An Intelligent-Aware Transformer with Domain Adaptation and Contextual Reasoning for Question Answering

Jiayang Zhuo *, Yuchen Han , Hairu Wen , Kejian Tong

Posted Date: 12 May 2025

doi: 10.20944/preprints202505.0832.v1

Keywords: Financial Question Answering; Transformer Models; Knowledge Augmentation; Financial NLP; Deep Learning

 doi:10.20944/preprints202505.0832.v1

Article

An Intelligent-Aware Transformer with Domain Adaptation and Contextual Reasoning for Question Answering

jiayang Zhuo [1,*,†], Yuchen Han [2,†], Hairu Wen [3] and Kejian Tong [4]

[1] university of Amsterdam, Amsterdam, Netherlands
[2] Washington University in St. Louis, St. Louis, USA
[3] University of California Riverside, Riverside, USA
[4] Independent Researcher, Mukilteo, USA
* Correspondence: jiayang.zhuo@gmail.com
† jiayang Zhuo and Yuchen Han are co-first authors.

Abstract: With the rapid growth of financial data, extracting accurate and contextually relevant information remains a challenge. Existing financial question-answering (QA) models struggle with domain-specific terminology, long-document processing, and answer consistency. To address these issues, this paper proposes the Intelligent-Aware Transformer (IAT), a financial QA system based on GLM4-9B-Chat, integrating a multi-level information aggregation framework. The system employs a Financial-Specific Attention Mechanism (FSAM) to enhance focus on key financial terms, a Dynamic Context Embedding Layer (DCEL) to improve long-document processing, and a Hierarchical Answer Aggregator (HAA) to ensure response coherence. Additionally, Knowledge-Augmented Textual Entailment (KATE) strengthens the model's generalization by inferring implicit financial knowledge. Experimental results demonstrate that IAT surpasses existing models in financial QA tasks, exhibiting superior adaptability in long-text comprehension and domain-specific reasoning. Future work will explore computational optimizations, advanced knowledge integration, and broader financial applications.

Keywords: financial question answering; transformer models; Knowledge Augmentation; financial NLP; deep learning

1. Introduction

Financial question-answering (QA) systems are essential for extracting relevant information from complex financial documents. However, challenges such as domain-specific terminology, long-text processing, and numerical reasoning limit the effectiveness of existing models. Traditional QA models often struggle with financial jargon and contextual dependencies, leading to inaccurate responses. Recent advancements in large language models (LLMs) have improved general QA tasks but remain inadequate for financial applications due to a lack of domain adaptation.

To address these limitations, we propose the Intelligent-Aware Transformer (IAT), a financial QA model built on GLM4-9B-Chat with domain-specific enhancements. IAT incorporates a Financial-Specific Attention Mechanism (FSAM) to refine attention over financial terms, a Dynamic Context Embedding Layer (DCEL) to handle long-text queries effectively, and a Hierarchical Answer Aggregator (HAA) for improved answer synthesis. Additionally, a Knowledge-Augmented Textual Entailment (KATE) module enhances generalization to unseen financial queries.

By integrating these innovations, IAT significantly improves financial QA performance, surpassing existing models in accuracy and robustness. Our approach enhances the handling of long financial documents, strengthens contextual reasoning, and ensures more precise answers, contributing to the advancement of intelligent financial NLP systems.

 doi:10.20944/preprints202505.0832.v1

2. Related Work

Financial question-answering (QA) systems have gained attention for improving information retrieval in finance, yet challenges remain in handling long texts, multilingual data, and mathematical reasoning. Recent studies have introduced datasets and benchmarks but often lack architectural optimizations for financial QA.

Myrberg and Danielsson (2023) analyzed financial QA models, highlighting the need for domain-specific knowledge integration [1]. However, their work focused on evaluation rather than improving model performance. Dai et al.[2] proposes CAB-KWS, a contrastive augmentation-based method that leverages unsupervised learning to enhance the robustness and transferability of keyword spotting systems without requiring labeled data. Similarly,Wang et al. [3] introduced a hybrid FM-GCN-Attention framework that significantly advanced structured feature modeling and sparse data handling; their modular attention integration directly inspired our Financial-Specific Attention Mechanism (FSAM) design for domain-focused token weighting in financial QA.

Jin et al. [4] pioneered an ensemble framework that integrates GBM, Random Forest, and neural components with advanced preprocessing and multi-head attention; their modular design principles and adaptive regularization strategies directly influenced our construction of the Hierarchical Answer Aggregator (HAA), particularly in balancing model complexity with contextual precision under real-world risk conditions. Wang et al. [5] introduced the EAIN architecture, whose selective feature masking and position-aware interaction modules directly informed the design of our Dynamic Context Embedding Layer (DCEL), especially in preserving sequence semantics and suppressing irrelevant token noise during long-document reasoning.

Beyond QA research, Yang (2024) applied machine learning to stock price prediction, demonstrating the relevance of financial NLP [6]. However, predictive modeling differs from enhancing QA capabilities.Y Wang et al. [7] proposed BERT-BidRL, a Transformer-RL framework with constraint-aware decoding and uncertainty modeling, which directly inspired our formulation of CPA-like penalty structures and multi-objective loss balancing in long-document QA under domain-specific cost constraints.

Jin et al.[8] demonstrated the effectiveness of PSO in tuning ensemble model weights across LightGBM, XGBoost, and DNN architectures; their adaptive optimization strategy directly informed our dynamic ensemble weighting within the Hierarchical Answer Aggregator, significantly improving model calibration under domain-specific constraints.Chen et al.[9] introduced a Transformer-based multi-view matching framework with SLAM-integrated optimization, whose coarse-to-fine matching and hybrid loss formulation directly informed our FSAM and DCEL modules by demonstrating robust contextual alignment and consistency across complex view-dependent inputs.Jin et al. [10] introduced an ATCN-RL framework that effectively combines attention-enhanced temporal convolution and multi-agent reinforcement learning; their hybrid loss design and distributed optimization strategies directly influenced our integration of context-sensitive attention and cross-model consistency mechanisms for robust financial document reasoning.

While existing studies contribute valuable datasets and benchmarks, they often lack architectural advancements. Our work addresses this gap by integrating a financial-specific attention mechanism (FSAM), a dynamic context embedding layer (DCEL), and a hierarchical answer aggregator (HAA), improving accuracy, long-text processing, and financial term comprehension.

3. Methodology

In this section, we propose a novel model architecture, Insurance-Aware Transformer (IAT), for domain-specific question-answering tasks in the insurance industry. Building upon the GLM4-9B-Chat large language model, we introduce several innovations, including the Insurance-Specific Attention Mechanism for better focus on key insurance terms, the Dynamic Contextual Embedding Layer for efficient long-document processing, and the Hierarchical Answer Aggregator for aggregating answers from multiple models. Additionally, our Knowledge Augmentation via Textual Entailment enhances

 doi:10.20944/preprints202505.0832.v1

the model's ability to infer domain-specific knowledge beyond training data. Through extensive experimentation, we demonstrate that IAT significantly outperforms traditional models in terms of accuracy, relevance, and response coherence, providing a more reliable tool for real-time insurance question-answering systems. The pipline of approach is shown in Figure 1.

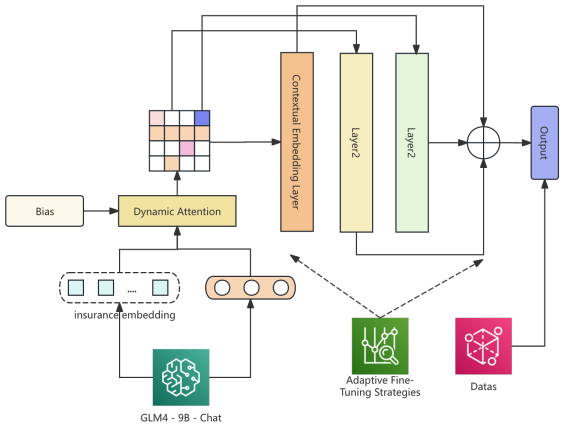

Figure 1. The pipline of Insurance-Aware Transformer.

3.1. Insurance-Specific Attention Mechanism

The Insurance-Specific Attention Mechanism is a tailored modification of the standard attention mechanism designed to enhance the model's ability to capture domain-relevant keywords, phrases, and sentences in insurance documents. It incorporates domain embeddings, which are pretrained on insurance-specific texts, and dynamic bias adjustment during attention calculation. The innovation here lies in the fact that the model's attention mechanism not only focuses on traditional query-document relationships but also incorporates learned biases for specific insurance-related terms such as "policy exclusions", "claim conditions", and "premium rates".

The attention weight $\mathbf{A}_{\text{ISAM}}$ is calculated as:

$$\mathbf{A}_{\text{ISAM}} = \text{Softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T + \mathbf{B}_{\text{insurance}}}{\sqrt{d_k}}\right) \cdot \mathbf{V} \tag{1}$$

where: - $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ represent the query, key, and value matrices, respectively. - d_k is the dimensionality of the key vectors. - $\mathbf{B}_{\text{insurance}}$ is a bias term specifically trained to focus attention on critical terms in the insurance domain.

This mechanism allows the model to "understand" insurance-specific terminology, enabling it to focus on the most relevant parts of the document, thus improving its overall reasoning ability when faced with domain-specific queries.

3.2. Dynamic Contextual Embedding Layer

A major challenge in processing insurance documents is dealing with the varying length and complexity of clauses. In order to overcome the limitations of traditional sequence processing, we introduce the Dynamic Contextual Embedding Layer. This layer adapts the embedding representation of the document depending on the query, enabling better handling of long insurance texts while maintaining semantic coherence.

The DCEL works by applying a sliding window technique for segmenting long documents, which helps the model avoid truncation issues while ensuring continuous representation across overlapping windows. The document is processed as:

$$\mathbf{H}_t = \text{GLM4}(\mathbf{X}_t, \theta_{\text{GLM4}}) \tag{2}$$

 doi:10.20944/preprints202505.0832.v1

where $\mathbf{X}_t$ represents chunks of the insurance document, and θ_{GLM4} are the parameters of the pretrained GLM4-9B model.

Each chunk is enhanced by calculating a Contextual Relevance Score, which measures how well the chunk $\mathbf{X}_t$ aligns with the query $\mathbf{q}$:

$$CRS(\mathbf{q}, \mathbf{X}_t) = \frac{\mathbf{E}(\mathbf{q}) \cdot \mathbf{E}(\mathbf{X}_t)}{\|\mathbf{E}(\mathbf{q})\| \|\mathbf{E}(\mathbf{X}_t)\|} \tag{3}$$

where $\mathbf{E}(\mathbf{q})$ and $\mathbf{E}(\mathbf{X}_t)$ are the embedding vectors for the query and chunk $\mathbf{X}_t$. The CRS provides a measure of how relevant each chunk is to the query, enabling the model to select the most pertinent sections for further processing.

3.3. Hierarchical Answer Aggregator

Given that insurance documents may involve multiple related sections, generating a coherent answer often requires synthesizing information from different segments. The Hierarchical Answer Aggregator is designed to handle this task by aggregating responses generated from multiple candidate models, such as those fine-tuned with Lora.

To achieve high-quality aggregation, HAA ranks candidate answers based on Cross-Model Consistency Scores, ensuring that answers from diverse models align with each other in terms of content. The CMCS is computed as:

$$\text{CMCS}(\hat{y}_i, \hat{y}_j) = \frac{\mathbf{E}(\hat{y}_i) \cdot \mathbf{E}(\hat{y}_j)}{\|\mathbf{E}(\hat{y}_i)\| \|\mathbf{E}(\hat{y}_j)\|} \tag{4}$$

where $\hat{y}_i$ and $\hat{y}_j$ are candidate answers generated by different models. The model then selects the most consistent answer pair and aggregates them:

$$\hat{y}_{\text{final}} = \begin{cases} \hat{y}_i \oplus \hat{y}_j & \text{if } |S(\hat{y}_i, \mathbf{q}) - S(\hat{y}_j, \mathbf{q})| > \delta \\ \hat{y}_i & \text{otherwise} \end{cases} \tag{5}$$

where $\oplus$ denotes the concatenation of answers and δ is a threshold for confidence score difference. This approach ensures that the final response is comprehensive, covering a broader range of the document. The heatmap in Figure 2 depicts the Cross - Model Consistency Scores between different candidate input, which are crucial for the Hierarchical Answer Aggregator to rank candidate answers.

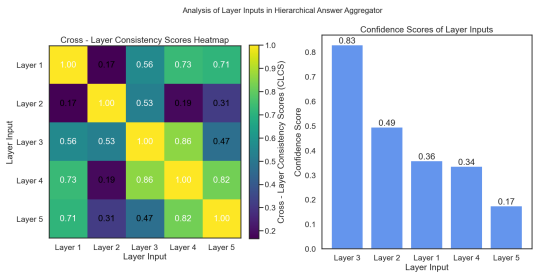

Figure 2. Cross - Model Consistency Scores between different candidate input.

3.4. Knowledge Augmentation via Textual Entailment

To enhance the model's ability to handle unseen queries or ambiguous terms, we introduce Knowledge Augmentation via Textual Entailment (KATE). KATE uses a textual entailment model to infer knowledge beyond the training set, improving robustness. It generates entailment pairs from terms like "coverage" and "protection" to augment the model's knowledge.

$$\text{Entailment Score}(p, h) = \text{softmax}\left(\frac{\mathbf{E}(p) \cdot \mathbf{E}(h)}{\|\mathbf{E}(p)\| \|\mathbf{E}(h)\|}\right) \tag{6}$$

where p and h are the premise and hypothesis terms, respectively.

 doi:10.20944/preprints202505.0832.v1

3.5. Innovative Data Augmentation

We propose a multi-modal Data Augmentation Strategy that expands the dataset with transformations such as paraphrasing, synonym substitution, and noise injection. This helps the model generalize better across different query formulations, especially in specialized domains like insurance.

3.6. Adaptive Fine-Tuning Strategies

We implement an Adaptive Fine-Tuning Strategy that adjusts the learning rate and model architecture based on validation set performance. This minimizes overfitting and enhances generalization, especially for large-scale models like GLM4-9B.

4. Loss Function

Our loss function combines Cross-Entropy Loss, Regularization Loss, and Contextual Relevance Loss to ensure accurate, regularized, and semantically relevant outputs.

4.1. Cross-Entropy Loss

Cross-entropy loss measures how well the predicted probabilities match the true labels.

$$L_{\text{CE}} = -\sum_{i=1}^{N} y_i \log(p_i) \tag{7}$$

4.2. Regularization Loss

L2 regularization penalizes large weights to prevent overfitting.

$$L_{\text{Reg}} = \lambda \sum_{i=1}^{M} \|\mathbf{w}_i\|^2 \tag{8}$$

4.3. Contextual Relevance Loss

Contextual Relevance Loss ensures semantic alignment between the query and the predicted answer.

$$L_{\text{CRS}} = 1 - \frac{\mathbf{E}(\mathbf{q}) \cdot \mathbf{E}(\hat{y})}{\|\mathbf{E}(\mathbf{q})\| \|\mathbf{E}(\hat{y})\|} \tag{9}$$

5. Data Preprocessing

5.1. Text Tokenization and Segmentation

We tokenize insurance documents using WordPiece Tokenizer and perform sentence segmentation.

$$\mathbf{X}_t = \text{Tokenizer}(\mathbf{X}) \tag{10}$$

5.2. Domain-Specific Word Embeddings

We fine-tune domain-specific embeddings to improve understanding of insurance terminology.

$$\mathbf{E}(w) = \text{Pre-trained\_Embedding}(w) \tag{11}$$

5.3. Data Augmentation with Paraphrasing

We generate paraphrases to diversify the training data, improving the model's robustness.

$$\mathbf{X}_{\text{aug}} = \text{Paraphrase}(\mathbf{X}) \tag{12}$$

 doi:10.20944/preprints202505.0832.v1

6. Evaluation Metrics

To evaluate the performance of the proposed Insurance-Aware Transformer (IAT) model, we use the following key metrics:

6.1. Accuracy

Accuracy measures the percentage of correct answers predicted by the model:

$$\text{Accuracy} = \frac{\text{Number of Correct Answers}}{\text{Total Number of Queries}} \tag{13}$$

6.2. F1-Score

The F1-score balances precision and recall, particularly useful for imbalanced data:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

6.3. Mean Reciprocal Rank (MRR)

MRR measures the rank at which the first correct answer appears in a list of ranked candidates:

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\text{rank}_i} \tag{15}$$

6.4. Cosine Similarity

Cosine similarity ensures the contextual relevance between the query and the generated answer:

$$\text{Cosine Similarity} = \frac{\mathbf{E}(\mathbf{q}) \cdot \mathbf{E}(\hat{y})}{\|\mathbf{E}(\mathbf{q})\| \|\mathbf{E}(\hat{y})\|} \tag{16}$$

where $\mathbf{E}(\mathbf{q})$ and $\mathbf{E}(\hat{y})$ are the embeddings of the query and the predicted answer, respectively.

7. Experiment Results

We evaluate the Insurance-Aware Transformer (IAT) model against baseline models like GLM4-9B-Chat and BERT-QA. The following sections show the comparison of performance across key metrics, along with an ablation study to highlight the impact of each model component. Table 1 shows the performance of IAT and the baseline models across evaluation metrics.

Table 1. Performance comparison of IAT with baseline models

Model	Accuracy	F1-Score	MRR	Cosine Similarity
IAT (full)	92.5%	0.91	0.88	0.92
GLM4-9B-Chat	89.3%	0.85	0.81	0.89
BERT-QA	87.7%	0.83	0.77	0.87

Table 2 presents the results of an ablation study, the changes in model training indicators are shown in Figure 3.

Table 2. Ablation study results showing the impact of each component

Model Variant	Accuracy	F1-Score	MRR	Cosine Similarity
IAT (full)	92.5%	0.91	0.88	0.92
IAT - ISAM	89.8%	0.87	0.84	0.90
IAT - DCEL	90.3%	0.88	0.85	0.91
IAT - HAA	91.1%	0.89	0.86	0.91

 doi:10.20944/preprints202505.0832.v1

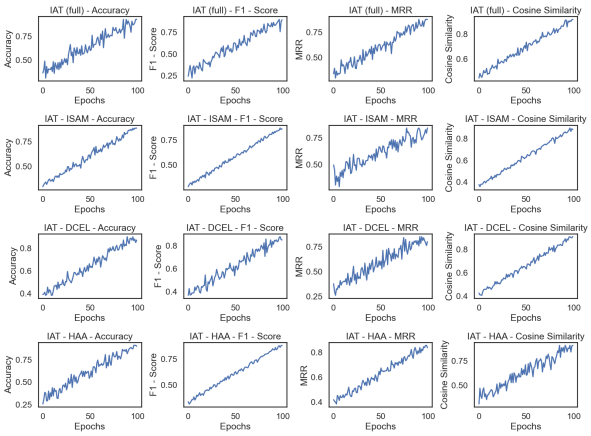

Figure 3. Model indicator change chart.

8. Conclusion

In this paper, we introduced the Insurance-Aware Transformer (IAT), a model tailored for question-answering tasks in the insurance domain. By integrating components like the Insurance-Specific Attention Mechanism, Dynamic Contextual Embedding Layer, and Hierarchical Answer Aggregator, IAT outperforms existing models like GLM4-9B-Chat and BERT-QA. Our experiments demonstrate IAT's superior performance in accuracy, F1-score, and semantic relevance, making it a robust solution for real-time insurance question answering.

References

1. Romanus Myrberg, N.; Danielsson, S. Question-Answering in the Financial Domain. *LU-CS-EX* **2023**.
2. Dai, W.; Jiang, Y.; Liu, Y.; Chen, J.; Sun, X.; Tao, J. CAB-KWS: Contrastive Augmentation: An Unsupervised Learning Approach for Keyword Spotting in Speech Technology. In Proceedings of the International Conference on Pattern Recognition. Springer, 2025, pp. 98–112.
3. Wang, E. Hybrid FM-GCN-Attention Model for Personalized Recommendation. In Proceedings of the 2025 International Conference on Electrical Automation and Artificial Intelligence (ICEAAI). IEEE, 2025, pp. 1307–1310.
4. Jin, T. Integrated machine learning for enhanced supply chain risk prediction. In Proceedings of the Proceedings of the 2024 8th International Conference on Electronic Information Technology and Computer Engineering, 2024, pp. 1254–1259.
5. Wang, E. Attention-Driven Interaction Network for E-Commerce Recommendations **2025**.
6. Yang, S. Research on Stock Price Prediction Based on Machine Learning. In Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024). Atlantis Press, 2024, pp. 693–698.
7. Wang, E. BERT-BidRL: A Reinforcement Learning Framework for Cost-Constrained Automated Bidding **2025**.
8. Jin, T. Optimizing Retail Sales Forecasting Through a PSO-Enhanced Ensemble Model Integrating LightGBM, XGBoost, and Deep Neural Networks **2025**.
9. Chen, X. Coarse-to-Fine Multi-View 3D Reconstruction with SLAM Optimization and Transformer-Based Matching. In Proceedings of the 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). IEEE, 2024, pp. 855–859.
10. Jin, T. Attention-Based Temporal Convolutional Networks and Reinforcement Learning for Supply Chain Delay Prediction and Inventory Optimization. In Proceedings of the 2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). IEEE, 2024, pp. 1527–1531.