International Journal of Financial Engineering
Vol. 12, No. 4 (2025) 2550016 (35 pages)
© World Scientific Publishing Company
DOI: 10.1142/S2424786325500161

World Scientific
www.worldscientific.com

Contract sentence-level evaluation (Con-SEN): A sentence-level semantic engine for accurate recognition of financial contract clauses

Zelin Li[*,¶], Zhiyong Li[*,†,‖], Tao Bai[‡,**], Yin Huang[‡,††] and Shaolei Chen[§,‡‡]

[*]*School of Finance, Southwestern University of Finance and Economics*
555 Liutai Avenue, Wenjiang District, Chengdu, Sichuan 611130, P. R. China

[†]*Random Forest Technology Co., Ltd.*
B0412 Jiaozi Fintech Center, 1677 Tianfu Avenue North
Gaoxin District 610043, P. R. China

[‡]*Xuyitong (Chengdu) Intelligent Technology Co., Ltd.*
1-3-1205, 1700 Tianfu Avenue North
Gaoxin District 610043, P. R. China

[§]*Sichuan XW Bank Co., Ltd.*
1-1-26, 8 Jitai Sanlu, Gaoxin District, Chengdu 610043, P. R. China

Received: 6 July 2025; Accepted: 30 July 2025
Published: 5 September 2025

Abstract

Financial contracts under regulatory review are often characterized by excessive length, intricate clause nesting, implicit negations, and cross-sentence legal dependencies — posing significant challenges for automated compliance systems. To address these issues, we propose Contract Sentence-level Evaluation (Con-SEN), a sentence-level semantic framework designed for fine-grained clause recognition and risk-oriented interpretation in financial documents. This work introduces three core innovations. First, we construct Con-SEN-Corpus, a domain-specific dataset spanning over 30 financial sub-industries from 2010 to 2024. Each sentence is annotated along three legal dimensions — clause type, legality status, and risk level — enabling multi-dimensional supervision. Second, we develop a structure-aware encoder based on an enhanced Longformer, incorporating a sliding-window mechanism and a novel clause-anchor global attention module to capture long-range dependencies and structural hierarchies across chapters. Third, we introduce a negation polarity and regulatory keyword injection module, which improves the model's ability to

[‖]Corresponding author.
Email addresses: [¶]14739184010@163.com, [‖]li@credit.li, [**]baitao@xuyitong.com.cn, [††]huangyin@xuyitong.com.cn, [‡‡]chenshaolei@xwbank.com

resolve adversarial logic, implicit obligations, and exemption clauses — often overlooked by general-purpose LLMs. To ensure coherent predictions across the three dimensions, we propose a multi-task consistency learning strategy that jointly optimizes clause classification, legality assessment, and risk estimation. Extensive experiments on real-world contract datasets show that Con-SEN significantly outperforms leading LLMs such as ChatGPT and Claude, achieving 18–25 percentage points higher document-level accuracy, up to 23% greater sentence coverage, and up to 6% improvements in clause-level classification. Moreover, it reduces volatility across evaluation metrics by an order of magnitude. These results position Con-SEN as a precise, regulation-aware framework for contract analysis, capable of handling the linguistic and structural complexity inherent in financial compliance tasks.

Keywords: Sentence-level semantic modeling; structure-aware encoding; multi-task learning; long-document processing; financial contracts.

1. Introduction

A recent report highlights that in current financial contract management processes, over 60% of the review work still depends on manual input, while version tracking largely relies on email or paper-based communication.[1] In high-frequency financial operations such as credit issuance, derivatives trading, and asset securitization, contracts serve as the primary vehicles for risk control and the allocation of rights and obligations, with their scale and complexity continuously increasing. Financial contracts typically feature lengthy sentences, deeply nested syntactic structures, frequent use of negation, and dense cross-references, all of which significantly complicate the review process. As a result, clause-by-clause manual review remains the dominant method, which is not only inefficient but also prone to human error and misjudgment, posing serious compliance and operational risks.

In recent years, Large Language Models (LLMs) have made significant advancements, demonstrating strong potential in tackling complex tasks across various domains, including medical diagnosis, educational support, and protein structure prediction (Kasneci *et al.*, 2023; Sallam, 2023). LLMs pre-trained on massive corpora have shown remarkable capabilities in semantic understanding, reasoning, and knowledge integration (Hao *et al.*, 2023). At the same time, LLMs have demonstrated fairly strong applicability in legal compliance review. Katz *et al.* (2024) found that GPT-4 with no fine-tuning can pass the United States Uniform Bar Examination with a score around the 90th percentile, approaching or even surpassing the level of junior lawyers in several subjective questions and case

[1] *Contract Lifecycle Management Market — Global Trends and Forecasts*. ResearchAndMarkets.com. Web. Accessed 30 June 2025 and available at https://www.researchandmarkets.com

analyses. In the layer of text classification, Guha *et al.* (2023) and Rodríguez *et al.* (2020) conducted a systematic evaluation of over 20 open-source and commercial LLMs, proving that after fine-tuning of examples in a small number of fields, the models can produce structured and interpretable legal Q&A and fact-legal provision matching results.

Existing general LLMs and multilingual legal models have made significant progress in tasks such as judgment prediction and clause classification, but there are still gaps and technical challenges in the sentence-level refined review of Chinese financial contracts: (1) Insufficient corpus in certain domains and the granularity of sentences is not fine enough. Current publicly available datasets mainly focus on English or Spanish commercial contracts, lacking a sentence-level, multi-dimensional annotated corpus covering the Chinese financial domain. (2) Ultra-long texts and hierarchical nesting create reasoning bottlenecks. Although the context window of mainstream LLMs has expanded to tens of thousands of tokens, when faced with complex syntactic structures and frequent cross-chapter references in financial contracts, models are still prone to issues such as context truncation or attention dilution, affecting reasoning accuracy. (3) Weak ability to parse negation semantics and cross-references. Financial contracts often include "unless…otherwise, you can't…", "without any event of default" and other nested negative expressions, as well as a large number of cross-references between terms. The current model shows low accuracy in dealing with such complex sentences. (4) The multi-task output lacks consistency constraints. Financial review requires the simultaneous generation of a ternary label: "Clause Type — Legality — Risk Level". However, mainstream models often treat these three tasks separately, failing to perform unified modeling, which leads to semantic conflicts in the results of outputs.

To fill the aforementioned gap, this paper proposes the Contract Sentence-level Evaluation (Con-SEN) model system tailored for financial contract scenarios, and trains and validates it based on the self-built Chinese financial sentence-level contract corpus Con-SEN-Corpus. The core contributions of this system include: (1) a structure-enhanced long text encoding mechanism: based on the improved Longformer, it is the first time that the "sliding window and clause-level anchor global attention" mechanism is applied to Chinese financial contract texts, effectively enhancing the model's ability to capture long-range dependencies across chapters and model the contract structures with coherence; (2) negation polarity and a global injection mechanism for keywords: a lightweight polarity annotator is designed to automatically mark negation trigger words and reference pointers, and injects them into the global attention pool during the encoding phase, enhancing the model's ability to parse complex negation sentence structures and cross-referencing structures; (3) a framework for a three-task consistency judgment head: using a shared sentence vector + parallel Softmax multi-task output structure, combined with a joint loss function, explicitly constraining the consistency of the “clause type — legality — risk level” ternary label, avoiding common semantic conflicts and category drift issues in traditional serial reasoning.

2. Literature Review

2.1. Long-text understanding models

In recent years, the Transformer architecture has reshaped the Natural Language Processing (NLP) landscape thanks to its powerful representation capability, yet its $O(n^2)$ complexity keeps the default context window capped at 512 tokens. For financial contracts that routinely span tens of thousands of words, this limit not only truncates semantics but also weakens the stability of downstream reasoning chains. Consequently, researchers have explored multiple technical routes to process ultra-long texts both efficiently and structurally.

Segment/recurrence paradigm. Transformer-XL (Dai *et al.*, 2019) and XL-Net (Yang *et al.*, 2019) propagate hidden states across segments via recurrent memory or permutation sampling, extending the input length from 512 tokens to several thousand. However, their memory windows still slide in fixed steps, leaving the chapter-clause-sentence hierarchy unmodeled. Compressive Transformer (Rae *et al.*, 2019) compresses stale memories to cut storage costs, but clustering inevitably sacrifices fine-grained details. Overall, these methods capture long-range dependencies better than pure sparse-attention models, yet they offer no direct modeling of article hierarchies or cross-section references.

Sparse/linear-attention paradigm. Longformer (Beltagy *et al.*, 2020) and BigBird (Zaheer *et al.*, 2020) reduce complexity to $O(n^2)$ by combining local windows with learnable global tokens. Performer (Choromanski *et al.*, 2022) linearizes attention via FAVOR+ kernel approximation. While they “make computation affordable”, their global tokens are usually limited to CLS or the first sentence, lacking explicit domain-specific anchors — hence they cannot pinpoint risk semantics in highly structured financial contracts.

Chunked/hierarchical models. Hierarchical Transformer (HT) (Pappagari *et al.*, 2019) and HTM (Zhu and Soricut, 2021) split documents into paragraphs and sentences, then aggregate across layers to capture discourse-level meaning. DocFormer (Appalaraju *et al.*, 2021) targets PDFs by fusing visual-text embeddings. Yet these approaches rely on fixed-granularity chunks and struggle with multi-line numbering and embedded tables common in financial-contract layouts.

Retrieval/augmented-memory models. RETRO (Borgeaud *et al.*, 2022) and KNN-LM (Khandelwal *et al.*, 2020) append externally retrieved passages back into the encoding sequence, excelling in open-domain Q&A. However, high domain specificity and copyright restrictions leave little retrievable Chinese financial-regulation text; moreover, the “retrieve-then-re-encode” pipeline incurs heavy overhead and has yet to be adopted in legal-contract settings.

In short, existing long-text models trade off computational complexity against explicit structural modeling: sparse attention prioritizes efficiency, hierarchical models focus on structure, and neither offers sentence-level multilabel supervision; handling of negation polarity and cross-clause references remains largely blank.

2.2. Domain adaptation and legal/financial long-text applications

The rapid iterative development of artificial intelligence technology has led to a growing interest in its application in the legal field. Legal Artificial Intelligence (LegalAI) aims to enhance the automation level of tasks such as legal document processing, compliance review, and legal consulting by leveraging methods like NLP. In recent years, LegalAI has gained widespread attention from both the NLP community and the legal practice community as an important direction of interdisciplinary integration (Chalkidis *et al.*, 2020; Hendrycks *et al.*, 2021; Ma *et al.*, 2021).

Meanwhile, the Transformer architecture, centered around the Self-Attention mechanism, has driven the rapid rise of Pre-trained Language Models (PLMs) (Vaswani *et al.*, 2017), becoming one of the key technologies for achieving breakthrough progress in the field of NLP. Typical models such as ELMo (Beltagy *et al.*, 2020), BERT (Devlin *et al.*, 2019), RoBERTa (Xiao *et al.*, 2021), and GPT-3 (Brown *et al.*, 2020) have successfully learned rich linguistic knowledge through unsupervised pre-training on large-scale unlabeled corpora, resulting in significant performance improvements across various downstream tasks such as text classification, question answering, and abstract generation. Inspired by the success of PLMs in general domains, an increasing number of studies have begun to focus on their potential in Legal AI, transferring and applying powerful language modeling capabilities to legal tasks, thereby promoting the development of several subfields, including legal clause classification, judgment prediction, and Legal Q&A (Chalkidis *et al.*, 2020; Ma *et al.*, 2021).

Driven by general PLMs, the application of NLP technology in the legal field continues to deepen. Early research primarily involved the direct transfer of models pre-trained in open domains, such as BERT and RoBERTa, to legal tasks, applied to typical tasks like text classification, legal Q&A, and judgment prediction (Chalkidis *et al.*, 2020; Shaghaghian *et al.*, 2020; Elwany *et al.*, 2019; Ma *et al.*, 2021). As the complexity of language structure increases and the complexity of syntactic nesting rises, these models often experience understanding breakdowns and reasoning errors due to a lack of legal context knowledge, exposing the semantic gap of open-domain models in the transfer of legal corpora.

To narrow the semantic gap between general models and legal tasks, researchers have proposed the Domain-Adaptive Pre-training strategy. Zhong *et al.* (2019) were the first to explore this path, followed by Chalkidis *et al.*, who released the LegalBERT model. Zhong *et al.* (2019) continued training BERT-base on documents such as EU and UK legal regulations and judgments from the US and the European Court of Human Rights (ECtHR), significantly improving the model's performance on tasks like text classification, abstract generation, and named entity recognition.

Building on continued pre-training, research further developed large legal-specific models trained from scratch or through deep adaptation. Zheng *et al.* (2021) pre-trained the model on 53,000 US case documents, optimizing the model's performance on legal reasoning datasets like CaseHOLD; Henderson *et al.* (2022) constructed a legal corpus Pile of Law, of approximately 256GB, covering regulations and contract documents from the US, Canada, and the EU, and conducted deep pre-training on BERT-large based on this corpus; Xiao *et al.* (2021) introduced a sparse attention mechanism (Longformer architecture) based on Chinese legal documents, combining sliding windows and global tokens to effectively model long legal documents, improving the model's performance on Chinese judgment-based tasks.

In recent years, Legal NLP has shown distinct shortcomings in data resource construction within the high-risk application scenario of financial contracts. Current mainstream contract datasets mainly consist of English commercial contracts, such as CUAD and LEDGAR. However, these datasets almost do not include Chinese contract samples and do not cover key contents in the financial sector. The gap in the corpus has led to a lack of systematic training for existing legal pre-trained models (Chalkidis *et al.*, 2020; Henderson *et al.*, 2022; Xiao *et al.*, 2021; Zheng *et al.*, 2021) on financial contract tasks, limiting their actual generalization ability in this subfield.

In recent years, the representation capability of long texts has gradually become one of the key bottlenecks for whether legal PLMs can be applied in real review scenarios. Current mainstream Transformer architecture models typically use self-attention mechanisms, with time and memory complexity of $O(n^2)$, and in practical applications, the input length is generally limited to within 512 tokens (Devlin *et al.*, 2019). However, a large number of legal documents far exceed this length limit. Existing studies have shown (Ma *et al.*, 2021; Zhong *et al.*, 2019) that when dealing with such ultra-long texts, models often fail to capture complete semantic dependencies due to context truncation, leading to the missing of important information and the overlooking of risk factors, which affects the understanding of sentence-level clause and the accuracy of accountability assessment.

To address the above challenges, researchers have proposed various sparse attention mechanisms to reduce computational costs and expand the effective receptive field of the models. Beltagy *et al.* (2020) combined sliding window attention with a learnable global token mechanism to achieve efficient long text modeling under complexity $O(n^2)$; however, the global nodes designed in such models are usually limited to the first sentence of the document, the CLS token, or a few manually selected points. For legal texts like financial contracts, which have clear structures and hierarchical clarity, the hierarchical structure of chapters-clauses-articles, frequent cross-clause references, and complex negation expressions have not yet been systematically modeled and utilized.

Although Xiao *et al.* (2021) introduced the Longformer architecture into this field for Chinese legal documents for the first time, significantly enhancing the model's ability to handle long documents, its global attention mechanism still primarily focuses on title tokens and has not been specifically designed for the risk factors in financial contract texts.

2.3. Sentence-level multi-task and consistency modeling

Despite the significant progress made by the current legal PLM, there are still issues such as a narrow range of corpus types that do not cover financial contract texts, insufficient modeling granularity, a lack of fine-grained analysis capabilities at the sentence level, limited semantic adaptation, and difficulties in accurately capturing the unique expressions and risk terminology specific to the financial industry. To address these issues, at the corpus level, we have constructed a large-scale Chinese financial contract dataset, Con-SEN-Corpus, which includes a total of 84,000 de-identified financial contracts and regulatory texts from 2010 to 2024. We have also introduced a ternary labeling system of "Clause Type — Legality — Risk Level" to provide a solid data foundation for fine-grained modeling of financial legal texts. At the model level, we propose a structure-enhanced Longformer encoder architecture, which adds "Clause Number and Title" as anchor points and injects these points into the global attention on the basis of the original sliding window (512 tokens) mechanism. This also integrates negation polarity and cross-clause reference pointers to achieve precise modeling of long-range dependencies and inverse semantic structures in financial contracts. Furthermore, we inject negation polarity markers and clause reference pointers into the input representation to explicitly mark reverse expressions such as "unless… Otherwise", "should not…" to enhance the model's ability to discern risk semantic reversals.

3. Con-SEN Method System and Its Implementation

3.1. The holistic design of methodology

To tackle the challenges in understanding contract texts in terms of complex structures, nested semantics, and ambiguous expressions at the sentence level granularity, this paper proposes a method for sentence-level identification for long texts — Con-SEN. The method treats the single sentence in the contract as its fundamental processing unit and designs three core functional modules, consisting of long-text encoding, semantic enhancement, and multi-task label prediction, enabling the structured understanding of the sentences in the contract and the generation of relevant labels. The overall processing workflow is as follows:

$$\mathcal{D} \xrightarrow{\mathcal{P}} \{s_i\}_{i=1}^{N} \xrightarrow{\mathcal{E}} \{h_i\}_{i=1}^{N} \xrightarrow{\mathcal{C}} \{(\tau_i, \lambda_i, \rho_i)\}_{i=1}^{N}. \tag{1}$$

Among the letters in the formula above, $\mathcal{D}$ refers to the original contract document, $\mathcal{P}$ the sentence-level preprocessing module, $\mathcal{E}$ is the encoding and semantic enhancement module, and $\mathcal{C}$ is the multi-task label prediction module. Every one of the sentences in the contract is finally reflected as a triples $(\tau_i, \lambda_i, \rho_i)$, while τ_i represents the categories of terms, λ_i is the legality tags, and ρ_i is the levels of risk.

Built upon an enhanced Longformer architecture, the Long Text Encoding Layer integrates sliding window attention with sparse global attention mechanisms. By configuring appropriate window sizes, this module effectively captures contextual relationships between contract statements and supports long-range dependencies that span across paragraphs and chapters. Additionally, the incorporation of anchor tokens — such as titles, numbering, and other structural markers — into the attention mechanism further strengthens the model's ability to represent the hierarchical structure inherent in legal and financial texts.

To enhance the model's understanding of sentence structure and semantic reversal, the semantic augmentation module injects multi-layered position encoding information such as "section-term-paragraph" along with negative polarity indication markers (such as signals for logical negation or condition inversions and other language phenomena) into the model on the basis of sentence vector encoding. This module provides explicit structure guidance with semantic markup support, enhancing its ability to model complex sentence structures such as syntactic nesting and conditional expressions.

The Multi-task Prediction Layer, adopting a parallel architecture, concurrently performs three classification prediction tasks on each sentence: clause type (multi-classification), legality assessment (binary classification), and risk level (multi-classification). Model outputs consist of three Softmax branches, which are combined with a cross-entropy loss function to construct an overall loss.

Through joint optimization, cooperative learning is achieved across three label dimensions. The overall procedure is depicted in Fig. 1.

3.2. Contract text encoding

To adapt to the financial contract texts' traits of deep structural hierarchy, lengthy length, and frequent semantic references across sentences, Con-SEN introduced a structure-aware long-text modeling mechanism during the text encoding phase. Specifically, the original contract documents would initially be represented as a Toke sequence $X = [x_1, x_2, \ldots, x_T]$ of length T, with x_i representing the ith token. The sequence was subsequently fed into an encoder based on an improved Longformer architecture, which employed a sliding window attention mechanism with a window width set as $w = 512$, enabling the construction of local context dependencies and the generation of hidden layer representations sequence:

$$H = \text{Longformer}_w(X). \tag{2}$$

To enhance the model's perception of the hierarchical structure of contracts, Global Structural Anchors (GSAs) was further introduced into the attention mechanism. To be concrete, tokens of certain clause titles in the contract, and of their corresponding hierarchical indicators (such as "Chapter X, Article Y"), will be explicitly annotated as structure anchors, forming an anchor set $\mathcal{G}$. In the process of attention weight computation, for the query vector Q_i at position i and the key vector K_j at position j, their attention weights A_{ij} are defined as

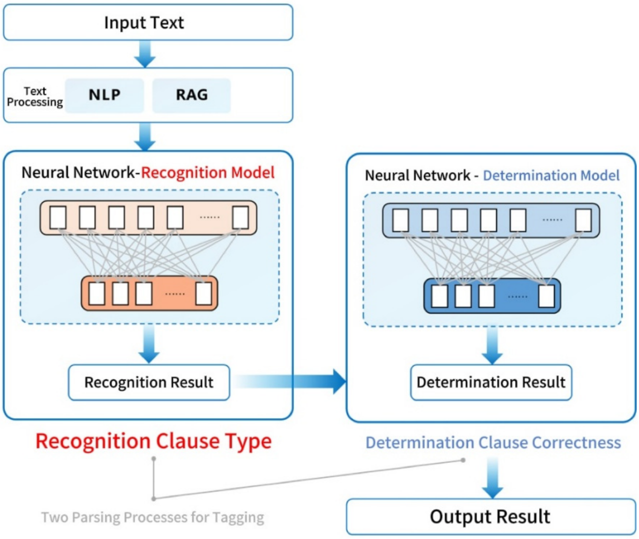
Input Text
Text Processing
NLP
RAG
Neural Network-Recognition Model
Recognition Result
Neural Network - Determination Model
Determination Result
Recognition Clause Type
Determination Clause Correctness
Two Parsing Processes for Tagging
Output Result

Fig. 1. An illustration of legal clause type recognition and correctness determination.

follows:

$$A_{ij} = \begin{cases} \mathrm{Softmax}(Q_i^\top K_j), & j \in [i-w, i+w] \cup \mathcal{G}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where A_{ij} indicates the attention strength of the ith position to the jth, $Q_i \in R^d$ is the query vector positioned at i, $K_j \in R^d$ is the key vector positioned at j, w denotes the radius size of the sliding window, while $\mathcal{G}$ represents the global structure anchor set. The mechanism enables the model to explicitly attend to hierarchical anchor points within the contract while modeling local context, thereby enhancing its ability to perceive cross-term and cross-section structural patterns.

3.3. Sentence segmentation and standardized processing

To ensure that financial contract corpora possess a consistent and controllable sentence-level granularity during the input phase, Con-SEN designed a mixed sentence segmentation mechanism that combines rule-based and learning-driven approaches. The module is divided into three steps: paragraph-level initial segmentation, rule candidate extraction, and boundary learning determination.

First, the system will segment the original contract document D into several natural paragraphs:

$$\mathcal{D} \Rightarrow \{p_1, p_2, \ldots, p_m\}, \tag{4}$$

where p_m denotes the mth paragraph. For each paragraph, the system utilizes rule-based sentence segmenter $\mathcal{R}(\cdot)$ to identify the potential sentence boundary candidate points in each paragraph. Built upon Chinese punctuation marks (such as periods, semicolons, colons) and clause numbering patterns (e.g., “Article X”, “Clause Y”), this rule-based sentence segmenter outputs a set of candidate segmentation position:

$$\mathcal{S}^{(0)} = \mathcal{R}(p_m). \tag{5}$$

Given the issues with financial contracts that include format nesting, abuse of numbering, and incomplete semantics, relying solely on rules could lead to errors in segmentation of sentences. In this regard, Con-SEN introduced a boundary binary classification model $\mathcal{F}_\theta$ fine-tuned with RoBERTa, further judged each candidate boundary position as to whether it constitutes a genuine sentence boundary. The process of modeling proceeds as follows:

$$y_i = \mathcal{F}_\theta(x_{b_i-k:b_i+k}) \in \{0, 1\}, \tag{6}$$

where x_{b_i-k}, x_{b_i+k} represents context fragments centered around candidate boundaries b_i with window size of $2k+1$; model output y_i serves as boundary labels.

If $y_i = 1$, it is judged to be the true end of a sentence, otherwise, it is considered a continuous text.

Upon completing boundary detection and forming sentence sequences, the system assigns unique identifiers to each sentence to support subsequent sentence-level tracking and alignment. Each sentence's identifier key is defined as follows:

$$\mathrm{ID}_j = (\mathrm{DocID}, \mathrm{SectionPath}, j), \tag{7}$$

where DocID denotes the document number of the original contract; while the SectionPath indicates the hierarchical path of the sentence in the contract structure (for instance, "3-1-2" denotes "Chapter 3, Article 1, Paragraph 2"); and j is the intra-sentence number of the same level under the hierarchy.

3.4. Training strategy and convergence mechanism

To ensure model efficiency and stability during long-text corpus training, Con-SEN introduced a series of strategies into its training process, including input sequence trimming, optimizer selection, batch configuration, and training epochs control. Conspicuously, the above mechanism constitutes a complete training configuration system from multiple aspects, including resource consumption counting, gradient convergence speed, and performance saturation control.

Given that financial contract texts commonly tend to be lengthy documents, directly inputting the entire text would significantly increase memory load and lower training efficiency. Thus, in the training phase, Con-SEN introduced a dynamic truncation strategy: a randomly sampled subsequence fragment of appropriate length from the original document was used as input for each round of training. The length of segments $L^{(t)}$ and the original position $s^{(t)}$ are uniformly sampled from the following intervals respectively:

$$L^{(t)} \sim \mathcal{U}(1024, 2048), \tag{8}$$

$$s^{(t)} \sim \mathcal{U}(1, T - L^{(t)}), \tag{9}$$

$$X^{(t)} = [x_{s^{(t)}}, x_{s^{(t)}+1}, \ldots, x_{s^{(t)}+L^{(t)}}], \tag{10}$$

where $\mathcal{U}(a, b)$ denotes the uniform sampling over the closed interval $[a, b]$, with T representing the total length of the contract document, and $X^{(t)}$ indicating the current training segment's token sequence.

In model parameter updating, Con-SEN compared the performance of Adam, AdamW, and Adafactor optimization algorithms in Masked Language Modeling (MLM). Experimental results indicate that AdamW can achieve faster convergence of loss under the same learning rate and maintain a lower perplexity. The updating rules read as follows:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \cdot \theta_t, \tag{11}$$

where θ_t refers to the parameter vector at the tth training round, η represents the learning rate, while $\hat{m}_t$ and $\hat{v}_t$ respectively denote the bias correction terms for first-order and second-order gradient estimates; ϵ is a small constant introduced to avoid zero errors, λ is the L2 regularization weight decay coefficient, which is used to suppress overfitting.

In order to balance computational efficiency with model stability, experiments were conducted to compare models across different batch sizes (Batch Size, BS), with values selected including $\mathrm{BS} \in \{16, 32, 64\}$. The results show that when $\mathrm{BS} = 32$, the model has the most stable performance, specifically reflected in the following aspects: utilization control rate maintained within 90%, balancing resource consumption; loss curve converges smoothly; gradient variance $\mathrm{Var}(\nabla_\theta \mathcal{L})$ is minimized, significantly reducing the oscillation caused by small batch noise.

In terms of model training rounds control, Con-SEN employs the Early Stopping strategy based on performance monitoring. The system records in real-time the changes in the perplexity (PP) covering the language model and in the structure reference accuracy A_{ref}. Experiments revealed that as the training rounds exceeded 2.5, the perplexity curve flattened out, with a less than 0.1% increase in structural reference accuracy, that is $\Delta A_{\mathrm{ref}} < 0.001$, indicating that the model had entered a saturation zone in terms of performance. Therefore, the final number of training rounds was set as 3, while an Early Stopping mechanism was introduced to judge the progress: if no significant performance improvement is observed in two consecutive epochs, the training process would be terminated ahead of schedule to prevent overfitting and resource wastage.

3.5. Building a multi-task assessment system

In order to achieve a comprehensive understanding and risk management of contract sentences, Con-SEN introduced a Multi-Task Inference Head (MTIH) with a parallel structure during its output phase, supporting synchronous reasoning across different semantic dimensions. The system, directing against each sentence that has finished semantic augmentation, executes three independent yet structurally consistent tasks: identification of clause type, legality status assessment, and risk level classification. This design aims to enhance the system's ability to express semantic attributes of sentences and support downstream compliance assessments through multi-label modeling.

In the task of clause type identification, the model must determine the specific category to which the current sentence belongs from among predefined financial clause types $C_{\tau} = 360$. The specific method involves feeding the semantic representation vector $h_i \in R^d$ into a fully connected layer, and then calculating the probability distribution for each class via Softmax. The final category labels are determined by selecting the item with the highest probability, as calculated by the following formula:

$$\tau_i = \arg \max_{c \in \{1,\ldots,C_{\tau}\}} \text{Softmax}(W_{\tau} \cdot h_i + b_{\tau})_c, \tag{12}$$

where $W_{\tau} \cdot h_i$ is classifier weight matrix, and b_{τ} is bias term.

Subsequently, the model performs legitimacy status assessment on sentences to identify any potential risks that may arise from violation of regulatory provisions or contractual constraints. This task is a binary classification problem, in its output labels $\lambda_i \in \{0, 1\}$, “1” indicates “violation” and “0” indicates “legality”. The calculation format shall be in accordance with the above text. The details are as follows:

$$\lambda_i = \arg \max_{y \in \{0, 1\}} \text{Softmax}(W_{\lambda} \cdot h_i + b_{\lambda})_y, \tag{13}$$

where $W_{\lambda} \in R^{2\times d}$ and $b_{\lambda} \in R^2$ are the parameter matrix and biased vector of this task, respectively.

For sentences deemed as violations, the system further categorizes them based on their risk level, distinguishing between their potential severity (high/mid/low). The task was modeled as a three-class classification problem, with output labels $\rho_i \in \{1, 2, 3\}$ corresponding to the “high risk”, “medium risk”, and “low risk” three levels. The calculation method being employed is as follows:

$$\rho_i = \arg \max_{r \in \{1, 2, 3\}} \text{Softmax}(W_{\rho} \cdot h_i + b_{\rho})_r, \tag{14}$$

where $W_{\rho} \in R^{3\times d}$ and $b_{\rho} \in R^3$ are model parameters in this risk evaluation module.

To ensure that the three tasks optimize collaboratively and remain consistent throughout training, Con-SEN introduces a joint loss function, which weighs the cross-entropy losses of the three branches tasks. The total loss function is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\tau} + \mathcal{L}_{\lambda} + \alpha \cdot \mathcal{L}_{\rho}. \tag{15}$$

Specifically, $\mathcal{L}_{\tau}, \mathcal{L}_{\lambda}, \mathcal{L}_{\rho}$ represent the loss functions for the three tasks, and α serves as a weight factor, which is used to adjust the contribution degree of each risk level task to the total loss. This design ensures that the model not only learns each task individually during optimization but also effectively prevents overfitting toward a particular subtask, thus enhancing the balance and overall effectiveness of judgment. In the end, the system outputs three-dimensional semantic label triplets for each sentence as the basis units for subsequent structural refinement and compliance risk assessment.

In addition to the architectural design of the multi-task framework, Con-SEN also embeds domain-level semantic attention to reflect behavioral and risk-related characteristics inherent in financial contracts. Specifically, the clause type identification task captures not only the functional role of each sentence (e.g., payment, guarantee, or termination), but also provides contextual cues for downstream assessments of legality and risk. The legitimacy assessment task goes beyond surface-level linguistic patterns by targeting clauses that may indicate improper behaviors — such as ambiguous obligations, imbalanced rights and duties, or potential violations of regulatory provisions. Finally, the risk level classification task stratifies those potentially noncompliant clauses based on the severity of their legal or financial consequences, such as the likelihood of default, litigation exposure, or reputational impact. Through the synergistic interaction among these tasks, the model constructs a tri-dimensional semantic understanding of each sentence — linking clause function, behavioral compliance, and risk intensity — to support refined contract interpretation and compliance risk analysis.

4. Data

4.1. Data sources

To support the model training for sentence-level review tasks in financial contracts, this paper constructs the Chinese financial corpus Con-SEN-Corpus, which includes two core sources: original financial contracts and regulatory normative texts. The corpus includes over 84,000 desensitized original financial contracts, spanning from January 2010 to December 2024, covering more than thirty financial sub-industries, including commercial banking, insurance, trust, securities, and asset management. Additionally, regulations issued by the China Banking and Insurance Regulatory Commission, the China Securities Regulatory Commission, and exchanges, as well as internal compliance manuals and risk lists from various institutions, were collected to provide an institutional basis for legitimacy labeling. After sentence-level segmentation and preliminary labeling, the corpus generated over 84,000 ten thousand sentences, covering a wide range of industries with balanced label categories, effectively supporting subsequent model training and evaluation. Detailed statistical information on the proportion of specific contract categories and the distribution of risk annotations can be found in the charts. To ensure annotation consistency across a large volume of data, we developed detailed labeling guidelines aligned with regulatory texts. Annotators underwent training sessions and the inter-annotator agreement was periodically measured to maintain high consistency. Disagreements were resolved through consensus or expert adjudication.

4.2. Date cleaning

This study designed and implemented a four-level preprocessing pipeline for the structured cleaning and sentence-level extraction of original financial contract images, in order to construct a high-quality corpus. First, a TrOCR-based text recognition engine was used to transcribe characters from PDF scans. Through multiple rounds of experimentation, we set the character confidence threshold at 0.98, which allowed us to control the OCR residual error rate to below 0.15% on the validation set, while avoiding excessive discarding of valid characters and ensuring maximum retention of original semantics. The second step is to perform structural standardization on the transcribed text. This includes: unifying the format of clause-level numbering, removing noise contents such as header watermarks and annotations, and operating linearization on the embedded tables in the contract and transforming these tables into row-column text that conforms to the internal JSON Schema, thereby ensuring the hierarchical consistency of clause semantics and machine readability. The third step is to implement a dual deduplication mechanism at the document and sentence levels. At the document level, the 64-bit SimHash algorithm is used, with a similarity threshold set at 0.92 (determined through grid search), which eliminates over 96% of highly templated contracts while ensuring 98% semantic coverage; at the sentence level, sentences shorter than 5 characters or longer than 120 characters are filtered out to balance information completeness and linguistic granularity. Finally, automatic sentence segmentation is performed. We combine a rule-driven sentence segmenter with a fine-tuned RoBERTa-large binary classification model to accurately assess candidate boundaries. On 5,000 manually reviewed samples, the $F1$ score for sentence boundary recognition reached 0.982, demonstrating high accuracy. After completing the above processing, the corpus retained a total of 118,406 contract texts, with a total of 2,106,589 extracted sentences and an average sentence length of 48 words, including 12.4% negation structures and 7.9% explicit cross-reference structures. All texts generated SHA-256 fingerprints and were written into a versioned storage system; simultaneously, they were divided into training, validation, and test sets in an 8:1:1 ratio.

4.3. Unified composite labeling

To address the practical demands of financial contract review, which include the structured understanding of sentence semantics, legality assessment, and risk categorization, Con-SEN treats each sentence as a fundamental processing unit and jointly models three core tasks: clause type classification, legitimacy determination, and risk level prediction. Considering the structural complexity, logical nesting, and semantic sparsity inherent in financial texts, the model is designed not only to achieve precise sentence-level classification but also to produce a structured, semantically coherent, and compliance-oriented output that facilitates downstream tasks such as automated auditing, responsibility attribution, and regulatory verification.

Concretely, based on the sentence representation h_i, Con-SEN produces for each sentence a ternary labels consisting of the clause type τ_i, the legitimacy status λ_i, and the risk level ρ_i, jointly forming the semantic-compliance output unit:

$$y_i = (\tau_i, \lambda_i, \rho_i), \quad i = 1, 2, \dots, N, \tag{16}$$

where τ_i encodes the semantic role of the sentence within the contractual structure, λ_i indicates whether the sentence violates regulatory or contractual rules, and ρ_i quantifies the severity of such risk. This ternary labels not only provide actionable classification results, but also enables interpretable structural reconstruction and semantic tracing at the document level, aligning with the paper's core vision of structured modeling, hierarchical semantic mapping, and risk-aware output. Accordingly, the complete model output is formally defined as

$$Y = \{y_1, y_2, \dots, y_N\} = \{(\tau_1, \lambda_1, \rho_1), \dots, (\tau_N, \lambda_N, \rho_N)\}. \tag{17}$$

This output formulation offers a unified representation framework for constructing sentence-level semantic graphs of contracts, facilitating the standardization of automated compliance workflows. It thereby reflects both the practical applicability and theoretical significance of the proposed model.

During the construction of the Con-SEN-Corpus, particular attention was paid to embedding behavioral compliance signals into the annotation schema. Rather than relying solely on syntactic or surface-level textual features, annotators incorporated regulatory guidelines, industry standards, and practical criteria for assessing contractual conduct. Sentences that demonstrated well-defined obligations, balanced responsibilities, and regulatory alignment were labeled as "legal" with low risk. In contrast, clauses exhibiting ambiguous liability, unilateral exemptions, or potentially deceptive phrasing were labeled as "illegal" and assigned moderate to high risk levels, depending on the severity of potential consequences.

Within the unified composite labeling framework, the legitimacy dimension λ explicitly captures behavioral compliance, while the risk level ρ reflects the extent of deviation or the potential legal and financial impact. The clause type τ provides structural context that situates these judgments within the broader contractual function. This behavior-aware annotation strategy ensures that Con-SEN's output triples (τ, λ, ρ) are not merely classification artifacts, but semantically rich and actionable representations that support downstream tasks such as liability attribution, risk monitoring, and automated regulatory auditing.

4.4. Evaluation metrics

To systematically evaluate the effectiveness of the Con-SEN approach in the task of sentence-level review of financial contracts, this study constructs a unified evaluation framework based on a representative and structurally diverse sample set. The evaluation centers on three core metrics: Document-level Accuracy (Doc-Acc), Sentence-level Coverage (Sent-Cov), and Sentence-level Accuracy (Sent-Acc), which together provide a comprehensive and multidimensional assessment of the model's performance.

The rationale behind the design of these three metrics lies in their ability to capture complementary aspects of model competence: global decision consistency, semantic activation capability, and fine-grained classification precision.

Specifically, Doc-Acc measures the model's ability to generate accurate and complete structured outputs at the document level. In real-world compliance workflows, contract review systems must deliver holistic, document-level assessments rather than isolated sentence labels. Thus, this metric reflects the model's capacity to transform localized predictions into coherent and reliable contract-level judgments, aligning with practical deployment requirements.

Sent-Cov, by contrast, evaluates whether the model successfully identifies and activates the semantically relevant sentences within a contract, irrespective of whether those sentences are classified correctly. This metric focuses on the model's semantic coverage and its ability to attend to regulatory hotspots or domain-critical clauses. It is particularly important for evaluating performance on documents with uneven content density or loosely structured narratives, where critical information may be sparsely distributed.

Lastly, Sent-Acc captures the sentence-level classification accuracy with respect to clause type, legality, and risk level. This metric reflects the model's fine-grained semantic understanding and its capacity to produce reliable predictions at the linguistic unit level. It serves as a foundational indicator of representation quality and label discrimination.

Considering that current generalized LLMs have demonstrated excellent performance in generic semantic understanding tasks and have received widespread attention in preliminary legal review applications, two representative types of generalized models (ChatGPT and Claude) are selected as baseline controls in this paper. However, since generic models often suffer from omission, illusion generation or logical misinterpretation in financial contract scenarios due to the lack of corpus adaptation and structural modeling capabilities in the financial domain, it is difficult to meet the prudent requirements of high-risk text processing.

Therefore, by introducing mainstream LLMs as the baseline for comparison, we aim to verify whether Con-SEN achieves substantial improvement in overall accuracy, information coverage and risk identification stability under the design of structural optimization and semantic enhancement for the financial contract context. In order to ensure the comparability and explanatory power of the evaluation results, this paper uniformly adopts the above three core indicators for experimental evaluation.

Accuracy of review results (Doc-Acc):

$$\text{Doc-Acc} = \frac{\text{Number of Correctly Audited Contracts}}{\text{Total Number of Contracts}}. \tag{18}$$

Statement Coverage (Sent-Cov):

$$\text{Sent-Cov} = \frac{\sum_j |S_j^{\text{hit}}|}{\sum_j |S_j^{\text{gold}}|}. \tag{19}$$

Statement accuracy (Sent-Acc):

$$\text{Sent-Acc} = \frac{\sum_j |S_j^{\text{hit}}|}{\sum_j |S_j^{\text{pred}}|}. \tag{20}$$

Built on the standardized processed financial contract dataset, this paper constructs a unified evaluation process, which includes the following steps: first, take the contract number (doc_id) as the unit to determine whether the overall review of each contract is completely correct, and accordingly calculate the contract-level accuracy index Doc-Acc; second, take the sentence number (sentence_id) as the unit to count the coverage and the prediction accuracy of each model at the sentence level, and then calculate the statement coverage (Sent-Cov) and the statement accuracy (Sent-Acc). The above calculations are all done automatically based on the unified evaluation script, which ensures the consistency of the evaluation process and the reproducibility of the experimental results.

Importantly, the Sentence-level Accuracy (Sent-Acc) metric inherently captures the model's capability to classify not only clause types and legitimacy status, but also the risk level dimension (ρ), which serves as a semantic proxy for potential contractual risk behaviors. In the context of the ternary labeling framework, risk level prediction reflects the model's ability to identify and differentiate clauses that exhibit varying degrees of risk exposure — ranging from regulatory violations to ambiguous liability or asymmetric obligations.

This evaluative dimension is directly aligned with the multi-task learning strategy outlined in Sec. 3.5, as well as the behavior-oriented composite labeling schema discussed in Sec. 4.3. Together, they constitute a coherent semantic structure through which the model internalizes and expresses contract-level compliance sensitivity. By incorporating risk-level classification into the metric design, the evaluation framework extends beyond surface-level predictive accuracy to also assess the model's capacity for risk-aware semantic representation and actionable clause-level judgment.

Accordingly, the metrics proposed in this section not only quantify performance but also serve as an empirical lens for examining the model's ability to extract, interpret, and structure risk-related features embedded in financial contractual language.

5. Results

5.1. Overall accuracy analysis

This round of experiments is based on the complete contract sample set and standardized sentence-level segmentation, focusing on three core metrics: Document Accuracy (Doc-Acc), Sentence Coverage (Sent-Cov), and Sentence Accuracy (Sent-Acc). The experimental metrics aim to systematically evaluate the comprehensive capabilities of the proposed model across practical financial contract review scenarios, with results shown in Table 1.

In the overall contract review task, Con-SEN achieved an accuracy of 86.87%, surpassing ChatGPT and Claude by 18.49 and 23.23 percentage points, respectively, with relative improvements of 27.05% and 36.49%. This demonstrates that Con-SEN is more effective at capturing contract-level risk assessment logic, accurately identifying compliance clauses and potential risk factors. In contrast, general-purpose language models often struggle with long, structurally complex, and densely cross-referenced financial contracts, leading to frequent misjudgments. By incorporating clause-level encoding and negation polarity recognition,

Table 1. Overall accuracy analysis.

	Doc-Acc	Sent-Cov	Sent-Acc
ChatGPT	68.38%	74.02%	89.65%
Claude	63.64%	69.64%	91.52%
Con-SEN	86.87%	92.02%	94.37%

Con-SEN establishes clearer and more stable decision boundaries, significantly enhancing overall accuracy.

For sentence-level coverage, Con-SEN reached 92.02%, outperforming ChatGPT and Claude by 18.00 and 22.38 percentage points, respectively, corresponding to relative improvements of 24.30% and 32.13%. Higher coverage indicates that Con-SEN can identify a greater proportion of valid sentences within contracts, thereby reducing the risk of omissions. This advantage stems from its strict sentence boundary segmentation during preprocessing and the integration of a long-text window mechanism during encoding, enabling consistent performance even on ultra-long contract documents. In contrast, general models often skip or truncate segments once the input exceeds their context window, resulting in lower coverage rates.

In terms of sentence classification accuracy, Con-SEN achieved 94.37%, outperforming ChatGPT and Claude by 4.72 and 2.82 percentage points, respectively. While covering more sentence units, Con-SEN also provides more precise classifications of clause type, legality, and risk level. Its multi-task consistency modeling mitigates semantic conflicts and label drift, while the use of negation polarity markers helps correct errors caused by semantic inversion, leading to a substantial improvement in sentence-level classification accuracy.

To ensure the reliability and fairness of the comparative analysis across contract types, we partitioned the dataset into training (80%), validation (10%), and test (10%) subsets. A stratified sampling strategy was employed based on the contract-type labels, so as to preserve proportional representation of the three categories — standardized, structurally complex, and risk-sensitive — in each subset. In the final test set, each contract type accounts for approximately one-third of the samples, allowing for balanced and meaningful performance comparisons. The full dataset comprises 1,216 contracts and 96,528 annotated sentences, with the distribution of clause types, legality labels, and risk levels consistently maintained across the training and evaluation splits.

5.2. Overall fluctuation analysis

To further evaluate the stability performance of each model in the financial contract sentence-level review task, this experiment calculated the sample variance (VAR) of three evaluation metrics (review result accuracy, sentence coverage rate, sentence accuracy rate) based on a complete contract sample set, with results shown in Table 2.

In terms of sample variance in contract review accuracy, Con-SEN achieved a low variance of 0.00105%, substantially outperforming ChatGPT (0.00937%) and Claude (0.00459%) with reductions of 88.80% and 77.13%, respectively.

Table 2. Overall volatility analysis.

	Doc-Acc volatilities	Sent-Cov volatilities	Sent-Acc volatilities
ChatGPT	0.94%	1.55%	0.69%
Claude	0.46%	0.43%	0.49%
Con-SEN	0.10%	0.06%	0.03%

This indicates that Con-SEN delivers significantly more stable performance across diverse contract samples. Its hierarchical encoding and contextual enhancement mechanisms help mitigate reasoning uncertainty caused by structural variations, avoiding the performance fluctuations commonly observed in general-purpose models.

For sentence-level coverage stability, Con-SEN reported a variance of only 0.00056%, compared to 0.01553% for ChatGPT and 0.00426% for Claude — representing reductions of 96.39% and 86.86%, respectively. This high level of consistency can be attributed to the unified sentence boundary segmentation during preprocessing and the incorporation of long-range cross-chapter dependency modeling during encoding, enabling robust performance across documents of varying length and structural complexity.

Regarding the variance in sentence-level classification accuracy and error rates, Con-SEN again achieved a minimal variance of 0.00031%, versus 0.00687% for ChatGPT and 0.00494% for Claude — reflecting a fluctuation reduction of over 90%. The model's multi-task consistency framework and negation polarity recognition module enhance its ability to adapt to fine-grained semantic shifts, ensuring robust and reliable classification even under complex contextual conditions.

5.3. Accuracy analysis of contracts by type

After confirming the performance advantages of the Con-SEN method on the overall contract sample set, this paper further reveals the specific differences in model adaptability across different contract scenarios by categorizing all contract samples into three major categories based on the structural characteristics and semantic complexity of the contract texts: standardized structural contracts, high-complexity structural contracts, and risk-sensitive contracts. The classification criteria are mainly based on the following considerations.

First, standardized structured contracts (such as guarantee contracts and loan contracts) typically have a unified template, fixed clause logic, and minimal cross-chapter dependencies, making them suitable for testing the model's basic inductive ability in standardized scenarios. Second, high-complexity structured contracts (such as fund supervision agreements and trust contracts) generally feature long text nesting, conditional layering, and dynamic responsibility transfer, which are suitable for examining the model's ability to capture long-range dependencies and complex logical reasoning. Finally, risk-sensitive contracts (such as asset transaction contracts and mortgage contracts) are characterized by implicit risk obligations and dense negative expressions, emphasizing the model's capability in fine-grained risk identification and polarity understanding.

Standardized structured contracts usually adopt a unified template, with standardized clause arrangement and clear boundaries of responsibilities and obligations, commonly found in traditional credit and financing fields. The characteristics of this type of text include low semantic redundancy, clear hierarchical divisions, and a format that leans toward consistency, which poses basic yet comprehensive requirements for sentence-level positioning and category discrimination in models.

In such contracts, Con-SEN has achieved overall stable performance advantages compared to general large models, with an overall review accuracy improvement generally exceeding 20 percentage points, and significant enhancements in sentence coverage and sentence accuracy. Its main advantage comes from the deep modeling of hierarchical information in contract clauses and explicit encoding of negation polarity. By structurally injecting hierarchical information of contract chapters, sections, clauses, and items, Con-SEN can accurately locate the semantic position of sentences within the contract topology, thereby reducing the classification drift issues encountered by general models in template variants or clause reorganization scenarios. At the same time, by clearly marking reversal trigger words, Con-SEN effectively suppresses compliance misjudgments caused by negation expressions, ensuring highly consistent output of the model in standard scenarios. The results are shown in Table 3.

High-complexity structured contracts are widely present in scenarios such as fund supervision, trust arrangements, and inter-agency cooperation. Such contract texts generally exhibit characteristics such as extremely long lengths, frequent cross-chapter references, deep nesting levels, and dense conditional constraints, which impose higher requirements for contextual modeling and logical reasoning capabilities on the model.

In this category, Con-SEN demonstrates a more significant improvement in adaptability, with a general increase in review accuracy of about 23–25 percentage points compared to general large models, and notable optimization in both statement coverage and statement accuracy. Unlike standardized contract scenarios, the challenge posed by high-complexity contracts to the model mainly lies in maintaining long-range dependencies and understanding dynamic conditions.

Table 3. Standardized contract accuracy analysis.

Standardized structural contracts	Model	Doc-Acc (%)	Sent-Cov (%)	Sent-Acc (%)	Percentage increase (%)
Guarantee contract	Con-SEN	85.02%	90.46%	93.74%	/
	ChatGPT	60.04%	69.80%	85.60%	24.98%
	Claude	60.18%	67.92%	88.17%	24.84%
Equity-type contracts	Con-SEN	86.55%	92.04%	94.03%	/
	ChatGPT	80.40%	85.26%	94.42%	6.15%
	Claude	78.32%	85.24%	92.13%	8.23%
Channel co-operation agreement for finance	Con-SEN	87.45%	92.47%	94.57%	/
	ChatGPT	80.09%	86.42%	92.83%	7.36%
	Claude	62.46%	66.66%	94.04%	24.99%
Factoring contract	Con-SEN	86.18%	91.40%	94.29%	/
	ChatGPT	61.85%	68.66%	90.03%	24.33%
	Claude	60.20%	68.34%	88.20%	25.98%
Loan contract	Con-SEN	87.21%	91.97%	94.82%	/
	ChatGPT	79.24%	85.23%	93.18%	7.97%
	Claude	79.44%	84.15%	94.60%	7.77%
Deposit agreement	Con-SEN	86.83%	92.16%	94.22%	/
	ChatGPT	59.05%	68.22%	86.72%	27.78%
	Claude	60.31%	69.02%	87.54%	26.52%
Insurance agency delegation agreement	Con-SEN	87.06%	91.96%	94.67%	/
	ChatGPT	79.60%	85.02%	93.77%	7.46%
	Claude	61.23%	65.99%	93.05%	25.83%
Delegated loan borrowing rollover agreement	Con-SEN	87.06%	92.50%	94.12%	/
	ChatGPT	62.45%	67.09%	93.40%	24.61%
	Claude	62.84%	65.89%	95.59%	24.22%

Con-SEN introduces a long-text encoding strategy using a sliding window and sparse global attention, allowing it to effectively retain semantic connections between distant clauses even when the document length exceeds the conventional processing limits of Transformers, thus avoiding local overfitting and logical short-circuiting in long-text environments typical of general models. At the same time, the preprocessing strategy of table linearization and dynamic numbering normalization significantly reduced structural ambiguity issues caused by table nesting and nonlinear layouts. The results are shown in Table 4.

Risk-sensitive contracts are commonly found in areas such as asset transactions, financing guarantees, and mortgage arrangements. Such texts often contain a large number of implicit obligation clauses, negative constraint expressions, liability

Table 4. Accuracy analysis of high-complexity contracts.

High-complexity structured contracts	Model	Doc-Acc (%)	Sent-Cov (%)	Sent-Acc (%)	Percentage increase (%)
Bank settlement account Management agreement	Con-SEN	87.20%	92.40%	94.37%	/
	ChatGPT	63.25%	71.05%	89.15%	23.95%
	Claude	64.18%	70.94%	90.65%	23.02%
Funds monitoring agreement	Con-SEN	87.07%	92.06%	94.57%	/
	ChatGPT	63.29%	69.12%	91.60%	23.78%
	Claude	63.25%	69.13%	91.60%	23.82%
Tripartite supervision Agreement for proceeds	Con-SEN	87.04%	92.45%	94.15%	/
	ChatGPT	62.23%	72.91%	85.54%	24.81%
	Claude	62.23%	72.91%	85.54%	24.81%
Agreement for supervision of pre-sale funds of commercial properties	Con-SEN	87.90%	92.41%	95.12%	/
	ChatGPT	63.56%	69.57%	91.38%	24.34%
	Claude	62.59%	71.61%	87.50%	25.31%
Agreement for supervision of second-hand property funds	Con-SEN	87.29%	92.44%	94.42%	/
	ChatGPT	61.72%	69.87%	88.42%	25.57%
	Claude	62.04%	67.26%	92.34%	25.25%
Bond tripartite supervisory agreement	Con-SEN	85.77%	91.31%	93.93%	/
	ChatGPT	61.61%	72.09%	85.60%	24.16%
	Claude	61.57%	70.28%	87.73%	24.20%
Cooperation agreement between a bank and a guarantee company	Con-SEN	87.06%	91.93%	94.70%	/
	ChatGPT	80.44%	85.15%	94.60%	6.62%
	Claude	78.82%	84.18%	93.76%	8.24%
Custody agreement for trust scheme funds	Con-SEN	87.13%	91.96%	94.73%	/
	ChatGPT	80.09%	83.80%	95.70%	7.04%
	Claude	78.23%	83.78%	93.51%	8.90%

exemption clauses, and risk transfer mechanisms. Their significant characteristics include dense negative statements, ambiguous liability attribution, and clause interpretation that relies heavily on contextual understanding.

In these highly sensitive scenarios, the advantages of Con-SEN are particularly prominent. The overall review accuracy has improved by more than 22 percentage points compared to general large models, and the accuracy of sentences is also significantly ahead. Con-SEN's superior performance in this scenario stems from its specialized optimization for negative polarity, conditional reversal, and ambiguous sentence structures.

By systematically injecting negative polarity markers during the pre-training and fine-tuning phases, and supplementing with small sample supervision targeting negative obligations and exemptions during the fine-tuning phase, Con-SEN can more accurately distinguish the positive and negative directions of rights and obligations in sentences, avoiding semantic direction reversal errors that general models encounter with complex expressions such as "unless… shall not…". At the same time, in asset transaction contracts involving dynamic obligation transfers among multiple parties, Con-SEN can perform fine-grained inference on the division of responsibilities among multiple parties by integrating the contextual hierarchy of the contract, significantly surpassing general large models in fine-grained risk identification metrics. The results are shown in Table 5.

To ensure the reliability and fairness of the comparative analysis across contract types, we partitioned the dataset into training (80%), validation (10%), and test (10%) subsets. A stratified sampling strategy was employed based on the contract-type labels, so as to preserve proportional representation of the three categories — standardized, structurally complex, and risk-sensitive — in each subset. In the final test set, each contract type accounts for approximately one-third of the samples, allowing for balanced and meaningful performance comparisons. The full dataset comprises 1,216 contracts and 96,528 annotated sentences, with the distribution of clause types, legality labels, and risk levels consistently maintained across the training and evaluation splits.

Table 5. Accuracy analysis of risk-sensitive contracts.

Risk-sensitive contracts	Model	Doc-Acc (%)	Sent-Cov (%)	Sent-Acc (%)	Percentage increase (%)
Asset trading contract	Con-SEN	87.83%	92.59%	94.85%	/
	ChatGPT	65.40%	73.05%	89.62%	22.43%
	Claude	63.56%	73.24%	87.01%	24.27%
Mortgage contract	Con-SEN	86.51%	92.10%	93.93%	/
	ChatGPT	79.12%	85.52%	92.63%	7.39%
	Claude	79.61%	84.86%	93.93%	6.90%
Asset trading contract	Con-SEN	87.83%	92.59%	94.85%	/
	ChatGPT	65.40%	73.05%	89.62%	22.43%
	Claude	63.56%	73.24%	87.01%	24.27%

5.4. Volatility analysis of contracts by type

Based on the overall performance evaluation, to further verify the stability of the model in the standardized financial contract review scenario, this section systematically compares the output consistency performance of Con-SEN with that of general LLMs (ChatGPT, Claude) around the categories of standardized structured contracts, based on three indicators: the volatility of review result accuracy, the volatility of statement coverage, and the volatility of statement accuracy.

In standardized structured contracts, Con-SEN maintains a review accuracy volatility within 0.0008, while ChatGPT and Claude fluctuate around 0.0010, with some subcategories exceeding 0.002. This indicates that general large models still suffer from prediction inconsistency under standardized formats, where minor expression changes can shift overall judgments. Con-SEN, by leveraging clause-level encoding and semantic enhancement, significantly reduces output fluctuation through a stronger grasp of contract structure and function.

For statement coverage volatility, Con-SEN remains within 0.003, compared to 0.006–0.008 for ChatGPT and Claude. A similar pattern is observed in statement accuracy, where Con-SEN consistently achieves low variance across contract subtypes. These metrics reflect the model's ability to maintain stable sentence identification across varying formats. General models tend to introduce omissions or redundancies when handling line breaks, tables, or irregular numbering, increasing fluctuation. In contrast, Con-SEN's preprocessing — featuring unified numbering, table linearization, and positional prompts — helps suppress layout-related noise and enhances sentence stability.

Finally, the low fluctuation in statement accuracy highlights Con-SEN's robustness in fine-grained clause classification. Its multi-task decision mechanism jointly optimizes clause type, legality, and risk level, reducing local drift caused by single-task outputs. Results are summarized in Table 6.

In high-complexity contract categories, Con-SEN maintains the overall volatility of review result accuracy within 0.0002, while ChatGPT and Claude generally exceed 0.001, with some subcategories surpassing 0.0029. Especially in typical ultra-long text scenarios such as bank account management agreements and fund supervision agreements, the accuracy volatility of Con-SEN is far lower than that of general models, with a difference exceeding an order of magnitude of 10. When faced with complex texts characterized by cross-chapter logical jumps, nested tables, and dense cross-references, the reasoning chain of general models is prone to interruption or local drift, leading to a significant decrease in output consistency. In contrast, Con-SEN effectively maintains the continuity of long-range dependency paths through a sliding window and sparse global

Table 6. Analysis of volatility of standardized contracts.

Standardized structural contracts	Model	Doc-Acc volatilities	Sent-Cov volatilities	Sent-Acc volatilities
Guarantee contract	Con-SEN	0.81%	0.37%	0.20%
	ChatGPT	0.99%	0.60%	0.77%
	Claude	1.00%	0.58%	0.86%
Equity-type contracts	Con-SEN	0.04%	0.02%	0.01%
	ChatGPT	0.10%	0.20%	0.38%
	Claude	0.16%	0.19%	0.59%
Channel co-operation agreement for finance	Con-SEN	0.03%	0.02%	0.01%
	ChatGPT	0.08%	0.20%	0.48%
	Claude	0.20%	0.18%	0.36%
Factoring contract	Con-SEN	0.03%	0.02%	0.01%
	ChatGPT	0.22%	0.11%	0.20%
	Claude	0.18%	0.18%	0.34%
Loan contract	Con-SEN	0.04%	0.02%	0.01%
	ChatGPT	0.13%	0.25%	0.30%
	Claude	0.12%	0.21%	0.28%
Deposit agreement	Con-SEN	0.04%	0.03%	0.01%
	ChatGPT	0.12%	0.14%	0.33%
	Claude	0.13%	0.16%	0.34%
Insurance agency delegation agreement	Con-SEN	0.04%	0.02%	0.01%
	ChatGPT	0.19%	0.14%	0.64%
	Claude	0.14%	0.18%	0.42%
Delegated loan borrowing rollover agreement	Con-SEN	0.02%	0.02%	0.02%
	ChatGPT	0.04%	3.96%	1.02%
	Claude	0.16%	0.17%	0.42%

attention mechanism, significantly suppressing reasoning fluctuations caused by local information loss, demonstrating a strong ability to maintain cross-text consistency.

In terms of sentence coverage fluctuation metrics, Con-SEN consistently remains between 0.0002 and 0.0005, while both ChatGPT and Claude fall within the range of 0.002–0.005, with the highest even approaching 0.0057. In various contract instances, Con-SEN can reliably identify and cover effective sentences of contract clauses, avoiding the phenomenon where general models miss key information in long texts or heterogeneous layouts. In the context of complex contract structures, diverse responsible parties, and highly coupled clause logic, general large models struggle to stably identify and classify clauses due to a lack of structural awareness; however, Con-SEN significantly enhances the model's semantic continuity perception ability under complex structures by introducing chapter-section-clause-item hierarchical encoding and conditional coherence constraints during the training phase, thereby maintaining high consistency in sentence-level judgment outputs. The results are shown in Table 7.

In the risk-sensitive sample, Con-SEN maintained a fluctuation range of accuracy in review results between 0.0003 and 0.0006, while ChatGPT and Claude were generally above 0.0015, with some samples exceeding 0.0020. In financial texts characterized by dense negation and complex liability allocation, general LLMs struggle to consistently capture the subtle differences between risk obligations and exemption clauses, leading to significant fluctuations in overall review results as contract samples vary. In contrast, Con-SEN effectively enhanced its stable judgment capability regarding the direction of clause responsibilities and risk boundaries through a negation polarity recovery mechanism and multi-task consistency optimization, significantly reducing the inter-sample fluctuation of overall accuracy output.

Table 7. Analysis of volatility of contracts for high-complexity structures.

		Doc-Acc	Sent-Cov
High-complexity structured contracts	Model	volatilities	volatilities
Bank settlement account management agreement	Con-SEN	0.02%	0.02%
	ChatGPT	0.16%	0.12%
	Claude	0.20%	0.21%
Funds monitoring agreement	Con-SEN	0.06%	0.05%
	ChatGPT	0.08%	0.10%
	Claude	0.13%	0.05%
Tripartite supervision agreement for proceeds	Con-SEN	0.05%	0.03%
	ChatGPT	0.36%	0.57%
	Claude	0.06%	0.26%
Agreement for supervision of pre-sale funds of commercial properties	Con-SEN	0.02%	0.02%
	ChatGPT	0.13%	0.08%
	Claude	0.20%	0.14%
Agreement for supervision of second-hand property funds	Con-SEN	0.04%	0.02%
	ChatGPT	0.22%	0.16%
	Claude	0.21%	0.11%
Bond tripartite supervisory agreement	Con-SEN	0.03%	0.02%
	ChatGPT	0.29%	0.96%
	Claude	0.13%	0.12%
Cooperation agreement between a bank and a guarantee company	Con-SEN	0.04%	0.02%
	ChatGPT	0.09%	0.12%
	Claude	0.09%	0.15%
Sample custody agreement for trust scheme funds	Con-SEN	0.07%	0.03%
	ChatGPT	0.09%	0.17%
	Claude	0.14%	0.13%

The sentence variations arising from changes in the responsible parties and clause references in risk-sensitive contract texts pose challenges to the sentence extraction capabilities of general LLMs, resulting in significant fluctuations in coverage. In contrast, Con-SEN significantly reduced the probability of sentence-level recognition fluctuations due to text changes during the preprocessing stage by enhancing sentence segmentation strategies and hierarchical positioning prompts, ensuring a high level of stable coverage performance. The results are shown in Fig. 2.

Based on the previous performance evaluation of contract types and text complexity, this section further introduces the number of contract statements as a stratification criterion to systematically examine the stability and adaptability of Con-SEN from the perspective of text scale. As the number of sentences in the contract text increases, the challenges for the model in terms of contextual continuity understanding, cross-sentence reasoning, and maintaining logical consistency will also intensify. The samples are divided into three categories based on the number of sentences: short text, medium text, and long text, with proportions of (30%, 40%, 30%), respectively assessing the changes in accuracy of review results, statement coverage, statement accuracy, and statement error rate for each model at different scales.

In the short text contract subset, Con-SEN achieved a review result accuracy of 85.93%, with a statement coverage of 91.51%, a statement accuracy of 93.84%,

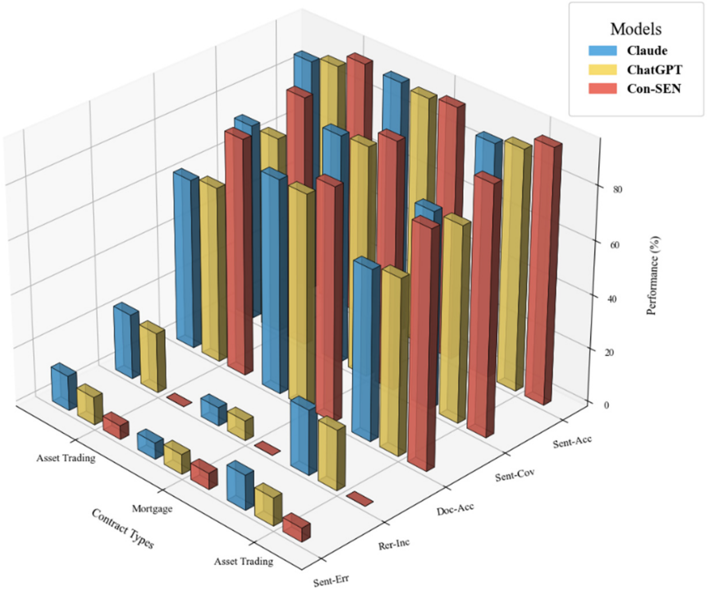

Fig. 2. Adaptive analysis of contract sentence scale.

and a statement error rate controlled at a low level of 6.16%. In contrast, ChatGPT's review accuracy was only 61.42%, and Claude's was 60.73%, with both showing significantly lower statement coverage and accuracy, and error rates reaching 14.42% and 9.77%, respectively. In contract texts with a limited number of statements and a small context span, general LLMs still face issues with insufficient understanding of fine-grained clause features and ambiguous risk classification boundaries. In contrast, Con-SEN, through clause-level encoding and a negation polarity prompting mechanism, can accurately identify clause attribution and compliance status in short texts, significantly improving the overall consistency and accuracy of review and sentence classification.

As the text size increases to a medium scale, the accuracy of Con-SEN's review results rises to 86.93%, with a statement coverage rate of 92.06%, a further increase in statement accuracy to 94.44%, and a reduction in statement error rate to 5.56%. Meanwhile, ChatGPT's accuracy slightly improves to 70.69%, but Claude's accuracy is only 63.77%, and both still exhibit significant disadvantages in coverage and accuracy. As the text size expands, logical jumps and cross-clause references in contracts become more frequent, placing higher demands on the model's continuous reasoning capabilities. General LLMs have begun to exhibit issues of information omission and broken reasoning chains in medium-sized texts due to a lack of structural awareness and dynamic context modeling capabilities, leading to increased performance fluctuations. In contrast, Con-SEN introduces long text mask training during the pre-training phase, and combines a sliding window with a global anchor mechanism, allowing it to maintain complete semantic chain capture and clause decision boundaries in medium text environments, thereby stably supporting improvements across various metrics.

In the subset of long text contracts, Con-SEN further demonstrates exceptional scalability and adaptability. Its review accuracy reaches 87.68%, with a statement coverage rate of 92.48%, and statement accuracy climbing to 94.81%, while reducing the statement error rate to just 5.12%. In comparison, ChatGPT's accuracy drops to 72.62%, and Claude's to 66.23%, with both coverage and accuracy declining simultaneously, and error rates rising to 8.20% and 7.05%, respectively. Long text contracts contain numerous cross-chapter references, negation trigger conditions, and complex responsibility attribution chains. General LLMs experience a significant decline in reasoning continuity due to limitations such as context window constraints and unstable polarity reversal understanding, resulting in a sharp degradation of overall performance as text length increases. Con-SEN utilizes a sparse attention mechanism combined with global anchor point localization, enabling stable modeling of cross-sentence dependencies and semantic coherence in ultra-long texts. Additionally, through multi-task consistency training, it effectively constrains the drift in risk classification, ensuring that even in ultra-long text

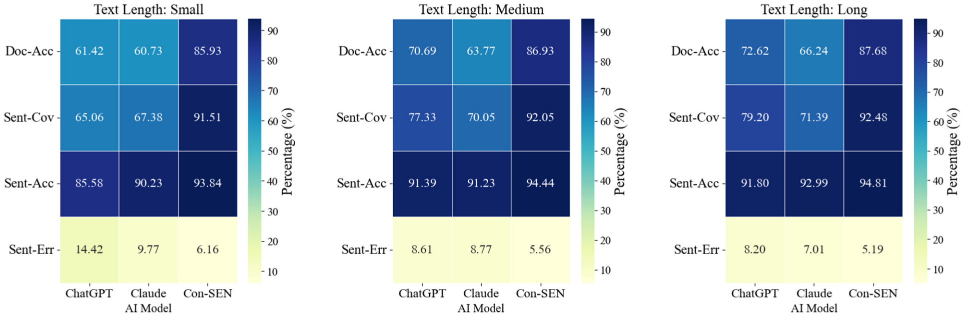

Fig. 3. Contract analysis metrics by model and text length.

environments, it can still produce high accuracy and low volatility in judgment results. The results are shown in Fig. 3.

This section further introduces a volatility metric to assess the trend of output stability across different models when faced with diverse contract samples. By examining the fluctuations in review result accuracy, statement coverage, statement accuracy, and statement error rates, we can more precisely measure the model's robustness and adaptability to changes in input scale and structure.

In the short text subset, Con-SEN exhibits significantly lower fluctuations across all four metrics compared to ChatGPT and Claude. In contract scenarios with smaller text sizes and relatively simple sentence structures, general LLMs still show instability in review outputs. Due to the disruption of reasoning chains and context drift caused by minor structural changes, ChatGPT and Claude are prone to consistency fluctuations. In contrast, Con-SEN enhances the local stability of input representations through clause-level and polarity encoding, effectively improving the robustness of output results against minor text perturbations.

In medium-length contract samples, the stability advantage of Con-SEN further expands, with the fluctuation in review result accuracy dropping to 0.00035, significantly lower than ChatGPT (0.0084) and Claude (0.0035). The fluctuation in statement coverage is only 0.00023, which is one-tenth of that of general models. As text length increases, the logical structure and clause intersection density within contracts rise sharply, leading to rapid deterioration in output stability for general models due to limitations in context window sliding and attention dilution. Con-SEN achieves consistency in long-distance semantic capture through a sliding window and a sparse global attention mechanism, maintaining continuous stability in clause recognition and risk classification decision-making even during cross-segment reasoning and multi-layer conditional nesting.

In the long text subset, Con-SEN demonstrates an almost scale-invariant stability characteristic, with the accuracy fluctuation of review results maintained at

Table 8. Analysis of the volatility of short, medium and long term contract statements.

Number of contractual statements	Model	Doc-Acc volatilities	Sent-Cov volatilities	Sent-Acc volatilities
Textbook	ChatGPT	0.68%	2.25%	1.14%
	Claude	0.36%	0.29%	0.59%
	Con-SEN	0.22%	0.12%	0.06%
Medium text	ChatGPT	0.84%	0.71%	0.34%
	Claude	0.35%	0.35%	0.46%
	Con-SEN	0.04%	0.02%	0.01%
Long text	ChatGPT	0.60%	0.59%	0.37%
	Claude	0.50%	0.55%	0.41%
	Con-SEN	0.04%	0.02%	0.02%

0.00044, statement coverage fluctuation at 0.00025, and both statement accuracy and error rate fluctuations at approximately 0.00016. In contrast, during the same period, the fluctuation levels of ChatGPT and Claude were between 0.0050 and 0.0060, exceeding those of Con-SEN by more than an order of magnitude. In ultra-long text environments, contracts exhibit large-scale cross-references, complex nested logic, and multiple conditional constraints, which greatly test the model's reasoning consistency and classification stability. Due to the lack of explicit hierarchical awareness and stable reasoning mechanisms, ChatGPT and Claude experience a rapid amplification of decision chain fragility as input length increases, leading to increased output result fluctuations. Con-SEN successfully mitigates performance degradation caused by scale growth through its deeply optimized long-text structure awareness and multi-task consistency judgment head design, maintaining high consistency in risk classification performance under varying length conditions. The results are shown in Table 8.

6. Conclusion

This research proposes Con-SEN, a structure-aware and semantically enhanced multi-task classification framework tailored to address the structural complexity and semantic ambiguity inherent in sentence-level financial contract review. By treating each sentence as an atomic analysis unit, Con-SEN integrates a domain-adaptive long-text encoder, a polarity-oriented semantic injection module, and a unified multi-task optimization objective to jointly predict clause type, legality status, and risk level. Built upon the Con-SEN-Corpus, a specialized dataset for this task, we establish a comprehensive evaluation protocol covering both document-level and sentence-level performance. Experimental results demonstrate that

Con-SEN consistently outperforms leading LLMs (e.g., ChatGPT, Claude) in accuracy, semantic coverage, and metric stability, particularly under high-risk real-world conditions — validating its robustness and practical suitability for compliance-critical contract analysis.

Nevertheless, limitations remain. The current model primarily operates at the sentence level and lacks explicit mechanisms for modeling discourse structure or inter-sentential logic, which may hinder reasoning in context-heavy scenarios. Its generalization also diminishes in the presence of semantic ambiguity, nested negation, or cross-sentence references. Additionally, the static prediction framework does not incorporate user interaction or feedback signals, limiting its adaptability to evolving compliance semantics.

Future work may proceed in several directions. First, structured discourse modeling — such as graph-based encoders or contrastive inter-sentence learning — could enhance contextual representation. Second, hybrid reasoning architectures that combine rule-based compliance logic with deep contextual modeling may improve the interpretation of normative and exception clauses. Third, the construction of datasets and evaluation protocols aligned with real-world auditing workflows would support dynamic annotation, continual learning, and feedback-driven optimization.

In summary, Con-SEN bridges structural modeling, semantic precision, and regulatory interpretability within a unified framework, filling a critical gap in the ability of general-purpose language models to handle high-risk financial contract review tasks. It offers a deployable and reliable pathway toward intelligent contract auditing in compliance-sensitive environments.

ORCID

Zhiyong Li https://orcid.org/0000-0001-9307-8453

References

Appalaraju, S, B Jasani, BU Kota, Y Xie and R Manmatha (2021). DocFormer: End-to-end transformer for document understanding, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 973–983, IEEE.

Beltagy, I, ME Peters and A Cohan (2020). Longformer: The long-document transformer, arXiv:2004.05150.

Borgeaud, S, A Mensch, J Hoffmann, T Cai, E Rutherford and K Millican (2022). Improving language models by retrieving from trillions of tokens, *International Conference on Machine Learning*, pp. 2206–2240, PMLR.

Brown, TB *et al.* (2020). Language models are few-shot learners, *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, pp. 1877–1901.

Chalkidis, I, M Fergadiotis, P Malakasiotis, N Aletras and I Androutsopoulos (2020). LEGAL-BERT: The muppets straight out of law school, arXiv:2010.02559.

Choromanski, K, V Likhosherstov, D Dohan, X Song, A Gane, T Sarlos, P Hawkins, J Davis, A Mohiuddin, L Kaiser, D Belanger, L Colwell and A Weller (2022). Rethinking attention with performers, arXiv:2009.14794.

Dai, Z, Z Yang, Y Yang, J Carbonell, QV Le and R Salakhutdinov (2019). Transformer-XL: Attentive language models beyond a fixed-length context, arXiv:1901.02860.

Devlin, J, MW Chang, K Lee and K Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, ACM.

Elwany, E, D Moore and G Oberoi (2019). BERT goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding, arXiv:1911.00473.

Guha, N, MF Chen, K Bhatia, A Mirhoseini and F Sala (2023). Embroid: Unsupervised prediction smoothing can improve few-shot classification, *Advances in Neural Information Processing Systems*, 36, 63259–63291.

Hao, S, Y Gu, H Ma, JJ Hong, Z Wang, DZ Wang and Z Hu (2023). Reasoning with language model is planning with world model, arXiv:2305.14992.

Henderson, P, MS Krass, L Zheng, N Guha, CD Manning, D Jurafsky and DE Ho (2022). Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset, *Advances in Neural Information Processing Systems*, 35, 29217–29234.

Hendrycks, D, C Burns, A Chen and S Ball (2021). CUAD: An expert-annotated NLP dataset for legal contract review, arXiv:2103.06268.

Kasneci, E, K Sessler, F Fischer, U Gasser and G Groh (2023). ChatGPT for good? On opportunities and challenges of large language models for education, *Learning and Individual Differences*, 103, 102274.

Katz, DM, MJ Bommarito, S Gao and P Arredondo (2024). GPT-4 passes the bar exam, *Philosophical Transactions of the Royal Society A*, 382(2270), 20230254.

Khandelwal, U, O Levy, D Jurafsky, L Zettlemoyer and M Lewis (2020). Generalization through memorization: Nearest neighbor language models, arXiv:1911.00172.

Ma, Y, Y Shao, Y Wu, Y Liu, R Zhang, M Zhang and S Ma (2021). LeCaRD: A legal case retrieval dataset for Chinese law system, *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2342–2348, ACM.

Pappagari, R, P Żelasko, J Villalba, Y Carmiel and N Dehak (2019). Hierarchical transformers for long document classification, arXiv:1910.10781.

Rae, JW, A Potapenko, SM Jayakumar and TP Lillicrap (2019). Compressive transformers for long-range sequence modelling, arXiv:1911.05507.

Rodríguez, P, I Laradji, A Drouin and A Lacoste (2020). Embedding propagation: Smoother manifold for few-shot classification, arXiv:2003.04151.

Sallam, M (2023). The Utility of ChatGPT as an Example of Large Language Models in Healthcare Education, Research and Practice: Systematic Review on the Future Perspectives and Potential Limitations, MedRxiv:2023-02.

Shaghaghian, S, LY Feng, B Jafarpour and N Pogrebnyakov (2020). Customizing contextualized language models for legal document reviews, *2020 IEEE International Conference on Big Data (Big Data)*, pp. 2139–2148, IEEE.

Vaswani, A, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez and I Polosukhin (2017). Attention is all you need, *Advances in Neural Information Processing Systems*, 30.

Xiao, C, X Hu, Z Liu, C Tu and M Sun (2021). Lawformer: A pre-trained language model for Chinese legal long documents, *AI Open*, 2, 79–84.

Yang, Z, Z Dai, Y Yang, J Carbonell, RR Salakhutdinov and QV Le (2019). XLNet: Generalized autoregressive pretraining for language understanding, *Advances in Neural Information Processing Systems*, 32.

Zaheer, M, G Guruganesh, A Dubey, J Ainslie, C Alberti, S Ontanon, P Pham, A Ravula, Q Wang, L Yang and A Ahmed (2020). Big bird: Transformers for longer sequences, *Advances in Neural Information Processing Systems*, 33, 17283–17297.

Zheng, L, N Guha, BR Anderson, P Henderson and DE Ho (2021). When does pretraining help?: Assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pp. 159–168. ACM.

Zhong, H, Z Zhang, Z Liu and M Sun (2019). Open Chinese language pre-trained model zoo, Technical report, NLP Lab, Department of Computer Science, Tsinghua University.

Zhu, Z and R Soricut (2021). H-transformer-1D: Fast one-dimensional hierarchical attention for sequences, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Vol. 1, pp. 3801–3815