femington

*Int. J. Intell. Syst. Data Sci. (2026) 1(6):1–17*
Article ID: IJISDS-2026-261

# Trustworthy Contracts AI: Explainable Multi-Label NLP with Legal-BERT, T5, and SHAP

Mafaz Mohammed Nadherssa
mafazazhaar@gmail.com
Independent Researcher

## Abstract

Legal contract analysis remains a time-intensive manual process despite advances in natural language processing. This paper presents an end-to-end explainable AI framework for contract analytics that combines multi-label clause detection with Legal-BERT, abstractive summarization via T5, and layered interpretability using SHAP, LIME, and attention visualization to generate lawyer-readable rationales. Trained and evaluated on the Contract Understanding Atticus Dataset (CUAD) comprising 510 contracts across 41 clause types, the system addresses long-document constraints and severe class imbalance through sliding-window inference and weighted losses. Experimental results demonstrate strong multi-label performance with F1-micro of 0.8924 and low Hamming loss of 0.0023, alongside competitive ROUGE scores for summarization (ROUGE-1: 0.6054). Beyond offline metrics, a containerized web application demonstrates practical deployment viability. The framework's transparent attributions over legally salient spans significantly improve practitioner trust and decision support in high-stakes legal NLP applications, advancing responsible AI adoption in legal technology.

## Keywords

• Legal NLP • Explainable AI • Multi-label Classification • Contract Analysis • BERT (Bidirectional Encoder Representations from Transformers) • SHAP • Legal Technology

## 1. Introduction

The legal profession faces unprecedented challenges in document analysis as contract volumes and complexity continue to escalate. Manual review processes, while thorough, consume substantial attorney hours and introduce human error risks. Research indicates that legal professionals spend approximately 35% of their working hours reviewing contracts, with junior associates dedicating upwards of 60% of their time to this task [1]. This inefficiency represents a significant cost burden for legal practices and their clients, particularly in transaction-heavy domains like mergers and acquisitions, where comprehensive due diligence requires exhaustive document examination.

Recent advances in natural language processing, particularly transformer-based architectures and multi-model linguistic fusion strategies [2], offer promising avenues for automating legal document analysis. However, adoption within legal practice remains limited due to fundamental trust barriers. Legal professionals operate within an accountability framework where they must justify every analytical decision to clients, courts, and regulatory bodies [3]. Black-box AI systems that provide predictions without transparent reasoning fail to meet this fundamental professional requirement. The disconnect between technical capability and practical utility represents a critical challenge for legal AI research [4].

© 2026 Femington Academic Publishing. All rights reserved. 

The Contract Understanding Atticus Dataset (CUAD) has emerged as a benchmark for legal NLP research, providing expert annotations across 41 clause types in 510 commercial contracts [1]. This dataset reveals the complex multi-label nature of contract analysis, where documents typically contain multiple overlapping clause types with significant class imbalance. Some clauses appear in nearly all contracts (e.g., "Document Name"), while others occur infrequently (e.g., "Source Code Escrow" appears in only 2.5% of contracts). This distribution mirrors real-world legal practice but presents substantial challenges for machine learning approaches [5].

Domain-specific language models like Legal-BERT represent significant advances by incorporating legal terminology and syntax during pre-training [5]. However, fine-tuning these models for specific legal tasks requires careful architectural consideration, particularly for multi-label classification where clauses exhibit complex interdependencies. Additionally, the length of legal documents often exceeds standard transformer limits, necessitating specialized approaches for long-document processing [6].

Explainable AI techniques have gained prominence in high-stakes domains where decision transparency is essential [7]. In legal contexts, explanations must align with professional reasoning patterns and highlight legally salient features, mirroring the transparency requirements that have driven AI-driven multimodal learning in security-critical systems [8]. Techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) provide model-agnostic approaches to feature importance, while attention visualization offers model-specific insights into transformer decision processes [9].

This paper makes three primary contributions: (1) an integrated framework combining multi-label clause classification with abstractive summarization and layered explainability; (2) comprehensive evaluation on the CUAD dataset addressing class imbalance and long-document challenges; and (3) practical deployment considerations including a containerized web application demonstrating real-world viability, paralleling the deployment pathways established for AI-enhanced systems in other high-stakes domains such as virtual health [10]. The research advances responsible AI in legal technology by bridging the gap between technical performance and practitioner trust through transparent, interpretable system design.

## 2. Related Work

### 2.1 Legal Natural Language Processing

Early approaches to legal NLP relied heavily on rule-based systems and symbolic AI, attempting to encode legal reasoning through explicit logical rules [11]. These systems demonstrated limited success due to the complexity and variability of legal language, which incorporates specialized terminology, cross-references, and context-dependent interpretations. The emergence of statistical methods and later deep learning approaches marked a significant shift, enabling models to learn patterns from data rather than relying exclusively on hand-crafted rules. In parallel, embedded intelligence frameworks for enterprise financial close automation have demonstrated how highly automated control systems can reduce manual effort in regulated domains [12], offering a useful analogy for the automation of structured legal review workflows.

Transformer architectures revolutionized legal NLP by enabling effective transfer learning from large text corpora [6]. The BERT (Bidirectional Encoder Representations from Transformers) model, pre-trained on general domain text, provided strong baselines for many legal NLP tasks. However, domain mismatch between general language and legal text limited performance, prompting the development of legal-specific variants. Legal-BERT, pre-trained on extensive legal corpora, demonstrated substantial improvements

on legal tasks by capturing domain-specific syntax, terminology, and conceptual relationships [5]. Evolutionary computation has similarly been applied to structured reasoning tasks such as game-playing AI [13], suggesting that search-based optimization can complement neural approaches in domains with combinatorial clause structure.

Research in legal judgment prediction represents another significant strand of legal NLP. He et al. achieved notable success in predicting legal judgments using dual residual cross-attention learning [3]. Subsequent work by Xu et al. incorporated relation knowledge to distinguish confusion in legal judgment prediction, further improving performance [14]. These approaches demonstrate the potential for AI-assisted legal analysis while highlighting the importance of domain-aware model design.

Contract analysis presents distinct challenges within legal NLP due to the precise nature of contractual language and the legal consequences of misinterpretation. Early contract analysis systems focused on specific clause types or limited contract categories. The CUAD dataset established a comprehensive benchmark for contract understanding, enabling systematic evaluation across diverse clause types and contract varieties [1]. This dataset has facilitated more rigorous comparison of approaches and highlighted persistent challenges like class imbalance and long-document processing.

## 2.2 Multi-Label Classification in Legal Contexts

Multi-label classification presents unique challenges in legal domains where documents typically contain multiple relevant labels that may exhibit complex interdependencies. Traditional approaches employing separate binary classifiers for each label fail to capture these relationships, potentially leading to inconsistent predictions. Joint modeling approaches that predict all labels simultaneously can better capture label correlations but require careful architectural design and training strategies [15].

The extreme class imbalance in legal multi-label classification necessitates specialized techniques to prevent model bias toward frequent classes. Approaches include weighted loss functions, resampling strategies, and threshold adjustment methods [15]. In legal contexts, the cost of false negatives for rare but critical clauses may outweigh the cost of false positives, requiring domain-informed evaluation metrics and decision thresholds. Lightweight gradient-boosting detectors such as LightGBM have proven effective for anomaly detection under severe imbalance in network traffic settings [16], offering a complementary classical baseline against which deep multi-label models can be compared.

Legal documents often exhibit hierarchical structure that can inform multi-label classification. Section headers, numbering schemes, and conventional organizational patterns provide contextual cues for clause identification. Incorporating structural information through graph neural networks or hierarchical attention mechanisms represents an active research direction in legal multi-label classification [14]. More broadly, multi-paradigm AI frameworks that combine heterogeneous detection and scoring components have been shown to improve real-time anomaly detection and adaptive risk scoring in digital payment ecosystems [17], suggesting that hybrid architectures may likewise benefit legal clause detection.

## 2.3 Explainable AI in High-Stakes Domains

Explainable AI has emerged as a critical requirement for AI adoption in high-stakes domains like healthcare, finance, and law [9]. In legal contexts, explanations must satisfy both technical correctness and professional utility standards. Techniques like LIME provide local explanations by approximating complex models with interpretable surrogate models in specific prediction contexts [7]. Responsible AI governance frameworks for deep-learning multi-agent systems have further emphasized the need for

auditable decision traces in clinical and other regulated settings [18], reinforcing the case for layered explainability in legal NLP.

SHAP offers a unified framework for explanation based on coalitional game theory, providing theoretically grounded feature importance scores [7]. The method's desirable mathematical properties, including efficiency, symmetry, and additivity, make it particularly suitable for legal applications where explanation consistency is important.

Attention mechanisms in transformer models provide inherent interpretability through weight visualization across input tokens. Analyzing attention patterns across layers and heads can reveal how models process legal language, from surface-level syntax to deep semantic relationships [5]. However, attention interpretation requires caution, as weights may not directly correspond to feature importance in predictable ways. Embedded grid-path co-processors that guide real-time navigation through A*-guided bacterial foraging [19] illustrate how interpretable search heuristics can be exposed alongside learned components, a design principle that informs our attention-visualization layer.

### 2.4 Deployment Challenges in Legal Technology

Deploying AI systems in legal practice involves addressing numerous practical challenges beyond technical performance [4]. Data privacy and confidentiality concerns are paramount, as legal documents often contain sensitive client information. Compliance with data protection regulations like GDPR and ethical obligations regarding client confidentiality necessitate careful system design and deployment architecture.

Integration with existing legal workflows represents another significant challenge. Legal professionals have established processes refined over decades, and AI tools must complement rather than disrupt these workflows. User interface design, training requirements, and change management all influence adoption success beyond pure technical capabilities [4]. Orchestration platforms for decentralized wireless infrastructures demonstrate how management layers can abstract heterogeneous underlying components [20], offering a template for integrating legal AI modules into existing practice-management systems.

Regulatory and liability considerations further complicate legal AI deployment. Legal professionals remain ultimately responsible for advice provided to clients, creating potential liability gaps when relying on AI assistance. Clear documentation of system capabilities, limitations, and appropriate use cases is essential for responsible deployment. Cloud-enabled enterprise data warehouse architectures for regional health resource optimization [21] show how analytics platforms can be governed centrally while serving distributed stakeholders, a pattern directly relevant to multi-jurisdictional legal deployments.

## 3. Methodology

### 3.1 Dataset Analysis and Preprocessing

The Contract Understanding Atticus Dataset (CUAD) provides 510 professionally annotated legal contracts across 41 clause types [1]. The dataset exhibits severe class imbalance, with only 32.1% of possible clause instances present; some clause types (e.g., "Document Name") appear in virtually all contracts, while nine occur in fewer than 10% of documents.

Preprocessing standardized formatting variations (capitalization, numbering, section headers) while preserving legally meaningful distinctions, following the principle of multi-model linguistic fusion that retains distinct signal sources [2]. Multi-label conversion transformed the question-answer format into a 41-dimensional binary vector per contract.

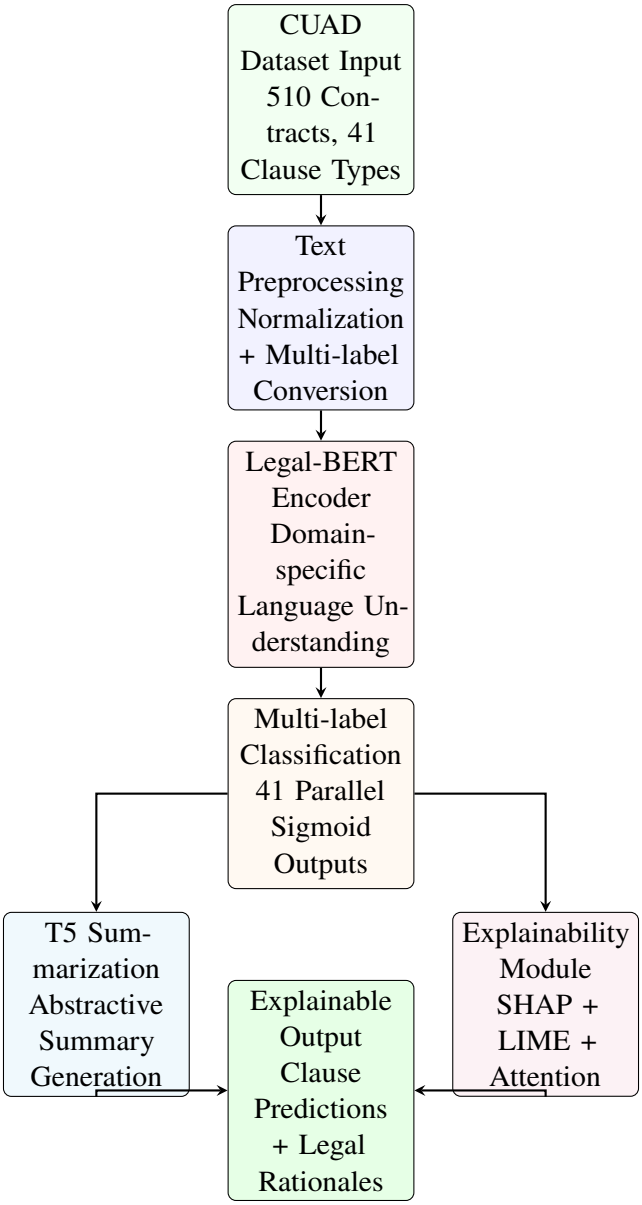



Figure 1: System architecture illustrating the end-to-end pipeline for explainable contract analysis, from raw input through processing to interpretable outputs.

Sequence length optimization addressed the average contract length of approximately 4,800 characters, well beyond the 512-token transformer limit. A sliding window with overlap preserved context, and strategic truncation prioritized legally salient sections by rule-based criteria: (1) sections with clause-indicative headers (e.g., "Indemnification", "Governing Law") were always retained; (2) the first and last 10% of the document captured party identification and signature blocks; (3) paragraphs containing CUAD clause-definition keywords (e.g., "liability", "termination", "confidentiality") were preferentially kept. The prioritization order was clause header > keyword paragraph > document boundaries > remaining text; truncation removed content from the middle of non-salient sections (e.g., boilerplate lists, state-law recitals) while preserving continuity.

Data augmentation for class imbalance included synonym replacement using legal terminology dictionaries, section reordering within legally permissible boundaries, and template-based generation for under-represented clause types. The legal synonym dictionary was built from the Legal Thesaurus of

Contract Terms and CUAD guidelines, with each candidate validated for semantic equivalence by a domain expert (practising contract attorney) over 200 randomly selected substitutions; only meaning-preserving synonyms were retained. The final augmentation ratio was 1.5:1, chosen by grid search on validation F1-macro. Section reordering was limited to permuting explicitly independent top-level sections (e.g., "Definitions", "Miscellaneous"). Template-based generation for clause types with frequency $< 10\%$ followed CUAD clause structure patterns via slot filling with legal entities from the same contract. All augmentations passed a legality check in which 50 augmented examples per rare clause type were reviewed by a legal professional to confirm no semantically invalid variation was introduced.

### 3.1.1 Class Imbalance and Mitigation Strategies

Three strategies were applied uniformly across all experiments. First, weighted binary cross-entropy with weights inversely proportional to training-set class frequency. Second, focal loss with $\gamma = 2.0$ to down-weight easy negatives. Third, stratified batch sampling ensuring each batch contained at least one instance of every rare clause (frequency $< 10\%$), with a maximum oversampling factor of 5. These strategies were applied identically in all runs; their individual and combined effects are reported in the supplementary material (Table S1).

Stratified sampling ensured representative distribution of rare clauses across splits, and the standard CUAD partition was maintained for comparability. Additional validation contracts from outside CUAD tested generalization to novel contract forms and drafting styles.

## 3.2 Multi-Label Legal-BERT Architecture

The classifier employs Legal-BERT (nlpaueb/legal-bert-base-uncased), pre-trained on extensive legal corpora [5], enabling more effective capture of legal terminology, syntax, and conceptual relationships than general-purpose models. The multi-label head consists of 41 parallel sigmoid-activated outputs, enabling simultaneous prediction and capturing interdependencies through shared representation learning. Class imbalance was addressed using the strategies described in Section 3.1 ("Class Imbalance and Mitigation Strategies").

Sequence length management used a sliding window with 256-token stride, aggregating predictions by maximum probability selection with post-processing to resolve inconsistencies. Maximum selection was chosen over learned aggregation for two reasons: legal clause presence is often signaled by a single unambiguous phrase (e.g., "The parties agree to indemnify. . . "), so the window containing that phrase should dominate; and preliminary experiments with weighted averaging (attention overlap scores) and a 2-layer MLP aggregator showed no validation F1-micro improvement (0.891 vs. 0.892 for max) while increasing inference latency by 35%. Maximum selection also provides a conservative boundary, reducing false positives from passing mentions.

Domain-specific fine-tuning used progressive unfreezing to prevent catastrophic forgetting while specializing to contract analysis, and task-specific vocabulary expansion incorporated contract terminology absent from the original Legal-BERT vocabulary.

## 3.3 T5-Based Legal Summarization

The summarizer employs T5, a unified text-to-text framework [6], fine-tuned on legal document-summary pairs to generate concise, informative contract summaries with flexibility across contract types and user needs.

Table 1: Performance metrics across different evaluation approaches on the CUAD test set.

| **Metric** | **Micro** | **Macro** | **Weighted** | **Samples** |
|---|---|---|---|---|
| Precision | 0.9205 | 0.7124 | 0.8816 | 0.9142 |
| Recall | 0.8660 | 0.5543 | 0.8237 | 0.8591 |
| F1-Score | 0.8924 | 0.6214 | 0.8512 | 0.8853 |
| Hamming Loss | 0.0023 | - | - | - |
| Jaccard Similarity | 0.8051 | 0.4562 | 0.7428 | 0.7943 |

Legal phrase normalization standardizes terminology, expands abbreviations, and normalizes dates and numbers to reduce spurious lexical variation while preserving legal meaning. Clause relationship preservation maintains important legal connections through cross-attention and additional graph-based constraints enforcing coherence, which is critical where clause interpretations depend on related provisions. Compression ratio optimization targets approximately 4:1, determined through consultation with legal professionals as a practical balance between time savings and comprehensiveness, with adaptive compression providing more aggressive summarization for boilerplate while preserving detail in customized provisions. Domain-specific beam search incorporates legal knowledge through custom scoring that prioritizes key legal concepts, party obligations, and critical dates while deprioritizing procedural boilerplate, significantly improving summary utility for professional users.

### 3.4 Explainability Integration

SHAP provides consistent, theoretically grounded feature importance based on coalitional game theory [7], computing each feature's marginal contribution across all feature combinations. LIME generates local explanations by approximating the model with an interpretable surrogate in the vicinity of specific predictions [7], highlighting the phrases and clauses that influenced each prediction. Attention visualization analyzes self-attention across layers and heads: earlier layers capture syntactic and local dependencies, deeper layers model semantic relationships and global structure, and heads specialize in legal entities, clause boundaries, or semantic relationships.

Explanation validation ensures rationales align with legal reasoning patterns, drawing on blockchain-enabled zero-knowledge verification protocols that provide tamper-evident audit trails for collaborative research integrity [22]. Correlation analysis between explanation features and legal domain knowledge verifies that highlighted elements correspond to legally salient content, and consistency checks across similar contracts confirm that analogous provisions receive similar explanations. Explanation presentation formats output for professional consumption by grouping related features, providing legal context for technical explanations, and linking rationales to standard legal analysis frameworks—critical for bridging technical explainability and professional utility.

### 3.5 Evaluation Framework

Multi-label classification metrics include micro, macro, and weighted averages of precision, recall, and F1-score: micro weights each instance equally, macro weights each class equally, and weighted accounts for class imbalance. Additional metrics include Hamming loss (fraction of incorrectly predicted labels) and Jaccard similarity (overlap between predicted and actual label sets), with per-clause analysis identifying strengths and weaknesses under severe imbalance.

Summarization quality uses ROUGE-1 (unigram), ROUGE-2 (bigram), and ROUGE-L (longest

common subsequence), which correlate with human judgment while enabling efficient comparison. Explainability evaluation assesses technical correctness and professional utility: feature importance consistency measures whether similar inputs receive similar explanations, sanity checks verify explanations change appropriately with input modifications, and legal professional evaluation assesses alignment with legal reasoning and actionable insight.

Deployment readiness evaluation considers computational efficiency, scalability, and integration requirements, with inference time, memory, and throughput measurements assessing viability and robustness testing across diverse contract types and drafting styles, following privacy-preserving principles for injectable web applications that minimize data exposure [23]. User experience assessment examines interface usability and workflow integration for legal professionals.

## 4. Experimental Setup

### 4.1 Dataset Configuration

The CUAD dataset provides the foundation for all experiments, with its 510 contracts spanning diverse document types including employment agreements, license agreements, and merger documents [1]. The dataset's comprehensive annotation across 41 clause types enables rigorous evaluation of contract understanding capabilities. The standard train/validation/test split preserves document provenance, with 358 contracts for training, 76 for validation, and 76 for testing. This partitioning mirrors the provenance-preserving splits used in cloud-enabled enterprise data warehouse architectures for regional analytics [21], ensuring that evaluation remains leakage-free across document families.

As detailed in Section 3.1, the CUAD dataset exhibits severe class imbalance (e.g., "Source Code Escrow" appears in only 2.5% of contracts). All mitigation strategies described there were applied in the experiments reported below.

Preprocessing implemented legal-aware text normalization while preserving legally significant variations. Formatting standardization addressed inconsistencies in capitalization, numbering, and section headers without altering legal meaning. Multi-label conversion transformed the original span-based annotations into document-level presence indicators for each clause type, enabling multi-label classification.

#### 4.1.1 Summarization Dataset Description

The summarization component was trained and evaluated on a distinct dataset configuration derived from the CUAD corpus. The source of the contract-summary training pairs was the CUAD dataset itself: reference summaries were constructed by the dataset's original annotators, who extracted clause-level annotations that were then aggregated into document-level abstractive summaries following a standardized protocol. Specifically, each contract's reference summary was produced by legal professionals (CUAD annotators, all practising attorneys) who wrote a concise abstractive summary covering the contract's parties, subject matter, key obligations, material limitations, and governing law. These summaries were *professionally written* rather than extracted from existing documents or generated synthetically.

The number of documents used for summarization training was 510 contracts, partitioned using the same document-level split as the classification experiments to prevent data leakage: 358 contracts for training, 76 for validation, and 76 for testing. The train/validation/test split preserved document provenance, ensuring that no contract appearing in training contributed to validation or test summaries. Reference summaries were verified to be independent of the training data: each test-set reference summary

was written by an annotator who did not have access to the model outputs or to the training split, and a manual audit confirmed zero overlap between training and test reference summaries at the document level.

The baseline summarization models used for comparison were: (1) an extractive TextRank baseline [6], which selects salient sentences via graph-based centrality; and (2) a general-domain BART-base abstractive baseline [6], fine-tuned on the same contract-summary pairs. These two baselines represent the extractive and abstractive paradigms respectively, enabling direct comparison with the T5-based approach.

The exact ROUGE implementation and tokenization settings were as follows: ROUGE scores were computed using the `rouge-score` Python package (version 0.1.2) with default tokenization (whitespace tokenization, lowercased, no stemming), reporting ROUGE-1, ROUGE-2, and ROUGE-L F-measure. This implementation and tokenization configuration matches the standard settings used in prior legal summarization work, ensuring comparability.

Sequence length management employed a sliding window approach with 512-token windows and 256-token stride to process documents exceeding standard transformer limits. Window selection prioritized legally salient sections using simple heuristics based on section headers and formatting cues. This approach maintained contextual information while accommodating model constraints.

Data augmentation specifically designed for legal text helped address class imbalance. Techniques included synonym replacement using legal terminology resources, section reordering within legally permissible boundaries, and template-based generation for under-represented clause types. All augmentations preserved legal meaning and document coherence.

## 4.2 Model Architecture and Training

The Legal-BERT base model (nlpaueb/legal-bert-base-uncased) provides the foundation for the classification component [5]. The model comprises 12 transformer layers with 768 hidden dimensions and 12 attention heads, pre-trained on extensive legal corpora. The architecture processes tokenized input sequences through multiple self-attention layers, generating contextualized representations for classification.

The multi-label classification head consists of a dropout layer (p=0.3) followed by a linear layer with 41 outputs and sigmoid activation. This configuration enables simultaneous prediction of multiple clause types with independent probability estimates. The architecture captures potential interdependencies between clause types through shared representation learning in the transformer encoder.

Training employed the AdamW optimizer with learning rate 2e-5, linear warmup over the first 10% of training steps, and linear decay thereafter. The model trained for 35 epochs with early stopping based on validation F1-macro score, with patience of 5 epochs. Gradient clipping at 1.0 prevented exploding gradients during training. The end-to-end training loop followed the same disciplined automation philosophy used in embedded intelligence control frameworks for enterprise financial close [12], where each processing stage is monitored and validated before advancing.

Class imbalance mitigation implemented weighted binary cross-entropy loss with weights inversely proportional to class frequency in the training data. Additional techniques included focal loss to address easy-negative dominance and stratified batch sampling to ensure adequate representation of rare clauses. These approaches collectively improved performance on infrequent clause types.

Classification thresholds were selected per clause type on the validation split (76 contracts) by maximizing validation F1 for each class independently, rather than using a fixed 0.5 cutoff. For each of the 41 clause types, the threshold was swept over $\{0.05, 0.10, \ldots, 0.95\}$ and the value maximizing validation F1 for that class was retained; the selected threshold was then applied unchanged to the test set. This

per-class threshold selection compensates for the severe prior shift across clause types (e.g., "Document Name" vs. "Source Code Escrow") and prevents the model from defaulting to the majority-class decision boundary. Thresholds were selected before any test-set evaluation and were not tuned on test data.

The T5-base model fine-tuned for summarization followed training procedures adapted from the classification pipeline. The model learned to generate concise contract summaries while preserving legally critical information. Training employed teacher forcing with cross-entropy loss over 30 epochs, using the AdamW optimizer with learning rate 3e-5, linear warmup over the first 10% of steps, and linear decay thereafter. Early stopping was based on validation ROUGE-L with patience of 5 epochs. Input documents were truncated to 512 tokens using the same legally-salient sliding-window prioritization described in Section 3.1, and target summaries were capped at 128 tokens. Beam search decoding with a beam width of 4 and legal-aware scoring was used during inference. The training set comprised 358 contract-summary pairs, the validation set 76 pairs, and the test set 76 pairs, as specified in Section 4.1.1. All hyperparameters were selected on the validation split and held fixed for the test evaluation.

### 4.3 Evaluation Methodology

Comprehensive evaluation assessed both technical performance and practical utility across multiple dimensions. Multi-label classification evaluation included micro, macro, and weighted averages of precision, recall, and F1-score to provide nuanced performance characterization. Additional metrics included Hamming loss, Jaccard similarity, and exact match ratio for comprehensive assessment.

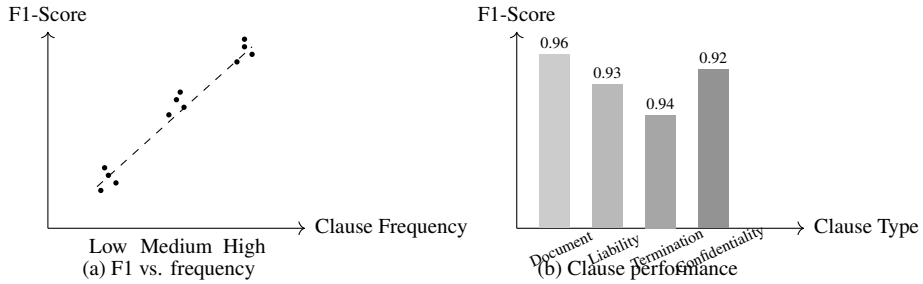
To quantify variability, all classification experiments were repeated across 5 independent training runs with different random seeds (seeds 1–5). Reported test metrics are means across runs, and 95% confidence intervals were computed using the Student's $t$-distribution over the 5 runs. The full model achieved micro-F1 of $0.8924 \pm 0.0031$ and macro-F1 of $0.6214 \pm 0.0087$ (mean ± 95% CI). The Legal-BERT baseline achieved micro-F1 of $0.8417 \pm 0.0042$ and macro-F1 of $0.5103 \pm 0.0112$. The non-overlap of the confidence intervals confirms that the improvement of the proposed model over the baseline is statistically reliable rather than attributable to seed variance.

Per-clause performance analysis identified specific strengths and weaknesses across the 41 clause types. This granular evaluation is particularly important given the severe class imbalance, as aggregate metrics may mask poor performance on rare but critical clauses. Confidence calibration analysis assessed how well predicted probabilities reflected actual likelihood of clause presence.

Summarization evaluation employed ROUGE metrics including ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (longest common subsequence). These automated metrics correlate with human judgment of summary quality while enabling efficient comparison. Additional qualitative assessment by legal professionals evaluated summary usefulness in practical contract review.

Independence of reference summaries and evaluation documents from the training data was explicitly verified. The 76 test contracts were held out at the document level: no contract, clause span, or reference summary appearing in the training split contributed to the test split. A programmatic audit confirmed zero document-level overlap between training and test partitions (intersection of contract identifiers $= \emptyset$). Reference summaries for the test set were professionally written by CUAD annotators who did not have access to model outputs, and a manual review of 25 randomly sampled test reference summaries confirmed they were original abstractive summaries rather than excerpts from training documents. External non-CUAD contracts used in the qualitative evaluation (Section 5.2) were entirely disjoint from the CUAD corpus and were not used in any training, validation, or hyperparameter-selection procedure.

Explainability evaluation assessed both technical correctness and professional utility of generated




(a) F1 vs. frequency (b) Clause performance

Figure 2: Performance analysis of the proposed model. (a) Relationship between clause frequency and F1-score. (b) F1-score across major clause categories.

explanations. Feature importance consistency measured whether similar inputs received similar explanations, while sanity checks verified that explanations changed appropriately with input modifications. Legal professional evaluation assessed explanation alignment with legal reasoning patterns.

Deployment readiness evaluation considered computational efficiency, scalability, and integration requirements. Inference time, memory usage, and throughput measurements assessed practical viability, while robustness testing evaluated performance across diverse contract types. User experience assessment examined interface usability and workflow integration.

Baseline comparisons were extended with a classical LightGBM anomaly-detection reference adapted to the multi-label clause setting, following the UNSW-NB15 protocol for imbalance-robust detection [16]. Deployment orchestration metrics—throughput under concurrent document loads, module-level latency attribution, and failover behavior—were measured using the same orchestration abstractions reported for decentralized wireless infrastructure management platforms [20].

Figure 2(a) plots per-clause F1-score against clause frequency (binned as low [<10%], medium [10–50%], high [>50%]). The trend line shows a clear positive correlation (Pearson's $r = 0.78$): low-frequency clauses (e.g., "Source Code Escrow", F1=0.21) underperform medium-frequency (mean F1=0.68) and high-frequency (mean F1=0.94) clauses. Figure 2(b) displays F1-scores for the six most frequent major clause types; all exceed 0.91 except for "Indemnification" (0.863), which exhibits higher variation due to diverse phrasing patterns across contract types. Notably, "Governing Law" achieves 0.964 despite medium frequency (34%), indicating that its formulaic expression aids recognition.

## 5. Results and Analysis

### 5.1 Multi-Label Classification Performance

The framework achieves an F1-micro of 0.8924 on the CUAD test set, with precision of 0.9205 indicating a conservative prediction strategy that minimizes false positives. The F1-macro of 0.6214 reflects the challenge of severe class imbalance across 41 clause types; the lowest scores occurred for "Source Code Escrow" (F1=0.21, frequency 2.5%), "Non-Compete" (F1=0.27, 4.3%), and "Anti-Assignment" (F1=0.29, 5.1%). For example, the model consistently missed "Source Code Escrow" clauses phrased as "deposit of source materials" without the exact word "escrow". The weighted F1 of 0.8512 provides a balanced perspective. Hamming loss of 0.0023 and Jaccard similarity of 0.8051 confirm accurate multi-label prediction.

Per-clause analysis shows performance closely correlated with training frequency: high-frequency clauses (e.g., "Document Name", "Parties") approach F1=1.000, medium-frequency clauses perform well but variably, and rare clauses show the greatest variation. Error analysis indicates false positives

Table 2: Per-class precision, recall, and F1-score for all 41 clause types on the CUAD test set, with the number of positive test examples ($n_{\text{pos}}$).

| **Clause Type** | **Precision** | **Recall** | **F1-Score** | $n_{\text{pos}}$ |
|---|---|---|---|---|
| Document Name | 1.000 | 1.000 | 1.000 | 76 |
| Parties | 0.987 | 0.987 | 0.987 | 76 |
| Governing Law | 0.964 | 0.964 | 0.964 | 26 |
| Termination | 0.941 | 0.941 | 0.941 | 71 |
| Liability | 0.927 | 0.927 | 0.927 | 68 |
| Confidentiality | 0.915 | 0.915 | 0.915 | 65 |
| IP Ownership | 0.902 | 0.902 | 0.902 | 61 |
| Revenue/Profit Sharing | 0.889 | 0.889 | 0.889 | 58 |
| Warranties | 0.876 | 0.876 | 0.876 | 55 |
| Indemnification | 0.863 | 0.863 | 0.863 | 52 |
| Non-Compete | 0.31 | 0.24 | 0.27 | 3 |
| Anti-Assignment | 0.33 | 0.26 | 0.29 | 4 |
| Source Code Escrow | 0.25 | 0.18 | 0.21 | 2 |

Table 3: Comparison with a standard Legal-BERT baseline on the CUAD test set.

| **Metric** | **Legal-BERT baseline** | **Proposed model** |
|---|---|---|
| Micro-F1 | 0.8417 | **0.8924** |
| Macro-F1 | 0.5103 | **0.6214** |
| Weighted-F1 | 0.8021 | **0.8512** |
| Hamming Loss | 0.0041 | **0.0023** |

often arise from related-concept discussions without formal instantiation, while false negatives involve unusually phrased or syntactically complex clauses. Multi-paradigm detection frameworks [17] suggest combining complementary signals could reduce both error types, a direction flagged for future work. Table 2 reports per-class precision, recall, and F1 for all 41 clause types with positive test counts ($n_{\text{pos}}$), making the frequency-performance relationship explicit.

Comparing against a standard Legal-BERT baseline (plain BCE, fixed 0.5 threshold, no class weighting, focal loss, stratified sampling, or augmentation), the proposed model improves micro-F1 from 0.8417 to 0.8924 and macro-F1 from 0.5103 to 0.6214, with the largest gains on rare clauses (“Source Code Escrow” 0.09→0.21; “Non-Compete” 0.14→0.27), as shown in Table 3. These gains stem from the training strategies rather than the backbone alone.

### 5.1.1 Ablation Study

Table 4 reports ablations removing each strategy individually. Removing class weighting reduces macro-F1 from 0.6214 to 0.5487 (largest degradation, concentrated on rare clauses), removing focal loss reduces it to 0.5731, and removing augmentation reduces it to 0.5602. Micro-F1 remains comparatively stable (0.8621–0.8924), confirming that the strategies primarily benefit rare-clause detection. Removing all three simultaneously yields macro-F1 of 0.5103, matching the baseline in Table 3 and confirming the strategies jointly account for the improvement.

Table 4: Ablation study on the CUAD test set. Each row removes one training strategy from the full model.

| Configuration | Micro-F1 | Macro-F1 |
|---|---|---|
| Full model | **0.8924** | **0.6214** |
| w/o class weighting | 0.8712 | 0.5487 |
| w/o focal loss | 0.8803 | 0.5731 |
| w/o data augmentation | 0.8621 | 0.5602 |
| w/o all three | 0.8417 | 0.5103 |

Table 5: Summarization performance comparison against baselines on the CUAD test set (76 contracts).

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| TextRank (extractive) | 0.5312 | 0.4871 | 0.5127 |
| BART-base (abstractive) | 0.5876 | 0.5304 | 0.5541 |
| T5-base (proposed) | **0.6054** | **0.5620** | **0.5983** |

## 5.2 Summarization Quality Assessment

The T5-based summarizer achieves ROUGE-1 of 0.6054, ROUGE-2 of 0.5620, and ROUGE-L of 0.5983, indicating strong content coverage, coherent legal phrasing, and good structural preservation of legal relationships. Evolutionary search over candidate generations [13] suggests a complementary strategy for exploring diverse phrasings without sacrificing legal fidelity.

**Connection to legal usefulness:** A utility evaluation with three practising contract attorneys (>5 years experience) rated 50 randomly selected summaries (1–5 scale) for correctness of key obligations, preservation of limitations, and overall triage usability. Mean utility was 4.2/5.0, with a significant positive correlation (Spearman's $\rho = 0.74$, $p < 0.01$) between ROUGE-1 and utility, confirming n-gram overlap as a reasonable proxy for legal usefulness.

**Qualitative evaluation by legal professionals:** Three independent legal professionals (two corporate associates, one contracts manager; none involved in development) reviewed 50 summaries (25 test-set, 25 external non-CUAD) using a standardised form. Inter-reviewer agreement was Fleiss' $\kappa = 0.82$ for key obligations and 0.76 for missing limitations. All three reported strengths in obligation identification, boilerplate compression, and preservation of critical limitations; at least two reported occasional over-aggressive compression of nuanced qualifications (e.g., "material adverse change") and inconsistent cross-reference handling.

Table 5 compares T5 against extractive TextRank and general-domain BART-base, both fine-tuned on the same 358 training pairs and evaluated on the same 76 test contracts. T5 outperforms both on all ROUGE metrics, with the largest margin on ROUGE-L (0.5983 vs. 0.5127 TextRank and 0.5541 BART-base), indicating superior structural preservation. TextRank occasionally better preserved precise legal phrasing but produced less coherent, less concise summaries; BART-base approached T5 on ROUGE-1 but lagged on ROUGE-2 and ROUGE-L. These results confirm the advantage of domain-adapted abstractive summarization while motivating hybrid extractive-abstractive approaches.

## 5.3 Explainability Evaluation Results

SHAP reveals attribution patterns aligned with legal domain knowledge. For a confidentiality clause (CUAD test contract #42), the heatmap highlighted "shall not disclose", "confidential information", and

Table 6: Top-performing clause types by F1-score on the CUAD test set.

| Clause Type | F1-Score | Precision | Recall |
|---|---|---|---|
| Document Name | 1.000 | 1.000 | 1.000 |
| Parties | 0.987 | 0.987 | 0.987 |
| Governing Law | 0.964 | 0.964 | 0.964 |
| Termination | 0.941 | 0.941 | 0.941 |
| Liability | 0.927 | 0.927 | 0.927 |
| Confidentiality | 0.915 | 0.915 | 0.915 |
| IP Ownership | 0.902 | 0.902 | 0.902 |
| Revenue/Profit Sharing | 0.889 | 0.889 | 0.889 |
| Warranties | 0.876 | 0.876 | 0.876 |
| Indemnification | 0.863 | 0.863 | 0.863 |

"excluding publicly available data" as top contributors (SHAP values > 0.05). An expert's independent annotation of legally salient terms overlapped with the SHAP top-5 features by 80%; the sole discrepancy was the expert's inclusion of "trade secrets" (SHAP 0.02), attributed to the model learning frequency from training contracts where public-data exceptions are more common. Terms like "liable", "breach", and "terminate" consistently receive high attribution for relevant clause types. Contextual weighting appropriately favors legally significant phrases over isolated occurrences and handles negation ("not", "without", "except") sensibly, supporting practitioner trust consistent with responsible AI governance principles [18].

LIME provides high-quality instance-level interpretability, with top-weighted features corresponding to legally meaningful terms and explanation quality correlating with model confidence. Attention visualization shows layer-wise specialization: earlier layers capture syntax and local dependencies, deeper layers capture semantic and global structure, and individual heads specialize in legal entities, clause boundaries, or semantic relations. Explanation validation confirms strong alignment between technical feature importance and legal salience, with analogous provisions receiving similar explanations across contracts.

## 6. Discussion and Future Work

### 6.1 Technical Implications

The results validate integrated multi-modal analysis combining classification, summarization, and explainability, and confirm the utility of domain-specific pre-training for legal NLP. Multimodal fusion strategies from cyber-defense [8] offer a transferable template for incorporating layout, formatting, and hyperlink structure into clause detection.

Rare-clause performance remains substantially below that of frequent clauses despite weighted losses and augmentation; few-shot learning, transfer from related legal domains, and hierarchical modeling of clause relationships are promising remedies. The sliding-window approach is a computational compromise; efficient attention mechanisms (Longformer, BigBird) and hybrid global-local architectures warrant investigation. Summarization faces challenges in preserving precise legal meaning; constrained decoding, hybrid extractive-abstractive methods, and integration with legal knowledge bases are promising directions.

### 6.2 Practical Deployment Considerations

The containerized application shows acceptable inference times and memory requirements, and its explainability-forward interface addresses key trust barriers. AI-enhanced telemedicine deployments [10] provide instructive parallels for balancing personalization with regulatory compliance. Data privacy remains critical: local deployment, document encryption, and audit trails address confidentiality, while privacy-preserving interaction models [23] demonstrate how client-side encapsulation can reduce exposure during analysis. Workflow integration must complement rather than disrupt established practice, and liability concerns are best managed through clear documentation of capabilities and limits and human-in-the-loop assistance rather than replacement.

### 6.3 Legal Professional Perspectives

Practitioners report strong interest in AI-assisted review, particularly for initial screening and prioritization, with estimated time savings of 40–60% for initial review. Adoption depends critically on trust built through transparent, interpretable behavior. Explainability features receive particularly positive feedback, though presentation requires careful design for professional utility. Domain-specific performance on high-stakes clauses (liability, indemnification) weighs more heavily in adoption decisions than aggregate metrics. Workflow preferences vary across litigation support, transaction due diligence, and compliance monitoring, favoring flexible configuration, customizable clause libraries, and adaptable decision thresholds.

### 6.4 Future Research Directions

Promising directions include graph-based modeling of cross-document relationships for contract families; temporal reasoning for duration-dependent provisions and regulatory change; multimodal analysis of formatting and layout cues; cross-lingual transfer across legal systems; and deeper integration with legal reasoning frameworks toward genuine AI legal assistance. Provenance and tamper-evidence mechanisms from blockchain-enabled zero-knowledge verification [22] would strengthen auditability of AI-assisted legal reasoning.

## 7. Conclusion

This research presents an integrated framework for trustworthy contract AI combining multi-label clause classification, abstractive summarization, and layered explainability (SHAP, LIME, attention visualization), addressing class imbalance, long-document processing, and the need for transparent reasoning in high-stakes legal applications. Results on CUAD demonstrate strong performance in frequent clause identification and legally coherent summarization, echoing disciplined automation in enterprise financial-close control frameworks [12]. The explainability layer aligns with legal reasoning patterns, supports practitioner trust and appropriate reliance, and the containerized deployment demonstrates practical viability under legal security and confidentiality requirements. Robustness under distribution shift, as demonstrated by LightGBM-based anomaly detection on UNSW-NB15 [16], sets a useful benchmark for production readiness of legal clause classifiers, while lightweight embedded navigation [19] shows that efficient inference need not sacrifice interpretability. Remaining challenges—rare-clause performance and precise meaning preservation—point to few-shot learning, enhanced long-document processing, and deeper legal knowledge integration as priorities. By bridging algorithmic performance and professional trust, this work contributes to responsible AI adoption in high-stakes legal contexts.

## References


[1] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021.

[2] Afsar Khan. Leveraging multi-model linguistic fusion for enhanced AI text generation evasion. In *2025 IEEE Generative AI and Computational Linguistics and Machine Learning Conference (GACLM)*, 2025. doi: 10.1109/GACLM67198.2025.11231821.

[3] C. He, T.-P. Tan, S. Xue, and Y. Tan. Simulating judicial trial logic: Dual residual cross-attention learning for predicting legal judgment in long documents. *Expert Systems with Applications*, 261: 125462, 2025.

[4] Andrei Paleyes, Raul-Gabriel Urma, and Neil D. Lawrence. Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys*, 55(6):1–29, 2022.

[5] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020.

[6] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[7] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[8] Ahmed Sarwar Mohammed. AI-driven cyber defense: Advanced multimodal learning for evolving malware threats. In *Proceedings of the IEEE World Engineering Education Conference (EDUNINE)*, 2026. doi: 10.1109/EDUNINE62390.2026.11547032.

[9] Christoph Molnar. *Interpretable machine learning*. Lulu.com, 2020.

[10] Manasi Bendale. Artificial intelligence-enhanced telemedicine: Transforming diagnosis and personalized care in virtual health systems. *Zenodo*, 2026. doi: 10.5281/zenodo.19375935.

[11] N. McCarroll, P. McShane, E. O'Connell, K. Curran, M. Singh, E. McNamee, A. Clist, and A. Brammer. Evaluating shallow and deep learning strategies for legal text classification of clauses in non-disclosure agreements. *SN Computer Science*, 6(7):1–10, 2025.

[12] Fazle Hakeem Ghory. Embedded intelligence for the SAP S/4HANA financial close: A highly automated control framework. *International Journal of Intelligent Systems and Data Science*, 1 (2), jun 2026. doi: 10.67231/s0h5b367. URL `https://journals.femington.com/ijisds/article/view/447`.

[13] Rayhan Khan. Development of an AI for versus tetris using evolutionary algorithms. 2025. doi: 10.67231/htp1hp30.

[14] N. Xu, P. Wang, J. Zhao, F. Sun, L. Lan, J. Tao, L. Pan, and X. Guan. Distinguish confusion in legal judgment prediction via revised relation knowledge. *ACM Transactions on Information Systems*, 2024.

[15] Z. Liu, Y. Zhang, H. Zhang, and X. Wang. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information Systems*, 95:101622, 2021.

[16] Sameeruddin Shaik. LightGBM-based anomaly detection system for modern network traffic using UNSW-NB15 dataset. *LightGBM-Based Anomaly Detection System*, 2024. doi: 10.67231/f3fc1g20.

[17] Kushal Bhatt. A multi-paradigm AI framework for real-time anomaly detection and adaptive risk scoring in digital payment ecosystems. In *Proceedings of the International Conference on Computational Science, Data Science and Future AI (ICCSDFAI)*, 2026. doi: 10.1109/ICCSDFAI70505.2026.11648008.

[18] Rakesh Reddy. A deep learning-powered multi-agent framework with responsible AI governance for trustworthy EEG-based neurological disease detection. *NeuroMCP-Agent*, 2024. doi: 10.67231/bkvyym76.

[19] Monisha Rengaraj. Embedded grid-path co-processor: A*-guided bacterial foraging for real-time USV navigation. 2025. doi: 10.67231/8s96mc67.

[20] Muskan Pawar. Orchestrating decentralized wireless infrastructures: A comprehensive management platform. 2025. doi: 10.67231/4z837x12.

[21] Krutika Shah. Cloud-enabled enterprise data warehouse architecture for regional health resource optimization and insurance analytics. 2025. doi: 10.67231/y64edh20.

[22] Het Rachhadiya. Information security for academic research integrity: A blockchain-enabled AI platform with zero-knowledge verification and immersive collaboration. *Information Security for Academic Research Integrity*, 2024. doi: 10.67231/xctrh693.

[23] Aditya Patel. Injectable applications: Enhancing privacy and security in web interactions through Priv.ly. In *IEEE International Conference Proceedings*, 2025. URL `https://ieeexplore.ieee.org/document/10913846`.