LITERATURE REVIEW & EXPLORATORY DATA ANALYSIS

LLM-Powered Contract Risk Detection and Clause-Based Inquiry System

Supervisor- Dr.Sharareh Taghipour

Nikhil Srinivasan

501286481

Abstract

Manual contract review remains a significant bottleneck in modern business operations. The traditional approach of having legal teams manually parse through hundreds of contracts is not only expensive, but often costing firms thousands of dollars per agreement and also introduces substantial human error, particularly when dealing with complex risk allocations or time sensitive transactions.

This project tackles these challenges by developing a comprehensive pipeline that leverages large language models for automated contract analysis. Rather than treating contract review as a monolithic task, we break it down into two core functions: identifying potential legal risks within specific clauses and enabling natural language queries about contracts.

Our approach draws on three established datasets that each address different aspects of contract analysis. We use CUAD primarily for training models to recognize and categorize risky clause types, while ContractNLI helps us build systems that can reason about legal relationships and implications. LEDGAR provides the large scale data needed to handle the diverse range of contract provisions encountered in practice.

What sets this work apart is the emphasis on practical deployment considerations. We conducted extensive preprocessing and analysis of each dataset to understand not just their strengths, but also their limitations and biases. The resulting system doesn't just classify clauses it also provides explanations for its risk assessments that legal professionals can actually use and verify.

The end goal is a tool that can realistically assist law firms handling high volume contract review, where the current manual approach simply doesn't scale. While we're not claiming to replace human legal judgment, this system could significantly reduce the time lawyers spend on routine risk identification tasks, allowing them to focus on more complex strategic analysis.

1. Literature Review

1.1 Methodologies

The methodologies described across the literature shows a diverse and evolving set of practices aimed at improving automated legal contract analysis. Rather than following a single roadmap, researchers have adopted varying strategies depending on the complexity of the task, resource availability, and specific legal contexts. Three common methodological themes emerge, the design of domain specific annotated datasets, the integration of techniques from other fields particularly software engineering and the adaptation of general purpose language models through fine tuning for legal tasks.

A.Annotated Datasets

Numerous studies underscore the importance of developing well labeled datasets as a foundation for model training and evaluation. Unlike general purpose corpora, these datasets are tailored to the structure of legal documents. For example, CUAD [1] comprises thousands of clause examples annotated with finely segregated legal categories, often collected through collaboration with legal practitioners. Similarly, LEDGAR [2] enables multi label classification of legal provisions across large scale contracts. This careful attention to domain relevance ensures that models trained on such data are more likely to reflect actual legal reasoning. However, these datasets tend to focus on specific contract types, which may limit the generalizability of resulting models to broader legal domains.

B.Interdisciplinary Approaches

Borrowing from software engineering, some researchers have experimented with techniques that identify "contract smells" [4], recurring patterns that signal potential problems in legal text. While originally used to highlight inefficiencies in code, this concept has been adapted to flag unclear, redundant, or even risky clauses. For example, Dechtiar et al. [4] implemented contract smell detection using LLMs to guide legal risk identification. Integrating such pattern based insights into contract analytics allows for a more fine tuned approach that goes beyond surface level text classification. These hybrid strategies attempt to bridge the gap between traditional rule based systems and more recent statistical methods, although their reliance on predefined patterns can be limiting in less structured or non standard contract formats.

C.Domain Specific Fine Tuning

Another area of importance is the customization of language models to better align with legal content. This is commonly achieved by further training large models on legal corpora data or employing task specific prompting techniques. Few shot learning and prompt chaining [16]

have shown potential in extracting structured information from unstructured text, even in low resource settings.Prompt Chaining decomposes complex tasks into manageable steps, rather than asking the model to directly identify risk from a dense contract clause, prompt chaining divides the problem into subtasks such as first determining whether a clause imposes a duty, then classifying the type of duty, and finally assessing its risk level or enforceability. For instance, Roegiest and Chitta [16] proposed a multi stage prompting strategy for legal QA that breaks complex legal interpretation tasks into manageable steps. This structured sequence mirrors the way legal analysts often approach such texts by isolating questions, interpreting them in stages, and building on previous conclusions. Few-shot learning complements this approach by enabling the model to generalize from a small number of examples, reducing the need for large annotated datasets. These strategies reduce dependence on extensive labeled datasets and allow for more flexible model deployment. However, they also introduce variability in results depending on the quality of prompts and the complexity of the tasks.

1.2 Results

Customizing language models for legal applications has shown measurable improvements in clause classification and risk detection tasks. When models are fine tuned using legal corpora particularly on datasets like CUAD [1] and LEDGAR [2] , they tend to perform well on standard clauses such as indemnity and termination. These clauses are relatively consistent in structure, which helps models recognize them with greater accuracy. In comparison, provisions that are less common or highly dependent on context continue to present difficulties. For instance Moon et al. [5] found, work on construction contracts has highlighted the inconsistent performance of BERT-based models when applied to clause categories like performance delays or ambiguous risk language.

Evaluation methods in legal NLP are still evolving, and the most common metrics don't always capture the complexity of legal interpretation. Breton et al. [6] noted that traditional span matching metrics can under represent model utility, especially when partial matches still carry legal meaning. Some researchers have proposed alternative evaluation techniques such as scoring based on proximity or structured matching to reflect legal context more accurately.

There is also increasing interest in combining machine learning with domain informed guidelines. For instance, Kim et al. [14] ,developed a hybrid system that was used to pair dense embedding search with more interpretable keyword based retrieval. This setup gave legal reviewers greater transparency and control, which is often more useful than a high raw accuracy score in isolation.

Rather than focusing solely on model precision, many recent studies like *CUAD [1]*, *LawLLM [7]*, and *LEDGAR* [2] are beginning to weigh practical outcomes more heavily. This shift reflects growing awareness that benchmark improvements don't always translate into usable tools for legal professionals. Whether a model output aligns with how a contract is reviewed in practice especially when clauses are interconnected or open to interpretation is becoming a more important measure of effectiveness.

1.3 Strengths

One of the clearest strengths in recent research is the shift toward building models and datasets that actually reflect how legal work is done. Instead of relying on general purpose benchmarks, many studies have focused on creating datasets that are grounded in real contracts and legal tasks. This focus has made model results more useful and has helped build trust among legal professionals who are often skeptical of automated tools.

Researchers have also started exploring techniques that don't just follow standard machine learning paths. For example, Dechtiar et al. [4] borrowed ideas from software development like using "contract smells" to detect patterns that could signal issues in clauses. Others, such as Kim et al. [14], have combined language models with rule based systems or domain knowledge to make results more precise and easier to explain. These efforts show a move toward more thoughtful, problem specific solutions rather than simply chasing leaderboard metrics.

Another positive trend is the growing awareness that legal tools must work for people, not just on paper. Some researchers have started involving legal professionals directly in the evaluation process to better understand whether model outputs align with practical needs. For example, in the study by Breton et al. [6], domain experts provided feedback on the usefulness of extracted legal terms, helping validate the tool in real contract review workflows. This type of user based evaluation helps ensure that models supports legal decision making, rather than complicating it.

1.4 Limitations

Despite the progress made by these studies, several important challenges continue to hold back the wider use of large language models in contract analysis. A major issue is that most research still focuses on English language contracts, mainly from the U.S. and U.K. This narrow focus makes it hard to apply the same models in other countries, where legal systems and contract styles can be very different. Models trained on one legal tradition often don't work well when moved into a new legal or cultural context. LawLLM [7], for instance, is trained only on U.S. legal data, limiting its scope internationally.

Another problem is that many models analyze clauses as if they stand alone, without considering how they relate to the broader contract context. Xu et al. [11] emphasized the importance of capturing clause dependencies in their ConReader system, which identifies implicit relations between contract provisions. When models don't account for this, they risk misinterpreting what a clause actually means especially in longer or more complex agreements.

There's also the challenge of transparency. While many models now score well on accuracy, they don't always explain how or why they made a particular decision. Studies like those by Kim et al. [14] and Wong et al. [10] highlight the need for more interpretable systems that can provide justifications legal professionals can review and trust.

Also, studies in this space often use different evaluation methods and annotation guidelines. As Lai et al. [8] note in their survey, the absence of standard benchmarks complicates comparison and collaboration. This inconsistency makes it difficult to build on previous work or establish clear performance baselines across studies.

1.5 Conclusion

The research on using large language models for contract analysis has gained real traction in recent years. There's been solid progress in building high quality legal datasets, trying out different modeling strategies, and designing tools that better understand the structure and language of contracts. These efforts have helped create more accurate and legally aware systems that move beyond basic keyword search.

With that in mind, some key obstacles are still holding back wider use in real legal settings. Most tools today still rely on English language data, and many don't fully account for how clauses connect across a contract. Another challenge is that the reasoning behind model decisions isn't always clear, which makes it hard for legal professionals to trust or explain the results. Moving forward, solving these problems will likely depend on stronger collaboration between legal experts and technologists, and a shared focus on building datasets that are both diverse and easy to work with.

It's better to think of today's legal AI tools not as complete solutions, but as promising early stage systems.Future research should focus not just on boosting performance scores, but also on making the tools easier to understand, fairer across different use cases, and more aligned with how legal professionals actually work.

2. Data Description

2.1 Overview of Selected Datasets

This project uses the following legal datasets:

- **CUAD (Contract Understanding Atticus Dataset)**
  A QA-style dataset where each sample includes a contract excerpt, a legal question, and an annotated span that should be reviewed by a lawyer.

- **ContractNLI**
  A natural language inference dataset framing clause level contract analysis as an entailment problem. Each contract is paired with 17 fixed hypotheses.

- **LEDGAR**
  A large scale dataset of contract provisions labeled with functional tags from U.S. SEC filings.

2.2 Descriptive Statistics

CUAD Dataset

- **Train set**
  - Total examples: 22,450
  - Context length: min 1,081 – max 338,211
  - Question length: min 143 – max 518
- **Test set**
  - Total examples: 4,182
  - Context length: min 645 – max 300,768
  - Question length: min 143 – max 518

ContractNLI Dataset

- **Train set**
  - Documents: 423
  - Label distribution: Entailment = 3,530; NotMentioned = 2,820; Contradiction = 841
- **Dev set**
  - Documents: 61
  - Label distribution: Entailment = 519; NotMentioned = 423; Contradiction = 95

- **Test set**
    - Documents: 123
    - Label distribution: Entailment = 968; NotMentioned = 903; Contradiction = 220

LEDGAR Dataset

- **Total provisions**: 846,274
- **Unique labels**: 12,608
- **Label distribution**
    - 1 label: 707,151 (83.56%)
    - 2 labels: 118,525 (14.01%)
    - 3–8 labels: <2.5%
- **Length buckets**
    - ≤200 chars: 94,781 provisions (11.2%)
    - 201–500: 271,290 (32.1%)
    - 501–1,000: 270,109 (31.9%)
    - 1k–2k: 167,733 (19.8%)
    - 2k–5k: 41,541
    - 5k–10k: 813
    - 10k+: 7 provisions

3. Exploratory Data Analysis

3.1 Data Preprocessing

The preprocessing phase focused on preparing three distinct datasets CUAD, ContractNLI, and LEDGAR for clause level contract risk analysis. Each dataset required domain specific handling due to their unique formats, annotation schemes, and document lengths.

CUAD Dataset

The CUAD dataset included complex legal texts with annotated clause spans. I first corrected inconsistencies in line breaks and ensured uniform formatting across contexts and question answer pairs. A key challenge was aligning answer spans as several annotations had index mismatches due to encoding or tokenization drift. To resolve this, I implemented a fuzzy matching routine that verified the correctness of each span. Cases with misalignment were corrected by locating the nearest plausible match in the original context. I also retained multiple answer spans when applicable, and structured the final records to support multi span retrieval during inference.

ContractNLI Dataset

ContractNLI required unification of multiple annotation sets. The raw data consisted of entailed, contradictory, or neutral hypothesis premise pairs derived from legal contract sections. Some entries had missing or outdated span mappings due to schema version drift. To address this, I reconstructed the span index pairs from metadata, normalized the hypothesis templates, and filtered malformed entries. All samples were restructured to maintain consistency across premise hypothesis pairs.

LEDGAR Dataset

LEDGAR, comprising labeled legal provisions, was preprocessed by trimming whitespace, normalizing label lists, and removing documents with empty or malformed text. I ensured label sets were consistently formatted (as lowercase, sorted lists) and that documents were token clean for downstream model readiness.

Overall, the cleaning process ensured that all datasets conformed to expected structure and format, minimized annotation errors, and handled edge cases like multi span answers or legacy metadata fields. This step was critical for maintaining the quality and legal interpretability of downstream outputs.

3.2 Exploratory Data Analysis

The EDA process aimed to understand the structural and semantic properties of the three datasets, uncover label distributions, and identify challenges relevant to LLM based contract understanding.

CUAD

The CUAD dataset exhibited wide variability in context length, with clause passages ranging from a few words to over a thousand tokens. Despite this, most question types had a relatively balanced answer span distribution. Multi span answers were more frequent in the test set, suggesting increased complexity during evaluation. I also observed a concentration of certain clause types , which may create label imbalance during fine tuning.

ContractNLI

In ContractNLI, every contract sample was evaluated against a fixed set of 17 hypotheses, creating a uniform sample structure. However, label distribution was skewed, most premise hypothesis pairs were labeled as "neutral," with "entailment" and "contradiction" underrepresented. This imbalance suggests the need for weighted loss functions or sampling strategies during model training. I also explored the diversity of hypotheses and noticed that certain legal entailments had higher disagreement rates, indicating potential annotation ambiguity.

LEDGAR

The LEDGAR dataset covered over 80 clause labels, but the distribution was highly long tailed. A few provision types such as "Confidentiality" and "Governing Law" dominated the dataset, while many categories had fewer than 20 samples. This imbalance poses a challenge for multi class classification and may benefit from hierarchical label grouping or data augmentation. Document lengths were generally short, and each provision was single labeled, making the dataset suitable for lightweight classification tasks.

4.Project Approach

To achieve the goal of automated contract risk detection and clause specific inquiry, this project adopts a modular architecture driven by large language models (LLMs), combining span extraction, risk classification, entailment validation, and question answering. Each module is aligned with a specific legal reasoning task and will be iteratively developed and evaluated.

4.1 Clause Risk Detection

The first step is to build a clause level classification system capable of identifying high risk provisions in full length contracts. A domain adapted LLM (such as Legal-BERT or RoBERTa-Legal) will be fine tuned to detect clauses related to categories like indemnification, limitation of liability, and regulatory compliance. A span based extraction head will be used to support both single span and multi span clause detection, with special care taken to handle long contexts using techniques like sliding windows or truncated attention.

4.2 Entailment Based Clause Validation

Once clauses are extracted, a second module will assess whether specific legal hypotheses are entailed by the identified text. This module will be trained as a sequence pair classification task,

leveraging legal entailment datasets. It will help filter out irrelevant or ambiguous clauses and confirm the presence of certain legal obligations or protections. We will explore models like RoBERTa-Legal and Longformer to manage the input size constraints typical in legal documents.

4.3 Clause Based Legal Question Answering

To support human in the loop legal review, the next component will enable users to ask natural language questions about specific clauses. A fine tuned QA model will retrieve relevant sections and provide span based or generative answers. This module will use question context pairs derived from contract analysis datasets and may incorporate retrieval augmented generation (RAG) for better generalization to novel queries and contract formats.

4.4 Evaluation Framework

Each module will be evaluated independently using task specific metrics:

- **Clause Detection:** Precision, recall, and F1 score based on overlap and exact match.
- **Entailment:** Accuracy and macro-F1 across entailment, contradiction, and neutral classes.
- **Question Answering:** Exact Match (EM), token-level F1, and answerability scores.

We will benchmark both general purpose LLMs and legal domain adapted variants to assess the benefit of legal fine tuning.

4.5 Final Integration and Interface

The final step involves integrating all modules into a seamless contract analysis pipeline with a user friendly interface. Users will be able to:

- Upload contracts for automated clause risk detection.
- Verify clause entailments and assumptions.
- Pose clause specific legal questions interactively.
- Receive a risk labeled contract summary with explanations.

5.GitHub Repository Link

https://github.com/NikStar2/MRP

6.References

[1] 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.

[2] D. Tuggener, P. von Däniken, T. Peetz, and M. Cieliebak, “LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts,” in *Proc. LREC*, 2020, pp. 1235–1241.

[3] I. Chalkidis et al., “LexGLUE: A Benchmark Dataset for Legal Language Understanding in English,” in *Proc. ACL*, 2022, pp. 4310–4330.

[4] M. Dechtiar, D. M. Katz, and H. Wang, “Software Engineering Meets Legal Texts: LLMs for Auto Detection of Contract Smells,” *Patterns*, vol. 6, no. 5, 2025.

[5] S. Moon, S. Chi, S. Chi, and S. Im, “Automated Detection of Contractual Risk Clauses from Construction Specifications using BERT,” *Automation in Construction*, vol. 135, pp. 104095, 2022.

[6] J. Breton et al., “Leveraging LLMs for Legal Terms Extraction with Limited Annotated Data,” *Artificial Intelligence and Law*, 2025.

[7] D. Shu et al., “LawLLM: Law Large Language Model for the US Legal System,” *arXiv preprint arXiv:2407.21065*, 2024.

[8] J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu, “Large Language Models in Law: A Survey,” *AI Open*, vol. 5, pp. 185–199, 2024.

[9] A. B. Candaş and O. B. Tokdemir, “Automating Coordination Efforts for Reviewing Construction Contracts with Multilabel Text Classification,” *J. Legal Affairs and Dispute Resolution in Engineering and Construction*, 2022.

[10] S. Wong, C. Zheng, X. Su, and Y. Tang, “Construction Contract Risk Identification Based on Knowledge-Augmented Language Models,” *arXiv preprint arXiv:2309.12626*, 2024.

[11] W. Xu et al., “ConReader: Exploring Implicit Relations in Contracts for Contract Clause Extraction,” in *Proc. EMNLP*, 2022, pp. 2420–2434.

[12] V. Aggarwal et al., “ClauseRec: A Clause Recommendation Framework for AI-aided Contract Authoring,” in *Proc. EMNLP*, 2021, pp. 8741–8754.

[13] P. G. Bizzaro, E. D. Valentina, N. Mana, M. Napolitano, and M. Zancanaro, “Annotation and Classification of Relevant Clauses in Terms-and-Conditions Contracts,” in *Proc. LREC*, 2024.

[14] E. W. Kim, Y. J. Shin, K. J. Kim, and S. Kwon, “Development of an Automated Construction Contract Review Framework Using LLM and Domain Knowledge,” *Buildings*, vol. 15, no. 6, p. 923, 2025.

[15] I. Dikmen et al., “Automated Construction Contract Analysis for Risk and Responsibility Assessment using NLP and ML,” *Computers in Industry*, vol. 155, 2025.

[16] A. Roegiest and R. Chitta, “Answering Questions in Stages: Prompt Chaining for Contract QA,” in *Proc. Natural Legal Language Processing Workshop at EMNLP*, 2024.