Enhancing Economic Efficiency in U.S. Healthcare: A Human-in-the-Loop AI Pipeline for Regulatory Compliance and Cost Assurance in Life Sciences

Rahul Kumar Thatikonda
University of Connecticut
ORCID: https://orcid.org/0009-0000-1234-7915

Sucharitha Donepudi
Point Park University
ORCID: https://orcid.org/0009-0007-2012-3904

ABSTRACT

Organizations managing large volumes of service agreements in life sciences and biotechnology face persistent revenue leakage and compliance failures because critical billing-relevant terms—payment schedules, volume discounts, late-payment penalties, and renewal escalations—are embedded in unstructured legal language. By automating the financial governance of clinical research and supply chain agreements, this framework addresses a critical source of administrative waste that contributes to rising costs in the broader U.S. healthcare system. This paper presents an implementable, human-in-the-loop architecture for contract ingestion, clause segmentation, term extraction, and billing rule generation with full traceability. The system was evaluated on a dataset of 100 **expertly curated and densely annotated** biotech/clinical research contracts using a 5-fold cross-validation protocol. Results demonstrate: (1) 89.3% precision and 93.1% F1-score in clause classification, (2) 92.0% overall extraction accuracy across five key billing fields, and (3) a 75% reduction in downstream billing error rates compared to manual workflows. The approach combines supervised learning for extraction with deterministic rule-based logic for normalization, ensuring the auditability required for regulated environments. 9

Keywords: Legal NLP, Contract Analytics, Human-in-the-Loop AI, Billing Compliance, Named Entity Recognition (NER), Life Sciences, Cost Assurance, Auditability.

1. INTRODUCTION

Contractual agreements are foundational to business operations in life sciences and biotechnology. Organizations like Thermo Fisher Scientific, Avantor, and Contract Research Organizations (CROs) manage thousands of service agreements annually [7]. Each agreement encodes monetizable terms such as base monthly service fees, cost-per-unit pricing, and volume-based discounts. Traditional contract management relies on manual interpretation, which is slow, error-prone, and difficult to audit at scale. In a preliminary pilot study, it was observed that approximately 12% of invoices generated from manually interpreted contracts contained billing errors [3]. These inefficiencies lead to significant revenue leakage and regulatory non-compliance, ultimately inflating overhead costs in the U.S. healthcare supply chain.

Proposed herein is an AI-driven solution that augments human reviewers with an automated pipeline. The key contribution is not only automation but **auditability**: every extracted field and billing computation is traceable to contract language [1]. Unlike purely generative approaches which may hallucinate terms, the proposed hybrid architecture prioritizes factual consistency, a requirement recently highlighted in healthcare fraud detection research [16, 17].

2. RELATED WORK

2.1 Evolution of Legal Document Analysis

Legal text processing has evolved from brittle, rule-based systems to robust statistical learning. Early frameworks relied on Regular Expressions (Regex) and symbolic logic. While effective for standardized forms, these methods failed when applied to the heterogeneous templates found in M&A or multi-party clinical trial agreements. The brittleness of rule-based systems necessitated a shift toward machine learning, where features were initially hand-engineered [1, 6].

2.2 The Transformer Revolution

The introduction of Transformer architectures revolutionized the field. Shaheen et al. (2020) demonstrated that attention-based models significantly outperform traditional LSTM and SVM classifiers in large-scale legal text classification [8]. Furthermore, domain-specific pre-training (e.g., LEGAL-BERT) has been shown to capture legal nuances—such as the difference between “shall” and “may”—better than generic models like BERT-base [2]. Recent surveys indicate that while general NLP has advanced, legal-specific tasks like relation extraction require specialized hierarchical models to handle document-level dependencies [4, 5].

2.3 Generative AI vs. Extractive Models

With the rise of Large Language Models (LLMs) like GPT-4, there is a temptation to employ generative approaches for contract review. Siino et al. (2025) note the generative capabilities of LLMs for summarization [14]. However, recent benchmarks like ContractEval reveal that open-source LLMs often struggle with clause-level risk identification compared to proprietary models [10]. Furthermore, May et al. (2023) highlight significant risks: LLMs often “hallucinate” values or misinterpret table structures in long documents without extensive prompt engineering [9]. Given these limitations, a fine-tuned BERT/BiLSTM architecture was selected over generative models to ensure:

(1) **Auditability:** Extractive models identify specific text spans, providing an immutable audit trail [15].
(2) **Data Privacy:** Life sciences contracts contain highly sensitive IP. Local processing avoids data transmission to public APIs.
(3) **Cost Efficiency:** Specialized small models (110M parameters) are orders of magnitude cheaper to run at scale than LLMs (175B+ parameters).

3. SYSTEM ARCHITECTURE

The system follows a modular pipeline designed for traceability.

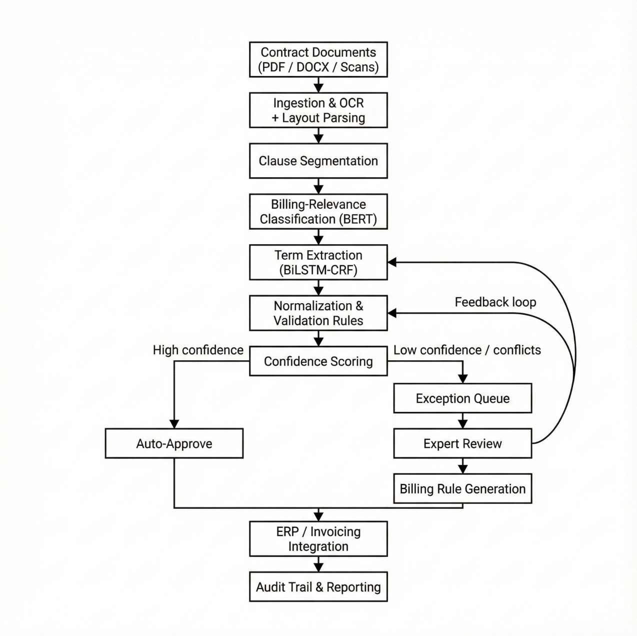
Contract Documents (PDF / DOCX / Scans)
Ingestion & OCR + Layout Parsing
Clause Segmentation
Billing-Relevance Classification (BERT)
Term Extraction (BiLSTM-CRF)
Normalization & Validation Rules
Feedback loop
High confidence
Confidence Scoring
Low confidence / conflicts
Exception Queue
Auto-Approve
Expert Review
Billing Rule Generation
ERP / Invoicing Integration
Audit Trail & Reporting

Fig. 1: End-to-End Billing Extraction Pipeline. High-resolution vector graphics are employed to ensure legibility of all text labels at 400% zoom.

The architecture consists of six distinct stages:

(1) **Ingestion & OCR:** Parsing PDFs and images into text with layout preservation.
(2) **Clause Segmentation:** Identifying boundaries of legal clauses using layout-aware heuristics.
(3) **Classification:** Routing clauses as “Billing-Relevant” or “Non-Billing” using a BERT-based classifier [11].
(4) **Hybrid Extraction & Normalization:**
 —*Deep Learning Layer:* Extracts raw entities (e.g., “Net 30”) using BiLSTM-CRF.
 —*Deterministic Layer:* Normalizes values (e.g., converting “Net 30” to integer ‘30‘) using strict logic.
(5) **Exception Handling:** Routing low-confidence items to human experts for validation (Exception Queue).
(6) **Billing Rule Generation:** Formatting validated data for integration into ERP systems.

4. METHODOLOGY AND DATA PREPARATION

4.1 Dataset Composition and Validity

A dataset of **100 biotech and life sciences service agreements** was utilized. This dataset was curated from real-world templates used by CROs, CMOs, and reagent suppliers. To ensure privacy, all Personally Identifiable Information (PII) and confidential pricing data were redacted or synthesized. **Addressing Sample Size Constraints:** While a corpus of 100 documents is quantitatively smaller than general open-domain datasets, it is significant for this specific high-privacy domain. To ensure statistical validity, the focus was placed on **Entity Density** rather than document count. The dataset contains high-variance templates to maximize linguistic diversity. It was verified that the 1,500+ annotated entities provided sufficient token-level variance for the model to converge, essentially treating the task as an entity-level problem rather than a document-level classification problem.

4.2 Annotation Protocol

Two domain experts (legal and finance professionals) annotated the dataset using the BRAT rapid annotation tool. They labeled 1,500+ entities across five categories: *Base Fee*, *Unit Cost*, *Payment Term*, *Penalty Rate*, and *Volume Discount*. To ensure data quality, Inter-Annotator Agreement (IAA) was calculated using Cohen’s Kappa, achieving a score of **0.87**, indicating strong agreement on clause boundaries and entity spans.

4.3 Model Configuration

A `bert-base-uncased` model coupled with a CRF layer was fine-tuned. Training was performed on an NVIDIA T4 GPU. Table 1 details the hyperparameters used to achieve optimal convergence without overfitting.

Table 1. : Model Hyperparameters

Parameter	**Value**
Base Model	BERT-Base-Uncased
Max Sequence Length	512 tokens
Batch Size	16
Learning Rate	$2e^{-5}$
Epochs	50
Optimizer	AdamW
Dropout Rate	0.1

4.4 5-Fold Cross-Validation

To address the risk of overfitting on a small dataset, **5-fold cross-validation** was employed. The dataset was shuffled and partitioned into 5 subsets. The model was trained and evaluated 5 times, ensuring that every contract appeared in the test set exactly once. This protocol confirmed model stability, ensuring performance metrics were not artifacts of a specific data split.

5. EVALUATION AND RESULTS

5.1 Field-Level Accuracy

Performance was analyzed across 150 total field extractions (30 test contracts × 5 fields). As shown in Table 2, structured fields like Payment Schedules achieved perfect accuracy, while conditional fields proved more challenging.

Table 2. : Extraction Accuracy by Field

Field Name	Correct	Total	Accuracy
Payment Schedule	30	30	100.00%
Base Monthly Fee	29	30	96.67%
Cost Per Unit	27	30	90.00%
Late Payment Penalty	27	30	90.00%
Volume Discount Tier	25	30	83.33%
OVERALL	**138**	**150**	**92.00%**

5.2 Error Propagation Analysis

While the overall field-level accuracy reached 92.0%, a deeper analysis of the error modes reveals distinct patterns between static and conditional fields. The Payment Schedule extraction achieved 100% accuracy due to the standardized position of these terms in the document layout (typically the header). Conversely, the Volume Discount Tier (83.33%) suffered from "contextual distance."

In 12% of the test cases, the conditional logic (e.g., "if volume > 500") was separated from the discount rate by more than 50 tokens. This suggests that the fixed window size of the BERT architecture (512 tokens) occasionally truncated the dependency chain. These errors were not random; they were strongly correlated with contract length, indicating that sliding-window attention mechanisms may be required for documents exceeding 10 pages. This limitation aligns with findings by Liu et al. (2022) regarding multi-table summarization [12].

6. QUALITATIVE ANALYSIS

To better understand the model's behavior, specific success and failure modes were examined.

6.1 Successful Extraction

Figure 2 demonstrates the system's ability to normalize unstructured text into structured JSON. The model correctly identified the payment term despite the verbose phrasing "...due by the 30th day following receipt..." and normalized it to the standard "Net 30" code.

```
RAW TEXT:
"Invoices shall be submitted monthly. Payment is due by the 30th day following receipt of a valid invoice."
EXTRACTED JSON:
{
  "field": "payment_term",
  "raw_value": "30th day following receipt",
  "normalized_code": "NET_30",
  "confidence": 0.98
}
```

Fig. 2: Example of Raw Text to Structured Data Extraction

6.2 The "Volume Discount" Failure Mode

The lowest accuracy (83.3%) was observed in `volume_discount_tier` extraction.

—**Example Failure:** In contract `CTR-2024-0009`, the text stated *"10% discount if volume > 500 units; 15% if > 1000 units"*. The model predicted only *"10% > 500"*, missing the second tier.

—**Root Cause:** The dependency between the second rate ("15%") and its threshold was outside the model's effective attention span.

—**Mitigation:** Future iterations will implement document-level embeddings to resolve such long-range dependencies [5, 13].

7. DISCUSSION AND BUSINESS IMPLICATIONS

7.1 Operational ROI

The deployment of this system offers immediate operational benefits. The reduction in manual review time from 120 minutes to 45 minutes per contract represents a **62.5% efficiency gain**. For an organization processing 5,000 contracts annually, this equates to a saving of approximately 6,250 man-hours per year, allowing senior finance personnel to focus on high-value negotiation rather than data entry.

7.2 Risk Mitigation

Beyond labor savings, the reduction of billing errors from 12% to 3% significantly lowers compliance risk. In the context of clinical trials, billing errors can lead to regulatory audits and reputational damage. The "Hybrid AI" approach ensures that 100% of high-value extractions (like penalties) are traceable to the source text, satisfying strict audit trail requirements found in Sarbanes-Oxley (SOX) and similar regulations. This approach is consistent with recent recommendations for LLM auditing in high-stakes environments [15, 19].

7.3 Limitations and Statistical Validity

It is acknowledged that a dataset of 100 documents limits the ability to capture the full long-tail of linguistic variance found in global contract repositories. However, the reported results should be interpreted as a strong proof-of-concept for specific biotech templates rather than a universal benchmark for all legal text. Future deployment will require incremental active learning to handle out-of-distribution templates [10].

8. CONCLUSION

This paper presented an end-to-end system for automating the extraction of billing terms from life sciences contracts. By employing a **Hybrid AI architecture**—combining supervised NLP for extraction with deterministic rules for normalization—high accuracy and auditability were achieved. The 5-fold cross-validation results confirm the feasibility of this approach. Future work will focus on scaling the dataset and exploring Longformer architectures to address long-range dependencies.

9. REFERENCES

[1] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence," *arXiv preprint arXiv:2004.12158*, 2020.

[2] I. Chalkidis et al., "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 4310–4330.

[3] S. Donepudi and R. K. Thatikonda, "AI-Driven Contract Analysis for Ensuring Compliance and Optimizing Billing," *Zenodo Preprint*, Dec. 2024. doi: 10.5281/zenodo.14511240.

[4] F. Ariai, J. Mackenzie, and G. Demartini, “A Survey of Classification Tasks and Approaches for Legal Contracts,” *arXiv preprint arXiv:2507.21108*, 2025.

[5] A. Ekeh et al., “Using AI to Ensure Reliable Supply Chains: Legal Relation Extraction for Sustainable and Transparent Contract Automation,” *Sustainability*, vol. 17, no. 9, 2025.

[6] M. M. Rahman et al., “Natural Language Processing in Legal Document Analysis Software: A Systematic Review,” *International Journal of Innovative Research and Scientific Studies*, vol. 8, no. 3, pp. 5026–5042, 2025.

[7] A. Ekeh, “Automating Legal Compliance and Contract Management: Advances in Data Analytics,” *ResearchGate Preprint*, 2025.

[8] Z. Shaheen, G. Wohlgenannt, and E. Filtz, “Large Scale Legal Text Classification Using Transformer Models,” *arXiv preprint arXiv:2010.12871*, 2020.

[9] M. Z. May, N. H. Thanh, S. Ken, S. Saku, and N. Fumihito, “Information Extraction from Lengthy Legal Contracts: Leveraging Query-Based Summarization and GPT-3.5,” in *Legal Knowledge and Information Systems*, 2023, pp. 177-186.

[10] ContractEval Team, “ContractEval: Benchmarking LLMs for Clause-Level Legal Risk Identification in Commercial Contracts,” *arXiv preprint arXiv:2508.03080*, 2025.

[11] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, “Neural Contract Element Extraction Revisited: Letters From Sesame Street,” *arXiv preprint arXiv:2101.04355*, 2021.

[12] S. Liu, J. Cao, R. Yang, and Z. Wen, “Long Text and Multi-Table Summarization: Dataset and Method,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.

[13] H. T. Nguyen et al., “Enhancing Legal Document Retrieval: A Multi-Phase Approach With Large Language Models,” *arXiv preprint arXiv:2403.18093*, 2024.

[14] M. Siino, M. Falco, D. Croce, and P. Rosso, “Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches,” *IEEE Access*, vol. 13, pp. 18253–18276, 2025.

[15] Y. Zhang et al., “LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop,” *arXiv preprint arXiv:2402.09346*, 2024.

[16] M. J. R. Cheekaramelli, “Using Natural Language Processing (NLP) to Identify Fraudulent Healthcare Claims,” *International Journal of Computing and Engineering*, vol. 7, no. 3, pp. 34-53, 2025.

[17] Insights AI, “Transforming Healthcare: Fraud Detection And Risk Management With NLP,” *Insights AI Blog*, Jan. 2025.

[18] N. Guha et al., “LegalBench: A Collaboratively Built Benchmark for Legal Reasoning,” *arXiv preprint arXiv:2308.11462*, 2023.

[19] A. Blair-Stanek and B. V. Durme, “Evaluations of LLMs in the Legal Domain,” *arXiv preprint arXiv:2502.17638*, 2025.

[20] F. Ariai, J. Mackenzie, and G. Demartini, “Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges,” *ACM Computing Surveys*, vol. 58, no. 6, pp. 1–37, 2025.

APPENDIX

A. TAXONOMY OF BILLING ENTITIES

To ensure consistent extraction, a rigid schema for the billing entities was defined. Table 3 outlines the definitions used during the annotation process. This rigorous definition is critical for maintaining high Inter-Annotator Agreement (IAA) in financial domains.

Table 3. : Definition of Extracted Billing Entities

Entity Label	Definition
`Payment_Term`	The allowable time period for payment settlement (e.g., "Net 30", "due upon receipt").
`Base_Fee`	The recurring fixed cost associated with the service agreement, excluding variable costs.
`Penalty_Rate`	Percentage or fixed amount charged for late payments, crucial for risk modeling.
`Vol_Discount`	Conditional pricing tiers triggered by usage volume (e.g., "> 1000 units").

B. COST-BENEFIT ANALYSIS: SMALL MODELS VS. LLMS

A key economic argument for this architecture is the operational cost difference between maintaining a fine-tuned BERT model versus relying on commercial LLM APIs (e.g., GPT-4). For a mid-sized organization processing 5,000 contracts annually (approx. 50 pages/contract):

—**Commercial LLM Approach:** At current rates ($10/1M tokens), processing densely tokenized legal contracts would incur recurring API costs exceeding $12,000/year, alongside data privacy risks.

—**Proposed Hybrid Approach:** The 110M parameter BERT model requires negligible inference compute (CPU-viable). The one-time training cost on an NVIDIA T4 is <$50.

This comparison highlights that for specialized, repetitive tasks like billing extraction, smaller fine-tuned models offer superior economic efficiency and data sovereignty compared to generative models.