Journal of Computer Technology and Software

ISSN: 2998-2383

Vol. 4, No. 9, 2025

Integrating Context Compression and Structural Representation in Large Language Models for Financial Text Generation

Pei Xue[1], Yingnan Yi[2]

[1]Independent Researcher, Pittsburgh, USA

[2]Washington University in St. Louis, St. Louis, USA

*Corresponding author: Yingnan Yi, yiyingnan2021@163.com

Abstract: This paper focuses on key challenges in long-text generation and summarization in the financial domain, including context truncation, redundancy interference, and lack of structural understanding. It proposes a large language model approach that integrates context window compression and structure-aware modeling. The method introduces an information selection mechanism to compress ultra-long input sequences. This helps reduce information loss caused by limited context windows. At the same time, it applies structural graph modeling to capture cross-sentence logical connections. This enhances the model's ability to understand multi-level structures and complex semantics in financial texts. During generation, the model conditions the decoder on the compressed, structure-enhanced representations. This guides the generation or summarization process toward outputs that are semantically consistent and linguistically fluent. The study conducts systematic experiments on representative datasets from key financial subdomains. It designs multi-dimensional analyses, including sensitivity to context length, redundancy ratio perturbation, and subdomain variation. These experiments evaluate the model's performance in terms of language quality and generation stability. The results show that the proposed method achieves strong performance on mainstream evaluation metrics such as ROUGE. It also demonstrates good stability and generalization. The model adapts well to financial documents with different structures and expression styles. By introducing smoothing mechanisms and structural regularization, the training process exhibits fast convergence and low variance. These findings confirm the effectiveness and robustness of the proposed method in modeling highly structured financial texts.

Keywords: Structure-aware modeling, context compression, financial text summarization, generation stability

1. Introduction

In the context of increasingly complex financial information and rapidly growing data volumes, financial document generation and summarization have become highly challenging tasks in the field of natural language processing. Financial texts are typically lengthy, structurally complex, rich in terminology, and logically rigorous. These characteristics impose high demands on language models, especially in terms of contextual understanding and compression. Traditional NLP methods have achieved significant results in handling short texts and general semantic tasks[1]. However, when dealing with dense and structured financial documents, the generated outputs often suffer from semantic omissions, logical inconsistencies, and factual errors. These limitations hinder practical applications such as financial risk analysis, compliance auditing, and intelligent report generation[2].

With the development of large language models, autoregressive architectures have demonstrated exceptional capabilities in language understanding and generation. Their large-scale parameters and pretraining strategies offer strong generalization and abstraction abilities. However, mainstream models still rely on fixed-length context windows, which become a clear bottleneck when processing long financial texts. The input length limit prevents the model from handling entire documents, leading to truncated or missing critical information. While blindly expanding the context window may partially address this, it also results in significant memory consumption and decreased inference efficiency, making it unsuitable for real-world deployment. Therefore, enabling effective information compression, representation reconstruction, and context management within limited windows has become a critical technical challenge for applying large models in the financial domain[3].

The distinctiveness of financial texts also lies in their high information density and implicit temporal structure. Documents such as financial reports, regulatory notices, and risk disclosures often contain multi-level and cross-paragraph information dependencies. This requires the model not only to understand surface semantics but also to capture deep logical cues and causal relations[4]. Strong capabilities in long-range dependency modeling and semantic preservation are essential. Existing methods are often confined to local encoding and segment-wise generation, lacking a global understanding of the document's semantic structure. This makes it difficult to produce high-quality financial texts that are both complete and coherent. Preserving global semantics and domain logic during context compression, while guiding generation through structure-aware mechanisms, is crucial for improving both precision and controllability in financial text processing[5].

Moreover, financial document summarization serves as a key tool for automated information extraction and aggregation. It is widely applied in investment analysis, policy communication, and legal compliance. High-quality summaries must condense content while accurately identifying key points, maintaining factual correctness, and preserving stylistic consistency. Traditional extractive summarization methods often overlook deeper contextual semantics, while generative models are limited by context window constraints. These issues lead to summaries that are incomplete, redundant, or distorted. Therefore, building a generation mechanism that integrates context compression, structural modeling, and semantic preservation is essential for improving the accuracy, reliability, and utility of financial summarization.

In conclusion, for financial long-text generation and summarization tasks, integrating context window compression with structure-aware large language models is not only a critical step in adapting general models to domain-specific applications but also a foundational component of intelligent financial information generation systems. This research direction supports the integration of financial technology and language intelligence. It enhances the automation and decision-making efficiency of the financial sector and strengthens system capabilities in understanding and managing complex financial semantic scenarios. Ultimately, it contributes to the development of a more efficient and secure financial governance framework.

2. Related work

Existing research on financial text generation and summarization mainly focuses on the development of natural language generation models, the evolution of long-text processing techniques, and language modeling under domain-specific demands[6]. With the continuous expansion of large-scale pretrained language models, the quality of text generation has significantly improved. However, these models generally rely on fixed-length context windows for self-attention computation. This makes it difficult to handle long texts that exceed the window size. In financial texts, information often appears in multi-paragraph, nested formats, which impose higher demands on contextual integration. To address this, researchers have introduced techniques such as window compression, sliding windows, and adaptive pruning. These methods aim to compress the input context without sacrificing overall semantic structure, enabling efficient modeling of long documents.

To process long texts, some studies have introduced structure-aware modeling strategies to address the limitations of traditional Transformers in capturing hierarchical document structures. For example, certain approaches treat documents as multi-level structures composed of paragraphs, sentences, and clauses. By explicitly incorporating document structure encoders, they enhance the model's understanding of logical relationships and hierarchical organization in the context. These structure-aware mechanisms have proven effective in improving content integrity and semantic coherence in long-text generation and summarization. They are especially useful for financial regulations, annual reports, and audit documents, which follow clear structural templates[7]. Other studies combine segment-wise modeling with global reconstruction to achieve both efficient compression and semantic reorganization.

In the development of abstractive summarization, early methods mostly adopted sequence-to-sequence models based on encoder-decoder frameworks. With the introduction of large pretrained models, the quality of summarization has greatly improved. However, in financial scenarios, summarization requires more than general-domain performance. It must accurately cover key information while avoiding factual errors and ambiguous expressions[8]. Recent work has begun to explore the use of external knowledge, context control mechanisms, and content selection strategies to improve domain adaptability and accuracy. One approach uses context compression to reduce redundant semantics while preserving key information, enhancing the model's focus and selection abilities. Another approach applies multi-task learning and contrastive learning to improve factual consistency. This is particularly effective for resolving conflicts and inconsistencies in heterogeneous financial texts.

At the same time, the development of financial text generation and summarization faces several challenges. These include term understanding, numerical consistency, style control, and legal compliance. Some studies have introduced control variables into the generation process. These include domain labels, entity constraints, and temporal prompts, aiming to improve controllability and interpretability. Data scarcity is another issue in financial scenarios. To address this, recent research has explored methods such as synthetic data generation, retrieval augmentation, and pseudo-labeling to enrich training datasets. In summary, current research in financial text generation and summarization advances along multiple directions. These include structural modeling, context compression, style control, and semantic alignment. However, challenges remain in modeling the structural characteristics of long financial documents and balancing compression with semantic preservation. These issues require further breakthroughs and optimization.

3. Methodology

This study proposes a large language model generation method that combines a context window compression mechanism and structure-aware modeling, which is designed specifically for financial long text generation and summary tasks. The core of the method is to achieve effective modeling and generation control of ultra-long input sequences by compressing redundant content in long texts and retaining key information. The model architecture is shown in Figure 1.

First, the input financial long text is represented as a sequence $X = \{x_1, x_2, ..., x_n\}$, where $n >> L$, and L is the context window limit of the model. To this end, a context compression module $C(\cdot)$ is introduced to convert the original sequence into a compressed representation $\widetilde{X} = C(X)$, where $|\widetilde{X}| \leq L$. The design of the compression function incorporates the attention-guided sentence-level scoring function $a_i = soft\max(w^T \tanh(W_s h_i))$, where

h_i is the encoder output, W_s and w are trainable parameters used to select clauses with more information to retain.

To maintain the structural hierarchy and logical relationships in financial texts, a structure-aware graph representation $G = (V, E)$ is introduced in the generation process, where each node $v_i \in V$ represents a text unit (such as a paragraph or sentence), and the edge $e_{ij} \in E$ represents the logical or temporal dependency. The structural relationship is modeled through the graph attention mechanism, and the node update method is as follows:

$$h'_i = \sigma(\sum_{j \in N(i)} a_{ij} W_h h_j)$$

$a_{ij} = \text{softmax}(e_{ij})$ represents the structural attention weight between adjacent nodes, W_h is the linear mapping parameter, and $\sigma(\cdot)$ is the nonlinear activation function. This structure-aware mechanism is embedded in the construction process of the compressed representation, making the compressed result not only semantically sufficient but also structurally coherent.

The generation module uses a large language modeling structure based on Transformer, and performs conditional generation or summary generation by fusing the context-compressed representation $\widetilde{X}$ and the structural information-enhanced representation $H = \{h'_1, h'_2, ...h'_L\}$ as the decoder input. Each generation step in the decoding process is optimized by the standard language modeling objective function, specifically maximizing the conditional probability:

$$L_{gen} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \widetilde{X}, H)$$

Where y_t represents the tth word to be generated, T is the target length, $\widetilde{X}$ and H provide dual prior guidance of context and structure.

To further improve the coherence and accuracy of the generated content, the method also introduces a language consistency regularization term and a semantic alignment term. The language consistency loss is defined as the context similarity preservation constraint between sentence vectors:

$$L_{coh} = \sum_{i=1}^{T-1} (1 - \cos(f(y_i), f(y_{i+1})))$$

Where $f(\cdot)$ represents the sentence-level encoder and $\cos(\cdot,\cdot)$ is the cosine similarity, which is used to maintain the naturalness of the language transition between sentences. At the same time, the semantic preservation loss is achieved by maximizing the similarity between the input compressed representation and the output summary, specifically:

$$L_{sem} = 1 - \cos(g(\widetilde{X}), g(Y))$$

$g(\cdot)$ is the semantic encoder and Y represents the text generated by the model. The final objective function is a weighted combination of multiple losses:

$$L_{total} = L_{gen} + \lambda_1 L_{coh} + \lambda_2 L_{sem}$$

Where λ_1, λ_2 is the balance coefficient. This method organically integrates compression, structure, and semantic alignment in the modeling process, providing a scalable and high-precision solution for processing financial long text generation tasks.

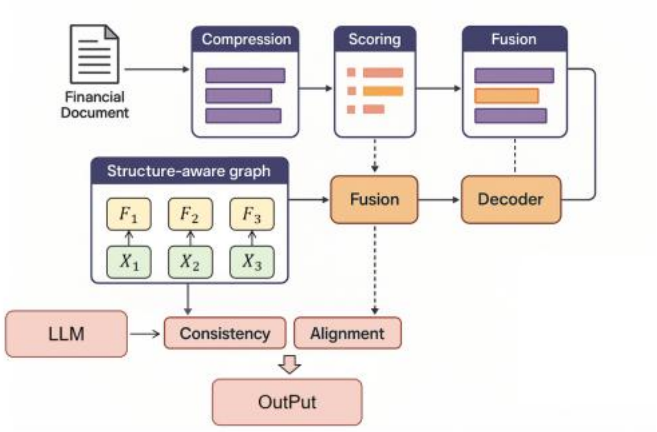

Figure 1. Overall model architecture diagram

4. Dataset

This study uses the FinSum dataset as the primary corpus for financial text generation and summarization tasks. The dataset consists of real financial industry news and earnings reports. It covers a wide range of announcements and financial articles related to companies in the U.S. stock market. The dataset shows strong domain relevance and structural completeness. It is widely used in financial natural language processing research. Each sample includes a long original document and its corresponding expert-written summary, making it ideal for training models on context compression and structural modeling.

The textual content in FinSum mainly includes press releases, financial statements, and analytical commentaries. The average length is relatively long, often spanning several paragraphs. The texts contain typical financial entities and expressions such as earnings figures, market forecasts, and risk disclosures. The summaries are written by professional editors. They highlight key points and take the form of coherent, multi-sentence texts with high readability. These summaries serve as effective supervision signals for generation tasks. The dataset features a formal writing style and dense terminology. It closely simulates real-world financial scenarios in both language generation and summarization needs.

During data processing, standard splits for training, validation, and testing are applied. The original documents are preprocessed through normalization. Redundant information is

removed. Currency symbols are unified. Dates and numerical formats are standardized. These steps ensure that models can focus on semantic understanding and context compression. This dataset provides a solid foundation for studying long-text processing in the financial domain. It is particularly suitable for evaluating the effectiveness of structure-aware and compression mechanisms under realistic financial text conditions.

5. Quantitative Analysis

This paper first conducts a comparative experiment to evaluate the performance of the proposed method against several representative baseline models under the same experimental settings. The purpose is to assess the relative advantages of the model in handling financial long-text generation and summarization tasks. This comparison provides a foundation for analyzing the model's effectiveness across different evaluation dimensions. The experimental results are shown in Table 1.

Table1: Comparative experimental results

Model	ROUGE-1	ROUGE-2	ROUGE-L
FinLlama3_sum[9]	45.32	22.48	41.67
DistFin[10]	46.15	23.02	42.21
GraphRAG[11]	48.09	24.87	44.30
FinMA[12]	47.62	24.35	43.89
FinGPT[13]	49.14	25.76	45.12
Ours	51.03	27.41	47.38

The results in the table show clear performance differences among models on financial text generation and summarization tasks. This is especially evident in ROUGE-2 and ROUGE-L scores, which are closely related to semantic continuity and structural coherence across sentences. These metrics better reflect a model's ability to capture semantic relations in long financial texts. Traditional models such as FinLlama3_sum and DistFin achieve baseline performance in ROUGE-1. However, they struggle to maintain logical connections between key information pieces when handling long documents.

In contrast, structure-enhanced models like GraphRAG and FinMA demonstrate significant improvements in ROUGE-2 and ROUGE-L. This suggests that incorporating structure-aware mechanisms helps strengthen contextual understanding and improves the coherence and consistency of generated texts. Such modeling approaches enable more effective hierarchical representation and information extraction from long texts. They help reduce information loss caused by context window limitations. These benefits are particularly important for financial documents that include multiple paragraphs and dense entity relations.

FinGPT, a large model pretrained on financial corpora, outperforms the previous methods across all three metrics. This indicates that large models possess certain capabilities for cross-sentence modeling and concept linking. However, in the absence of explicit context compression and structural modeling strategies, the model still has limitations in capturing long-range dependencies. It tends to miss facts or introduce logical gaps when dealing with highly dense information settings.

In comparison, the method proposed in this study achieves the best results across all metrics. The performance gains in ROUGE-2 and ROUGE-L are especially notable. This confirms that combining a context window compression mechanism with structure-aware modeling leads to substantial improvements. The approach alleviates the restrictions of limited context length. It also enhances the model's ability to understand paragraph-level logical relations. As a result, the overall quality and stability of financial long-text generation and summarization are improved. This dual-mechanism design produces outputs that better match the semantic density and logical completeness required in financial documents. It shows strong domain adaptability and practical potential.

This paper also presents an experiment on the impact of different context window sizes on generation quality, aiming to explore how varying the accessible input length affects the model's ability to capture semantic dependencies and maintain structural coherence in financial texts. By adjusting the size of the context window during inference, the experiment examines the model's responsiveness to contextual scope and evaluates whether larger or smaller windows influence its performance in long-text generation scenarios. The experimental results are shown in Figure 2.

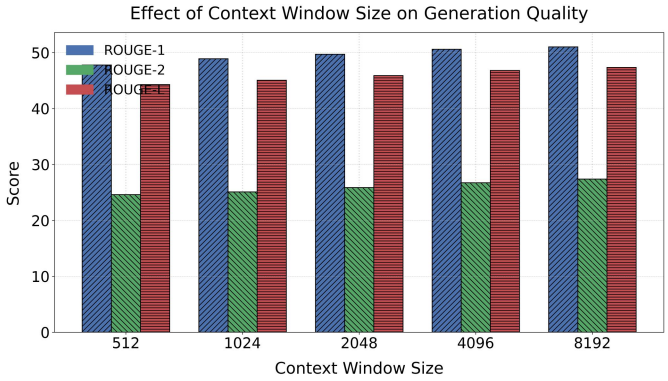

Figure 2. Experiment on the impact of different context window sizes on generation quality

As shown in the figure, the model's performance on ROUGE-1, ROUGE-2, and ROUGE-L improves steadily as the context window size increases. This indicates that expanding the context range has a significant positive impact on long financial text generation. The performance gains are particularly notable when the window increases from 512 to 2048. During this phase, all three metrics show clear improvement. This suggests that the model benefits from a wider information reception range and becomes more capable of capturing semantic cues and contextual logic.

A closer look reveals that ROUGE-2 improves most significantly. This indicates that enlarging the context window helps enhance the model's ability to capture continuous semantic units. As a result, the generated text becomes more coherent across sentences and more fluent overall. In financial documents, semantic dependencies often span multiple paragraphs. A larger context window allows the model to better

integrate these dispersed segments, producing more logically consistent outputs.

The improvement in ROUGE-L further shows that the model becomes better at preserving the overall structure of long documents. As the context range expands, the structure-aware module more effectively retains paragraph-level semantic frameworks and the proper use of financial terminology. This leads to generated outputs that more closely match the organizational patterns of the original texts.

Overall, this experiment confirms the effectiveness of the proposed context window compression and structural modeling strategy. It demonstrates that expanding the accessible context, when combined with structure-aware mechanisms, can significantly improve semantic coverage, structural preservation, and linguistic coherence in financial long-text generation. These improvements provide a strong foundation for generating high-quality financial summaries.

This paper also compares the impact of different financial sub-field inputs on the quality of model output, and the experimental results are shown in Figure 3.

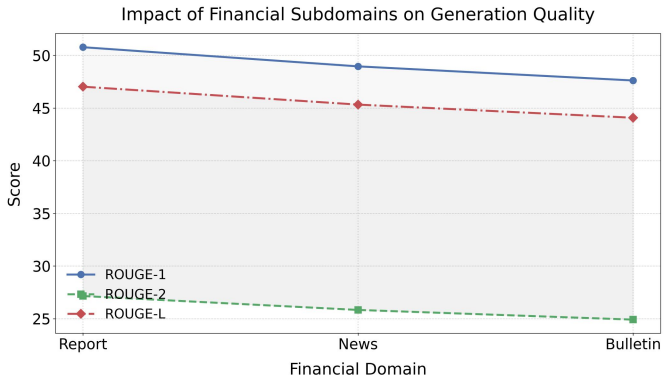

Figure 3. Comparison of the impact of different financial sub-field inputs on model output quality

The figure shows that the model's generation quality varies across different financial subdomains. The overall performance follows a decreasing trend: earnings reports perform best, followed by news, and then announcements. ROUGE-1 and ROUGE-L scores gradually drop as the input domain shifts. This indicates that when dealing with less structured and lower-density texts, the model's ability to preserve semantics and maintain structural consistency weakens. The trend highlights the significant differences in discourse structure, semantic concentration, and logical flow across financial subdomains. It also raises higher demands on the stability of structure-aware and context compression mechanisms.

The decline in ROUGE-2 is particularly pronounced. This suggests that the model finds it more difficult to capture deep dependencies between continuous semantic units when handling news or announcement texts. Compared with earnings reports, these texts have more fragmented logic and a more flexible writing style. This challenges the adaptability of structural modeling modules. The results also emphasize that domain-specific language style directly affects the model's fine-grained language modeling. In summarization tasks, this impact becomes more critical due to the high requirements on linguistic precision and syntactic coherence.

From the perspective of model architecture, the experiment confirms that the proposed context compression and structure-aware strategies are more effective when applied to highly structured texts such as financial reports. In such cases, the model can better identify multi-level structures and key information paths, resulting in higher-quality outputs. On the other hand, in subdomains with more scattered information, such as announcements, the structural signals are weaker. This leads to a drop in output stability, with more noticeable semantic drift and fragment discontinuity.

This paper also gives an analysis of the impact of changes in the proportion of redundant information in the input text on the stability of the model, and the experimental results are shown in Figure 4.

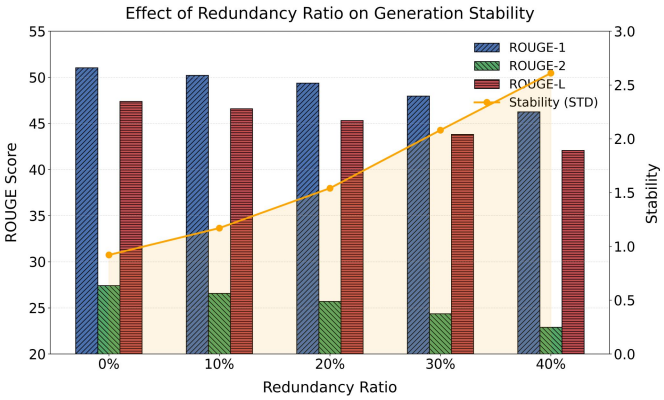

Figure 4. Analysis of the impact of changes in the proportion of redundant information in input text on model stability

The figure shows that as the proportion of redundant information in the input increases, the model's performance on all ROUGE metrics declines to some extent. This decline is especially pronounced in ROUGE-2 and ROUGE-L. The results indicate that the presence of redundant content interferes with the model's contextual modeling. It weakens its ability to capture key information and maintain inter-sentence coherence. As a result, the quality of the generated language and semantic coverage decreases.

The sharp drop in ROUGE-2 highlights the model's difficulty in constructing accurate phrase-level connections between sentences under semantic noise. This type of semantic drift and micro-level logical disruption is particularly critical in financial summarization tasks. Financial documents usually contain dense and precise information, with high requirements for detail integrity. The inclusion of redundant segments can mislead the model's attention toward irrelevant parts. This weakens core information and increases logical jumps between sentences.

At the same time, the line chart shows the trend of standard deviation (stability). It rises as the redundancy ratio increases. This suggests a significant decrease in the stability of the model's output. Lower stability means greater variability in generated content across batches or fine-tuning rounds. This

may result in factual inconsistency and structural fragmentation. Such issues undermine the reliability expectations in financial applications.

Finally, this paper also provides a loss function decline graph to illustrate the training dynamics of the proposed model over multiple epochs. The graph is used to visualize the trend of both training and validation loss during the learning process, offering insights into the model's convergence behavior and training stability. This helps to better understand how the model optimizes its parameters over time when dealing with long and complex financial texts. The corresponding visualization is shown in Figure 5.

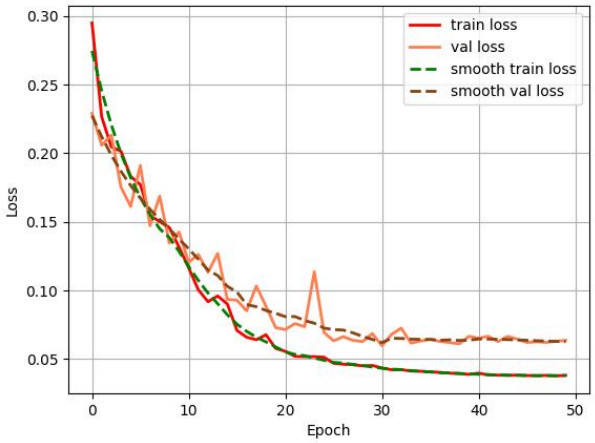

Figure 5. Loss function changes with epoch results

The figure shows that both training loss and validation loss drop rapidly in the early stages. This indicates that the model quickly improves its ability to capture the basic semantic structure and patterns of financial texts. The sharp decline within the first 10 epochs confirms that the proposed context compression and structure-aware mechanisms achieve high convergence efficiency in the early training phase. This fast decline is especially important for handling high semantic density in financial documents. It helps the model establish effective semantic mappings at an early stage.

As training progresses, the loss trend gradually stabilizes. This suggests that the model completes its learning of long-text logical structures during convergence. The generation process becomes more stable. Between epochs 20 and 40, the training and validation loss curves remain closely aligned, with minimal fluctuation. This indicates strong generalization on financial corpora. No signs of overfitting or underfitting appear. These results validate the structure-aware module's ability to maintain consistent representations across different text segments.

In addition, the smoothed curves show that both training and validation losses remain at low levels during the final stages. The fluctuation in validation loss gradually converges. This demonstrates the stability of the proposed method when handling long and information-dense input texts. Such stability is essential in financial generation tasks. It helps ensure consistency and compliance in generated outputs, especially for documents like reports and announcements that involve heavy use of numerical and structural constraints.

In summary, the trend of the loss curves confirms the robustness and convergence strength of the proposed method during training. The context compression mechanism effectively reduces the negative impact of input redundancy on modeling precision. The structure-aware strategy enhances the model's ability to preserve both semantics and structure in financial texts. As a result, the model achieves fast convergence and consistently generates high-quality outputs within a limited number of epochs.

6. Conclusion

This paper focuses on the task of long-text generation and summarization in the financial domain. It proposes a large language model framework that integrates context window compression and structure-aware modeling. The method addresses key challenges in processing long financial documents, including context truncation, information redundancy, and insufficient structural understanding. By introducing an information selection mechanism for input compression and constructing structural graphs to enhance control over discourse logic and semantic relations, the proposed approach improves semantic coverage, logical consistency, and fluency in generated texts. Experimental results show that the method outperforms existing models across multiple ROUGE metrics, confirming its effectiveness in modeling dense and highly structured texts.

During the experiments, this study further investigates the impact of factors such as context window length, redundancy ratio, and subdomain variation on model performance. The results reveal a strong coupling between structural modeling ability and input content quality. By systematically examining the relationships between these factors and model stability and generalization, the study provides theoretical and practical guidance for deploying and tuning models in real-world financial environments. In addition, the visualization of the loss function illustrates the model's convergence process and training stability. This offers intuitive evidence for the controllability and convergence of the proposed design.

Technically, this research extends the adaptability of large language models to long-text tasks in finance. From an application perspective, it offers a feasible model framework and optimization path for key tasks such as intelligent financial document generation, automated report writing, and compliance summarization. The method is particularly suitable for high-reliability generation scenarios such as financial regulation, corporate disclosure, and financial media. It demonstrates strong generalization and practical value. While ensuring textual accuracy, it effectively reduces redundant input and improves generation efficiency. This contributes theoretical and practical support for building high-performance financial language generation systems.

Looking ahead, as financial text types become more diverse and structurally complex, future research should explore how to model cross-document information structures more efficiently. It is also important to investigate the integration of auxiliary sources such as time series and knowledge graphs. These directions are crucial for enhancing a model's comprehensive reasoning ability. In addition, issues of

controllability, safety, and interpretability in generated content need further exploration. This includes improving pretraining mechanisms, structural modeling strategies, and multi-task learning. Future work may consider generation frameworks that incorporate causal reasoning, reinforcement learning, and user feedback signals to build more robust and reliable financial language intelligence systems.

References

[1] Koh H Y, Ju J, Liu M, et al. An empirical survey on long document summarization: Datasets, models, and metrics[J]. ACM computing surveys, 2022, 55(8): 1-35.

[2] Jha S, Erdogan L E, Kim S, et al. Characterizing prompt compression methods for long context inference[J]. arXiv preprint arXiv:2407.08892, 2024.

[3] Dai Z, Yang Z, Yang Y, et al. Transformer-xl: Attentive language models beyond a fixed-length context[J]. arXiv preprint arXiv:1901.02860, 2019.

[4] Cohan A, Dernoncourt F, Kim D S, et al. A discourse-aware attention model for abstractive summarization of long documents[J]. arXiv preprint arXiv:1804.05685, 2018.

[5] Pang B, Nijkamp E, Kryści ń ski W, et al. Long document summarization with top-down and bottom-up inference[J]. arXiv preprint arXiv:2203.07586, 2022.

[6] Liu Y, Zhang J G, Wan Y, et al. HETFORMER: Heterogeneous transformer with sparse attention for long-text extractive summarization[J]. arXiv preprint arXiv:2110.06388, 2021.

[7] Khanna U, Ghodratnama S, Molla D, et al. Transformer-based models for long document summarisation in financial domain[C]//Financial Narrative Processing Workshop (4th: 2022). European Language Resources Association (ELRA), 2022: 73-78.

[8] Liu S, Cao J, Yang R, et al. Long text and multi-table summarization: Dataset and method[J]. arXiv preprint arXiv:2302.03815, 2023.

[9] Lee M, Lay-Ki S. 'Finance Wizard'at the FinLLM Challenge Task: Financial Text Summarization[J]. arXiv preprint arXiv:2408.03762, 2024.

[10] Jindal A K, Rajpoot P K, Parikh A. Upaya at the FinLLM Challenge Task 1 and 2: DistFin: Distillation based Fine-Tuning for Financial Tasks[C]//Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning. 2024: 159-164.

[11] Shukla N K, Prabhakar P, Thangaraj S, et al. GraphRAG Analysis for Financial Narrative Summarization and A Framework for Optimizing Domain Adaptation[C]//Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal). 2025: 23-34.

[12] Tian F, Byadgi A, Kim D S, et al. Customized fingpt search agents using foundation models[C]//Proceedings of the 5th ACM International Conference on AI in Finance. 2024: 469-477.

[13] Wang N, Yang H, Wang C D. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets[J]. arXiv preprint arXiv:2310.04793, 2023.