LexiCheck: Semantic Analysis and Risk Assessment in Legal Contracts using Domain-Specific Transformers

Sarah L. Vos, Michael Chang
Center for Computational Law, Stanford University

July 12, 2025

Abstract

The manual review of legal contracts is a labor-intensive, expensive, and error-prone process. In high-volume environments such as procurement or mergers and acquisitions (M&A), legal teams often struggle to identify risky clauses hidden within thousands of pages of legalese. This paper introduces ”LexiCheck,” an automated contract review system based on the Longformer architecture. Unlike standard BERT models limited by a 512-token window, LexiCheck utilizes sparse attention mechanisms to process entire agreements (up to 4,096 tokens) in a single pass. We demonstrate its efficacy in extracting and classifying 12 distinct types of high-risk clauses, achieving an F1-score of 0.84 on the CUAD (Contract Understanding Atticus Dataset), comparable to junior legal associates.

1 Introduction

Contracts are the lifeblood of modern commerce. A typical Fortune 2000 company maintains between 20,000 and 40,000 active contracts. Reviewing these documents—specifically Non-Disclosure Agreements (NDAs), Master Services Agreements (MSAs), and Software Licensing Agreements (SLAs)—consumes significant legal resources.

The primary challenge in automating this process is the ”Semantic Gap.” A clause limiting liability might be phrased in endless variations. For example, ”Vendor's liability shall not exceed $50,000” and ”In no event shall the Service Provider be liable for damages surpassing the aggregate fees paid” mean semantically similar things (a Cap on Liability) but share very little lexical overlap.

Traditional keyword-based systems (”Ctrl+F” approaches) fail to capture these nuances. While recent advancements in Large Language Models (LLMs) have shown promise, general-purpose models often hallucinate legal interpretations or fail to cite the specific text responsible for a risk flag.

LexiCheck addresses these issues by treating contract review as a two-stage problem: 1. **Clause Extraction:** Identifying the specific span of text relevant to a legal topic. 2. **Risk Classification:** Categorizing that span as ”Standard,” ”Non-Standard,” or ”High Risk” based on a client's playbook.

2 Dataset Construction

Training models for the legal domain is notoriously difficult due to the scarcity of public data. Contracts are confidential by nature. We utilized the Contract Understanding Atticus Dataset (CUAD), an expert-annotated corpus of 510 legal contracts.

2.1 Annotation Taxonomy

We focused on 12 critical clause categories that most frequently trigger negotiation:

- **Indemnification:** Who pays if a third party sues?
- **Limitation of Liability:** Is there a cap on damages?
- **Governing Law:** Which jurisdiction (e.g., New York, Delaware) applies?
- **Termination for Convenience:** Can one party quit without cause?
- **Non-Compete:** Are parties restricted from working with competitors?

2.2 Data Augmentation

To improve robustness, we employed ”Legal Synonym Replacement.” We parsed a secondary corpus of 50,000 EDGAR filings to build a dictionary of interchangeable legal terms (e.g., substituting ”Vendor” with ”Contractor,” ”Supplier,” or ”Provider”). During training, we randomly swapped these entities to ensure the model learned the underlying legal concepts rather than overfitting to specific entity names.

3 Model Architecture

Legal documents are long. A standard MSA can easily exceed 20 pages (10,000+ words). Standard Transformer models like BERT have a quadratic complexity $O(n^2)$ with respect to sequence length, limiting them to 512 tokens. Truncating a contract risks losing critical context defined in the definitions section or addendums.

3.1 The Longformer Encoder

We adopted the Longformer architecture, which utilizes a ”Sliding Window Attention” mechanism. Instead of every token attending to every other token, each token attends only to its local neighborhood and a few global ”special tokens” (like the `[CLS]` token). This reduces complexity to $O(n)$, allowing us to process sequences up to 4,096 tokens.

3.2 Hierarchical Classification Head

The output embeddings from the Longformer are fed into a hierarchical classification head. Layer 1: Binary classifier (Is this sentence a risk clause?) Layer 2: Multi-class classifier (If yes, which type: Indemnity, Warranty, etc.?) Layer 3: Risk Scorer (Is this favorable to the Buyer or Seller?)

4 Experimental Results

We evaluated LexiCheck against two baselines: a standard BERT-base model (using a sliding window chunking approach) and a Rule-Based System (Regex).

Table 1: Performance on Clause Extraction (Exact Match)

Model	**Precision**	**Re**
Rule-Based (Regex)	0.92	0
BERT-base (Chunked)	0.76	0
LexiCheck (Longformer)	**0.81**	**0.**

4.1 Analysis of Errors

The Rule-Based system had high precision but terrible recall; it missed any clause that didn't match its rigid patterns. BERT-base suffered from "context fragmentation." In several cases, a definition appearing on Page 1 (e.g., "The term 'Services' excludes maintenance") altered the meaning of a clause on Page 15. The chunking approach severed this link. LexiCheck, with its larger context window, successfully maintained this dependency.

5 Risk Assessment and Playbooks

Identifying a clause is only half the battle. A lawyer needs to know *why* it matters. LexiCheck integrates a "Playbook Logic" engine.

5.1 Playbook Configuration

Users can define policies using natural language logic. *Example Policy:* "Flag any Limitation of Liability that is less than 2x the annual contract value."

The system extracts the `Liability Cap` (e.g., "$100,000") and the `Contract Value` (e.g., "$60,000/year"). It then performs symbolic reasoning:

$$100,000 < (2 \times 60,000) \rightarrow \text{True (Risk Flagged)}$$

5.2 Redlining Suggestions

For detected risks, the system suggests alternative wording from a "Preferred Clause Library."

Listing 1: Example Suggestion

```
ORIGINAL: "Vendor␣shall␣have␣no␣liability␣for␣indirect␣damages."
RISK: High. Unbalanced.
SUGGESTION: "Except␣for␣cases␣of␣Gross␣Negligence␣or␣Willful␣Misconduct,␣
    neither␣party␣shall␣be␣liable␣for␣indirect␣damages."
```

6 Ethical and Practical Considerations

6.1 Bias in Legal Data

The CUAD dataset is heavily skewed towards US Commercial Law. When tested on UK or EU contracts, performance dropped by 15%. The model struggled with concepts unique to GDPR (e.g., "Data Controller" vs. "Data Processor") that were underrepresented in the US-centric training data.

6.2 The ”Black Box” Problem

Lawyers are risk-averse. They are hesitant to trust a neural network that cannot explain its reasoning. To mitigate this, LexiCheck employs Attention Visualization. When a user clicks a risk flag, the UI highlights the specific words in the source text that the model ”looked at” most heavily (high attention weights) when making the classification.

7 Conclusion

LexiCheck demonstrates that domain-specific Transformer models can significantly accelerate the contract review lifecycle. By overcoming the sequence length limitations of standard BERT models, we can capture the long-range dependencies essential for legal interpretation.

However, we emphasize that this is an ”Assistive Intelligence,” not an autonomous lawyer. The system requires a human-in-the-loop to verify flags and handle edge cases. Future work involves integrating Generative AI (e.g., GPT-4) to draft novel clauses from scratch, rather than just selecting from a static library.

8 References

1. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*.
2. Hendrycks, D., et al. (2021). CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. *NeurIPS Datasets and Benchmarks*.
3. Chalkidis, I., et al. (2019). Neural Contract Element Extraction. *European Conference on Information Retrieval*.
4. Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.
5. Surden, H. (2014). Machine Learning and Law. *Washington Law Review*, 89, 87.
6. Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics*. Cambridge University Press.