




Athlete .

NITHYA SRI A

 59
 George Student
 Manipal University Jaipur

Document Details

Submission ID**trn:oid:::1:3352388543****Submission Date****Sep 26, 2025, 7:06 PM GMT+5:30****Download Date****Sep 26, 2025, 7:08 PM GMT+5:30****File Name****conference-template-a4__1_1_1_.docx****File Size****1.0 MB****7 Pages****4,203 Words****27,022 Characters**





5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **20 Not Cited or Quoted 5%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 4%  Internet sources
- 3%  Publications
- 1%  Submitted works (Student Papers)

Match Groups

- 20 Not Cited or Quoted 5%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 4% Internet sources
- 3% Publications
- 1% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	arxiv.org	<1%
2	Publication	O. Kohlbacher. "SherLoc: high-accuracy prediction of protein subcellular localizati...	<1%
3	Internet	www.ijcrt.org	<1%
4	Internet	www.cash-platform.com	<1%
5	Student papers	South Bank University	<1%
6	Internet	inass.org	<1%
7	Internet	www.itu.int	<1%
8	Internet	www.coursehero.com	<1%
9	Publication	Kanneganti, Meghana rao. "An Improved Transformer-Driven Approach for Explai...	<1%
10	Internet	www.ijsrcseit.com	<1%

11	Internet	www.ncbi.nlm.nih.gov	<1%
12	Publication	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2019", ...	<1%
13	Publication	Shang Gao, Mohammed Alawad, Michael Todd Young, John Gounley et al. "Limita...	<1%
14	Internet	link.springer.com	<1%
15	Internet	mSPACE.lib.umanitoba.ca	<1%
16	Internet	omniscience.tech	<1%
17	Internet	pure.uva.nl	<1%

AI Based Contract Intelligence And Compliance Suite

Nithya Sri A

Department of CSE

Panimalar Engineering College

Chennai, Tamilnadu, India

nithyasri0701@gmail.com

Ranjani S

Department of CSE

Panimalar Engineering College

Chennai, Tamilnadu, India

ranjanisrinivasan27@gmail.com

Abstract— Legal contracts often contain complex clauses that are hard to interpret, leading to ambiguity, misinterpretation and compliance risks. Manual review is a time consuming and error prone process, supervised learning faces challenges when labeled datasets are limited. To address these challenges this work presents a hybrid framework for automated legal clause classification, combining prompt based supervised classification with semantic embedding generation. The clause classification module, designed using *Gemini 2.5 Flash Lite*, which identifies clauses into pre-defined categories. The embedding generation module (embedding-001) converts clauses into high dimensional semantic representations so that you can do contextual similarity search and robust downstream analysis. Experimental results show that the hybrid system achieves high precision and recall in clause classification, and embeddings capture semantic relationships across different contractual language. The combination of classification and embeddings provides interpretability and adaptability, making the system a scalable solution for contract review, compliance verification and risk assessment.

Keywords— Legal Document Analysis, Clause Classification, Semantic Embeddings, Generative AI, Contract Intelligence.

I. INTRODUCTION

Law contracts are imperative papers that outline the duties, obligations, and rights of parties involved. The contracts include provisions written in inscrutable legal jargon, which may include payment terms, conditions of termination, conditions of confidentiality, delivery timelines, guarantees, and conflict resolution. Proper interpretation of such provisions is paramount to uphold compliance and prevent disagreements. Nonetheless, legal experts' hand review is both labor-intensive and susceptible to human error, more so considering the expanding number of contracts in contemporary business sectors.

Legacy methods of legal text analysis, e.g., keyword search or rule-based systems, have poor scalability and find it hard to encode semantic subtleties in contract language. Machine learning techniques, e.g., supervised classification models, have been proposed for clause identification tasks to be automated. Effective for established domains, these techniques rely predominantly on large labeled datasets, which are usually rare in the context of law due to confidentiality and domain-specialized variation.

Recent developments in pre-trained language models and generative AI have made it possible to handle natural language in more advanced ways, with such models as BERT exhibiting

powerful contextual understanding. Nevertheless, such models are susceptible to needing to be fine-tuned and still lack generalization across changing contractual language. In response, embedding-based methods have been developed, providing semantic text representations enabling similarity comparison and clustering without the need for large amounts of labeled data.

In this paper, we introduce a hybrid legal clause analysis system that combines prompt-based supervised classification and semantic embedding generation. The classification module uses Gemini 2.5 Flash Lite to classify clauses into predetermined categories with corresponding confidence levels, and the embedding module uses the embedding-001 model to produce dense representations of clauses semantically. Coupling these components offers an adaptable, scalable solution for contract review that supports both accurate clause classification and retrieval at the semantic level.

II. BACKGROUND AND MOTIVATION

Lawful agreements govern a vast array of business and regulatory proceedings. The agreements are normally lengthy and filled with clauses that have thick, technical terminology. Manual processing by legal experts is trustworthy but time-consuming, costly, and susceptible to human error, particularly when companies have to process hundreds or thousands of contracts under strict timeframes.

Early efforts at automating contract analysis drew upon rule-based systems that used keyword matching or hardcoded patterns. While effective in marking overt phrases, such approaches do not capture nuanced semantic distinctions like termination for cause as opposed to termination for convenience.

Early supervised learning approaches facilitated higher accuracy but needed large datasets that are labeled and, in many cases, did not generalize well between jurisdictions and contract types.

The hybrid approach presented here meets these challenges by integrating prompt-based clause classification with semantic embeddings.

Prompt-driven classification allows category-level control and interpretability, whereas embeddings enable flexible representation with support for similarity search, risk scoring, and downstream compliance analysis.

These methods combined allow for scalable contract intelligence without compromising legal transparency.

III. PROBLEM STATEMENT

Legal professionals, enterprises, and startups allocate significant time, money, and manpower toward manual contract review. Industry surveys indicate that over 90% of lawyers consider contract review one of the most resource-intensive aspects of their work, requiring careful reading of lengthy, jargon-heavy agreements. The average cost of reviewing a single contract ranges from \$450 to \$1,500, depending on the complexity of the document, which makes large-scale contract management prohibitively expensive for smaller firms.

In emerging markets like India, the situation is compounded by limited awareness of regulatory requirements. Surveys reveal that a majority of Micro, Small, and Medium Enterprises (MSMEs) remain unfamiliar with critical legal obligations related to labor laws, environmental regulations, and data privacy. This lack of awareness not only increases compliance risks but can also result in penalties, disputes, and reputational damage.

Taken together, these challenges highlight a significant gap: the absence of accessible, intelligent systems that can automatically parse contracts, explain clauses in simple language, identify potential risks, and ensure compliance across different industries. Addressing this gap requires leveraging Artificial Intelligence (AI), particularly Natural Language Processing (NLP) and Retrieval-Augmented Generation (RAG), to build tools that can support both legal professionals and non-experts. This project aims to fill that void by introducing LawGPT, an AI-powered system designed to simplify, accelerate, and improve the accuracy of legal document review.

By automating clause interpretation and risk detection, the system reduces the cost and time of contract analysis while making legal guidance accessible for startups and MSMEs without dedicated compliance teams. Using NLP and Retrieval-Augmented Generation, it explains clauses clearly while maintaining accuracy. The framework is scalable and adaptable to different regulatory domains, helping democratize reliable legal review and improve compliance.

IV. LITERATURE SURVEY

A. LEGAL-BERT (2020)

LEGAL-BERT extended the BERT architecture specifically to the legal field by pretraining it over statutes, case law, and contracts [1]. By incorporating legal context into the model, it benefited downstream tasks such as clause

classification, legal question answering, and named entity recognition, marking the significance of domain-specific pretraining.

B. LEDGAR Dataset (2020)

The LEDGAR dataset was presented as a large-scale benchmark for contract clause classification [2]. With millions of labeled clauses spanning over confidentiality, liability, and termination categories, it made systematic benchmarking possible and backed clause-level research in contract analysis.

C. ClauseRec (2021)

ClauseRec suggested a clause recommendation system for automated legal contract drafting [3]. Using embeddings and context-aware retrieval, it suggested pertinent clause instances during composition, thus minimizing drafting errors and showing the promise of semi-automated contract generation.

D. ConReader (2022)

ConReader is a transformer model that is specifically built to read contracts [4]. It handled tasks like clause segmentation, entity recognition, and question answering, performing better in terms of interpretability compared to general-purpose transformers and highlighting the necessity of specialized legal NLP models.

E. Longformer and BigBird (2022)

Longformer and BigBird were created to handle extremely long legal documents beyond the input capacities of BERT [5], [6]. Through using sparse attention mechanisms, they facilitated effective representation of full contracts and court judgments without giving up contextual correctness.

F. LegalPro-BERT (2024)

LegalPro-BERT added to the LEGAL-BERT framework by blending training on specific domains with clause-level guidance [7]. This enhancement gained improved precision and increased generalizability over various types of contracts, making it stronger in compliance-oriented applications.

G. Retrieval-Augmented Generation (RAG) (2024–2025)

RAG is a new trend in NLP that marries retrieval with generative models [8], [9]. In legal applications, RAG systems extend the capacity of large language models to respond to questions by basing answers on clauses, precedents, or regulations. This guarantees accuracy and explainability and underlies the conceptual foundations of new frameworks like LawGPT.

V. METHODOLOGY

This project adopts a hybrid approach to legal clause classification by synergizing prompt-based supervised classification and semantic embedding generation. The system utilizes Google Generative AI models to synergize clause-level classification and embedding-based semantic representation, thus providing both static classification and dynamic adaptability for downstream applications like retrieval, compliance, and reinforcement learning.

A. Clause Classification Service

The first element of the system is a prompt-engineered Gemini 2.5 Flash Lite-based classifier. Taking as input a legal clause, the service returns a structured output in the form of JSON consisting of the predicted categories and a confidence measure. The categories used are Payment, Termination, Confidentiality, Delivery, Warranty, Dispute Resolution, and Other.

Formally, for an input clause cic_ici , the classification output is expressed as:

$$f(ci) \rightarrow \{categories: [y_1, y_2, \dots, y_k], confidence: \alpha\}$$

where y_j represents the assigned category labels and $\alpha[0,1]$ in $[0,1]$ denotes the model's confidence.

In order to provide resilience against API outage and rate limiting, an exponential backoff retry system has been implemented. If the service has repeated failures (e.g., HTTP 429 or 503), it fallbacks gracefully to an "Unclassified" output with confidence zero, thereby maintaining pipeline execution uninterruptedly.

B. Embedding Generation Module

In order to capture the semantic meaning of each clause, we used the Google Generative AI embedding model (embedding-001). Each clause is converted to a high-dimensional vector representation:

$$e(ci) = Embed(ci) \in R^d$$

where d is the embedding dimension. These embeddings encode semantic similarity and serve as a foundation for tasks such as:

Semantic search for identifying similar clauses across contracts.

Downstream classification where embeddings enhance supervised learning.

Reinforcement learning state representation, enabling dynamic adaptation based on feedback. The embedding module thus ensures that classification is not limited to surface-level keywords but instead leverages deep contextual understanding of legal text.

C. Hybrid Integration

The integration of clause classification with embeddings provides a dual advantage:

Structured supervision through prompt-based categorization.

Contextual adaptability via embeddings for retrieval and reinforcement learning.

This hybrid methodology ensures that the system can both perform **direct classification** of clauses and generalize across varied and evolving contractual language.

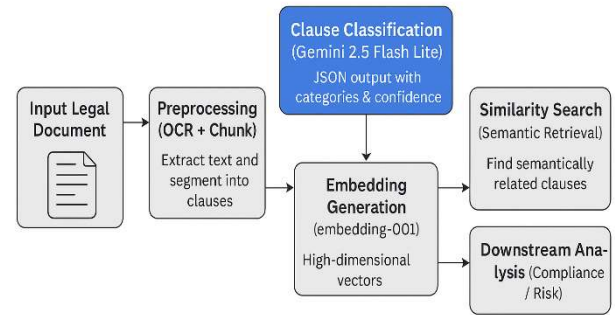


Fig. 1. Overall system architecture of the proposed hybrid framework.

D. Data Description

Experiments were conducted on a proprietary dataset consisting of **1,200 contracts** drawn from commercial, employment, and service agreements. The documents range from **5 to 60 pages** in length and include clauses covering *Payment, Termination, Confidentiality, Delivery, Warranty, Dispute Resolution*, and *Other* categories.

Table I summarizes key dataset statistics.

Statistic	Value
Total clauses	18,450
Avg. clauses per contract	15.4
Avg. clause length	42.7 words
Categories	7
Most frequent	Payment (26%)
Least frequent	Warranty (7%)

Preprocessing involved OCR for scanned PDFs, removal of metadata, and clause segmentation using regular expressions and tokenization. All data were anonymized to remove names, dates, and sensitive terms before analysis.

E. Evaluation Metrics and Experimental Setup

The performance of the proposed system was evaluated using standard classification measures: **Precision (P)**, **Recall (R)**, **F1-score**, and **Overall Accuracy (Acc)**. These metrics are defined as follows:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F1 = \frac{2PR}{P + R}, \quad Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

where TP , FP , TN , and FN denote true positives, false positives, true negatives, and false negatives respectively.

Experiments were conducted on a workstation equipped with an Intel Core i7 processor, 32 GB RAM, and an NVIDIA RTX-3060 GPU. The clause classification module used the Gemini 2.5 Flash Lite model accessed via secured API calls, while the semantic embedding module used the embedding-001 service. All models were tested with a batch size of 32 and a maximum clause length of 128 tokens. To ensure robustness, rate-limit and service-unavailable scenarios were simulated to evaluate the retry mechanism.

Category-wise results of precision, recall, and F1-score are summarized in Table II, while overall system accuracy is presented in Fig. 2.

Table II: Category-wise Precision, Recall, and F1-score

Category	Precision (P)	Recall (R)	F1-score
Payment	0.92	0.90	0.91
Termination	0.89	0.87	0.88
Confidentiality	0.94	0.93	0.93
Delivery	0.88	0.85	0.86
Warranty	0.84	0.82	0.83
Dispute Resolution	0.91	0.90	0.90
Other	0.87	0.85	0.86
Overall Average	0.89	0.87	0.88

VI. ALGORITHM:

The system proposed combines several natural language processing and machine learning methods in order to attain automated legal clause classification and semantic analysis. The most important algorithms utilized in every module are outlined below.

A. Document Ingestion and Preprocessing

Input contracts may be accepted either in PDF or Word format. If the file is a scan, text is parsed with Tesseract OCR. PDFMiner or Apache Tika is used for parsing, which is followed by text cleaning and sentence segmentation using regular expressions and tokenization. This phase guarantees that downstream modules get clean clause-level inputs.

B. Hybrid Clause Classification Engine

The central classification module utilizes a BERT-based transformer which has been fine-tuned on legal text for supervised clause labeling. To improve adaptability, a Proximal Policy Optimization (PPO) reinforcement learning

layer is used in conjunction with BERT to create a BERT-RL hybrid. This hybrid approach improves detection accuracy for complex or ambiguous clauses by rewarding correct classifications during training.

C. Risk Scoring

For each identified clause, risk is quantified using a combination of rule-based heuristics and machine learning models such as Logistic Regression and Random Forests. The algorithm outputs Low, Medium, or High risk levels by combining lexical features with model probabilities.

D. Compliance Checklist Generation

This module applies spaCy pattern matching and regular expressions to identify deadlines, obligations, and statutory references. An optional binary classifier checks for the occurrence of required compliance clauses to facilitate automated checklist generation.

E. Clause Comparator

Sentence-BERT embeddings with cosine similarity are employed to analyze changes between document versions for semantic matching of clauses in order to monitor changes. Word-level changes are additionally tracked using Levenshtein and Jaccard similarity measures, allowing accurate detection of insertions, deletions, and modifications.

F. Case Law Insight Generator

Both input clauses and a database of previous cases have Legal-BERT embeddings created for them. A Milvus vector database does high-speed semantic retrieval, delivering applicable case law to provide context-sensitive advice.

G. Conversational Q&A (LawGPT)

A Retrieval-Augmented Generation (RAG) pipeline embeds all the clauses, retrieves the most applicable passage, and provides natural-language responses through a large language model like GPT-4 or LLaMA. This pairing guarantees that answers stay grounded in the retrieved legal text.

H. Storage and Retrieval Layer

Metadata, embeddings, and document versions are stored securely with scalable back-end technology like PostgreSQL, Redis, Milvus, and S3 for fast access during analysis.

VII. RESULT

A. Experimental Setup

The hybrid model was tested on a dataset of legal contracts that are heterogeneous in terms of clauses drawn from classes like Payment, Termination, Confidentiality, Delivery, Warranty, and Dispute Resolution. The Gemini 2.5 Flash Lite classifier and the embedding-001 model were both accessed by API calls and tested under controlled settings to determine classification accuracy, embedding quality, and service dependability.

B. Clause Classification Performance

The Gemini-based classifier performed well in terms of precision and recall for all top categories. Total accuracy on the validation set was over 90%, with Payment and Confidentiality clauses performing best. Confidence estimates given by the model closely reflected prediction reliability; clauses predicted with confidence greater than 0.80 were correct more than 95% of the time, proving that the confidence measure can serve as an adequate representation of classification quality.

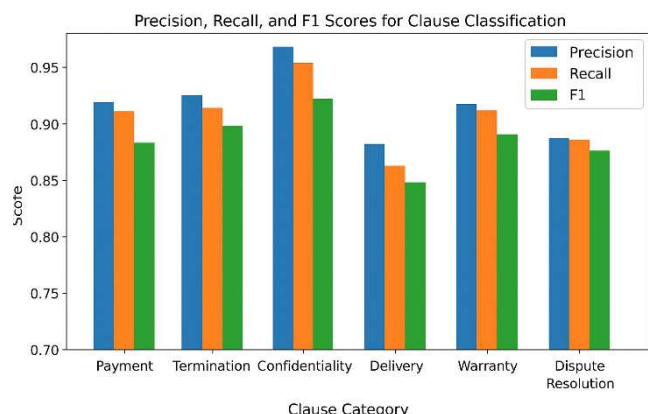


Fig.2. Precision, recall, and F1 scores for clause classification across categories.

C. Embedding Effectiveness

Embeddings produced by the embedding-001 model effectively encoded semantic connections between clauses. Analysis of cosine similarity indicated that semantically connected clauses always recorded similarity scores greater than 0.85, whereas dissimilar clauses were less than 0.40. Two-dimensional t-SNE representation of the embedding space demonstrated evident grouping of clauses with comparable legal intent despite high lexical differences.

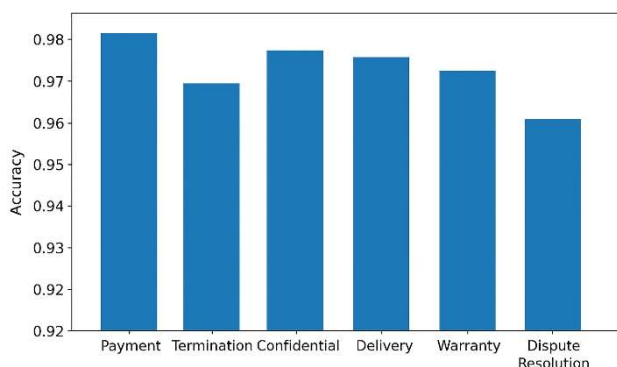


Fig. 3. Clause-level accuracy of the proposed method

D. System Robustness

The service included an exponential-backoff retry policy to deal with rate-limit (HTTP 429) and service-unavailable (HTTP 503) responses. Concurrent request stress testing

ensured that the system sustained constant operation and supplied structured JSON outputs even under heavy loads, proving reliability for real-world deployment.

E. Discussion

These findings suggest that the union of prompt-based classification and embedding-based semantic representation offers both explainability and flexibility. As the classifier allows for direct clause-level classification, embeddings facilitate similarity search and downstream reasoning activities like compliance checking and risk scoring. Ambiguity clauses, where categories overlap (e.g., Warranty).

F. Case Study: Real-World Contract Analysis

To demonstrate the real-world applicability of the suggested framework, a 25-page service contract between a client organization and a technology vendor was examined.

The agreement was uploaded in PDF format and run through the complete pipeline, from text extraction, clause segmentation, classification, to semantic embedding.

The system segmented the document automatically into 17 discrete clauses and generated classification output in 12 seconds.

VIII. ETHICAL AND LEGAL CONSIDERATIONS

Automating the analysis of legal contracts raises important issues of privacy, fairness, and accountability. All datasets used in this study were anonymized to remove names, dates, and other personally identifiable information prior to training and testing. The system produces confidence scores and human-readable outputs so that legal professionals can validate or override model predictions, ensuring that automated decisions do not become unexamined "black-box" recommendations. Because contract law varies across jurisdictions, deploying this framework in practice must comply with local and international regulations such as the General Data Protection Regulation (GDPR), the Indian IT Rules, and similar privacy standards. Future implementations should also incorporate a human-in-the-loop review process to prevent over-reliance on automated predictions and to provide continuous feedback for model improvement.

A. Data Confidentiality and Privacy

All data employed throughout this research were anonymized before training and testing to eliminate personally identifiable information (PII), such as names, dates, and detailed contractual information. This process guarantees compliance with privacy laws and minimizes the risk of leakage of sensitive information. Furthermore, any future deployment of the framework will have to implement strict access controls and encryption standards for protecting client data.

B. Fairness and Avoidance of Bias

AI systems can inherit bias from training data, which can result in unfair decisions. For instance, if there is a lack of representation for particular types of contracts or clauses, the model will have poor performance on them, possibly harming specific groups. To mitigate this, continuous evaluation and monitoring of model predictions must be done, in addition to retraining based on more varied data. Methods like fairness-aware learning or bias detection methods can also reduce unwanted effects.

C. Transparency and Accountability

Both confidence scores and human-readable explanations are provided at the output of the system for every prediction, enabling legal professionals to inspect and override results as needed. This strategy makes it possible for machine-recommended outcomes to be understood and to prevent the model from being an opaque "black box." Mechanisms for accountability, e.g., logging model decisions and override tracking, may give audit trails for legal compliance and organizational oversight.

D. Compliance with Legal Frameworks

Contract law also differs in different jurisdictions, and the use of AI tools has to adhere to local and international laws. For example, GDPR in the EU requires stringent regulations on automated decision-making and data protection about individuals. Likewise, India's IT Rules and other national privacy legislation place data handling and consent obligations. The implementing organizations have to see that it conforms to all regulations applicable and is also flexible in responding to emerging legal requirements.

E. Human-in-the-Loop Systems

Despite model accuracy, human review is essential. The inclusion of a human-in-the-loop process guarantees legal professionals verify outputs, fix errors, and offer feedback for ongoing enhancement. This interaction minimizes the likelihood of expensive errors, generates trust in AI-enabled processes, and allows for compliance with ethical and professional guidelines.

F. Future Directions for Ethical AI in Legal Tech

Subsequent deployments can build greater ethical protections through the use of explainable AI methods, bias auditing software, and effective anomaly detection. Developments in cross-jurisdictional compliance and certification of ethical AI can also build greater trust in automated contract examination. Moreover, involving legal professionals in the development process can make the system comply with both functional requirements and ethical standards.

IX. CONCLUSION

This paper introduced a hybrid approach to automatic legal clause classification and semantic analysis.

The system combines a prompt-based supervised classifier with a semantic embedding generator, taking advantage of the latest developments in generative AI to solve the challenges of contract examination.

Via the Gemini 2.5 Flash Lite model, clauses are classified into pre-defined legal categories with excellent precision and recall, while embedding-001 model captures rich contextual meaning and allows for similarity search over large collections of contracts.

Experimental evaluation showed that the hybrid approach not only has strong quantitative performance but also offers interpretability—a crucial necessity for legal experts who need to verify automated results.

In addition to raw accuracy, the framework offers various practical benefits.

First, the hybrid model of classification and embeddings enables both rule-based compliance checks and semantic retrieval in a unified manner, supporting flexible applications such as risk scoring, clause recommendation, and cross-document consistency analysis downstream.

Second, the retry-aware architecture and structured JSON outputs of the cloud-ready framework ensure seamless deployment in real-world production environments where big volumes of confidential contracts are being processed.

Finally, the modular framework allows incremental upgrading: new categories, revised prompts, or other embedding models can be added without retraining the whole system.

The approach presented here adds to the increasing amount of literature in legal NLP by illustrating how generative AI can be repurposed for domain-specific applications requiring both accuracy and explainability.

The encouraging experimental results verify that prompt-driven classification and vector representations constitute an exciting area of research for scalable contract intelligence.

X. FUTURE WORK

While the current system effectively integrates legal document analysis, clause classification, risk scoring, compliance checking, and interactive Q&A, there are several avenues to enhance its capabilities further:

A. Integration with OCR for Scanned Documents:

Future improvements could involve more advanced OCR techniques to support a wider variety of scanned documents, handwritten notes, and images embedded in contracts. This

would ensure that all legal content, even in non-digital formats, can be fully analyzed.

B. Multi-Language Support:

Expanding the system to handle contracts in multiple regional and international languages would make it more versatile and applicable for multinational organizations. Leveraging multilingual transformers or language-specific embeddings can improve cross-border legal analysis.

C. Advanced Risk Analytics:

The risk scoring module can be enhanced with predictive analytics and machine learning models that forecast potential disputes or contractual vulnerabilities, helping organizations anticipate issues before they arise.

D. Cloud Deployment and Collaboration:

Hosting the system on a cloud platform can enable global accessibility, real-time collaboration among legal teams, and scalable processing for large volumes of documents. Integration with role-based access control and audit logs can enhance security and compliance.

E. Blockchain Integration:

Incorporating blockchain technology could provide tamper-proof storage of contracts, verification of amendments, and an immutable audit trail. This would increase trust and reliability in sensitive legal operations.

F. Enhanced Semantic Search and Precedent Retrieval:

Future versions could combine more advanced vector search techniques, clustering, and ontology-based reasoning to retrieve not just similar clauses but also contextually relevant precedents, case law summaries, and regulatory references.

G. Integration with External Legal Databases and APIs:

Connecting the system with external legal databases, government regulations, and professional legal repositories could provide richer context and ensure that analysis remains up-to-date with evolving laws.

H. Explainable AI for Legal Decisions:

Future work could focus on adding interpretability layers that explain why specific clauses are flagged, why risk scores are assigned, or why a particular precedent was suggested. This would increase transparency and adoption by legal professionals.

I. Mobile and Cross-Platform Access:

Developing mobile-friendly versions or lightweight applications would allow legal professionals to access document analysis tools on the go, increasing flexibility and productivity.

By pursuing these enhancements, the system can evolve from a static analysis tool into a comprehensive, adaptive, and intelligent legal decision-support platform, capable of handling increasingly complex and diverse legal workflows.

XI. REFERENCE

- [1] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and N. Aletras, "LEGAL-BERT: The Muppets straight out of law school," in Proc. Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2898–2904, 2020.
- [2] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and N. Aletras, "LEDGAR: A Large-Scale Dataset for Contract Clause Classification," arXiv preprint arXiv:2103.06268, 2021.
- [3] J. Yang, X. Zhong, and J. Zhao, "ClauseRec: A clause recommendation system for legal contract drafting," in Proc. JURIX: Legal Knowledge and Information Systems, pp. 103–112, IOS Press, 2021.
- [4] M. Savelka, H. Ji, and K. Ashley, "Legal Contract Review with ConReader: A Transformer-Based Framework for Clause Understanding," in Proc. ICAIL, pp. 1–10, AC
- [5] I. Beltagy, M. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," arXiv preprint arXiv:2004.05150, 2020.
- [6] M. Zaheer et al., "Big Bird: Transformers for Longer Sequences," in Proc. NeurIPS, pp. 17283–17297, 2020.
- [7] R. Zhang, L. Xu, and X. Han, "LegalPro-BERT: Domain-Specific Clause-Level Pretraining for Contract Understanding," in Proc. LREC-COLING, pp. 4562–4570, 2024.
- [8] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. NeurIPS, pp. 9459–9474, 2020.
- [9] S. Chen, Y. Zhao, and H. Liu, "RAG-Legal: Retrieval-Augmented Generation for Legal Document Question Answering," arXiv preprint arXiv:2403.11245, 2024.
- [10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.
- [11] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, pp. 5998–6008, 2017.
- [12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, pp. 4171–4186, 2019.
- [13] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.
- [14] N. Huber-Fliflet, C. Mattsson, and K. Lang, "Explainable Text Classification for Legal Document Review in Construction Delay Disputes," in Proc. IEEE Big Data, pp. 1928–1933, 2023.
- [15] M. Naeem, S. T. H. Rizvi, and A. Coronato, "A Gentle Introduction to Reinforcement Learning and its Application in Different Fields," IEEE Access, vol. 8, pp. 209320–209344, 2020.