



Contract Sentinel: A Literature Survey on Transformer-Based Clause Extraction and Risk Detection in Legal Contracts

Vaishnavi Karpe¹, Tanishka Makode², Maria Shaikh³, P. J. Jambhulkar⁴

Student, Computer Engineering, Pune Institute of Computer Technology, Pune, India¹

Student, Computer Engineering, Pune Institute of Computer Technology, Pune, India²

Student, Computer Engineering, Pune Institute of Computer Technology, Pune, India³

Assistant Professor, Computer Engineering, Pune Institute of Computer Technology, Pune, India⁴

Abstract: Legal contracts such as non-disclosure agreements, lease agreements, and employment contracts often contain complex clauses that can create substantial financial, legal, and operational risk. Manually reviewing these documents takes time and requires domain expertise that individuals, startups, or small businesses do not always have, leaving them vulnerable to missing unfavourable or unusual terms. In this project, we present Contract Sentinel, an AI system that automatically reviews legal contracts and flags risky or unusual provisions for manual review. The system is based on natural language processing and machine learning algorithms trained on the CUAD (Contract Understanding Atticus Dataset), consisting of 510 real-world contracts with more than 13,000 expert labels across 41 clause types. The system takes a contract document as input, extracts clauses from it, and assigns each clause to a known risk category such as indemnification, liability caps, termination rights, non-compete terms, and auto-renewal conditions. Unusual or unfair clauses are marked with a description of the danger in plain English, allowing users to rapidly grasp and negotiate complex agreements without the need for legal expertise. Contract Sentinel uses transformer-based language models and a curated legal clause taxonomy to speed up and improve contract review, making it more accessible and reliable by reducing the need for expensive legal consultation for routine document checks.

Index Terms: Legal NLP, Contract Review, Clause Classification, Risk Detection, CUAD, Legal-BERT, Transformers, Explainable AI

I. INTRODUCTION

Contracts can mask unfavorable terms such as unlimited liability, one-sided indemnification, restrictive non-compete provisions or auto-renewal conditions. Manual review is slow and expensive and the field of Legal-NLP research has been moving to automate clause identification and risk assessment. Tugener et al. [1] introduced LEDGAR, a large corpus for provision classification. Hendrycks et al. proposed CUAD, an expert-annotated dataset that formulates contract review as an extractive question answering problem [2]. LexGLUE was proposed by Chalkidis et al. as a standard benchmark for legal language understanding [3].

Contract Sentinel builds upon these foundations. It takes a contract as input, pulls out clauses, assigns each to a known risk category (indemnification, liability caps, termination rights, non-compete terms, auto-renewal), and flags non-standard clauses with a plain-English explanation so non-experts can understand and negotiate them.

II. LITERATURE REVIEW

A. Datasets and Benchmarks for Legal Contract Understanding

SEC filings that are publicly available were used by Tugener et al. [1] to build LEDGAR. It has about 60,500 contracts, 846,000 provisions and about 12,600 labels when cleaned. They examined label-name matching, logistic regression, fastText based MLPs (with and without attention) and DistilBERT on four subsets of labels. DistilBERT performed best on the 13 most common provision types (micro-F1 0.95). Simple logistic regression fared at least as well for bigger low frequency label sets because DistilBERT's confidence scores for uncommon labels were unreliable. On a transfer test of 46 NDA labels annotated by lawyers, classifiers transferred well for some types of provision. Others scored around 0.30 F1 as the same label name has different meaning in different contract types. The study is about classification, not risk assessment.



CUAD introduced by Hendrycks et al. [2] contains 510 contracts from SEC EDGAR across 25 contract types, and 13,101 expert annotations across 41 clause categories. The categories include: General information Restrictive covenants Revenue risks The task is to extract the span of text relevant to each category from a model. We tested ten transformer models and found DeBERTa-xxlarge to perform best (AUPR 47.8%, precision at 80% recall 44.0%). Performance was highly dependent on the training set size and model architecture. Model size helped in a mixed way: ALBERT-xxlarge was only about 3% better than its smallest cousin. We pretrained RoBERTa-base on approximately 8 GB of unlabeled contracts which only increased AUPR by about 3%. Performance varied widely by category, reaching from close to 100% to about 20% AUPR. The authors specifically restrict the work to contract review and analysis, not “counseling” (assessing the business risk of clauses) which they characterize as very contextual.

Chalkidis et al. [3] introduced LexGLUE, a GLUE-style benchmark with seven English legal datasets on court cases, legislation and contracts from the US and EU. It covers multi-class, multi-label and multiple choice QA tasks. It compares general models (BERT, RoBERTa, Longformer) and legal models (Legal-BERT, CaseLaw-BERT). Pretraining on the domain was generally helpful, but legal documents that were long were still hard. Reproducibility is enhanced by public data, code, and consistent metrics. However, the benchmark is English-only, has limited task diversity and does not contain complex tasks like CUAD.

These works together offer the data and evaluation basis for Contract Sentinel. They show that rare clauses and extraction precision are still unresolved.

B. Domain-Adapted vs. General-Purpose Language Models

Singh et al. [4] benchmarked 13 legal-specific models vs 9 generalist models on UNFAIR-ToS, LEDGAR (100-class) and LEXDEMOD. The legal models consisted of encoders (Legal-BERT, Contracts-BERT, CaseLaw-BERT, LexLM and other), one encoder-decoder (LexT5) and two decoders (SaulLM-7B, AdaptLLM). Legal-specific models generally performed better than the generalist models, especially on rare long-tail classes. Contracts-BERT and Legal-BERT achieve new state-of-the-art results on UNFAIR-ToS and LEXDEMOD respectively, with approximately 69% fewer parameters than RoBERTa-large. RoBERTa-large remains best on LEDGAR. Encoders outperform larger decoders on rare classes. Newer models trained on broad legal mixtures performed worse than older models trained on focused contract data (Legal-BERT, Contracts-BERT). The authors blame this on a data-distribution mismatch. We find that the UNFAIR-ToS (few-shot micro-F1 64.7) is significantly better performed by fine-tuned models than by GPT-3.5 prompting (see appendix). The baselines are Legal-BERT, Contracts-BERT, CaseLaw-BERT and LexLM.

Instead of proposing a new architecture, Govea et al. [5] propose an evaluation-and-adaptation pipeline. It compares Legal-BERT, CaseLaw-BERT and LoRA/QLoRA/prompt-tuned BERT, T5 and BART against general models (BERT-base, GPT-3.5/GPT-4) for inference-only. The tasks are document classification (EURLEX, LEDGAR), clause extraction (five clause types), and regulatory summarisation (EUR-Lex-Sum). Metrics are Macro-F1, span-F1, exact match, IoU, ROUGE, BLEU and BERTScore over 20 runs with significance tests. Legal-adapted encoders performed better and more robustly on classification and extraction tasks. The results depend on the regime as GPT-3.5 had the best ROUGE-L in one table whereas GPT-4 was best in every metric in another. In a case study of contract updating, term-of-validity clauses were easiest (93.4% F1) and termination clauses hardest (85.9% F1, 6.4% semantic errors). Limitations: ~200 contracts in the corpus, no tests for OCR-noise, ROUGE does not measure legal validity.

These studies support Contract Sentinel's choice of legal-domain transformer encoders. They also show that the benefit depends on the task, the data distribution and the evaluation regime.

C. AI-Based Contract Risk Detection and Scoring

Chakrabarti et al. [6] (PwC, TENCON 2018) proposed “risk-o-meter”, an early ML framework for finding risk-prone paragraphs. It uses Paragraph Vectors (PV-DM with negative sampling) with SVM and Naïve Bayes classifiers, and a feedback loop that retrains on reviewer accept/reject decisions. The data were 1,382 training, 151 validation and 75 test paragraphs from request-for-proposal documents. A linear SVM performed best: Termination (499 training paragraphs) reached 91% accuracy and Indemnity 93%, but Indemnity recall was only 56%. Insurance (89 paragraphs) could not be optimized (0% recall). The framework treats all risky paragraphs equally and has no severity assessment. The authors list severity scoring as future work.

Gangadhar et al. [7] built an AI-powered legal contract risk analyzer using NLP and Gemini Flash 2.5. It accepts PDF, DOC and image inputs, uses OCR where needed, identifies clauses such as liability, indemnity and termination, and classifies them as high, medium or low risk. It also produces risk reports, mitigation recommendations, compliance checks and a dashboard with alerts. The evaluation is qualitative: it reports successful flagging of unlimited liability and ambiguous termination clauses, but no accuracy figures and little detail on the dataset or the criteria for assigning risk



levels. The authors acknowledge difficulty with complex or uncommon clauses and list explainable AI and multilingual support as future work.

Dhaksha Charan and others [8] have developed a serverless system (using React, AWS Lambda, S3, and Amazon Bedrock) which extracts ten specific types of clause by means of zero-shot generative-AI prompts, thereby eliminating the need for labelled training data. The clauses extracted are then compared with a standard template and any missing clauses are identified. The risk score is calculated as $\min(10H + 5M + wc, 100)$, H and M being the numbers of missing high- and medium-risk clauses respectively and wc representing a weight based on document complexity. The system also generates executive summaries and is able to compare different versions of contracts. It achieved 100% detection when tested on a single invoice, had an average extraction accuracy of 94.2% on 20 synthetic contracts, showed a correlation of 0.87 between the risk scores and the experts' ratings on 30 sample contracts, and detected 93.3% of the conflicts in 15 pairs of contracts. In this context, risk refers to the absence of a clause, not to the fairness of the way the clause is worded. The use of fixed weights, predefined categories and synthetic data restricts the system's ability to show nuance and generalise.

Prakash and others [9] have developed the Legal Document Analyser (LDA), which is the most comprehensive system examined. This system adjusts Legal-BERT in a multi-task configuration using four different heads: detection of clause boundaries (using a CRF), multi-label classification across 41 CUAD categories, NER for nine types of entities, and risk scoring (on a scale from 0 to 10 together with the categories Low, Moderate, High and Critical). A hybrid rule module (consisting of 3,200 gazetteer patterns and 180 regex expressions) intervenes when the neural NER predictions have low confidence (that is, when they are below 0.72). The training data consisted of CUAD, LEDGAR and 1,200 Indian commercial contracts. The system achieves an F1 score of 91.4% for clauses, 89.2% for NER, a risk MAE of 0.93 and a Cohen's Kappa of 0.87, outperforming rule-based methods, BiLSTM-CRF and BERT-base baseline models. Studies involving ablations indicate that the pretraining of Legal-BERT makes the greatest contribution (a decrease of 12.0 in clause F1 when it is replaced by BERT-base), then comes the rule module and multi-task training. In a simulation involving the review of 200 contracts, the median time required for the initial risk assessment dropped from 47 to 4.2 minutes. About 38% of the classification errors were caused by compound clauses. The limitations include coverage being limited to common law, a loss of information after 80 pages, lack of multilingual support, reliance on text-only risk indicators, and the use of a proprietary corpus which cannot be made available. It has been noted that explainability will be addressed in future work.

It is possible to achieve automated risk detection with these systems, but they vary considerably in terms of the realism of their datasets, the criteria they use for identifying risk, and the quality of the explanations they provide.

III.ANALYSIS AND SYNTHESIS

The literature provides a range of complementary features rather than a single comprehensive solution. Studies involving datasets and benchmarks supply annotated clauses together with evaluation procedures [1]–[3]. Investigations comparing different models indicate that legal-domain encoders, particularly Legal-BERT and Contracts-BERT, are good and efficient baseline options, especially when it comes to rare classes [4], [5]. Risk-detection systems go beyond clause understanding by adding categorisation, severity levels, scoring, and the generation of reports [6]–[9].

Several tensions appear across the papers:

- The value of domain pretraining. The authors of CUAD found only small improvements from using contract pretraining (approximately a 3% increase in AUPR) [2]. LexGLUE, Singh et al., Govea et al. and LDA all report clear advantages from the use of legal-domain models [3]–[5], [9]. The extent of the benefit seems to depend on the task, the data used, and whether the pretraining data match the target distribution [4].
- The weak points are rare and complicated clauses since they are consistently the weakest [1], [2], [4] to [9].
- It is difficult to compare the results given that there are differences in the data and the tasks:
 - The CUAD approach treats extraction as a span retrieval task, and its best model achieves a precision of 44.0% at 80% recall [2].
 - The LDA achieves an F1 score of 91.4% at the clause level when it comes to the classification of the annotated spans [9].
 - The evaluation of other systems involved 200 contracts [5], 20 synthetic contracts and one invoice [8], a small number of paragraphs from an RFP [6], or qualitative samples [7].
- The term "risk" is defined in different ways in various systems:
 - absence of expected clauses [8];
 - paragraph membership in a risk category [6];



- high/medium/low labels with undocumented criteria [7];
- A learned rating on a scale from 0 to 10 [9].

Contract Sentinel brings together all these elements. It makes use of transformer models that have been trained with CUAD for extracting clauses [2], [4], [9], a carefully selected taxonomy mapping clauses to risks [6]–[9], and plain-English explanations of why a clause is flagged, since [7] and [9] regard this as an area that requires further attention. The value that it offers must be established through baseline comparisons and ablation studies and not simply assumed on the basis of its architecture.

IV. RESEARCH GAP

“The studies we reviewed examined different points in the contract review process. Dataset papers are not about risk but about classification or extraction [1]–[3]. Papers comparing models evaluate classification and extraction not end to end review [4], [5]. There are four kinds of risk systems:

- those not measuring severity [6];
- those not having documented risk criteria or quantification [7];
- those with scores based only on missing clauses, with fixed weights and synthetic data [8];
- those giving scores without user-facing explanation, using text-only signals [9].

The gaps are:

- Rare and complex clause performance is still poor [1], [2], [4]–[9].
- Clear criteria for risk. Risk levels are assigned by fixed weights⁸, by presence alone⁶ or without documented criteria⁷.
- Content-sensitive risk. The scoring at hand penalizes missing clauses [8] instead of assessing if a clause is unusual or unfair.
- Realistic data evaluation. Studies use small synthetic or proprietary sets [5]–[9].
- Explainability for the layman Future work mentions explainable AI [7], [9]. Legal-validity metrics are missing [5].
- Handling of long documents is still hard [2, 3, 9].

Contract Sentinel addresses these by combining CUAD based clause extraction, a risk taxonomy and plain-English explanations in a single system. What are the research questions?

- Are domain-adapted transformers better than general models for clause extraction in CUAD in our setting?
- Does the mapping of clauses to a curated risk taxonomy improve the consistency of risk flags?
- Does plain-English explanation enhance understanding of flagged clauses by non-expert users?
- Does the system reduce the time for review compared to manual review?

Evaluation could involve precision/recall at fixed thresholds, AUPR and a Jaccard-overlap match criterion for extraction [2]. Micro/macro-F1 fit classification [1], [4]. Span-F1 and IoU fit boundary quality [5]. MAE and Cohen’s Kappa fit risk scoring [9]. Correlation with expert ratings fits scoring validity [8]. Review time saved fits efficiency [9].

V. CONCLUSION

The nine reviewed studies support expert-annotated clause datasets, legal-domain transformer models, and automated risk categorization and scoring for contract review. Contract Sentinel employs clause extraction trained by CUAD, a legal risk taxonomy, and plain-English explanations. The literature provides the technical motivation and evaluation principles for testing whether this integrated approach allows routine contract review to be faster, more accessible and more reliable

REFERENCES

- [1] D. Tuggenner, 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. 12th Lang. Resources and Evaluation Conf. (LREC)*, Marseille, France, May 2020, pp. 1235–1241.
- [2] D. Hendrycks, C. Burns, A. Chen, and S. Ball, "CUAD: An expert-annotated NLP dataset for legal contract review," in *Proc. 35th Conf. Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.
- [3] I. Chalkidis, A. Jana, D. Hartung, M. Bommarrito, I. Androutsopoulos, D. Katz, and N. Aletras, "LexGLUE: A benchmark dataset for legal language understanding in English," in *Proc. 60th Annu. Meeting Assoc. Computational Linguistics (ACL), Vol. 1: Long Papers*, 2022, pp. 4310–4330, doi: 10.18653/v1/2022.acl-long.297.



- [4] A. Singh, H. S. Karaca, A. Joshi, H. Paik, and J. Jiang, "Evaluating customized vs. generalist transformer-based models for legal contract classification," arXiv:2508.07849v2 [cs.CL], May 2026.
- [5] J. Govea, I. Ortiz-Gárce, P. Palacios, A. Maldonado Navarro, S. Acurio Del Pino, and W. Villegas-Ch, "Technical evaluation of language models adapted for the automation of legal contracts: Clause extraction, classification, and summarization," *Front. Artif. Intell.*, vol. 9, Art. no. 1782405, Mar. 2026, doi: 10.3389/frai.2026.1782405.
- [6] D. Chakrabarti, N. Patodia, U. Bhattacharya, I. Mitra, S. Roy, J. Mandi, N. Roy, and P. Nandy, "Use of artificial intelligence to analyse risk in legal documents for a better decision support," in *Proc. IEEE Region 10 Conf. (TENCON)*, Jeju, Korea, Oct. 2018, pp. 683–688.
- [7] K. D. P. Gangadhar, G. S. Vardhan, P. Kalyani, A. S. Kiran, and V. Jahnavi, "AI-powered legal contract risk analyzer," *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)*, vol. 14, no. IV, pp. 4465–4469, Apr. 2026, doi: 10.22214/ijraset.2026.80127.
- [8] R. Dhaksha Charan, S. Siddarth, and M. Jeevitha, "Automated legal clause extraction and risk scoring using NLP and generative AI," *IJRASET*, vol. 14, no. IV, pp. 687–695, Apr. 2026, doi: 10.22214/ijraset.2026.79369.
- [9] A. Antony Prakash, P. Arockia Joshwa, and S. Arunkumaran, "Legal Document Analyser: An intelligent NLP system for automated contract review, clause extraction, and legal risk evaluation," *Int. J. Sci. Innov. Eng. (IJSCI)*, vol. 3, no. 3, pp. 564–568, Mar. 2026, doi: 10.70849/ijsci03032695005.