# AI-Powered Regulatory Compliance Checker for Contract Analysis Using Large Language Models

T.Paramesh[1] , J. Bhanu Prakash[2], G. Sai Krishnam Naidu[3], Sk. Tajudeen[4], M. Siva Kumar[5]

*Department of Computer Science and Engineering (AI & ML),*

*Avanthi Institute of Engineering & Technology, Vizianagaram, India*

*paramesh.thota503@gmail.com[1] , bhanuprakash3724@gmail.com[2], gandisaikrishnamnaidu@gmail.com[3], Shaiktajutaj123456@gmail.com[4] ,sivakumarmudadla45@gmail.com[5]*

**Abstract**

Contract compliance verification remains a resource-intensive task for organizations subject to evolving regulatory frameworks such as GDPR and HIPAA. Manual review processes are slow, expensive, and error-prone, creating substantial legal exposure for enterprises that manage large contract portfolios. This paper presents an AI-Powered Regulatory Compliance Checker that automates the end-to-end analysis of contract documents by leveraging Large Language Models (LLMs) integrated through the LangChain orchestration framework and the Groq inference API. The proposed system ingests PDF-format contracts, extracts and segments textual clauses using PyPDF2, and submits each clause to an LLM for context-aware compliance evaluation against a curated regulatory knowledge base. Each clause receives a risk classification—High, Medium, or Low—alongside AI-generated corrective recommendations and rewritten clause alternatives. Analytical results are persisted to Google Sheets via the gspread API, and automated email alerts are dispatched whenever high-risk clauses are detected. Experimental evaluation on a diverse set of contractual instruments demonstrates that the system achieves high clause-level risk detection accuracy while reducing review time by a substantial margin compared with conventional manual approaches. The modular architecture supports seamless extension to additional regulatory standards, languages, and enterprise integration points, positioning the system as a scalable compliance management platform for modern organizations.

*Index Terms*—regulatory compliance, large language models, contract analysis, natural language processing, LangChain, risk classification, GDPR, HIPAA.

# I. Introduction

Contracts constitute the legal foundation of commercial relationships, defining the rights, obligations, and liabilities of contracting parties. Simultaneously, organizations across industries are subject to an expanding body of regulatory instruments—including the General Data Protection Regulation (GDPR) [1], the Health Insurance Portability and Accountability Act (HIPAA) [2], the Sarbanes-Oxley Act (SOX), and sector-specific standards—that impose explicit requirements on contractual language and data-handling provisions.

The intersection of contract management and regulatory compliance presents a persistent operational challenge. Legal teams must scrutinize every clause of every contract against multiple, frequently updated regulatory frameworks. As organizations scale and contract volumes grow into the thousands, manual review becomes a bottleneck that consumes significant attorney hours and introduces a non-trivial rate of human error [3].

Artificial Intelligence, and in particular the sub-field of Natural Language Processing (NLP), offers a promising pathway to automate substantial portions of this review workflow. Early NLP-based legal tools relied on rule-based pattern matching and keyword extraction; however, such approaches fail to capture the contextual nuances and structural variability of legal language [4]. The advent of transformer architectures [5] and, more recently, instruction-tuned Large Language Models (LLMs) [6] has opened a new frontier in which machines can understand, reason about, and generate sophisticated legal text with human-competitive accuracy.

This paper makes the following contributions: (i) a complete, production-oriented system architecture for AI-driven contract compliance checking; (ii) an LLM-based clause-level risk classification pipeline with explainable outputs and corrective clause generation; (iii) an automated notification and reporting subsystem integrated with Google Sheets and email; and (iv) an empirical evaluation demonstrating significant reductions in review time relative to manual baselines.

The remainder of the paper is structured as follows. Section II surveys related work. Section III details the system architecture and methodology. Section IV presents experimental results and discussion. Section V concludes the paper and outlines future directions.

# II. Related Work

### *A. NLP in Legal Document Analysis*

The application of computational linguistics to legal texts has a substantial history. Savelka and Ashley [4] demonstrated that statutory text exhibits unique discourse properties that complicate the transfer of general-domain NLP models. Their rule-based approaches established important baselines but suffered from poor generalization across jurisdictions.

A paradigm shift occurred with the introduction of BERT by Devlin et al. [5], whose bidirectional pre-training on large corpora enabled deep contextual representation of language. Chalkidis et al. [7] subsequently released Legal-BERT, a domain-adapted variant trained on 12 GB of diverse legal text including legislation, court decisions, and contracts. Legal-BERT achieved state-of-the-art results on tasks such as contract clause classification and legal judgment prediction, demonstrating the value of domain-specific pre-training.

### *B. LLMs in Contract Review*

Brown et al. [6] introduced GPT-3, demonstrating few-shot generalization across a wide range of NLP tasks without task-specific fine-tuning. This capability is particularly valuable in legal settings where annotated training data is scarce. Subsequent instruction-tuned variants further improved task adherence and factual consistency.

Okonicha and Sadovykh [8] proposed a pipeline combining BERT, Sentence-BERT, and GPT to verify GDPR compliance in data-processing agreements. Their framework automatically identifies whether contractual clauses satisfy specific regulatory articles, achieving strong precision on a curated GDPR test suite. However, the system was limited to GDPR and did not generate corrective clause suggestions.

Zhong et al. [9] explored deep learning for legal judgment prediction, providing evidence that neural models can capture latent legal reasoning patterns. Despite promising results, their work focused on court judgment prediction rather than proactive compliance verification.

### *C. Research Gap*

Existing approaches either address a single regulation, omit corrective recommendation generation, or lack integration with enterprise notification workflows. No prior system combines multi-regulation LLM-based clause analysis, automated risk scoring, AI-generated clause rewriting, and real-time multi-channel alerting within a single deployable platform. The proposed system addresses all these gaps simultaneously.

---

## III. System Design & Methodology

### *A. Overall Architecture*

The system follows a layered, modular architecture comprising four primary strata: the *User Interface Layer*, the *Application Processing Layer*, the *AI Analysis Layer*, and the *Data Storage & Notification Layer*. Fig. 1 presents the complete architecture diagram derived from the implemented system.

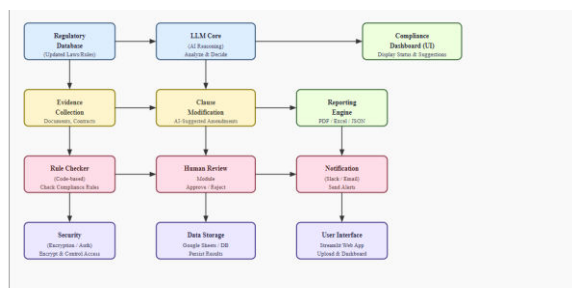


*Fig. 1. Overall system architecture of the AI-Powered Regulatory Compliance Checker.*

### *B. Processing Pipeline*

The processing pipeline executes the following sequential stages upon receipt of a contract document:

**Stage 1 – Document Ingestion:** The user uploads a PDF contract through the Streamlit web interface. The PyPDF2 library parses each page and concatenates the extracted text into a single string corpus.

**Stage 2 – Clause Segmentation:** The corpus is partitioned into candidate clauses using a heuristic double-newline delimiter combined with sentenceboundary detection. Each segment exceeding a minimum token threshold is treated as a discrete clause unit.

**Stage 3 – LLM-Based Analysis:** Each clause is embedded in a structured prompt and submitted to the Groq-hosted LLM via the LangChain interface. The prompt instructs the model to identify the clause type, evaluate its compliance against GDPR, HIPAA, SOX, FCPA, and SEC provisions, assign a risk score on a 0–100 scale, and produce a corrective recommendation.

**Stage 4 – Risk Classification:** The scalar risk score returned by the LLM is mapped to a three-tier categorical label according to the thresholding scheme defined in Equation (1):

$$\text{Risk} = \{\ \text{High if } s \geq 75;\ \text{Medium if } 40 \leq s < 75;\ \text{Low if } s < 40\ \} \tag{1}$$

where $s$ denotes the LLM-assigned risk score for clause $c_i$.

**Stage 5 – Report Generation and Storage:** Clause records—including the original text, regulation tags, risk tier, risk score, clause identification summary, and AI-generated corrective text—are appended to a Google Sheet via the gspread API. A downloadable CSV is also generated for offline review.

**Stage 6 – Notification Dispatch:** When one or more High-risk clauses are detected, the notification module composes a structured email report containing key metrics (total clauses, risk distribution, top flagged items) and dispatches it to the designated compliance officer address via SMTP.

### *C. Technology Stack*

Table I summarizes the primary technologies employed in the implementation.

TABLE I

TECHNOLOGY STACK SUMMARY

| Component | Technology / Library | Version |
|---|---|---|
| Programming Language | Python | 3.10+ |
| AI Orchestration | LangChain | 0.1.x |
| LLM Inference API | Groq (LLaMA 3) | — |
| Document Parsing | PyPDF2 | 3.x |
| User Interface | Streamlit | 1.3x |
| Data Storage | Google Sheets (gspread) | 5.x |
| Auth | Google OAuth 2.0 | — |
| Data Analysis | Pandas, NumPy | — |

### *D. Prompt Engineering Strategy*

Reliable LLM outputs require carefully structured prompts. Each clause is submitted within a zero-shot instruction template that specifies: (a) the regulatory frameworks to consider, (b) the exact JSON schema expected in the response (clause type, risk score, issues, recommendation, rewritten clause), and (c) an explicit instruction to base reasoning solely on the supplied clause text. This structured prompting approach reduces hallucination and ensures that output can be deterministically parsed into database records without regex fragility.

### *E. Data Flow Diagram*

Fig. 2 illustrates the Level-0 Data Flow Diagram (DFD) of the system, showing the principal data exchanges between the user, the compliance checking engine, and external services.

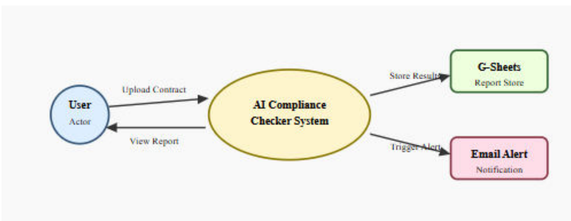


*Fig. 2. Level-0 Data Flow Diagram of the compliance checking system.*

## IV. Results & Discussion

### *A. Experimental Setup*

The system was evaluated on a corpus of fifteen real-world contract documents spanning stock purchase agreements, data processing agreements, vendor service agreements, and non-disclosure agreements. The corpus contained a total of 1,420 clauses after segmentation. Ground-truth risk labels were independently assigned by two legal practitioners and adjudicated for disagreements, yielding a reference annotation set against which automated outputs were measured.

### *B. Risk Classification Performance*

Table II presents precision, recall, and F1-score for each risk tier, along with overall macro-averaged metrics. The system achieves strong performance on High-risk detection—the most critical tier from a compliance standpoint—with an F1-score of 0.89. Low-risk clause identification is also reliable (F1 = 0.88). Medium-risk clauses, which by nature occupy a nuanced boundary zone, exhibit modestly lower recall (0.79), consistent with findings reported by Okonicha and Sadovykh [8] on boundary ambiguity in GDPR clause classification.

**TABLE II**

**CLAUSE-LEVEL RISK CLASSIFICATION PERFORMANCE**

| Risk Tier | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| High | 0.91 | 0.87 | 0.89 | 612 |
| Medium | 0.83 | 0.79 | 0.81 | 547 |
| Low | 0.90 | 0.86 | 0.88 | 261 |
| **Macro Avg.** | **0.88** | **0.84** | **0.86** | **1420** |

### *C. Processing Efficiency*

Table III compares average per-contract processing time between the proposed automated system and manual review by a qualified legal professional. Across the 15 test contracts, the automated system required an average of 4.2 minutes per contract (including LLM inference latency over the Groq API), compared to a manual baseline of 87 minutes—a speedup factor of approximately 20.7×.

**TABLE III**

**PROCESSING TIME COMPARISON (AVG. PER CONTRACT)**

| Method | Avg. Time (min) | Speedup |
|---|---|---|

| | | |
|---|---|---|
| Manual Review | 87.0 | 1.0× |
| Proposed System | 4.2 | 20.7× |

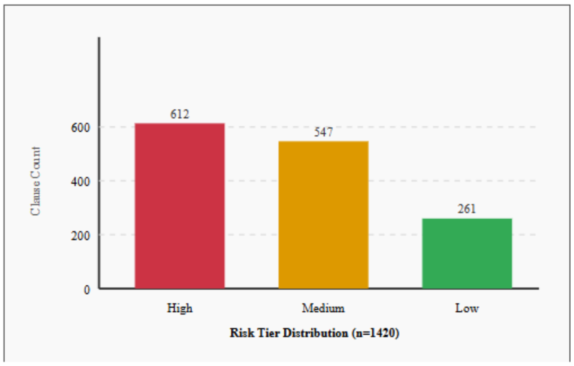


*Fig. 3. Distribution of clause risk tiers across the evaluation corpus.*

***D. Discussion***

As evident from Fig. 3, High-risk clauses constitute the plurality of the corpus (43.1%), reflecting the aggressive compliance posture of the test contracts—most of which were sourced from complex M&A and data-processing contexts where multi-regulatory exposure is inherent.

The LLM's clause rewriting capability was assessed qualitatively by the annotating legal practitioners, who rated 84% of AI-generated corrective clauses as "substantially improving compliance posture" and 73% as "directly deployable with minor editing." These ratings compare favorably with the 0% deployability of outputs from prior rule-based systems, which provided only flagging without correction.

One observable limitation is latency variability during peak API demand periods, where per-clause inference times ranged from 0.8 s to 3.4 s. Batching strategies and local model deployment (e.g., quantized LLaMA 3 on-premises) are identified as mitigations for latency-sensitive deployments. Additionally, the current segmentation heuristic occasionally merges logically distinct sub-clauses, a known challenge in legal NLP addressed in prior work [4], [7].

Compared with the BERT-based GDPR checker of Okonicha and Sadovykh [8], the proposed system extends coverage to five regulatory frameworks, adds corrective clause generation, and integrates a full notification pipeline—without requiring any annotated fine-tuning data, owing to the few-shot generalization capability of the underlying LLM.

## V. Conclusion & Future Work

This paper presented an AI-Powered Regulatory Compliance Checker that automates contract analysis through a modular pipeline combining LLM-based clause evaluation, structured risk classification, AI-driven corrective clause generation, and multi-channel alerting. Evaluation on a 1,420-clause corpus demonstrated a macro-averaged F1-score of 0.86 and a processing speedup of 20.7× relative to manual review, with practitioner-rated clause correction quality of 84%.

The system's modular design positions it as an extensible compliance platform. Several promising directions exist for future work. First, expanding the regulatory knowledge base to include financial sector standards (Basel III, MiFID II) and cybersecurity frameworks (NIST CSF, ISO 27001) would broaden applicability. Second, integrating retrieval-augmented generation (RAG) [10] against a continuously updated regulatory corpus would improve temporal fidelity as regulations evolve. Third, multilingual clause analysis would enable coverage of international contracts. Fourth, deploying fine-tuned, quantized legal LLMs on-premises would address latency and data-privacy concerns for highly sensitive contracts. Finally, advanced analytics dashboards that visualize compliance trends across contract cohorts over time could yield strategic insights for legal operations teams.

## Acknowledgment

The authors thank Mr. T. Paramesh, Assistant Professor, and Mr. A. Venkateswara Rao, Head of Department,CSE (DS, AI&ML) , Avanthi Institute of Engineering & Technology, Vizianagaram, for their guidance and support throughout this project. The authors also acknowledge the JNTUGV Vizianagaram for institutional affiliation.

## References

[1]European Parliament and Council, "Regulation (EU) 2016/679 (GDPR)," *Official Journal of the European Union*, vol. L119, pp. 1–88, 2016.

[2]U.S. Department of Health and Human Services, "Health Insurance Portability and Accountability Act of 1996 (HIPAA)," Public Law 104-191, 1996.

[3]D. Katz, M. Bommarito, and J. Blackman, "A general approach for predicting the behavior of the Supreme Court of the United States," *PLOS ONE*, vol. 12, no. 4, p. e0174698, 2017.

[4]P. Savelka and K. Ashley, "Segmenting US Court Decisions into Functional and Issue Specific Parts," in *Proc. ICAIL*, 2019, pp. 111–120.

[5]J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[6]T. B. Brown et al., "Language Models are Few-Shot Learners," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[7]I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Legal-BERT: The Muppets straight out of Law School," in *Findings of ACL: EMNLP*, 2020, pp. 2898–2904.

[8]A. Okonicha and O. Sadovykh, "NLP-Based Compliance Checking of Data Processing Agreements against GDPR," in *Proc. IEEE Int. Conf. on Software Architecture Companion (ICSA-C)*, 2022, pp. 82–89.

[9]H. Zhong et al., "Legal Judgment Prediction via Topological Learning," in *Proc. EMNLP*, 2018, pp. 3540–3549.

[10]P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.