, "ARTIFICIAL INTELLIGENCE AND THE LAW: AN OVERVIEW," *ABUAD Law Journal*, vol. 12, no. 1, 2024, doi: 10.53982/alj.2024.1201.01-j.

[17] M. Araszkiewicz, T. Bench-Capon, E. Francesconi, M. Lauritsen, and A. Rotolo, "Thirty years of Artificial Intelligence and Law: overviews," *Artif Intell Law (Dordr)*, vol. 30, no. 4, 2022, doi: 10.1007/s10506-022-09324-9.

[18] I. K. Nti, S. Boateng, J. A. Quarcoo, and P. Nimbe, "Artificial Intelligence Application in Law: A Scientometric Review," 2024. doi: 10.47852/bonviewAIA3202729.

[19] M. Abdel-Basset, R. Mohamed, S. A. A. Azeem, M. Jameel, and M. Abouhawwash, "Kepler optimization algorithm: A new metaheuristic algorithm inspired by Kepler's laws of planetary motion," *Knowl Based Syst*, vol. 268, p. 110454, May 2023, doi: 10.1016/j.knosys.2023.110454.

[20] J. Lee, *Artificial Intelligence and International Law*. 2022. doi: 10.1007/978-981-19-1496-6.

[21] A. V. Minbaleev, "THE CONCEPT OF 'ARTIFICIAL INTELLIGENCE' IN LAW," *Bulletin of Udmurt University. Series Economics and Law*, vol. 32, no. 6, 2022, doi: 10.35634/2412-9593-2022-32-6-1094-1099.

[22] D. J. Brand, "Algorithmic decision-making and the law," *eJournal of eDemocracy and Open Government*, vol. 12, no. 1, 2020, doi: 10.29379/jedem.v12i1.576.