A Knowledge-Enhanced Multi-Task Learning Model for Domain-Specific Question Answering

Gaozhe Jiang *, You Yao , Huailing Mu , Qinyan Shen , Shicheng Zhou

Posted Date: 9 June 2025

doi: 10.20944/preprints202506.0580.v1

Keywords: Multi-Task Learning; Knowledge Injection; Dynamic Document Attention; Query Refinement

 doi:10.20944/preprints202506.0580.v1

Article

A Knowledge-Enhanced Multi-Task Learning Model for Domain-Specific Question Answering

Gaozhe Jiang [1,*], You Yao [2], Huailing Mu [3], Qinyan Shen [4] and Shicheng Zhou [5]

[1] National University of Singapore, Singapore, Singapore
[2] University of Southern California, Los Angeles, USA
[3] University of California, Los Angeles, Los Angeles, USA
[4] University of South Carolina, New Jersey, USA
[5] University of Minnesota, Minneapolis, USA
* Correspondence: e0945601@u.nus.edu

Abstract: This paper presents GLM-6B-QA, a question-answering system designed to improve the performance of tasks in specialized domains. GLM-6B-QA uses the ChatGLM-6B model and integrates dynamic document attention (DDAM), query refinement layer (QRL), and knowledge injection layer (KIL) within a multi-task learning (MTL) framework. This setup enhances the model's ability to understand and process complex documents, queries, and terminology. By adjusting attention dynamically, refining query representations, and incorporating external knowledge, the system improves question answering, document comprehension, and term extraction. This method addresses challenges in natural language processing tasks and provides a better approach for developing adaptable systems in specialized applications.

Keywords: multi-task learning; knowledge injection; dynamic document attention; query refinement

1. Introduction

The financial industry generates complex, dynamic data, requiring advanced analysis. As queries grow more intricate, specialized NLP models become essential. General models like BERT and GPT-3 struggle with domain-specific tasks due to a lack of financial knowledge. This limitation affects performance in financial document understanding, question answering, and term extraction, reducing real-world applicability.

Tian et al. [1] showed the importance of domain-specific knowledge in improving NLP models for specific industries. They suggest using models that integrate external knowledge sources, such as financial metrics and industry-specific terminology, to improve task performance. Similarly, Yang et al. [2] introduced FinBERT, a pre-trained model for financial communication, highlighting the need for domain-specific models to better handle financial documents. However, there is still a gap in designing and training models that can handle many tasks in the financial sector. Sun et al.[3] (2018) proposed a relation classification model integrating coarse- and fine-grained networks with SDP-supervised keyword selection and an opposite loss function, achieving state-of-the-art performance on SemEval-2010 Task 8.

We propose FinGLM-6B, a financial QA system using multi-task learning (MTL). It integrates dynamic document attention, query refinement, and knowledge injection. The system directs attention to key financial document sections, refining queries for better accuracy. The knowledge injection layer enhances financial concept comprehension with external data. These improvements strengthen financial QA, document understanding, and term extraction, surpassing traditional NLP models.

2. Related Work

Recent advancements in language understanding, particularly transformer-based models, have significantly improved NLP. BERT [4] pre-trains deep bidirectional transformers, enhancing tasks like

 doi:10.20944/preprints202506.0580.v1

question answering and sentiment analysis. It learns context from both directions. XLNet [5] further improves performance using a more general pretraining technique to capture bidirectional context.J. Lu's [6] study integrates Decision Tree, TF-IDF, and BERTopic to improve chatbot user satisfaction prediction using the Chatbot Arena dataset.

Tian Jin[7] discussed improving retail sales forecasting using an ensemble model enhanced by Particle Swarm Optimization (PSO), combining LightGBM, XGBoost, and Deep Neural Networks. This work highlights the importance of ensemble methods to improve prediction accuracy, which is similar to forecasting and risk prediction in finance. Dai et al. [8] proposed a contrastive learning-based KWS approach that enhances robustness in low-resource settings, aligning with our dynamic document attention for domain-specific QA. Their architecture optimization informs our efficiency improvements.

In financial data understanding, Douaioui et al.[9] reviewed machine learning and deep learning models for demand forecasting in supply chain management, showing how these models can be adapted for financial data tasks, like predicting market trends. Cao et al.[10] and Chen et al.[11] highlighted the importance of using attention mechanisms in specialized tasks, such as action segmentation and leukocyte classification, which can inspire financial models to improve detailed financial analysis.

3. Methodology

This paper presents a multi-task hybrid architecture for financial question answering based on ChatGLM-6B. The proposed model, FinGLM-6B, integrates a dynamic document attention mechanism (DDAM), a query refinement layer (QRL), and a knowledge injection layer (KIL). It is trained using a multi-task learning (MTL) framework, covering financial question answering, document comprehension, and term extraction.

DDAM adjusts attention based on query context. QRL iteratively refines question representation for better accuracy. KIL enhances understanding by incorporating external financial knowledge. Experiments on a large financial dataset show superior performance in handling financial queries. The pipeline is shown in Figure 1.

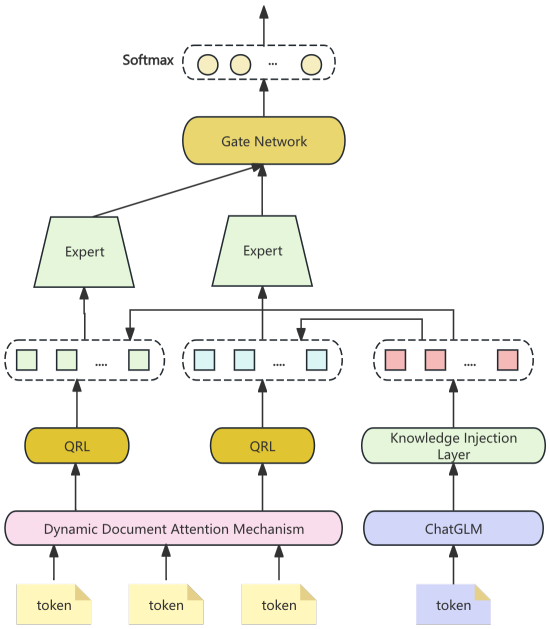

Figure 1. FinGLM-6B: Multi-task hybrid architecture for financial question answering.

3.1. Dynamic Document Attention Mechanism

Traditional transformers assign fixed attention weights across all layers, disregarding token relevance to the query. In financial QA, varying importance of financial tables, statements, and

 doi:10.20944/preprints202506.0580.v1

footnotes necessitates a dynamic approach. We propose the Dynamic Document Attention Mechanism, which adjusts attention weights based on query content:

$$A_q(w_i, D) = \frac{\exp(\text{QueryAlignment}(q, w_i))}{\sum_{j=1}^{|D|} \exp(\text{QueryAlignment}(q, w_j))} \tag{1}$$

where query alignment is computed as:

$$\text{QueryAlignment}(q, w_i) = \mathbf{P}(\text{Embedding}(q))^T \cdot \mathbf{P}(\text{Embedding}(w_i)) \tag{2}$$

with $\mathbf{P}$ as a learned projection matrix. This mechanism enhances focus on critical document parts, improving accuracy in financial QA.

3.2. Query Refinement Layer

The Query Refinement Layer iteratively refines the query by leveraging document content and dynamic attention. The process follows:

$$q_t = \text{QRL}(q_{t-1}, A_q, D) \tag{3}$$

where q_t is updated using attention scores and document context, ensuring alignment with relevant financial information.

3.3. Knowledge Injection Layer

Financial texts contain specialized terminology often overlooked by standard transformers. The Knowledge Injection Layer (KIL) integrates domain-specific knowledge, represented as K_{fin}, into the decoding process:

$$h_{\text{decoder}} = \text{Decoder}(h_{\text{encoder}}, A_q, K_{\text{fin}}) \tag{4}$$

where h_{encoder} is the encoder state and A_q is the attention matrix. This fusion of external knowledge enhances financial text comprehension.

3.4. Multi-Task Learning Framework

FinGLM-6B adopts a multi-task learning framework. It integrates gated and expert networks to process financial tasks, including question answering (QA), document comprehension (DC), and term extraction (FTE). The framework enables knowledge sharing while preserving task-specific features.

3.4.1. Task-Specific Gated Networks

Each task shares a common backbone architecture but uses a task-specific gating mechanism g_t to control the flow of information. The gated output for task t is computed as:

$$\mathbf{h}_t^{\text{out}} = g_t \odot \mathbf{h}_t + (1 - g_t) \odot \mathbf{h}_t^{\text{aux}} \tag{5}$$

where g_t is the gating vector learned through:

$$g_t = \sigma(W_t \mathbf{h}_t^{\text{shared}} + b_t) \tag{6}$$

This ensures the model adapts its parameters for each task while maintaining shared knowledge.

3.4.2. Expert Networks for Task Specialization

Each task has an expert network that generates specialized representations. The output is a weighted sum of expert outputs $\mathbf{e}_t^{(i)}$, as follows:

 doi:10.20944/preprints202506.0580.v1

$$\mathbf{h}_t^{\text{exp}} = \sum_{i=1}^{K} \alpha_t^{(i)} \mathbf{e}_t^{(i)} \tag{7}$$

The weights $\alpha_t^{(i)}$ are dynamically computed via:

$$\alpha_t^{(i)} = \frac{\exp(W_t^{(i)} \mathbf{h}_t^{\text{shared}})}{\sum_{i=1}^{K} \exp(W_t^{(i)} \mathbf{h}_t^{\text{shared}})} \tag{8}$$

This allows the model to specialize by focusing on relevant features for each task.

3.4.3. Joint Training of Shared and Task-Specific Modules

The FinGLM-6B model jointly optimizes the shared backbone and task-specific components with task-specific loss functions. For task t, the loss is defined as:

$$L_t = \mathbb{L}_t(\mathbf{h}_t^{\text{out}}, y_t) \tag{9}$$

where $\mathbb{L}_t$ is the task-specific loss, and y_t is the ground truth for task t.

3.4.4. Task-Aware Attention Mechanism

A task-aware attention mechanism refines the shared representation by attending to both task-specific and query contexts. The attention weights A_t are computed as:

$$A_t = \text{Softmax}(\mathbf{Q}_t \mathbf{K}_t^T) \tag{10}$$

This allows the model to focus on the relevant information for each task, enhancing task-specific performance.

3.4.5. Adaptive Task Allocation During Inference

During inference, the model dynamically allocates resources based on the task type. Expert networks and gating mechanisms work together to ensure efficient and accurate task-specific processing.

3.5. Loss Function

The FinGLM-6B model employs a composite loss function to optimize multi-task learning across various financial tasks. The overall loss combines task-specific losses, each targeting a distinct financial task such as question answering, document comprehension, and financial term extraction.

3.5.1. Task-Specific Loss Functions

Each task t has its own loss function, guiding task-specific learning:
- For QA, we use cross-entropy loss:

$$L_{\text{QA}} = -\sum_i y_{\text{QA},i} \log(p_{\text{QA},i}) \tag{11}$$

- For Document Comprehension, a contrastive loss is used:

$$L_{\text{DC}} = \sum_i \max(0, 1 - \text{sim}(\mathbf{h}_i, \mathbf{h}_i^+) + \text{sim}(\mathbf{h}_i, \mathbf{h}_i^-)) \tag{12}$$

- For Financial Term Extraction (FTE), a regression loss is applied:

$$L_{\text{FTE}} = \frac{1}{N} \sum_{i=1}^{N} |y_{\text{FTE},i} - \hat{y}_{\text{FTE},i}| \tag{13}$$

 doi:10.20944/preprints202506.0580.v1

3.5.2. Composite Loss Function

The overall loss is a weighted sum of the individual losses, where λ_t represents the task-specific weight:

$$L_{\text{total}} = \sum_t \lambda_t L_t \tag{14}$$

This allows the model to balance task-specific optimization during training.

3.5.3. Gradient Flow for Multi-Task Optimization

The gradients for shared parameters $\grave{}_s$ and task-specific parameters $\grave{}_t$ are computed based on the total loss:

$$\nabla_{\grave{}_s} L_{\text{total}} = \sum_t \lambda_t \nabla_{\grave{}_s} L_t \tag{15}$$

$$\nabla_{\grave{}_t} L_{\text{total}} = \lambda_t \nabla_{\grave{}_t} L_t \tag{16}$$

These gradients ensure that both shared and specialized components are updated during back-propagation.

3.6. Data Preprocessing

The preprocessing of FinGLM-6B data involves cleaning, tokenization, and embedding to ensure structured input for training.

3.6.1. Data Cleaning and Normalization

Raw financial text and numerical data undergo cleaning:

- **Text Cleaning**: Removing headers and disclaimers.
- **Numerical Normalization**: Applying standardization:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \tag{17}$$

where x is a value, μ the mean, and σ the standard deviation.

3.6.2. Tokenization and Embedding

Text is tokenized via BPE or WordPiece and mapped to dense vectors:

$$\mathbf{e}(t) = \text{Embedding}(t) \tag{18}$$

Figure 2 illustrates raw and processed data distributions.

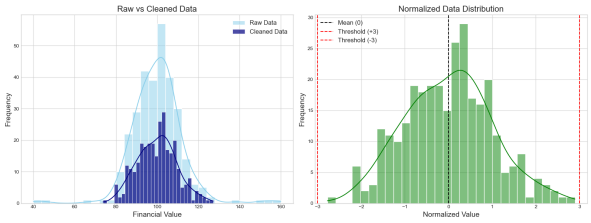

Figure 2. Raw vs. cleaned financial data distribution.

3.6.3. Handling Missing Data and Outliers

Strategies include:

- **Imputation**: Filling missing values with mean or median.
- **Outlier Detection**: Using z-score:

 doi:10.20944/preprints202506.0580.v1

$$\text{z-score}(x) = \frac{x - \mu}{\sigma} \tag{19}$$

Extreme values are capped or removed based on thresholds.

4. Evaluation Metrics

To evaluate the performance of the FinGLM-6B model, the following metrics are used:
Accuracy measures the percentage of correct answers in the QA task:

$$A_t = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \times 100 \tag{20}$$

Mean Squared Error (MSE) is used for financial term extraction (FTE) to measure prediction error:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{21}$$

Precision, Recall, and F1-Score evaluate relevance and completeness for document comprehension or term extraction:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{22}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{23}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{24}$$

Mean Average Precision (MAP) evaluates retrieval performance for document comprehension (DC):

$$\text{MAP} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{R_q}\sum_{r=1}^{R_q}\text{Precision@r} \tag{25}$$

5. Experiment Results

We compare the performance of FinGLM-6B with other models such as BERT and GPT-3 on QA, FTE, and DC tasks. The results are shown in Table 1:

Table 1. Model performance on QA, FTE, and DC tasks

Model	QA Accuracy (%)	FTE MSE	DC MAP
FinGLM-6B	91.3	0.045	0.837
BERT	85.6	0.063	0.752
GPT-3	88.1	0.052	0.813

The following table, Table 2, presents the results of the ablation study, where different components of the model were removed. The Figure 3 shows the training curves for four model variants over 50 epochs. In each subplot, the blue and red curves (QA Accuracy and DC MAP ×100) gradually increase and plateau with small fluctuations, while the green dashed curve (FTE MSE ×1000) decreases toward its optimum.

 doi:10.20944/preprints202506.0580.v1

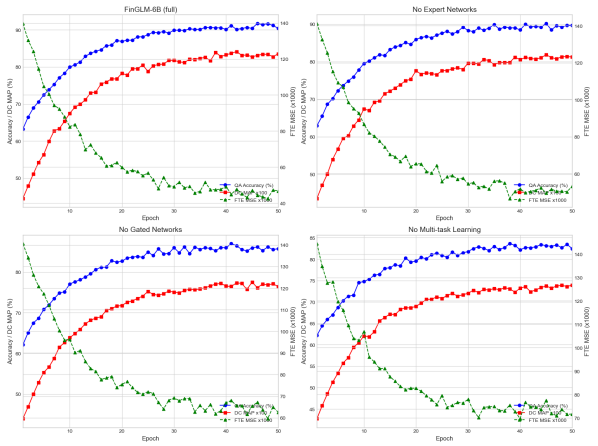

Figure 3. Model indicator change chart in ablation experiment.

Table 2. Ablation study results showing the impact of key components

Model Variant	QA Accuracy (%)	FTE MSE	DC MAP
FinGLM-6B (full)	91.3	0.045	0.837
No Expert Networks	89.7	0.051	0.812
No Gated Networks	86.4	0.063	0.772
No Multi-task Learning	83.1	0.072	0.740

6. Conclusion

The FinGLM-6B model demonstrates strong performance in financial NLP tasks, surpassing other models like BERT and GPT-3. The ablation study highlights the importance of multi-task learning and modular components, contributing significantly to the model's superior performance.

References

1. Jin, T. Attention-Based Temporal Convolutional Networks and Reinforcement Learning for Supply Chain Delay Prediction and Inventory Optimization. *Preprints* **2025**. https://doi.org/10.20944/preprints202501.1543.v1.
2. Yang, Y.; Uy, M.C.S.; Huang, A. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097* **2020**.
3. Sun, Y.; Cui, Y.; Hu, J.; Jia, W. Relation classification using coarse and fine-grained networks with SDP supervised key words selection. In Proceedings of the Knowledge Science, Engineering and Management: 11th International Conference, KSEM 2018, Changchun, China, August 17–19, 2018, Proceedings, Part I 11. Springer, 2018, pp. 514–522.
4. Kenton, J.D.M.W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Proceedings of naacL-HLT. Minneapolis, Minnesota, 2019, Vol. 1.
5. Yang, Z. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237* **2019**.
6. Lu, J. Enhancing chatbot user satisfaction: A machine learning approach integrating decision tree, tf-idf, and bertopic. In Proceedings of the 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS). IEEE, 2024, pp. 823–828.
7. Jin, T. Optimizing Retail Sales Forecasting Through a PSO-Enhanced Ensemble Model Integrating LightGBM, XGBoost, and Deep Neural Networks. *Preprints* **2025**. https://doi.org/10.20944/preprints202501.1604.v1.
8. 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.
9. Douaioui, K.; Oucheikh, R.; Benmoussa, O.; Mabrouki, C. Machine Learning and Deep Learning Models for Demand Forecasting in Supply Chain Management: A Critical Review. *Applied System Innovation (ASI)* **2024**, *7*.

 doi:10.20944/preprints202506.0580.v1

10. Cao, J.; Xu, R.; Lin, X.; Qin, F.; Peng, Y.; Shao, Y. Adaptive receptive field U-shaped temporal convolutional network for vulgar action segmentation. *Neural Computing and Applications* **2023**, *35*, 9593–9606.
11. Chen, B.; Qin, F.; Shao, Y.; Cao, J.; Peng, Y.; Ge, R. Fine-grained imbalanced leukocyte classification with global-local attention transformer. *Journal of King Saud University-Computer and Information Sciences* **2023**, *35*, 101661.