International Journal of Advance and Applied Research
Peer Reviewed | International Open Access Journal
ISSN: 2347-7075 | Impact Factor – 8.26 | *Website:* https://ijaar.co.in/
Volume-14, Issue-1 | September - October 2026

Original Article

ContractSense: A Literature Survey of AI-Assisted Contract Intelligence for Indian Legal Contracts

Prathamesh Bhushan Gawas, Soham Shridhar Kadam, Paras Anant Jadhav, Prof. Madhura K. Zagade

Department of Information Technology, Finolex Academy of Management and Technology, Ratnagiri, Maharashtra, India

Manuscript ID: IJAAR-140136

ISSN: 2347-7075
Impact Factor – 8.26

Volume – 14
Issue – 1
September- October 2026
Pp. 288 - 302

Submitted: 05 Oct 2026
Revised: 14 Oct 2026
Accepted: 20 Oct 2026
Published: 25 Oct 2026

Corresponding Author: ***Prathamesh Bhushan Gawas***

Quick Response Code:

Website: https://ijaar.co.in/

DOI: 10.5281/zenodo.23972100

DOI Link: https://doi.org/10.5281/zenodo.23972100

Creative Commons

OPEN ACCESS

Abstract:

The rapid expansion of legal documents, digital contracting, and artificial intelligence has increased interest in automated contract analysis systems. Research in contract intelligence has moved beyond rule-based information extraction and traditional natural language processing to transformer-based legal language models, contract-specific datasets, document-level inference, retrieval-augmented generation (RAG), and the incorporation of large language models (sLLMs). Nevertheless, the existing work remains scattered across topics such as clause classification, contract entailment, risk detection, information extraction, legal question answering, and explainability. Another limitation is jurisdiction: many benchmark datasets and systems were developed using contracts or legal requirements from jurisdictions other than India. This review examines relevant research to inform the design of ContractSense, an AI-assisted contract intelligence platform focused on India, specifically for non-disclosure agreements (NDAs), employment agreements, and vendor agreements. The review groups previous work into five complementary areas: legal natural language processing and contract information extraction, clause classification and datasets, document-level reasoning and risk analysis, retrieval-augmented legal question answering, and explainable or rule-supported legal AI. It then compares the literature according to task, data, model family, explainability, retrieval, jurisdictional specificity, and practical limitations. Drawing on the existing literature, the review points out a research gap consisting of the integration of LegalBERT-based clause classification, an India-specific deterministic legal rule engine, risk analysis, RAG combined with ChromaDB and LangChain, and LLM-assisted explanation and negotiation support into a single workflow. ContractSense is therefore proposed as a conceptual synthesis inspired by the literature and not as an empirically proven substitute for legal professionals. The paper ends with a list of open challenges concerning dataset governance, the provenance of legal sources, control of hallucinations, evaluation, privacy, explainability, and the necessity of qualified legal review.

Keywords: legal NLP, contract intelligence, LegalBERT, contract clause classification, RAG, large language models, explainable AI, Indian legal technology, contract analysis, literature survey

Creative Commons (CC BY-NC-SA 4.0)

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), *which permits others to remix, adapt, and build upon the work non-commercially, provided that appropriate credit is given and that any new creations are licensed under identical terms.*

How to cite this article:

Prathamesh Bhushan Gawas, Soham Shridhar Kadam, Paras Anant Jadhav, Prof. Madhura K. Zagade (2026). ContractSense: A Literature Survey of AI-Assisted Contract Intelligence for Indian Legal Contracts. International Journal of Advance and Applied Research, 14(1), 288 – 302. https://doi.org/10.5281/zenodo.23972100

International Journal of Advance and Applied Research
Peer Reviewed | International Open Access Journal
ISSN: 2347-7075 | Impact Factor – 8.26 | *Website:* https://ijaar.co.in/
Volume-14, Issue-1 | September - October 2026

Introduction:

Because business relationships are becoming more digital, there is now a large number of contracts that have to be examined, compared, understood and managed. Contracts such as non-disclosure agreements (NDAs), employment agreements, vendor agreements, procurement contracts and data-processing arrangements include clauses which can influence payment obligations, liability, intellectual property, confidentiality, termination, dispute resolution and data protection. (Arora, 2026) For organisations that do not have easy access to legal professionals, these clauses are difficult to understand since contractual language is usually long, technical and very contextual. (Dubey, 2026) Hence, natural language processing (NLP) has become a key area of research in the analysis of legal documents. Early legal NLP systems concentrated on information extraction, document classification, named-entity recognition, rule-based processing and the creation of structured representations of legal text. For example, LexNLP showed how useful NLP and information extraction techniques can be when applied to legal and regulatory text [1]. Since transformer architectures and domain-adapted language models have appeared, legal NLP has gradually shifted towards contextual representation learning. LegalBERT proved that pretraining on corpora from the legal domain can lead to better performance on downstream legal NLP tasks [2]. Contract analysis has now become a separate research field. The Contract Understanding Atticus Dataset (CUAD) provided expert annotations for a wide range of contract review questions and has therefore become a significant benchmark for contract clause extraction and review [3]. ContractNLI further looked into document-level natural language inference for contracts, using entailment, contradiction and not-mentioned labels to show the relationships between contract text and legal hypotheses [4]. These resources show that contract understanding goes beyond simple text classification: the models have to take into account the semantics of the clauses, the context of the document and the legally significant relationships. Recent research has also looked at the automatic detection of clauses that could be abusive or problematic, at consistency analysis and at contract question answering. (ContractNerd: An AI Tool to Find Unenforceable, Ambiguous, and Prejudicial Clauses in Contracts, 2025) These systems show the potential of machine learning and information retrieval to help with legal review, while at the same time highlighting the limitations connected with dataset coverage, jurisdiction, explainability and reliability [7, 8]. At the same time, retrieval-augmented generation (RAG) has appeared as a method of combining language generation with retrieved evidence [5]. In legal applications, retrieval is especially important since the responses may have to be based on a particular contract, on authoritative legal sources or on both. The Indian situation adds another dimension to the research. Indian contracts may be affected by the Indian Contract Act, 1872; the Arbitration and Conciliation Act, 1996; the Micro, Small and Medium Enterprises.

Depending on the agreement and the specific circumstances, the Development (MSMED) Act, 2006, the Digital Personal Data Protection Act, 2023, and the relevant state-specific stamp duty regulations may apply. A system which is mainly based on foreign contracts or foreign legal principles cannot

simply be assumed to offer the same level of coverage for contracts made in India. For this reason, a review of the relevant literature is needed that looks not only at the machine-learning methods used for contract intelligence but also at the jurisdictional and governance problems that occur when applying these methods in India. The present paper therefore looks at the literature that is pertinent to the design of ContractSense, an AI-assisted contract intelligence platform focused on India. The intended system will analyse NDAs, employment agreements, and vendor agreements by means of document extraction and OCR, clause segmentation, LegalBERT-based clause classification, the application of deterministic Indian legal rules, risk analysis, embeddings and retrieval using ChromaDB, RAG orchestration via LangChain, and LLM-assisted explanation and negotiation support. The platform is to be used as a decision-support tool and not as a replacement for professional legal advice.

Motivation and Research Gap:

The available literature shows that contract intelligence is not merely a single task; it consists of a series of steps including document processing, clause segmentation, classification, information extraction, legal reasoning, retrieval, risk identification, explanation, and user interaction. Most of the existing studies focus on one or only a small number of these components. (Singh et al., 2024) Although benchmark datasets like CUAD offer a solid basis for research into contract review, they do not in themselves form an India-specific legal knowledge base. (Hendrycks et al., 2021) In the same way, general-purpose large language models are able to produce fluent explanations but need grounding mechanisms and carefully designed evaluation methods if they are to be used with legal documents. (Mani & Chowdhury, 2026) The main research gap that this review has identified is therefore the lack of integration of the following elements within a single workflow that is focused on India: (i) contract-specific transformer classification, (ii) deterministic jurisdiction-specific legal checks, (iii) reproducible risk analysis, (iv) contract-grounded retrieval, and (v) LLM-based explanation and negotiation assistance. (Barbosa et al., 2025) ContractSense was developed in response to this integration gap.

The central research gap identified by this review is therefore the limited integration of the following components within a single India-focused workflow: (i) contract-specific transformer classification, (ii) deterministic jurisdiction-specific legal checks, (iii) reproducible risk analysis, (iv) contract-grounded retrieval, and (v) LLM-based explanation and negotiation assistance. ContractSense is motivated by this integration gap.

Research Questions:

This review is organized around the following research questions:

- RQ1: What NLP, machine-learning, and information-extraction techniques have been used for automated contract analysis?
- RQ2: What datasets and benchmark tasks are available to support contract clause classification, extraction, and document-level reasoning?
- RQ3: What role are transformer models, retrieval techniques, RAG, and LLMs

playing in supporting understanding of legal contracts and answering questions about them?

- RQ4: What limitations remain regarding explainability, jurisdictional adaptation, legal-source grounding, reliability, and dataset governance?
- RQ5: In what way can the literature gaps that have been identified serve as a basis for designing a system focused on India such as ContractSense?

Contributions of this Review:

This review is organised around the following research questions:

1. It brings together representative research in the area of legal NLP and contract intelligence into a single taxonomy that includes extraction, classification, reasoning, risk analysis, retrieval, generation, and explainability.
2. It looks at the key contract datasets and research tasks, such as CUAD and ContractNLI, in relation to their role in contract intelligence.
3. The text looks at the shift from traditional NLP and rule-based approaches to LegalBERT, transformer models, RAG, and legal systems assisted by LLMs.
4. It points out the limitations concerning jurisdictional coverage, explainability, evidence grounding, dataset governance, and evaluation that are especially pertinent to the analysis of contracts in India.
5. The conceptual research framework for ContractSense is obtained by combining the complementary capabilities that have been reported in the literature.

Organization of the Paper:

The rest of this paper is put together in the following way: Section 2 outlines the review method and the criteria for inclusion. Section 3 gives the technical background of contract intelligence. Section 4 sets out the literature taxonomy and looks at some representative studies. Section 5 gives a comparative analysis Section 5 provides a comparative analysis. Section 6 examines the research gaps and their implications for the Indian context. Section 7 sets forth the ContractSense conceptual framework, which has been derived from the reviewed literature. Section 9 addresses the challenges and suggests future research directions. Section 11 concludes the review.

Review Methodology:

The paper adopts a systematic literature survey method. Its aim is not to carry out a comprehensive and systematic review of all publications on legal AI, but rather to combine relevant and directly applicable research that has informed the design of ContractSense. In carrying out the review, preference is given to peer-reviewed research and to widely used benchmark resources in the areas of legal natural language processing, contract analysis, information retrieval, RAG, and explainable AI.

1. Review Scope:

The review covers the following research themes:

- legal NLP and legal information extraction;
- contract clause classification and extraction;
- contract-specific datasets and benchmarks;
- document-level contract reasoning and natural language inference;

- abusive, risky, or inconsistent clause detection;
- transformer-based legal language models;
- retrieval-augmented generation and legal question answering;
- explainability and evidence grounding; and
- jurisdiction-specific legal rule integration.

2. Representative Selection Criteria:

A source was regarded as relevant if it covered at least one of the following areas: contract document analysis, legal natural language processing, clause classification, legal-domain language models, contract inference, legal question answering, retrieval-augmented generation, or explainable legal AI. Preference was given to research that included a clearly described dataset, model, task, evaluation methodology, or system architecture. With respect to the Indian legal aspect, official legislation and reliable legal sources are kept separate from examples of contracts obtained through machine learning. Legal statutes and rules are seen as legal knowledge sources for a rule engine or for the RAG layer; they are not automatically used as training examples for contract clauses.

3. Literature Synthesis Dimensions:

Each reviewed work is considered along the following dimensions:

- Task: classification, extraction, inference, risk detection, question answering, or generation;
- Data: generic legal text, contracts, benchmark datasets, or domain-specific corpora;
- Model: rule-based NLP, conventional machine learning, transformer, retrieval, or LLM;
- Evidence grounding: whether answers are linked to source documents or retrieved evidence;
- Explainability: whether predictions or outputs have interpretable supporting information;
- Jurisdiction: whether the work is jurisdiction-neutral, jurisdiction-specific, or primarily outside India; and
- Practical limitation: dataset, evaluation, privacy, reliability, or deployment limitations.

Technical Background of Contract Intelligence:

1. Legal NLP and Contract Understanding:

Unlike ordinary text, legal documents rely on defined terms, cross-references, exceptions, obligations, conditions, and the relationships between clauses to determine their meaning. Since legal natural language processing involves not only linguistic processing but also interpretation specific to the domain, LexNLP demonstrated the benefits of using NLP and rule-based methods to extract structured information from legal text [1]. The use of these approaches is still valuable in current systems since deterministic rules can supplement statistical models when the requirements are explicitly stated.

2. Transformer-Based Legal Language Models:

By using contextual representations, Transformer models are able to capture the relationships between words and their surrounding text. LegalBERT has modified the BERT architecture to suit legal-domain corpora and has shown the benefits of domain-specific pretraining for legal NLP tasks [2]. The model intended for use with ContractSense is primarily

International Journal of Advance and Applied Research
Peer Reviewed | International Open Access Journal
ISSN: 2347-7075 | Impact Factor – 8.26 | *Website:* https://ijaar.co.in/
Volume-14, Issue-1 | September - October 2026

designed for the purpose of clause category classification, the candidate categories being Payment, Liability, Termination, Confidentiality, Intellectual Property, Data Protection, Dispute Resolution, Governing Law, Restrictive Covenant, Indemnification, Warranty, and Other.

3. Contract Datasets:

CUAD includes expert annotations that are geared towards answering questions about contracts and thus serves as a major benchmark for automated contract review [3]. ContractNLI deals with document-level inference by using the labels Entailment, Contradiction, and NotMentioned [4]. Since these label sets correspond to different tasks, they should not be quietly combined. Hence, if both a clause-category model trained on CUAD-type annotations and a ContractNLI-type inference model are used, the two should be regarded as separate components.

4. Retrieval-Augmented Generation:

By combining retrieval with language generation, RAG retrieves relevant information before producing a response [5]. In the case of a contract-intelligence application, retrieval can be carried out on the contract clauses that have been uploaded and on the curated legal sources. ChromaDB can function as the vector store for the embeddings, while LangChain can manage the processes of loading the documents, chunking them, creating the embeddings, carrying out the retrieval, constructing the prompt, and making the call to the LLM. It is important to note that a vector database is a retrieval component and not the main application database.

5. Explainability:

In the context of legal AI, explainability is of particular importance since users need to understand the reason why a clause has been flagged. SHAP offers a general framework for assigning the model's output to its input features [6]. In ContractSense, explainability integrates model-level explanations when it is technically appropriate, together with the clause text, deterministic rule identifiers, evidence sources, and explanations in plain language. An explanation must not be treated as legal proof merely because it has been generated by an explainability method.

Literature Review and Taxonomy:

1. Legal Information Extraction and Rule-Based NLP:

Early and hybrid legal NLP systems established the importance of extracting structured facts from unstructured legal text. LexNLP is representative of this direction, providing reusable tools for legal information extraction and text processing [1]. The key advantage of this type of system is that it provides determinism and transparency when it comes to explicitly defined patterns. Yet, rule-based systems can become hard to maintain if there is a large variation in legal language across different documents.

The literature indicates that rules and machine learning should not necessarily be seen as competing options; explicit legal requirements, for instance, such as verifying that a document includes a specific obligation or that a clause triggers a previously defined condition, can be expressed as deterministic rules. Statistical models, on the other hand, are more appropriate for tasks involving contextual classification and semantic similarity. This idea of combining both methods has a direct influence on the ContractSense architecture.

2. Contract Clause Classification and CUAD:

The CUAD benchmark includes a large number of annotations provided by experts for contract review tasks [3]. What makes it important is that it offers a shared resource for evaluation and presents contract review as involving structured question answering and information extraction, rather than just generic document classification.

For ContractSense, CUAD serves as a basic dataset for learning about clauses; yet it does not remove the need for a dataset from India. Contracts drawn from India might make use of different drafting practices, refer to different laws, follow different commercial customs, and make different assumptions regarding the jurisdiction. Therefore, the project follows a hybrid approach to its dataset: using CUAD as a benchmark-based foundation, and at the same time gathering Indian contract clauses with information about their source, permission or licence, integrity checks, and subsequent legal review.

3. Document-Level Contract Reasoning:

ContractNLI developed a benchmark for document-level natural language inference with regard to contracts [4]. The three-way labels it uses—Entailment, Contradiction, and NotMentioned—address a problem that is different from clause-category classification. It is important for the design of the system since merely identifying a clause as a ”Termination” clause does not in itself provide an answer to the question of whether the document entails a particular legal hypothesis.

The fact that document-level reasoning shows the importance of context means that a clause can refer to definitions, schedules, exceptions, or other sections of the agreement. A reliable contract intelligence system therefore needs to keep the document structure, the section information, and the references rather than treating each clause as a separate sentence.

4. Abusive Clause and Risk Detection:

Studies into abusive or possibly problematic clauses have looked at the automatic detection of clauses which might disadvantage consumers or give rise to other concerns in reviews. Dadas et al. examined support for the detection of abusive clauses in B2C contracts [7]. This kind of work is relevant to ContractSense since it demonstrates how automated systems can go beyond simple clause classification and move on to risk identification aimed at reviewing the contracts.

However, a ’risk’ designation is dependent on context. Just because a machine-learning model gives a clause a high risk score does not mean that the clause is legally invalid. The risk analysis should therefore be presented as a decision-support stage, making the specific clause, the reason, the rule or the evidence, as well as the applicable uncertainty, visible to the user.

5. Automated Contract Consistency Analysis:

Rather than assessing each clause on its own, automated consistency analysis looks at the relationships between different provisions. Khoja et al. carried out a study on automated consistency analysis for legal contracts [8]. This approach is significant since contractual obligations have the potential to conflict with one another or to become ambiguous when several clauses are taken into account.

For ContractSense, this body of literature provides a reason to extend the system beyond clause-level classification. The original system might concentrate on clause classification and certain rule checks, while

International Journal of Advance and Applied Research
Peer Reviewed | International Open Access Journal
ISSN: 2347-7075 | Impact Factor – 8.26 | *Website:* https://ijaar.co.in/
Volume-14, Issue-1 | September - October 2026

further work could then look at cross-clause consistency, contradiction detection, and the construction of dependency graphs.

6. LegalBERT and Domain Adaptation:

LegalBERT has shown that it is possible to obtain useful representations for legal natural language processing by adapting transformer pretraining to legal corpora [2]. For contract intelligence, this means that a domain-specific model is better than a general language model as a starting point for clause classification.

Yet LegalBERT by itself does not include Indian legal reasoning. Since a language model only picks up the statistical patterns from its training data, it does not automatically become an authoritative source of current Indian law. That is why ContractSense separates the LegalBERT-based classification from deterministic legal rules and the legal sources that are retrieved.

7. Legal question answering using RAG and LLMs:

The RAG approach developed by Lewis et al. involves combining a parametric generator with a non-parametric retrieval element [5]. In the context of legal applications, retrieval is useful since users often need answers that are based on a specific contract rather than on general legal knowledge. More recent research into legal agents and the answering of questions about contracts has looked at multi-stage or multi-agent retrieval and generation workflows. These systems show that document retrieval, planning, and the use of structured evidence all contribute to improved legal question answering. Nevertheless, the generated answers still have to be carefully assessed since fluent language does not ensure factual or legal accuracy.

For ContractSense, RAG is therefore seen as a means of grounding rather than as a standalone source of legal reasoning. The retrieval stage should return the relevant clauses from the contract and, where appropriate, also authoritative passages from legal sources before the LLM produces an explanation.

8. Explainable Legal AI:

There are several levels to explainability in legal AI. A classifier can show the text of the clause linked to a label; a rule engine can provide the rule identifier and the relevant legal source; a retrieval system can display the evidence that has been retrieved; and an LLM can generate a natural-language explanation. These different methods should not be mixed up.

ContractSense uses a layered approach to explanation comprising: *what was detected*, *why it was flagged*, *which clause or evidence supports the finding*, and *which legal source or rule was used*. This design is in line with the traceability requirement that has been noted in both legal NLP and explainable AI research

Comparative Analysis of Representative Literature:

Table 1 summarizes representative works relevant to ContractSense.

Work	Primary task	Data / domain	Main approach	Key contribution	Relevance / limitation for ContractSense
LexNLP [1]	Legal informa-tion extraction	Legal/regulatory text	NLP + rules	Structured legal text extraction	Supports hybrid rule/NLP design; not an India-specific contract platform.
LegalBERT [2]	Legal NLP representation	Legal corpora	BERT / domain adaptation	Legal-domain contextual representations	Suitable foundation for clause classification; not itself an Indian legal rule engine.

International Journal of Advance and Applied Research
Peer Reviewed | International Open Access Journal
ISSN: 2347-7075 | Impact Factor – 8.26 | *Website:* https://ijaar.co.in/
Volume-14, Issue-1 | September - October 2026

CUAD [3]	Contract review / extraction	Contracts	Expert annotations + NLP tasks	Large contract review benchmark	Useful training/ evaluation resource; jurisdiction and label coverage require adaptation.
ContractNLI [4]	Document-level inference	Contracts	NLI	Entailment, Contradiction, Not Mentioned	Motivates separate inference component; labels should not be merged with clause categories.
RAG [5]	Knowledge-grounded generation grounded	General knowl-edge tasks	Retrieval + generation	Combines retrieved evidence with generation	Provides grounding architecture; legal reliability depends on retrieval quality and source governance.
Abusive-clause detec-tion [7]	Clause risk de-tection	B2C contracts	ML/NLP	Automated support for detecting abusive clauses	Motivates risk layer; legal risk is jurisdiction and context dependent.
Consistency analysis [8]	Cross-clause consistency	Legal contracts	Automated reasoning	Detects inconsistencies among provisions	Motivates future cross-clause analysis; broader integration is still needed.

1. Synthesis of the Literature:

The comparison reveals that no single reviewed component addresses the complete ContractSense objective. LegalBERT addresses legal-language representation, CUAD provides contract-review data, ContractNLI provides document-level inference, RAG provides a retrieval-grounded generation paradigm, and rule-based approaches provide deterministic checks. These capabilities are complementary rather than interchangeable.

A key implication is that contract intelligence should be viewed as a layered architecture. The document layer converts heterogeneous files into structured text. The NLP layer classifies and extracts information. The rule layer evaluates explicit legal conditions. The retrieval layer provides contextual evidence. The LLM layer communicates results to the user. Finally, the application layer manages authentication, history, reports, and user interaction.

Research Gaps Identified from the Literature:

1. Jurisdictional Gap:

A substantial portion of contract NLP research is built around datasets and legal environments that are not specifically designed for India. A model trained on foreign contract conventions may still recognize general legal language, but this does not establish that its risk interpretation or legal compliance checks are appropriate for Indian contracts. ContractSense addresses this gap by introducing an India-specific rule layer and a governed collection of Indian contract clauses.

2. Integration Gap:

Research often evaluates one task at a time: classification, extraction, inference, or question answering. Practical contract review, however, requires these tasks to operate together. ContractSense proposes an integrated pipeline from document ingestion through clause analysis, rule checking, risk assessment, retrieval, explanation, and reporting.

3. Grounding and Hallucination Gap:

LLMs can generate fluent legal explanations, but generation without evidence can produce unsupported statements. RAG reduces this risk by retrieving relevant context, but retrieval itself can fail or return irrelevant material. Therefore, a legal contract system should preserve source metadata and provide evidence-aware responses rather than treating LLM output as authoritative.

International Journal of Advance and Applied Research
Peer Reviewed | International Open Access Journal
ISSN: 2347-7075 | Impact Factor – 8.26 | *Website:* https://ijaar.co.in/
Volume-14, Issue-1 | September - October 2026

4. Explainability Gap:

A classification label alone may not be sufficient for a contract user. The system should explain the detected issue using the original clause, a reason, relevant evidence, and, where applicable, a deterministic rule. Explainability should therefore be designed as part of the system architecture rather than added only after model development.

5. Dataset Governance Gap:

Public availability of a contract does not automatically imply permission to use the document for machine-learning training. A governed dataset must preserve provenance, source URL, license or permission status, access date, document hash, format, jurisdiction, and annotation status. Personal or confidential information must be handled according to applicable permissions and data-governance requirements. ContractSense therefore separates pending clauses from lawyer-approved clauses.

ContractSense: A Conceptual Framework Derived from the Literature:

ContractSense is proposed as a conceptual synthesis of the capabilities identified in the literature. It is not presented in this review as a completed or empirically validated system.

1. Target Users and Contract Types:

The intended users are Indian startups, freelancers, MSMEs, small businesses, and independent professionals who need assistance understanding selected contractual provisions. The initial supported contract types are:

- Non-Disclosure Agreement (NDA);
- Employment Agreement; and
- Vendor Agreement.

2. Proposed Workflow:

The proposed ContractSense workflow integrates document processing, clause-level classification, Indian legal rule checking, retrieval, and LLM-based assistance into a unified pipeline. The system first accepts a contract document and extracts its textual content using PDF/DOCX parsers and OCR where required. The extracted document is segmented into individual clauses and passed to the LegalBERT-based classification layer. The classified clauses are then processed by the risk engine and the Indian legal rule engine. Relevant contractual content is indexed and retrieved through embeddings and ChromaDB, after which the RAG pipeline implemented using LangChain provides contextual information to the selected LLM. The final outputs include plain-language explanations, contract chat, negotiation support, reports, and dashboard visualizations.

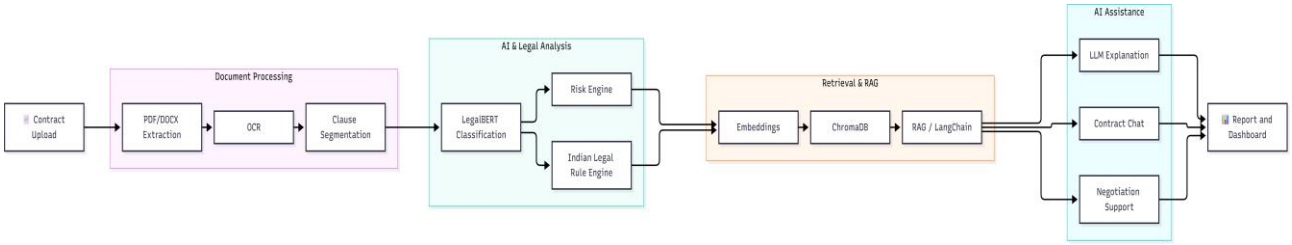


Figure 1 Proposed end-to-end workflow of the ContractSense platform.

International Journal of Advance and Applied Research
Peer Reviewed | International Open Access Journal
ISSN: 2347-7075 | Impact Factor – 8.26 | *Website:* https://ijaar.co.in/
Volume-14, Issue-1 | September - October 2026

As illustrated in Fig. 1, the workflow follows the sequence:

Contract Upload → PDF/DOCX Extraction + OCR → Clause Segmentation → LegalBERT Classification → Risk Engine + Indian Legal Rule Engine → Embeddings + ChromaDB → RAG/LangChain → LLM Explanation / Chat / Negotiation Support → Report and Dashboard.

MongoDB is conceptually used as the application data store for users, contracts, clauses, analyses, risk findings, legal findings, conversations, and reports. ChromaDB is used for semantic retrieval rather than as the primary transactional database.

3. Clause Classification Layer:

The LegalBERT component is intended to classify clauses into a controlled taxonomy. The initial categories are Payment, Liability, Termination, Confidentiality, Intellectual Property, Data Protection, Dispute Resolution, Governing Law, Restrictive Covenant, Indemnification, Warranty, and Other.

The purpose of this layer is classification, not final legal judgment. For example, identifying a clause as a "Termination" clause does not imply that the clause is legally defective. The classification provides structured input to subsequent risk and rule-analysis components.

4. Indian Legal Rule Engine:

The rule engine is intended to implement deterministic, versioned checks for selected Indian legal areas, including:

- Indian Contract Act, 1872;
- Micro, Small and Medium Enterprises Development Act, 2006;
- Digital Personal Data Protection Act, 2023;
- Arbitration and Conciliation Act, 1996; and
- applicable stamp-duty provisions.

These areas should not be represented as a universal checklist for every contract. Applicability depends on contract type, facts, parties, jurisdiction, transaction characteristics, and the current legal position. Stamp duty in particular can depend on the relevant state and document type. Therefore, the system should provide verification-oriented findings rather than unconditional legal conclusions.

A rule record can be represented conceptually as:

$$R_i = (id_i, law_i, condition_i, severity_i, explanation_i, source_i, version_i). \quad (1)$$

This structure allows each rule finding to be traced to a defined legal source and rule version.

5. Risk Analysis Layer:

ContractSense can represent findings using qualitative levels such as High, Medium, Low, and Informational. A reproducible scoring mechanism may combine model confidence, rule-triggered conditions, clause characteristics, and contextual evidence. However, a numerical risk score should be treated as a system-generated prioritization signal rather than a legal determination.

6. RAG and AI Assistance:

The RAG layer retrieves relevant contract clauses and approved legal-source material before the LLM generates a response. A conceptual query flow is:

1. receive the user's question;
2. identify the relevant contract and user context;
3. retrieve semantically relevant clauses from ChromaDB;
4. retrieve applicable legal-source material where required;
5. construct a grounded prompt;
6. generate an explanation or response through the selected LLM provider; and

International Journal of Advance and Applied Research
Peer Reviewed | International Open Access Journal
ISSN: 2347-7075 | Impact Factor – 8.26 | *Website:* https://ijaar.co.in/
Volume-14, Issue-1 | September - October 2026

7. preserve supporting evidence and response metadata where appropriate.

The system may support Gemini, Groq, or an Ollama-compatible provider through an abstraction layer. The provider is therefore an implementation choice rather than the defining research contribution.

7. Explainability and Negotiation Assistance:

For each finding, the system can present the original clause, category, risk level, reason, relevant legal rule or evidence, and a plain-language explanation. Negotiation assistance can then generate a proposed alternative wording. Such a rewrite should be clearly presented as a suggested draft for user review and not as guaranteed legally safer language.

8. Document Processing:

The document layer is intended to support PDF and DOCX files. PyMuPDF can be used for direct PDF text extraction, python-docx for DOCX extraction, and Tesseract OCR as a fallback for scanned or image-based documents. Page and section information should be retained where possible because document location can improve traceability and user review.

Dataset Governance for ContractSense:

A major implication of the literature survey is that model performance cannot be separated from dataset quality and provenance. ContractSense therefore follows a governed collection workflow for Indian contract clauses.

1. Source Eligibility:

Candidate sources should be publicly accessible Indian contract documents from authoritative or clearly licensed sources, such as government ministries, India Code-related public resources, official procurement portals, public-sector organizations, official contract templates, and open-license datasets. Documents behind logins or paywalls, private contracts, sources with unclear licensing, and documents whose provenance or text integrity cannot be verified should be rejected.

Public accessibility should not be treated as equivalent to machine-learning permission. Each document should record its source license or permission status and the basis on which the document is considered eligible.

2. Provenance Record:

Each document should maintain metadata including document identifier, title, source URL, direct download URL where available, issuing organization, contract type, jurisdiction, governing law if stated, license or permission status, license-verification notes, publication date, access date, SHA-256 hash, source format, language, document/template status, and machine-learning eligibility.

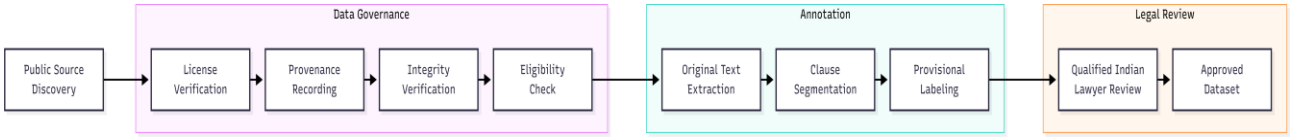


Figure 2. Governed clause annotation workflow used for the ContractSense dataset.

3. Clause Annotation:

Original clause text is preserved without rewriting. Each provisional clause is assigned one category from the controlled taxonomy and remains marked as *pending* until qualified legal review is completed. The governed annotation workflow is illustrated in Fig. 2.

The project should maintain separate storage for pending and approved records. Only records that have passed the defined governance

and legal-review process should be promoted to the approved training set.

4. Separation of Contract Data and Legal Sources:

Contract examples and legal-source references serve different purposes. Contract clauses provide examples for machine-learning annotation, whereas statutes, rules, notifications, and authoritative legal references support the rule engine and RAG. Mixing the two collections can introduce label leakage and conceptual confusion. Contract-Sense therefore maintains separate logical datasets for Indian contract examples and legal sources.

Challenges and Future Research Directions:

1. Legal and Regulatory Change:

Indian legal requirements can change through amendments, notifications, judicial interpretation, and changes in applicable rules. A static model cannot be assumed to remain current. Future systems should maintain versioned legal sources, rule histories, effective dates, and review procedures.

2. Evaluation of Legal Correctness:

Standard NLP metrics such as accuracy, precision, recall, and F1 are useful for clause classification but do not measure the full quality of legal assistance. Future evaluation should additionally consider retrieval relevance, faithfulness, evidence coverage, hallucination rate, rule-engine correctness, explanation quality, and human expert agreement.

3. Cross-Clause Reasoning:

Many contract risks arise from interactions between clauses. Future ContractSense versions could incorporate document-level inference, contradiction detection, obligation graphs, and cross-reference resolution inspired by contract reasoning research.

4. Multilingual and Indian-Language Contracts:

The initial scope is primarily English-language contract analysis. Indian commercial environments may contain multilingual documents or clauses with regional-language text. OCR and multilingual legal NLP therefore remain important future directions.

5. Privacy and Security:

Contracts can contain confidential business information, personal data, financial information, and intellectual property. A production system must use secure transport, password hashing, authentication and authorization, file validation, protected secrets, user-level data isolation, appropriate retention controls, and careful handling of third-party LLM APIs. Uploaded contract text should also be treated as untrusted input to reduce prompt-injection and data-exfiltration risks.

6. Human-in-the-Loop Legal Review:

The literature and the system requirements both indicate the importance of human oversight. A qualified Indian contract lawyer should be involved in validating the curated Indian dataset, reviewing legal-rule mappings, and evaluating high-impact outputs. Human review is especially important when the system moves from identifying potentially relevant provisions to suggesting contractual changes.

Discussion:

The literature supports a hybrid rather than purely generative approach to contract intelligence. Transformer models are well suited to contextual language understanding, while deterministic rules are useful for explicitly

defined legal checks. Retrieval provides evidence grounding, and LLMs provide a flexible natural-language interface. The practical value of these components depends on how they are combined and evaluated.

For ContractSense, the literature therefore motivates the following separation of responsibilities: LegalBERT performs clause-category classification; the rule engine performs deterministic selected-law checks; the risk engine prioritizes findings; ChromaDB stores embeddings for semantic retrieval; LangChain orchestrates retrieval and generation; the LLM explains findings and supports conversational interaction; and MongoDB stores application-level information. This separation also improves maintainability because individual components can be evaluated independently.

The review also indicates that an India-focused contract system should not claim comprehensive legal coverage. The initial project scope is deliberately limited to selected contract types and selected Indian legal areas. This controlled scope is preferable to presenting a general-purpose legal system whose coverage and reliability cannot be established.

Conclusion:

This literature survey examined research relevant to AI-assisted contract intelligence and used the findings to motivate the ContractSense project. The reviewed literature demonstrates a progression from rule-based legal information extraction toward transformer-based legal language models, contract-specific datasets, document-level reasoning, retrieval-augmented generation, and LLM-assisted legal interaction. At the same time, the literature exposes persistent challenges involving jurisdictional adaptation, evidence grounding, explainability, dataset governance, evaluation, and legal reliability.

The review identifies a practical integration opportunity for an India-focused system combining LegalBERT-based clause classification, deterministic Indian legal rules, risk analysis, embeddings and ChromaDB, RAG/LangChain, and LLM-based explanation and negotiation assistance. ContractSense is consequently framed as an AI-assisted decision-support platform for NDA, employment, and vendor agreements rather than as a replacement for lawyers or a source of guaranteed legal conclusions.

Future work should empirically evaluate each layer of the proposed architecture, construct and legally review a governed Indian contract dataset, measure classification and retrieval performance, evaluate rule-engine correctness, test hallucination and evidence-grounding behavior, and conduct expert assessment of the resulting explanations. Such evaluation is necessary before making claims about real-world legal reliability or deployment suitability.

Acknowledgements:

The author acknowledges the academic literature, benchmark datasets, open research resources, and publicly available legal materials that motivate this review. Any implementation of ContractSense should independently verify applicable licenses, permissions, and current legal requirements before using external data or deploying legal-information functionality.

References:

1. Michael J. Bommarito II and Daniel Martin Katz. LexNLP: Natural language processing and information extraction for

legal and regulatory texts. *Law, Technology and Humans*, 2018.

2. Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.
3. Dan Hendrycks, Collin Burns, Anu Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. CUAD: An expert-annotated NLP dataset for legal contract review. In *Advances in Neural Information Processing Systems*, 2021.
4. Kento Koreeda and Christopher D. Manning. ContractNLI: A dataset for document-level natural language inference of contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.
5. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rockta¨schel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.
6. Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.
7. S. Dadas et al. A support system for the detection of abusive clauses in B2C contracts. *Artificial Intelligence and Law*, 2024.

A. Khoja et al. Automated consistency analysis for legal contracts. *Artificial Intelligence and Law*, 2025.

A. Singh et al. A survey of classification tasks and approaches for legal contracts. *Artificial Intelligence Review*, 2025.

8. Maria J. Grant and Andrew Booth. A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, 2009.
9. Government of India. *The Indian Contract Act, 1872*.
10. Government of India. *The Micro, Small and Medium Enterprises Development Act, 2006*.
11. Government of India. *The Arbitration and Conciliation Act, 1996*.
12. Government of India. *The Digital Personal Data Protection Act, 2023*.