Transactions on Computational and Scientific Methods | Vol. 4, No. 07, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Intelligent Compliance Risk Detection in the Pharmaceutical Industry via Transformer-Driven Semantic Discrimination

Yuliang Wang
Imperial College London, London, United Kingdom
yuliang.wang20@alumni.imperial.ac.uk

Abstract: This study addresses the challenges faced by pharmaceutical enterprises in compliance risk identification, including complex multi-source heterogeneous data, deep semantic dependencies in text structures, and hidden risk cues. An intelligent semantic discrimination model based on the Transformer architecture is proposed. The model builds a multi-layer self-attention mechanism and a semantic-weighted pooling module to achieve global semantic modeling and dynamic feature aggregation across regulatory documents, audit reports, and corporate disclosure texts. The embedding and positional encoding layers serve as inputs, while the multi-head attention mechanism captures cross-sentence dependencies. Residual connections and layer normalization ensure stable semantic propagation. During feature aggregation, an adaptive weighting mechanism enhances the model ' s ability to focus on key risk-related semantic segments, improving both accuracy and robustness. Experimental results show that the proposed model outperforms traditional deep learning methods across multiple evaluation metrics and achieves higher accuracy, recall, and overall stability under complex text conditions. Validation on real-world pharmaceutical corpora confirms that the model effectively identifies potential compliance risk behaviors, demonstrating strong semantic understanding and risk recognition capabilities. This study provides a scalable technical framework for intelligent compliance analysis in the pharmaceutical industry and introduces a new algorithmic approach for semantic risk modeling in regulatory technology.

Keywords: Compliance risk identification; Transformer; semantic modeling; medical text

1. Introduction

The pharmaceutical industry is currently undergoing a critical phase of digital transformation alongside the advancement of global regulatory systems. Pharmaceutical companies must maintain technological innovation and market competitiveness across stages such as drug development, clinical trials, production, and distribution. At the same time, they face complex compliance risks and increasingly stringent regulatory frameworks[1]. In recent years, as data-driven decision-making has become essential for enterprise management, pharmaceutical compliance issues have exhibited greater levels of intelligence and systematization. Challenges such as drug safety, clinical data integrity, supply chain traceability, financial transparency, and intellectual property protection have become intertwined, creating unprecedented difficulties in compliance management. Particularly in the context of coexisting international regulatory standards and constantly evolving legal frameworks, the ability to identify and assess risks has become a crucial indicator of a company's governance capacity and market stability[2].

Traditional compliance risk assessment in the pharmaceutical industry often relies on manual review and static rule matching. These approaches show clear limitations when dealing with unstructured data, cross-

domain information flows, and dynamic risk scenarios. On one hand, pharmaceutical enterprises generate highly complex data, including research literature, clinical reports, regulatory documents, contractual clauses, and market announcements. On the other hand, risk events often display latency and multi-stage propagation, making rule-based systems ineffective in achieving early warning and precise semantic recognition. With the rapid progress of natural language processing technologies, extracting latent risk signals from massive textual and heterogeneous semantic data has become a key challenge in integrating enterprise governance with regulatory technology[3].

The emergence of the Transformer model provides a breakthrough for this field. Its strong contextual modeling and semantic understanding capabilities enable it to capture long-range dependencies across paragraphs and topics within pharmaceutical texts. This allows for more accurate risk identification in complex regulatory provisions, corporate reports, and policy documents. Compared with traditional machine learning methods, the Transformer-supported by multi-layer self-attention mechanisms detect potential non-compliance features within a global semantic space, promoting the transition from "rule-driven" to "semantics-driven" intelligent risk detection. This paradigm shift not only enhances the automation of risk recognition but also provides enterprises with interpretable and traceable decision-making support[4].

From a broader perspective, intelligent risk discrimination for pharmaceutical enterprises holds significant practical and strategic value. It strengthens the industry's risk prevention system and improves responsiveness to policy changes, market fluctuations, and cross-border regulations. At the same time, it accelerates the deep integration of artificial intelligence with corporate governance and social responsibility. By systematically learning from enterprise textual data, policy corpora, and compliance reports, intelligent discrimination models help build dynamic compliance monitoring systems, shifting the focus from post-event review to pre-event identification. This transformation aligns with the trend of digital governance and provides technological support for the sustainable development of the pharmaceutical industry.

Furthermore, developing Transformer-based discrimination models carries substantial theoretical and methodological significance. It promotes the application of natural language processing in high-risk and high-compliance domains, validating the transferability and robustness of deep semantic models in specialized corpora. It also introduces a new research perspective for regulatory technology by combining algorithmic reasoning with semantic understanding to achieve intelligent analysis of policy documents, audit reports, and corporate disclosures. As the industry transitions from experience-based management to data-driven governance, Transformer-based compliance risk discrimination not only enhances corporate risk management but also offers policymakers and regulators technical tools for assisted decision-making. This contributes to greater transparency and standardization across the pharmaceutical supply chain, underscoring the real-world relevance and forward-looking value of this research direction.

2. Related work

Existing research on compliance risk identification and discrimination in pharmaceutical enterprises can be divided into three main categories: rule-based traditional methods, statistical learning methods, and deep representation learning methods. Early studies relied on domain experts to design risk identification rules or keyword templates. Potential compliance issues, such as drug quality abnormalities, clinical trial violations, and improper financial disclosures, were identified through pattern-matching strategies[5]. These methods were interpretable and controllable in the early stages, but they showed limited generalization and high maintenance costs when dealing with large-scale unstructured texts, cross-context semantics, and frequent policy changes. As regulatory requirements became more dynamic and data sources more diverse, rule-based systems could no longer meet the needs of intelligent compliance analysis in complex pharmaceutical environments.

Subsequently, researchers began to introduce statistical learning and traditional machine learning models to process structured and semi-structured enterprise data. The main approaches at this stage included classification models such as logistic regression, support vector machines, and random forests based on

feature engineering. These models predicted risks and identified compliance issues by extracting key indicators or textual features. Some studies also incorporated natural language features such as word frequency, TF-IDF, or topic models to improve text comprehension. However, these methods relied heavily on manually designed features and could not capture contextual dependencies and deep semantic relations in pharmaceutical texts. When handling long texts such as regulatory reports, drug registration documents, and policy announcements, the models often failed to understand sentence logic and cross-paragraph relations, leading to limited accuracy in risk discrimination. Moreover, machine learning models usually assume balanced data distributions, while non-compliance cases in pharmaceutical contexts are rare and hidden, making the models sensitive to sample bias[6].

With the rise of deep learning, neural network models have become an important direction for pharmaceutical compliance risk research. The introduction of convolutional and recurrent neural networks facilitated automated modeling of tasks such as text classification, event detection, and entity recognition, enabling models to learn representations directly from raw corpora without relying on manual rules. In particular, some studies explored multimodal modeling by integrating financial data, clinical trial records, and regulatory texts to capture compliance risks from multiple perspectives. However, these models generally lacked strong contextual modeling capabilities. When facing long documents, complex dependency structures, or cross-file relationships, they still struggled to achieve global semantic understanding. More importantly, traditional neural networks often lack interpretability and traceability, which limits their practical application in highly regulated pharmaceutical domains[7].

In recent years, Transformer-based pretrained language models have shown remarkable advantages in risk discrimination and knowledge extraction. The multi-layer self-attention mechanism enables the model to capture semantic dependencies across sentences, paragraphs, and even documents. This helps address the challenges of professional terminology, complex syntax, and wide contextual span in pharmaceutical texts. Some studies have further integrated graph neural networks, knowledge graphs, or contrastive learning strategies to enhance structural understanding and entity association. These advancements allow for more precise identification of hidden risk cues in corporate documents. The shift from feature-driven to semantic-driven modeling marks the entry of pharmaceutical compliance risk analysis into the era of intelligent semantic modeling. Current research trends indicate that Transformer models can achieve advanced semantic representation and contextual comprehension. They can also work with external resources such as compliance knowledge bases and policy ontologies to support dynamic and interpretable risk identification systems. However, how to optimize pretraining for domain-specific pharmaceutical texts, balance model complexity and interpretability, and establish an effective mapping from semantic information to risk logic remain important topics for future research.

3. Method

This study constructs a Transformer model framework for pharmaceutical enterprise compliance risk identification. Taking multi-source semantic features as input, it uses a multi-layer self-attention mechanism to capture potential dependencies between texts, thereby achieving high-precision identification of risk categories. First, a sample sequence of pharmaceutical enterprise compliance text is represented as $X = [x_1, x_2, ..., x_n]$, where x_i represents the i-th word vector. The input sequence is mapped into a word vector matrix $E \in R^{n \times d}$ through an embedding layer, and a positional encoding P is added to preserve sequential information, forming the final input representation:

$$H_0 = E + P$$

This input is then fed into a multi-layer Transformer encoder to capture contextual dependencies and cross-semantic relationships. This encoding approach enables the model to identify implicit logical and compliance

relationships within different text paragraphs, providing high-level semantic features to support subsequent risk assessment. Figure 1 shows the overall architecture.

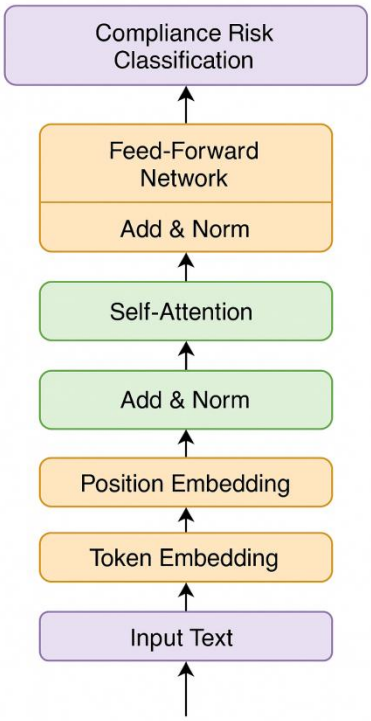


Figure 1. Overall model architecture

The core structure of each Transformer encoder layer is the Multi-Head Self-Attention (MHSA) mechanism. This mechanism uses different attention heads to learn dependencies in the feature space from multiple perspectives. For each attention head h, the query, key, and value matrices $Q = H_{l-1}W_Q, K = H_{l-1}W_K, V = H_{l-1}W_V$ are defined. The attention weight is calculated by the scaled dot product:

$$Attention(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

The results of multiple attention heads are concatenated and linearly mapped to obtain the layer output:

$$H_l = Concat(head_1,...,head_h)W_O$$

This multi-headed structure enables the model to simultaneously model the complex dependencies between regulatory terms, risk descriptions, and behavioral characteristics in different semantic subspaces, providing structural support for the global semantic modeling of risk characteristics.

During the feature aggregation phase, this study introduces a semantically weighted pooling mechanism to enhance the model's focus on key information. For the output A encoded by the multi-layer Transformer, a sentence-level representation is obtained using an adaptive weighting strategy:

$$z = \sum_{i=1}^{n} a_i h_i, \quad where \quad a_i = \frac{\exp(w^T h_i)}{\sum_{j=1}^{n} \exp(w^T h_j)}$$

Where h_i represents the hidden layer output at position i, and a_i represents the corresponding attention weight. This design automatically focuses on risk-related segments within the text, such as regulatory conflicts, descriptions of missing data, or unusual clauses, thereby improving the model's discriminative

power and semantic sensitivity. Through this weighted convergence, the model effectively maps local features to global risk representation at the semantic level.

Finally, the aggregated global representation z is fed into the classifier for risk discrimination. The classifier uses a feedforward neural network layer and a softmax activation function to output multi-category risk. The predicted probability is defined as:

$$\hat{y} = \text{softmax}(W_c z + b_c)$$

Where W_c and b_c are the classification layer weights and bias terms, respectively. The optimization goal of the model is to minimize the cross-entropy loss function:

$$L = -\sum_{k=1}^{K} y_k \log(\hat{y}_k)$$

This loss function ensures the model's ability to distinguish between different risk categories and maintains stable learning results even with imbalanced sample distributions. Through this approach, the entire model implements an end-to-end semantic discrimination process from raw text to risk categories, providing a structured algorithmic foundation and scalable semantic understanding framework for intelligent compliance analysis in pharmaceutical companies.

4. Experimental Results

4.1 Dataset

The dataset used in this study is from the Kaggle platform's publicly available Enterprise Operations and Risk Management Dataset. This dataset covers real-world operational, financial, compliance, and disclosure records from multiple companies in the pharmaceutical and related industries. The data includes annual reports, financial statements, audit opinions, policy compliance documents, and market announcements, comprehensively reflecting potential compliance risks and governance issues within companies' operations. The dataset is large and diverse in sample sources, encompassing both structured fields (such as assets and liabilities, revenue and expenditure, and employee numbers) and a significant amount of unstructured text (such as company descriptions, audit conclusions, and regulatory notices), providing a rich semantic foundation for compliance risk assessment.

During the data preprocessing phase, the study cleansed and standardized the raw enterprise text. First, content related to corporate compliance, risk disclosure, and governance information was screened, removing duplicate records and irrelevant noise. Second, the unstructured text was segmented into sentences and paragraphs, annotated with risk factors, and mapped into unified semantic units to ensure semantic consistency and model training stability. Furthermore, a domain-specific lexicon was introduced to standardize the use of specialized terminology and reduce ambiguity caused by industry-specific expression differences.

To ensure sample balance and representativeness, this study employed a stratified sampling strategy when constructing the task set, dividing the data into training, validation, and test sets in an 8:1:1 ratio. This dataset not only contains rich semantic information about corporate operations and risks but also reflects the multi-dimensional regulatory and compliance environment, providing reliable data support and validation for the Transformer-based semantic discrimination model.

4.2 Experimental Results

This paper first gives the results of the comparative experiment, as shown in Table 1.

Table1: Comparative experimental results

Model	ACC	F1-Score	Precision	Recall
1DCNN[8]	0.832	0.868	0.879	0.857
Transformer[9]	0.894	0.892	0.901	0.884
BILSTM[10]	0.882	0.880	0.887	0.873
BERT[11]	0.906	0.905	0.911	0.898
Ours	0.932	0.931	0.938	0.925

From the overall results, the proposed Transformer-based compliance risk discrimination model achieved significant advantages across all evaluation metrics. It performed particularly well in accuracy (ACC) and F1-Score, reaching 0.932 and 0.931, respectively. This indicates that the model not only achieves higher precision in overall classification but also maintains better balance across different risk categories. Compared with traditional deep learning structures such as 1DCNN and BiLSTM, the proposed model captures cross-sentence and cross-paragraph semantic dependencies more effectively. It enables a finer understanding of potential risk features. This improvement demonstrates the strength of multi-layer self-attention mechanisms in modeling complex semantic contexts, providing a more interpretable and stable technical foundation for intelligent compliance analysis in pharmaceutical enterprises.

Compared with standard Transformer and BERT models, the proposed model shows further improvement in Precision and Recall. This means it not only reduces false positives but also captures more potential high-risk samples. In practical pharmaceutical enterprise scenarios, relying on single semantic cues often leads to biased risk detection. The model in this study introduces a semantic-weighted pooling mechanism and a global context aggregation strategy, allowing it to automatically focus on key risk segments within long texts. Examples include non-compliant clauses, abnormal audit descriptions, or deficiencies in drug management. This structural enhancement allows the model to maintain high precision while improving recall, demonstrating strong practical usability and adaptability to industry needs.

From a comparative perspective, although BERT provides stronger semantic representation than traditional networks, its transferability to domain-specific corpora remains limited. The proposed model introduces structural optimization and domain adaptation to achieve a deeper semantic understanding of professional terminology and compliance logic in pharmaceutical texts. In contrast, BiLSTM and 1DCNN are constrained by their local feature extraction when modeling long sequences, making it difficult to capture cross-level dependencies effectively. As a result, their performance in Recall and F1-Score is weaker. This indirectly confirms the necessity and effectiveness of global attention mechanisms for semantic understanding in pharmaceutical compliance tasks.

Overall, the results of this study validate the comprehensive advantages of the Transformer-based compliance risk discrimination framework in semantic understanding, risk representation, and feature integration. By integrating domain semantics and compliance structure information at the feature level, the model achieves a transformation from "text understanding" to "semantic discrimination." It can accurately characterize the latent patterns of enterprise compliance risks. The superior performance of this approach not only provides a new technical pathway for intelligent compliance in the pharmaceutical industry but also offers a practical modeling paradigm for the application of regulatory technology (RegTech) in high-risk sectors. It highlights the real-world value and development potential of deep semantic models in complex governance scenarios.

This paper also presents an experiment on the sensitivity of the learning rate to ACC, and the experimental results are shown in Figure 2.

As shown in Figure 2, the variation trend of model accuracy with respect to the learning rate exhibits a clear nonlinear pattern. When the learning rate is low, the model converges slowly, but the parameter updates are

stable and show strong generalization ability. As the learning rate gradually increases, the model achieves its best performance in the range of 1×10^{-5} to 5×10^{-5}, with an ACC peak of 0.932. At this stage, the model effectively captures cross-paragraph semantic associations in pharmaceutical texts within a shorter training period, enabling precise modeling of risk features. This indicates that an appropriate learning rate plays a crucial role in ensuring stable optimization of the Transformer-based discrimination model in pharmaceutical semantic environments.

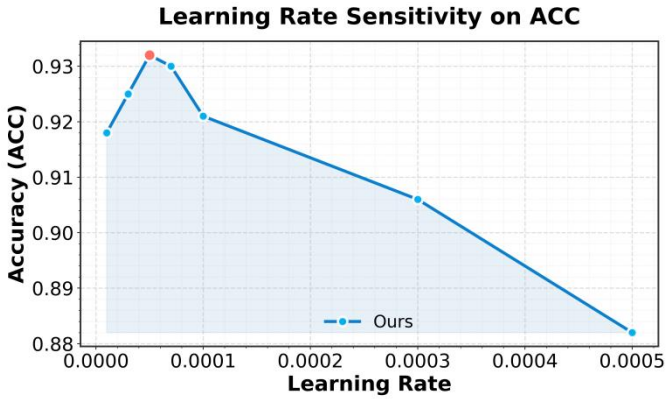


Figure 2. Experiment on the sensitivity of the learning rate to ACC

When the learning rate increases further to 7×10^{-5} and above, model performance shows a slight decline. This trend suggests that an excessively high learning rate causes large parameter update steps, leading to oscillations in the high-dimensional semantic space. As a result, the model struggles to maintain stable convergence in complex semantic structures of risk-related content. For pharmaceutical compliance texts, which feature deep semantic hierarchies and intricate entity relationships, overly rapid parameter adjustments may impair the model's ability to capture long-term dependencies, reducing the precision of classification boundaries. In such cases, the model can quickly learn surface patterns but fails to maintain deep semantic consistency and structural stability.

From the perspective of risk discrimination tasks, even small variations in the learning rate can significantly affect model performance on high semantic density texts. A lower learning rate helps preserve the semantic consistency of fine-grained features, while an excessively high learning rate may cause abnormal shifts in attention distribution, weakening the model's focus on key risk segments. Especially when processing drug registration reports, regulatory documents, or financial disclosure texts, the model's convergence stability and semantic fidelity directly determine the accuracy and reliability of risk identification. Therefore, learning rate control is not only a matter of hyperparameter tuning but also a critical factor in ensuring the model's sensitivity to pharmaceutical knowledge structures.

In summary, the experimental results confirm the strong sensitivity of the Transformer model to learning rate in pharmaceutical compliance risk discrimination tasks. An appropriate learning rate balances stable training and fast convergence, allowing the model to achieve better robustness and interpretability in complex semantic contexts. This finding provides an important reference for subsequent model optimization and highlights that, in pharmaceutical semantic understanding and risk classification tasks, the learning rate influences not only model performance but also its ability to learn reliably in high-risk textual environments.

This paper also gives the impact of the hidden layer dimension on experimental results, and the experimental results are shown in Figure 3.

As shown in Figure 3, the model exhibits significant differences in performance under different hidden layer dimension settings. This indicates that the hidden dimension plays a crucial role in compliance risk discrimination for pharmaceutical enterprises. When the hidden dimension is small, such as 128 or 256, the model's feature representation ability is limited. It cannot fully capture the multi-layer semantic relationships

in complex texts, resulting in relatively lower performance in metrics such as ACC and F1-Score. As the dimension increases to 512, the semantic representation space expands significantly. The model can better capture deep dependencies among pharmaceutical regulations, corporate reports, and policy clauses, thereby reaching its performance peak. This shows that a moderate hidden dimension can enhance global semantic modeling while maintaining good generalization ability.

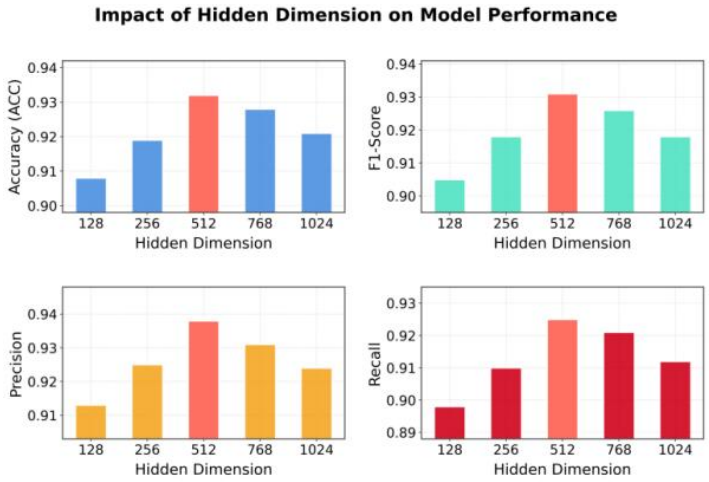


Figure 3. The impact of the hidden layer dimension on experimental results

When the hidden dimension continues to increase to 768 or 1024, the model's performance shows a slight decline. This is mainly due to parameter redundancy and the risk of overfitting caused by high-dimensional representations. In pharmaceutical texts, where semantics are dense and terminology distribution is uneven, an excessively large dimension may cause the model to focus too much on local features, weakening its ability to integrate global risk logic. Moreover, higher dimensions increase training instability, making the optimization process more sensitive to learning rate and regularization parameters. Therefore, the selection of hidden dimensions must balance representational capacity and stability to ensure robust performance under complex regulatory semantic conditions.

From the trend of Precision and Recall, the model with a moderate dimension (around 512) achieves the best results in both metrics. This indicates that such a setting can reduce both false positives and false negatives, effectively identifying potential high-risk samples. The improvement in Precision shows that the model can more accurately focus on genuine compliance risk signals. The increase in Recall reflects enhanced sensitivity to marginal risk cases. For pharmaceutical compliance analysis, this dual improvement is particularly important. It helps avoid the resource waste caused by excessive false alarms while ensuring that potential violations are detected in time.

Overall, the influence of hidden layer dimension on model performance reflects the balance between semantic modeling and parameter complexity. A well-designed hidden dimension strengthens the model's ability to abstract semantic features and identify risks in complex pharmaceutical texts. In contrast, dimensions that are too low or too high reduce the model's efficiency in capturing multi-level semantics. The experimental results confirm the flexibility and robustness of the proposed Transformer framework in feature space construction. They also provide a quantitative structural reference for future model optimization and a stable parameter design guideline for intelligent compliance discrimination in the pharmaceutical industry.

This paper also gives the influence of position encoding type on experimental results, and the experimental results are shown in Figure 4.

As shown in Figure 4, different types of positional encodings have a significant impact on the model's performance in pharmaceutical compliance risk discrimination tasks. This finding highlights the importance of temporal and structural positional information for semantic modeling within the Transformer framework.

The overall trend shows that traditional sinusoidal positional encoding provides fixed global order information but has limited expressive power when dealing with complex semantic hierarchies and cross-sentence dependencies. Therefore, it performs slightly worse in metrics such as ACC and F1-Score. In contrast, learnable positional encoding partially compensates for these limitations. It allows the model to adaptively adjust semantic weights between sequences, thereby improving the overall contextual modeling ability.

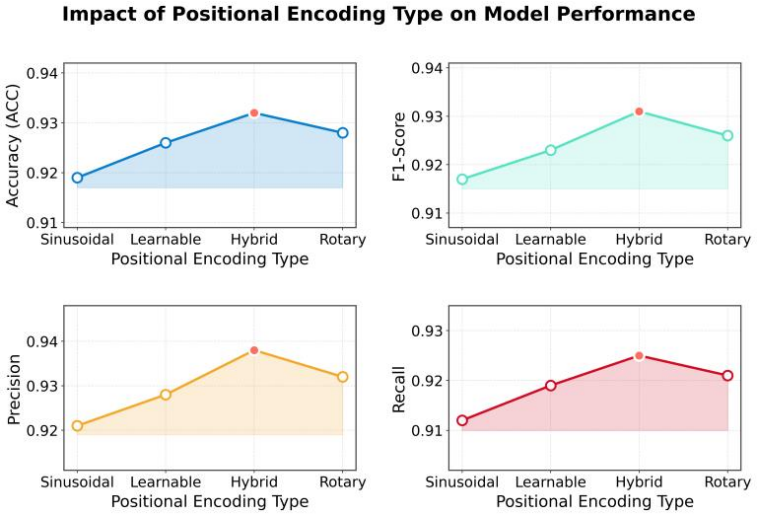


Figure 4. The impact of position encoding type on experimental results

Further observation reveals that hybrid positional encoding achieves the best performance across all four evaluation metrics. This result indicates that combining fixed positional patterns with learnable mechanisms effectively balances structural constraints and dynamic adaptability. The hybrid approach preserves the global structural advantages of fixed encodings while introducing learnable components for flexible adjustment of local dependencies. As a result, the model can better understand the hierarchical logic of pharmaceutical texts, such as regulatory clause references, risk descriptions, and approval conditions. This structural fusion greatly enhances the model's stability and generalization in high semantic complexity scenarios, demonstrating fine-grained modeling advantages for compliance semantics.

When using rotary positional encoding, the model's performance is slightly lower than that of the hybrid method. Although rotary encoding has certain advantages in capturing relative positional relationships, it tends to weaken long-range dependencies in pharmaceutical texts with long sentences and multiple co-occurring entities. In particular, when risk paragraphs involve contextual jumps or cross-document references, the smooth angular representation of the rotary mechanism may reduce semantic focus. This affects the model's ability to accurately identify key risk patterns. The results suggest that for pharmaceutical compliance tasks with large semantic spans, relying solely on rotational positional structures cannot fully ensure consistent propagation of global information.

Overall, the experimental results indicate that the design of positional encoding has a decisive impact on the compliance discrimination performance of Transformer models. The fixed, learnable, and hybrid mechanisms complement each other in terms of semantic constraint and information flexibility. The hybrid encoding achieves a balance between the two, enabling unified global dependency capture and local semantic reinforcement. For risk identification scenarios in pharmaceutical enterprises, this mechanism maintains high sensitivity to key semantic segments in complex documents. It effectively enhances the model's understanding depth and classification accuracy for compliance texts and provides an important direction for future model optimization.

This paper further presents an experiment on the sensitivity of the class imbalance ratio to recall, and the experimental results are shown in Figure 5.

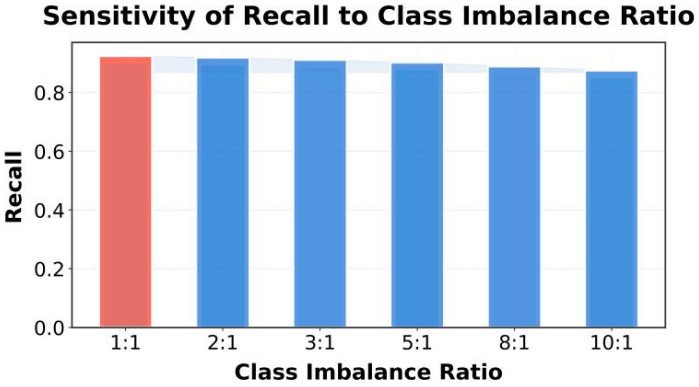


Figure 5. Sensitivity experiment of class imbalance ratio on Recall

As shown in Figure 5, the imbalance ratio between classes has a clear impact on the model's Recall. When the ratio of positive to negative samples is 1:1, the model achieves the highest recall rate and reaches optimal performance. This indicates that under balanced sample distribution, the model can effectively learn semantic features of both risk and non-risk samples and maintain high sensitivity to minority classes. For pharmaceutical compliance risk discrimination, this means the model can comprehensively identify both actual and potential violations, improving the completeness and accuracy of risk detection.

As the imbalance ratio increases, the model's Recall gradually decreases, showing that an unbalanced sample distribution weakens its ability to capture minority classes. When the ratio exceeds 5:1, the recall rate drops significantly, indicating that the model tends to overlook some high-risk samples when biased toward the majority class. Pharmaceutical compliance texts usually show a pattern of "many normal cases and few violations." When negative samples, representing risky behaviors, are scarce, the model is more likely to develop a "safety preference," ignoring potential violations. This bias amplifies compliance risks in real-world regulatory scenarios and reduces the model's ability to issue early warnings.

From a semantic perspective, class imbalance affects not only the optimization direction of the loss function but also the attention distribution of the Transformer model. As the gap between positive and negative samples widens, the model's attention weights become increasingly concentrated on the dominant normal samples. Consequently, the focus on key semantic cues in minority risk samples, such as regulatory conflicts, abnormal word usage, or non-compliant clauses, decreases, leading to a continuous decline in Recall. Therefore, maintaining the model's semantic sensitivity and generalization ability under imbalanced conditions is a key challenge in pharmaceutical risk identification tasks.

Overall, the experimental results reveal the systematic influence of class imbalance on the performance of compliance risk discrimination models. By controlling sample ratios, introducing loss-weighting mechanisms, or applying data augmentation strategies, the degradation caused by imbalance can be mitigated effectively. In pharmaceutical compliance analysis, a reasonable sample balancing mechanism helps the model identify low-frequency but high-risk violations more accurately. This ensures stable Recall performance while enhancing overall regulatory effectiveness, providing more robust technical support for risk management.

5. Conclusion

This study focuses on the compliance risk identification task in pharmaceutical enterprises and proposes an intelligent semantic model based on the Transformer architecture. The model is designed to address the challenges of complex semantic structures, diverse entity relationships, and hidden risk cues in pharmaceutical texts. By introducing a multi-layer self-attention mechanism and a semantic-weighted pooling

structure, the model effectively captures global dependencies within multi-level semantic spaces, achieving a shift from traditional rule-based matching to deep semantic understanding. Experimental results demonstrate the model's adaptability and robustness in multidimensional text environments, providing theoretical and algorithmic foundations for intelligent compliance analysis in the pharmaceutical industry.

The model achieves hierarchical modeling from lexical features to semantic relations and exhibits remarkable advantages in the structured understanding of long texts. By integrating positional encoding, contextual attention, and adaptive aggregation mechanisms, it enables unified modeling of heterogeneous sources such as regulatory clauses, audit reports, and risk descriptions. This facilitates precise risk identification and interpretable compliance judgment. Compared with traditional statistical or shallow learning methods, the proposed framework performs better in capturing global semantic dependencies and latent risk logic, proving the feasibility and value of deep language models in the pharmaceutical compliance domain.

From an application perspective, the findings of this study have significant implications for risk governance, regulatory technology development, and enterprise compliance management in the pharmaceutical sector. The model can be widely applied in areas such as drug registration review, production and distribution monitoring, and corporate information disclosure auditing, enabling a transformation from "post-event compliance" to "intelligent early warning." By introducing deep semantic discrimination mechanisms, enterprises can efficiently perform risk screening and decision support under complex policy contexts, enhancing regulatory transparency, reducing violations, and advancing the digital governance framework of the industry.

Future research can explore model interpretability, multimodal compliance data integration, and cross-lingual transfer learning. With the ongoing development of large language models and knowledge graph technologies, future compliance risk identification systems will be able to combine structured regulatory data with unstructured corporate documents to build more intelligent and traceable risk monitoring frameworks. In response to the needs of real-time compliance auditing and cross-regional regulation, lightweight and federated learning frameworks can also be explored to achieve privacy protection and distributed intelligent analysis, providing forward-looking technical support for the high-quality development of the pharmaceutical industry.

References

[1] Cejas, O. A., Azeem, M. I., Abualhaija, S. and Briand, L. C., "NLP-based automated compliance checking of data processing agreements against GDPR", IEEE Transactions on Software Engineering, vol. 49, no. 9, pp. 4282-4303, 2023.

[2] Guo, H., An, B., Guo, Z. and Su, Z., "Deep semantic compliance advisor for unstructured document compliance checking", Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, pp. 4446-4452, January 2021.

[3] Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T. and Leiserson, C. E., "Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics", arXiv preprint arXiv:1908.02591, 2019.

[4] Kakani A B, Nandiraju S K K, Chundru S K, et al. A Survey on Regulatory Compliance and AI-Based Risk Management in Financial Services[J]. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 2023, 4(4): 46-53.

[5] Dimlioglu T, Wang J, Bisla D, et al. Automatic document classification via transformers for regulations compliance management in large utility companies[J]. Neural Computing and Applications, 2023, 35(23): 17167-17185.

[6] Van Liebergen, B., "Machine learning: A revolution in risk management and compliance?", Journal of Financial Transformation, vol. 45, pp. 60-67, 2017.

[7] Spasic, I. and Nenadic, G., "Clinical text data in machine learning: Systematic review", JMIR Medical Informatics, vol. 8, no. 3, p. e17984, 2020.

[8] Han X. Application Research of Corporate Fraud Identification Model based on One-Dimensional Convolutional Neural Network[J]. Frontiers, 2021, 2(12).

[9] Li Y, Jiang X, Wang Y. TRAM-FIN: A transformer-based real-time assessment model for financial risk detection in multinational corporate statements[J]. Journal of Advanced Computing Systems, 2023, 3(9): 54-67.

[10] Chen, T., "XGBoost: A scalable tree boosting system", arXiv preprint arXiv:1603.02754, 2016.

[11] Li Z, Fan J, Zhang Y, et al. BERT-based intelligent text record mining for risk analysis of power system equipment[C]//3rd International Conference on Control Theory and Applications (ICoCTA 2023). IET, 2023, 2023: 12-15.