International Journal of Scientific Research Studies

ISSN (print): 3050-6905, ISSN (online): 3050-6913

DOI: 10.58806/ijsrs.2025.v2i11n01

Volume 02 Issue 11 December 2025

Page No: 115-123

Hybrid NLP Framework for Contract Risk Assessment-A Dual-Agent Approach Combining Roberta and Rule-Based Analysis with Unified Decision-Making

Sandhya B S[*1], Nishanth R[2]

[1]Student - DSML, PES University, Hosur Rd, Konappana Agrahara, Bengaluru, 560100, Karnataka, India.

[2]Data Scientist, Great Learning, PES University, Hosur Rd, Konappana Agrahara, Bengaluru, 560100, Karnataka, India.

ABSTRACT

Efficient contract analysis in the legal domain demands accuracy, traceability, and interpretability, yet existing NLP systems often rely on opaque neural models, limiting transparency critical for law and compliance. We propose a dual-agent NLP frame-work that integrates high-performance contract analysis with explainability. A Clause Extraction Agent employs fine-tuned RoBERTa ensembles with pattern matching and semantic validation to identify key clauses, while an Enhanced Risk Assess-ment Agent combines rule-based legal logic with LLM-driven insights for vagueness detection, contextual risk evaluation, and clause rewriting recommendations. Modular design enables interpretability, focused debugging, and avoids black-box limitations. A proof-of-concept demonstrates significant improvements in clause extraction and risk assessment over baseline methods. Stateless data handoffs ensure full transparency from extraction to actionable recommendations. A Streamlit interface provides inter-active dashboards for overview, detailed clause analysis, risk assessment, and rewrite guidance. This framework establishes a blueprint for explainable, LLM-integrated legal AI suitable for real-world deployment.

KEYWORDS: Legal AI, Contract Analysis, NLP, Roberta, Large Language Mod-Els (Llms), Explainable AI, Risk Assessment, Clause Extraction

ARTICLE DETAILS

Published On:
11 December 2025

Available on:
https://ijsrs.org/

1 INTRODUCTION

The increasing complexity, volume, and legal risks embedded in contractual documents have intensified the demand for automated solutions in legal text analysis. Traditional manual review, while comprehensive, is inherently time-consuming, costly, and prone to inconsistency. The integration of Natural Language Processing (NLP) into the legal domain has emerged as a viable **response** to these challenges, with recent advances in transformer-based models and large language models (LLMs) enabling significant improvements in clause extraction, risk identification, and contract interpretation [1][2][3]

Early innovations such as LegalBench [4] and CUAD [3] provided essential anno-tated datasets that accelerated the benchmarking of clause extraction systems, while transformer architectures such as RoBERTa [5] and LegalBERT [6] demonstrated superior adaptability to domain-specific tasks. Despite these advances, existing sys-tems are often fragmented: clause extraction is treated separately from risk assessment, and LLMs are either underutilized or deployed as black-box models, raising concerns regarding transparency and legal defensibility .

Recent empirical studies underscore the transformative potential of LLMs in contract analysis. For instance, GPT-4–based systems have achieved performance comparable to legal professionals while reducing review costs by up to 99.97% [7][8]. Similarly, benchmark initiatives such as ContractEval [5] and ACORD [9] have demon-strated that hybrid systems integrating rule-based logic with LLM capabilities yield substantial improvements in both accuracy and explainability. However, research consistently highlights the need for interpretable architectures that maintain traceabil-ity and accountability, particularly in high-stakes applications such as risk-sensitive contract review [10][11][12].

This research proposes a dual-agent framework that integrates RoBERTa-based clause extraction with LLM-driven risk assessment in a unified and transparent decision-making pipeline. The key contributions are:

i. A modular design combining transformer-based extraction, semantic similarity, and rule-based reasoning.
ii. Integration of LLMs into a traceable workflow for contract risk analysis.
iii. A comprehensive evaluation covering accuracy, calibration, and practical utility.
iv. A proof-of-concept system demonstrating end-to-end contract analysis.

2 LITERATURE SURVEY

2.1 Evolution of Legal NLP

The evolution of NLP in the legal domain has progressed from symbolic rule-based systems [13][14] to data-driven machine learning approaches, and more recently to transformer and LLM-based paradigms [15]. Recent surveys [16][1] provide a comprehensive overview of these advancements, highlighting diverse applications such as clause extraction and legal drafting, while emphasizing persistent challenges related to domain adaptation, explainability, and judicial defensibility.

Benchmarks such as LegalBench [4] and CUAD [3] have been instrumental in pro-viding standardized evaluation protocols, enabling transformer models like RoBERTa [5] to outperform domain-specific models in clause extraction tasks [17]. Domain adap-tation techniques such as FastDoc [18] and topic-driven summarization [15] further demonstrate that hybrid pretraining strategies enhance efficiency without excessive computational cost.

2.2 Transformer-Based Contract Analysis

Transformers remain central to legal clause analysis. Comparative studies (Table 1) demonstrate that RoBERTa outperforms domain-specific LegalBERT when fine-tuned on CUAD [15], [17]. Subsequent applications, including LegalBERT for small business contracts [6] and ensemble transformer models for patent analysis [19], further validate the adaptability of transformer architectures. Empirical evidence shows that ensemble approaches integrating statistical, pattern-based, and semantic components achieve superior precision and recall over standalone models [18], [19].

Table 1: Comparative Approaches for Legal NLP and Contract Analysis

Approach	Strengths	Weaknesses	Applications
Rule-based Systems	High interpretability; deterministic outputs; easy to audit	Lack scalability; brittle to linguistic variation; limited generalization	Regulatory compliance, standardized contracts, legal checklists
Transformer-only Models (e.g., RoBERTa, LegalBERT)	Strong contextual understanding; superior performance on clause extraction tasks	Require large training data; domain adaptation challenges; limited transparency	Clause extraction, legal summarization, entity recognition
LLM-only Systems (e.g., GPT-3.5, GPT-4)	High adaptability; can perform zero/few-shot tasks; strong reasoning abilities	Opaque decision-making; high computational cost; risk of hallucinations; limited defensibility	Clause summarization, drafting assistance, legal Q&A
Hybrid Systems (Rule + Transformer + LLM)	Balances adaptability with traceability; improved accuracy and explainability; scalable for diverse contracts	Increased architectural complexity; dependency on model coordination	Risk-sensitive contract review, compliance automation, multi-agent legal workflows

2.3 LLMs for Legal Risk Assessment

The integration of LLMs in contract analysis extends beyond automation to advanced decision support. Empirical studies show that GPT-3.5 and GPT-4 achieve human-comparable performance in information extraction, clause summarization, and risk detection [7], [20], [21]. Benchmarks such as ContractEval **Liu2025** and ACORD [9] further demonstrate that hybrid rule–LLM frameworks outperform monolithic models. Comparative analyses also underscore the need for transparency and interpretability in legal AI systems [2], [22], [23]..

Hybrid neural-symbolic systems are increasingly proposed to bridge the gap between LLM adaptability and rule-based interpretability [13][24][25]. These frame-works allow LLMs to provide nuanced reasoning while rules ensure traceability, meeting explainability standards demanded in legal contexts [10][11][12].

2.4 Multi-Agent and Industry Implementations

Multi-agent architectures are increasingly adopted for modular and interpretable legal AI. Systems such as ChatLaw [2], LAWMA [13], and Syllo [26] demonstrate that agent specialization enhances both accuracy and transparency. Industry implementations, including eBay's NDA review and Codvo.ai's RoBERTa-based pipeline, report up to 10× efficiency gains and reduced approval latency, validating the practicality of dual-and multi-agent designs in operational settings.

2.5 Specialized Datasets and Benchmarking Efforts

High-quality datasets remain foundational to progress in legal NLP. Beyond CUAD [3] and LegalBench [4], specialized resources such as LAWMA [13] enhance LLM interpretability through annotated corpora. Recent benchmarks—ContractEval [5] and ACORD [9]—enable systematic clause-level and retrieval evaluations for con-tract analysis and risk assessment. Complementary corpora like FastDoc [18] leverage metadata-driven pretraining for efficiency gains. Collectively, these initiatives advance standardization and accelerate hybrid framework adoption.

2.6 Explainability, Transparency, and Legal Compliance Explainability remains a central challenge in deploying AI for legal workflows. Schol-ars emphasize that legal AI systems must provide auditable decision paths consistent with regulatory and professional standards [10], [11]. In outcome prediction, explain-ability is viewed as essential for judicial defensibility rather than a mere technical enhancement [12]. Neural-symbolic approaches with structured prompting further aim to balance interpretability and accuracy [25]. Collectively, these works stress that LLM-based systems must justify decisions in forms acceptable to courts, regulators, and practitioners.

2.7 Cross-Domain Inspirations

Insights from adjacent domains further inform legal NLP. In patent analysis, ensemble BERT-based models enhance semantic similarity across technical texts [19]. In health-care compliance, AI-enabled blockchain frameworks integrate legal reasoning with secure auditability [23]. Financial regulation studies similarly report efficiency gains under strict oversight [27]. Collectively, these cross-domain findings underscore that hybrid AI architectures—combining statistical, symbolic, and semantic reasoning—are essential for accuracy and regulatory compliance in sensitive legal applications.

2.8 Emerging Directions in Legal AI

Recent research introduces new paradigms in legal AI deployment. Multi-agent frame-works such as ChatLaw [2] and Syllo [26] demonstrate that modularity enhances scalability and interpretability. Developments in contract summarization and simplifi-cation expand accessibility for non-legal users [15], [21]. The convergence of blockchain and explainable AI enables immutable, auditable contract review records [23]. Emerg-ing surveys further anticipate deeper integration of agentic reasoning, explainability, and domain-adapted pretraining aligned with evolving technical and legal standards [16], [27].

2.9 Research Gap and Study Positioning

Despite significant progress, key challenges persist: integration architectures remain fragmented [16], confidence calibration is seldom addressed [27], and evaluation metrics often emphasize accuracy over legal utility and compliance [5][9][14]. Few frameworks achieve both transparency and adaptability. This study addresses these limitations by introducing a dual-agent, LLM-integrated framework that unifies rule-based and semantic reasoning for clause extraction, risk assessment, and explainable decision-making. Unlike prior systems [4][22][7], it ensures traceable, interpretable, and analytically robust outputs.

3 MATERIALS AND METHODS

3.1 Datasets

Three datasets were used for training, evaluation, and demonstration of the system (Table 2):

- **Clause Extraction (CUAD-based):** A curated subset of 7,000 examples cov-ering 41 clause types was used for training and validation (6,500 train, 500 validation). Tokenization expanded these to approximately 65,000 and 5,000 samples, respectively.
- **Clause Extraction Test Set:** 20 synthetic contracts in CUAD-style JSON, manually annotated for 25+ clause types, to evaluate extraction accuracy and robustness.
- **Risk Assessment Test Set:** Clause-level annotations for 20 contracts, with expected risk levels (Low, Medium, High) and numerical score ranges, used to assess the risk evaluation agent.
- **Streamlit Demo Dataset:** PDF contracts uploaded manually to demonstrate the end-to-end workflow, including preprocessing, extraction, and risk assessment.

Table 2: Dataset summary used in training, testing, and demonstration.

Task	Format	Size	Description
Clause Extraction (train/val)	CUAD JSON	7,000 examples (6,500 train, 500 val; expanded to 70k after tokenization)	CUAD subset used for fine-tuning
Clause Extraction (test)	CUAD-style JSON	20 contracts	Synthetic contracts with manually annotated clauses (25+ types)
Risk Assessment (test)	Clause-level JSON	20 test cases	Clauses labeled with expected risk levels and score ranges
Streamlit Demo	PDF contracts	N/A (manual uploads)	End-to-end flow for clause extraction and risk evaluation

3.2 Workflow Integration and System Architecture

The framework follows a dual-agent modular workflow illustrated in Figure 1. :

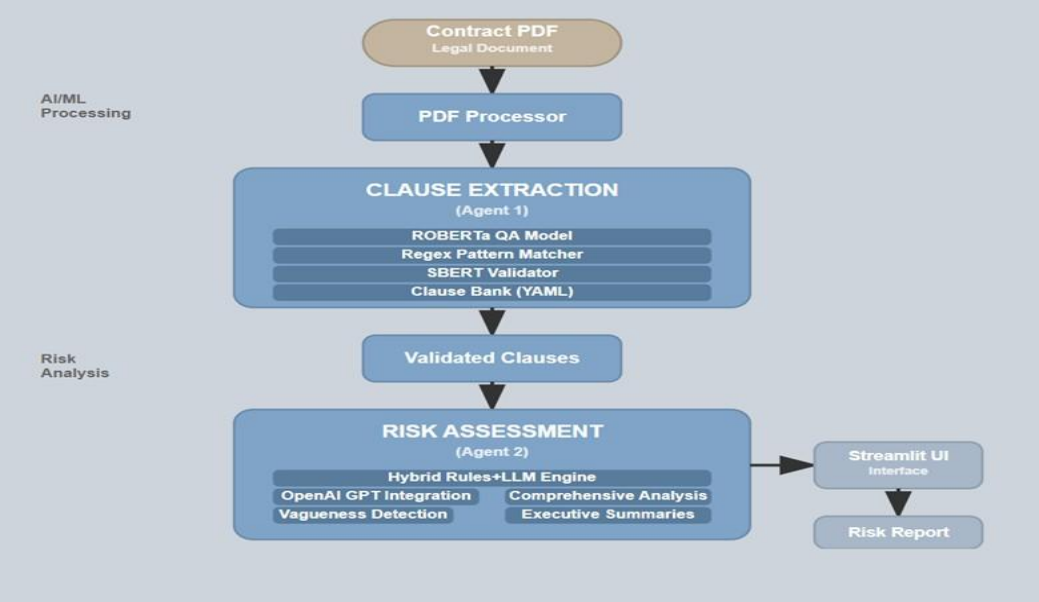

Fig. 1: System architecture of the proposed dual-agent framework.

1. **PDF Preprocessing Agent:** Extracts and normalizes contract text using direct parsing with fallback OCR.
2. **Clause Extraction Agent (Agent 1):** Identifies clauses using an ensemble of:

i. Fine-tuned RoBERTa QA model (CUAD subset),
ii. Regex-based deterministic clause patterns,
iii. SBERT embeddings for semantic similarity.

The ensemble confidence is defined as:

$$C_{final} = (0.5 \cdot C_{QA}) + (0.3 \cdot C_{regex}) + (0.2 \cdot C_{sem}) \quad (1)$$

where C_{QA} is the transformer confidence, C_{regex} is the regex match score, and C_{sem} is the semantic similarity score.

3. **Risk Assessment Agent (Agent 2):** Combines rule-based and LLM-derived scoring. The final risk score is given by:

$$R_{final} = (0.6 \cdot R_{rule} + 0.4 \cdot R_{LLM}) \cdot \quad (2)$$

where R_{rule} is the rule-based score, R_{LLM} is the GPT-derived contextual score, and γ is a confidence adjustment factor.

4. **Unified Decision Engine:** Integrates agent outputs, determines risk levels, flags vague terms, and triggers clause rewriting.
5. **User Interface Layer:** A Streamlit-based dashboard for visualization, analysis, and reporting.

3.3 User Interface Implementation

A Streamlit dashboard was developed with four main tabs:

i. Dashboard Tab: High-level contract metrics and summaries.

ii. Clause Details Tab: Extracted clauses with confidence scores and semantic validation.
iii. Risk Analysis Tab: Clause-level risk scores, vagueness detection, and AI explanations.
iv. Rewrite Analysis Tab: LLM-generated clause revisions with justifications and comparisons.
The interface supports real-time progress indicators and exports in CSV, JSON, and Markdown.

3.4 Implementation Details

The system was implemented in Python 3.10 and deployed locally using Streamlit.A detailed summary of the environment and configurations is provided in Table 3.:

Table 3: Implementation environment and setup.

Component	Specification
Programming	Python 3.10, VSCode IDE
Libraries	transformers, datasets, sentence-transformers, scikit-learn, Streamlit
Model	deepset/roberta-base-squad2 (CUAD fine-tuned)
Training	Hugging Face Trainer + LoRA adapters
GPU	NVIDIA Tesla T4
Deployment	Streamlit dashboard (localhost)

3.5 Evaluation Metrics

Evaluation was carried out at two levels: clause extraction and risk assessment. Stan-dard measures such as Precision, Recall, F1 score, ROUGE-n, and Mean Absolute Error (MAE) follow their conventional definitions and are therefore not repeated here. To ensure completeness, the task-specific metrics used in this work are formally defined below.

3.5.1 Clause Extraction

Exact Match (EM): Measures the proportion of predicted clauses that exactly match the gold-standard annotations.

$$EM = \frac{\text{Number of exact string matches}}{\text{Total number of clauses}}$$

Model Confidence (MC): Captures the average confidence level of the model across predictions.

$$MC = \frac{1}{N} \sum_{i=1}^{N} \max_{c} p_i(c)$$

3.5.2 Risk Assessment

Vagueness Detection Rate (VDR): Proportion of vague clauses correctly identi-fied by the system.

$$VDR = \frac{\text{Correctly detected vague clauses}}{\text{Total vague clauses in gold standard}}$$

Rewrite Generation Rate (RGR): Proportion of vague clauses for which the system successfully generated a rewrite.

$$RGR = \frac{\text{Vague clauses with generated rewrites}}{\text{Total vague clauses detected}}$$

Coverage (COV): Proportion of clauses in the document that were assigned a risk assessment by the system.

$$COV = \frac{\text{Clauses with assigned risk assessment}}{\text{Total number of clauses}}$$

3.5.3 Testing Procedure

A three-stage testing protocol was followed:

i. **Clause Extraction Testing:** Performance measured on 20 CUAD-style syn-thetic contracts with manually annotated clauses (25+ types).

ii. **Risk Assessment Testing:** Evaluated on 20 annotated clauses with predefined risk levels and score ranges, measuring classification accuracy, MAE, and rewrite correctness.

iii. **End-to-End System Testing (Streamlit Demo):** Full workflow tested with uploaded PDF contracts, validating preprocessing, extraction, unified scoring, vagueness detection, rewriting, and reporting.

Baseline comparison was performed against a rule-based + SBERT model for risk assessment, with improvements attributed to LLM integration and unified scoring.

4. RESULTS

4.1 Clause Extraction Performance

Metric	Score
Token F1	0.823
Token Precision	0.913
Token Recall	0.870
ROUGE-1	0.865
ROUGE-2	0.847
ROUGE-L	0.860
Confidence	0.742
Exact Match (%)	54.8

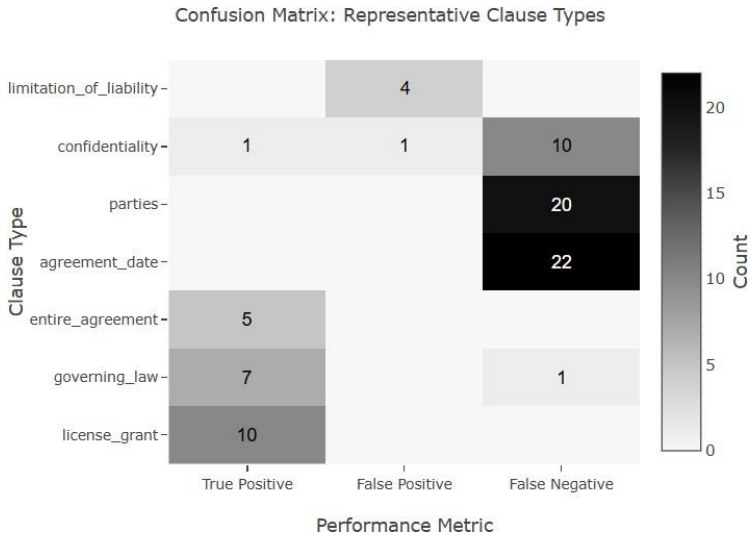

Fig. 2: Confusion matrix (subset) showing reliable detection for clauses such as governing law and license grant, and higher error rates for agreement date and parties.

The ensemble clause extraction agent achieved strong overall performance as shown in Table 4.. Token-level evaluation yielded a precision of 0.913, recall of 0.870, and F1 score of 0.823. Overlap-based metrics (ROUGE-1 = 0.865, ROUGE-2 = 0.847, ROUGE-L = 0.860) confirmed robust span coverage. Exact match accuracy was 54.8%, while the average model confidence was 0.742.

Clause-level confusion matrix shown in Figure 2 shows High-performing clauses included license grant (10 TP, 0 FP/FN), governing law (7 TP, 1 FN), and entire agreement (5 TP, 0 FP/FN), demonstrating reliable extraction capabilities. However, several clause types showed systematic detection failures, including agree-ment date (0 TP, 22 FN), parties (0 TP, 20 FN), and confidentiality (1 TP, 10 FN). Additionally, limitation of liability exhibited false positive issues (0 TP, 4 FP), suggesting potential misclassification with similar clause structures.

4.2 Risk Assessment Performance

A comparative analysis between the baseline SBERT+Rules system and the proposed LLM+Rules framework is presented in Table 5, highlighting improvements in accuracy, error reduction, and functionality.

Metric	Baseline (SBERT+Rules)	Enhanced (LLM+Rules)
Risk Level Accuracy	50%	90%
Mean Absolute Error	1.37	0.79
Coverage Rate	100%	100%
Rewrite Generation	None	90% clauses
Vagueness Detection	None	0.4 avg terms/clause

Table 5: Performance comparison of baseline and enhanced risk assessment systems. The hybrid risk assessment system significantly outperformed the baseline (rule-based + SBERT). Risk-level classification accuracy improved from 50% to 90%, and mean absolute error (MAE) decreased from 1.37 to 0.79. Enhanced functionality included automated rewrite generation for 90% of cases and vagueness detection at a rate of 0.4 terms per clause.

4.3 Unified Decision-Making Insights

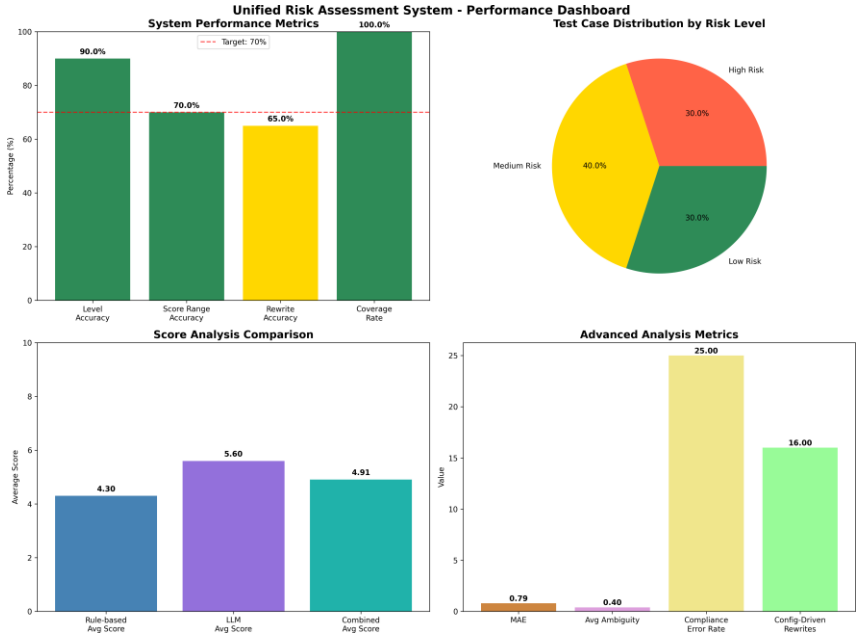

Fig. 3: Unified Risk Assessment System – Performance Dashboard showing overall accuracy, coverage, and risk distribution across test cases.

The unified scoring mechanism enabled accurate integration of rule-based and LLM-derived insights. In 18 of 20 test cases (90%), the system correctly determined whether clause rewriting was required. Vagueness detection flagged terms such as reasonable, appropriate, and substantial.

5 CONCLUSION AND FUTURE WORK

This study introduced a dual-agent NLP framework for contract risk assessment, com-bining RoBERTa-based clause extraction with LLM-enhanced risk evaluation through a unified decision-making process. By integrating transformer-based understanding with rule-based reliability, the framework addresses key limitations in existing legal AI systems, ensuring interpretability, transparency, and legal defensibility. The mod-ular dual-agent design effectively balances specialization and coordination: Agent 1 leverages RoBERTa QA, rule-based pattern matching, and SBERT validation for robust clause extraction across diverse contract types, while Agent 2 employs uni-fied risk assessment, combining rule-based analysis with LLM-driven reasoning for adaptive scoring and explainable decision-making. End-to-end evaluation on contract PDFs confirmed the framework's practical viability for real-world legal workflows.

PERFORMANCE METRICS:

- **Clause extraction F1 score:** 0.823
- **Risk assessment accuracy:** Improved from 50% to 90%
- **Mean Absolute Error (MAE):** Reduced from 1.37 to 0.79

Additional capabilities, including vagueness detection and intelligent clause rewriting, further enhance the system's practical utility, establishing it as a robust and interpretable solution for automated legal contract analysis.

Future work will focus on refining the clause extraction component through hyperparameter tuning, Low-Rank Adaptation (LoRA), and expanded evaluation on large-scale real-world contracts to strengthen generalization. Incorporating LLM-based fallback mechanisms is envisioned to address missed extractions and enhance seman-tic consistency. The risk assessment module will be extended with dynamic confidence scoring, adapting thresholds by clause type and risk severity to minimize false positives and negatives. In the longer term, the modular architecture may evolve into multi-agent frameworks where specialized sub-agents autonomously perform cross-contract consistency checks and multi-document analysis. These directions aim to reinforce reliability, scalability, and explainability, ultimately enabling broader deployment of neural-symbolic systems for contract analysis in professional legal practice.

REFERENCES

1) F. Ariai and G. Demartini, "Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges," arXiv preprint arXiv:2410.21306, 2024.
2) T. Cui, Y. Wang, C. Fu, et al., "Risk taxonomy, mitigation, and assessment benchmarks of large language model systems," arXiv preprint arXiv:2401.05778, 2024.
3) D. Hendrycks, C. Burns, A. Chen, and S. Ball, "Cuad: An expert-annotated nlp dataset for legal contract review," arXiv preprint arXiv:2103.06268, 2021.
4) N. Guha, J. Nyarko, D. Ho, et al., "Legalbench: A collaboratively built bench-mark for measuring legal reasoning in large language models," Advances in Neural Information Processing Systems, vol. 36, pp. 44 123–44 279, 2023.
5) Y. Liu, M. Ott, N. Goyal, et al., "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
6) K. Shanker, M. Maurya, P. Mayekar, and K. Shah, "Legalbert for small business contract analysis: An advanced nlp tool for identifying high risk clauses," in 2025 International Conference on Innovation in Computing and Engineering (ICE), IEEE, 2025, pp. 1–6.
7) L. Martin, N. Whitehouse, S. Yiu, L. Catterson, and R. Perera, "Better call gpt, comparing large language models against lawyers," arXiv preprint arXiv:2401.16212, 2024.
8) Y. Xi and Y. Zhang, "Measuring time and quality efficiency in human-ai collab-orative legal contract review: A multi-industry comparative analysis," Annals of Applied Sciences, vol. 5, no. 1, 2024.
9) S. H. Wang, M. Zubkov, K. Fan, et al., "Acord: An expert-annotated retrieval dataset for legal contract drafting," arXiv preprint arXiv:2501.06582, 2025.
10) V. Papadouli, "Transparency in artificial intelligence: A legal perspective," Journal of Ethics and Legal Technologies, vol. 4, pp. 25–40, 2022.
11) K. M. Richmond, S. M. Muddamsetty, T. Gammeltoft-Hansen, H. P. Olsen, and T. B. Moeslund, "Explainable ai and law: An evidential survey," Digital Society, vol. 3, no. 1, p. 1, 2024.
12) J. Valvoda and R. Cotterell, "Towards explainability in legal outcome prediction models," arXiv preprint arXiv:2403.16852, 2024.
13) R. Dominguez-Olmedo, V. Nanda, R. Abebe, et al., "Lawma: The power of specialization for legal annotation," arXiv preprint arXiv:2407.16615, 2024.
14) S. Moon, S. Chi, and S.-B. Im, "Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (bert)," Automation in Construction, vol. 142, p. 104 465, 2022.
15) K. Harikrishnan, M. Malathi, and K. Sundharakumar, "Topic-driven contractual language understanding and summarization: An integrated approach for sim-plifying legal documents," in 2023 4th International Conference on Intelligent Technologies (CONIT), IEEE, 2024, pp. 1–6.
16) M. Siino, M. Falco, D. Croce, and P. Rosso, "Exploring llms applications in law: A literature review on current legal nlp approaches," IEEE Access, 2025.
17) G. S. Hartz, "Exploring cuad using roberta span-selection qa models for legal contract review," English, 2022.

18) A. Nandy, M. Nitin Kapadnis, S. Patnaik, Y. Parag Butala, P. Goyal, and N. Ganguly, "FastDoc: Domain-specific fast pre-training technique using document-level metadata and taxonomy," arXiv e-prints, arXiv–2306, 2023.
19) L. Yu, B. Liu, Q. Lin, X. Zhao, and C. Che, "Semantic similarity matching for patent documents using ensemble bert-related model and novel text processing method," arXiv preprint arXiv:2401.06782, 2024.
20) A. Kwak, C. Jeong, G. Forte, D. Bambauer, C. Morrison, and M. Surdeanu, "Information extraction from legal wills: How well does gpt-4 do?" In Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 4336–4353.
21) M. M. Zin, H. T. Nguyen, K. Satoh, S. Sugawara, and F. Nishino, "Information extraction from lengthy legal contracts: Leveraging query-based summarization and gpt-3.5," in Legal Knowledge and Information Systems, IOS Press, 2023, pp. 177–186.
22) Y. He, Y. Tang, and T. Chen, "A study on large language model-based approach for construction contract risk detection," in Proceedings of the 2024 International Conference on Big Data and Digital Management, 2024, pp. 136–141.
23) D. Ressi, R. Romanello, C. Piazza, and S. Rossi, "Ai-enhanced blockchain tech-nology: A review of advancements and opportunities," Journal of Network and Computer Applications, vol. 225, p. 103 858, 2024.
24) K.-Y. Lam, V. C. Cheng, and Z. K. Yeong, "Applying large language models for enhancing contract drafting.," in LegalAIIA@ ICAIL, 2023, pp. 70–80.
25) A. Sadowski, J. Chudziak, et al., "Explainable rule application via structured prompting: A neural-symbolic approach," arXiv preprint arXiv:2506.16335, 2025.
26) Syllo.ai, Agentic legal review platform: Multi-agent ai for clause tagging, risk assessment, and summarization, https://syllo.ai/white- paper- 2025/, White Paper, 2025. [Online]. Available: https://syllo.ai/white-paper-2025/.
27) Y. Saleh, M. Abu Talib, Q. Nasir, and F. Dakalbab, "Evaluating large language models: A systematic review of efficiency, applications, and future directions," Frontiers in Computer Science, vol. 7, p. 1 523 699, 2025