FinMoE: A MoE-based Large Chinese Financial Language Model

Xuanyu Zhang
Du Xiaoman Financial
Beijing, China
zhangxuanyu@duxiaoman.com

Qing Yang
Du Xiaoman Financial
Beijing, China
yangqing@duxiaoman.com

Abstract

Large-scale language models have demonstrated remarkable success, achieving strong performance across a variety of general tasks. However, when applied to domain-specific fields, such as finance, these models face challenges due to the need for both specialized knowledge and robust general capabilities. In this paper, we introduce FinMoE, a MOE-based large-scale Chinese financial language model that bridges the gap between general language models and domain-specific requirements. FinMoE employs a dense MoE architecture, where all expert networks are simultaneously activated and dynamically combined to effectively integrate general linguistic understanding with domain-specific financial expertise. Experimental results demonstrate that FinMoE achieves state-of-the-art performance on both general-purpose and financial benchmarks at a comparable scale, validating its ability to balance domain specialization with general knowledge and reasoning.

1 Introduction

In recent years, large-scale language models such as InstructGPT, ChatGPT, and GPT-4 (Radford et al., 2018; Ouyang et al., 2022; OpenAI, 2022, 2023) have demonstrated impressive advancements in conversational and generative AI. These models excel in understanding complex language structures and engaging in natural, coherent interactions, pushing the boundaries of natural language processing (NLP) applications. Their development has profoundly influenced industries, enabling innovations in areas such as virtual assistants and intelligent customer support systems. However, applying such models effectively to domain-specific contexts, such as finance, remains a challenging task due to the need for domain expertise and specialized data.

A significant challenge in building a financial large language model lies in achieving strong financial task performance while maintaining robust general capabilities. General-purpose models often lack sufficient domain-specific knowledge to address the intricacies of financial problems effectively. At the same time, models tailored exclusively to finance risk losing their ability to perform well on tasks requiring broader reasoning and general linguistic understanding. Therefore, there is a critical need for financial models that combine strong general knowledge with specialized financial capabilities to ensure they can handle complex real-world financial scenarios effectively.

Moreover, answering financial questions often requires integrating financial concepts, domain-specific methodologies, and general reasoning frameworks. Financial tasks are not only about understanding specialized terminology but also about applying general reasoning skills, contextual awareness, and common-sense knowledge to solve problems. This interplay between domain-specific and general-purpose knowledge is essential for tasks such as financial analysis, decision support, and risk assessment, where precise and reliable insights are paramount.

To address these challenges, we propose FinMoE, a large-scale Chinese financial language model built on a dense Mixture-of-Experts (MoE) architecture (Jordan and Jacobs, 1994; Collobert et al., 2001; Ma et al., 2018). Unlike conventional models, FinMoE employs a dense MoE approach where all expert networks are activated simultaneously during each forward pass, and their outputs are dynamically combined through a weighted summation mechanism. This design ensures that FinMoE leverages both domain-specific and general-purpose knowledge, effectively capturing the intricate relationships between financial concepts and broader reasoning capabilities.

By combining financial expertise with robust general abilities, FinMoE bridges the gap between general and financial domain requirements. Built

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 48–53
January 19–20, 2025.

upon a dense MOE architecture and carefully designed training strategies, FinMoE achieves state-of-the-art performance on both general-purpose and financial-specific benchmarks at a comparable scale. This demonstrates its ability to effectively balance domain-specific knowledge and general reasoning capabilities, providing accurate and reliable insights for financial institutions, investors, and researchers.

2 Related Work

The Mixture of Experts (MoE) architecture (Jacobs et al., 1991; Cai et al., 2024) has a long and storied history in the field of deep learning, dating back to its introduction as a method to enhance predictive performance by combining multiple expert models. This architecture was initially conceived as a way to address the limitations of single-model approaches, which often struggle to capture the complexity and diversity of real-world data. At its core, the MoE framework is built on the principle of specialization, where each expert network is designed to focus on specific aspects of the data or tasks at hand. This modular approach allows the model to leverage the strengths of multiple specialized networks, each contributing unique insights to the overall prediction process.

2.1 Composition of MOE

The MoE architecture consists of two key components: the expert networks and the gating network. The expert networks are the backbone of the system, each possessing specialized knowledge that allows it to excel in a particular domain or task. The gating network, on the other hand, plays a crucial role in orchestrating the interaction between the experts and the input data. Its primary function is to intelligently route the input to the most appropriate expert network based on the characteristics of the input. The combination of specialized expert networks and an intelligent gating mechanism allows the MoE architecture to handle diverse inputs and tasks with remarkable flexibility.

2.2 Sparse MoE

The sparse MoE model (Shazeer et al., 2017) is a common type of MoE model. It activates only a small portion of experts in each forward pass, thus significantly reducing the computational load. This model typically uses a top-k gating mechanism to select the most relevant experts, where k is a relatively small integer. For example, the Switch Transformer (Fedus et al., 2022) successfully expanded the model parameters to trillions while maintaining computational efficiency by sparsely activating experts when processing large-scale language models. However, the sparse MoE model also faces some challenges. For instance, there are issues with training stability. Due to the unbalanced load of experts, some experts may be overused while others are underutilized, which can affect the model's performance and generalization ability. Additionally, the non-uniformity of sparse operations on hardware accelerators makes it difficult to fully realize the theoretically computational efficiency advantages in practical applications.

2.3 Dense MoE

In contrast to the sparse MoE, the dense MoE model (Nie et al., 2021; Wu et al., 2024) activates all experts in each forward pass and then combines their outputs through weighted summation. This approach ensures that every expert contributes to the final output, leveraging the collective knowledge of all specialized networks. Although this method has a relatively large computational cost, it can provide higher prediction accuracy in some cases, particularly when the task requires a comprehensive understanding of diverse data domains or when the model needs to handle complex, multifaceted problems. Additionally, the stability of dense MoE during training is another significant advantage. By activating all experts uniformly, the model avoids the potential instability caused by uneven expert utilization in Sparse MoE, ensuring more reliable performance.

In the financial domain, the dense MoE model offers several unique advantages. Financial tasks often require a deep understanding of both general knowledge and domain-specific expertise. The dense MoE model is particularly well-suited for this dual requirement, as it allows the integration of specialized financial knowledge with broader, general-purpose capabilities. This combination enables the model to handle the complexity and diversity of financial tasks more effectively, making it a powerful tool for financial modeling.

3 Model Structure

In this section, we describe the architecture of FinMoE to address the unique challenges of financial tasks. Unlike traditional sparse MoE approaches,

FinMoE adopts a dense MoE structure, which activates all experts simultaneously during each forward pass and combines their outputs through a weighted summation. This design choice ensures that every expert contributes to the final prediction, leveraging the capabilities of different experts in general and financial fields. The dense MoE model in FinMoE can be formally described as:

$$M(\mathbf{x}; \theta, \{W_i\}_{i=1}^{N}) = \sum_{i=1}^{N} G(\mathbf{x}; \theta)_i f_i(\mathbf{x}; W_i), \quad (1)$$

where $f_i(\mathbf{x}; W_i)$ is the gating value produced by a gating network parameterized by W_i, and $G(\mathbf{x}; \theta)_i$ denotes the gating weight assigned to the i-th of N experts. The gating weight is obtained through the softmax operation over the gating values $g(\mathbf{x}; \theta)$, as defined below:

$$G(\mathbf{x}; \theta)_i = \frac{\exp(g(\mathbf{x}; \theta)_i)}{\sum_{j=1}^{N} \exp(g(\mathbf{x}; \theta)_j)}, \quad (2)$$

where $g(\mathbf{x}; \theta)$ is the gating value produced by a gating network parameterized by θ. In this formulation, the gating network dynamically assigns weights to the outputs of all experts based on the input $\mathbf{x}$, ensuring that all experts contribute to the final prediction in a weighted manner.

Specifically, each expert network f_i in FinMoE adopts the same multi-layer MLP architecture as the attention block. These expert networks are parameterized independently by W_i and share a common input $\mathbf{x}$, producing the corresponding output $f_i(\mathbf{x}; W_i)$. By activating all experts simultaneously, FinMoE ensures that both domain-specific and general-purpose knowledge are combined effectively during each forward pass.

The gating network G, parameterized by θ, plays a crucial role in MOE. It determines the contribution of each expert network to the final output through a softmax-based gating mechanism. Given an input $\mathbf{x}$, the gating network first generates a gating value $g(\mathbf{x}; \theta)$, which is then passed through a softmax function to produce the gating weights $G(\mathbf{x}; \theta)_i$. These weights determine how much emphasis each expert network f_i receives during the summation. The gating network consists of a linear layer and a softmax layer, making it computationally efficient and effective in dynamically adjusting the expert contributions based on the input. This dynamic gating mechanism enables FinMoE to adaptively integrate the outputs of all experts, ensuring that the model can effectively capture the complex relationships present in financial and general data.

4 Model Training

4.1 Pre-training

Large-scale pretraining is fundamental to building high-performing language models, as it allows the model to acquire a general understanding of language and knowledge representations through unsupervised learning. For FinMoE, we adopt an autoregressive language modeling approach, where the model predicts the next token given the previous tokens in a sequence. Formally, the joint probability of tokens in a text is expressed as:

$$p(\mathbf{x}) = p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t | x_{<t}) \quad (3)$$

where $\mathbf{x}$ represents the input, x_t is the t^{th} token, and $x_{<t}$ represents all preceding tokens. T denotes the total number of tokens in the sequence.

We adopt the decoder-only architecture of LLaMA (Touvron et al., 2023a,b) rather than encoder architecture (Zhang et al., 2023b; Zhang and Yang, 2021a), which has been widely recognized for its efficiency and effectiveness in large language models. To incorporate positional information, we utilize RoPE (Su et al., 2024) as a position embedding technique. The activation function employed in our model is SwiGLU (Shazeer, 2020), and we use RMSNorm (Zhang and Sennrich, 2019) for normalization purposes. The pretraining corpus includes both general-domain data and a substantial amount of financial-domain data, which enables FinMoE to build strong general-purpose language capabilities while simultaneously learning specialized financial knowledge.

4.2 Hybrid-turning

The finetuning phase plays a crucial role in aligning the pretrained FinMoE model with task-specific instructions and domain knowledge. We employ a hybrid-tuning strategy following XuanYuan (Zhang and Yang, 2023c), which addresses limitations observed in conventional two-stage domain-specific training methods. Specifically, we construct a unified dataset by randomly shuffling pretraining data and supervised fine-tuning instruction data. The pretraining data includes both general-domain and financial-domain corpora, while the instruction data consists of general instruction data and financial instruction data.

Model	Size	Language	Knowledge	Reasoning	Subject	Code	Finance	Average
BlueLM	7B	66.4	66.4	52.9	54.4	20.7	55.3	52.7
Yi	6B	62.9	67.6	51	61.3	19.4	62.4	54.1
Qwen	7B	79	67.6	59.1	56.7	30.2	52.4	57.5
FinMOE	7B	76.7	70.6	58.5	68.5	20.7	80	62.5

Table 1: Results of different large language models.

To generate high-quality instruction-tuning data, we leverage Self-QA (Zhang and Yang, 2023b), a method that addresses the challenges of constructing supervised fine-tuning datasets. Unlike approaches such as Self-Instruct (Wang et al., 2022), which rely on a small set of manually created seed instructions, Self-QA generates instruction-tuning data from large-scale unsupervised knowledge sources. This method not only reduces the reliance on human annotation but also enables the generation of accurate and diverse customized instruction data tailored to the financial domain.

5 Experiments

5.1 Datasets

To evaluate our model, we constructed a comprehensive benchmark that includes both general and financial scenarios. The evaluation set consists of six main categories: Language, Knowledge, Reasoning, Subject, Code, and Finance, each containing multiple sub-datasets. For example, the Knowledge category comprises datasets like CommonsenseQA (Talmor et al., 2018), TriviaQA (Joshi et al., 2017), and OpenbookQA (Mihaylov et al., 2018), which assess the model's ability to apply general world knowledge and commonsense question answering (Zhang, 2019; Zhang and Yang, 2021b; Zhang, 2020; Zhang and Wang, 2020; Zhang and Yang, 2023a). The Finance category includes datasets like FinanceIQ and CGCE (Zhang et al., 2023a), which test the model's financial reasoning and understanding in a domain-specific context. Each of these sub-datasets is designed to evaluate different capabilities of the model, allowing for a thorough and multi-dimensional assessment across a range of tasks and domains.

5.2 Results

We compare our model, FinMOE, with several baseline models across multiple domains mentioned above. The models evaluated include BlueLM, Yi (Young et al., 2024), and Qwen (Bai et al., 2023). These models are chosen for their strong performance in recent evaluations, representing state-of-the-art architectures at a comparable scale. As shown in Table 1, FinMOE achieves strong performance, particularly in the Finance domain, where it outperforms all other 6B or 7B models with a score of 80. In comparison, FinMOE generally demonstrates a more balanced and robust performance across domains than the other models, especially when it comes to tasks requiring financial expertise and domain-specific knowledge. While Qwen shows strength in certain areas like Language, it struggles in Finance. And Yi delivers a more consistent performance but does not outperform FinMOE in the critical areas of Finance and Subject tasks. Overall, FinMOE stands out due to its targeted design for finance-related tasks, as well as its general versatility in handling a broad range of domain-specific and reasoning challenges. This demonstrates the effectiveness of the Mixture of Experts approach in addressing both general-purpose and specialized evaluation benchmarks.

6 Conclusion

In this paper, we introduced FinMoE, a Mixture-of-Experts-based large Chinese financial language model designed to address the limitations of general-purpose language models in financial tasks. FinMoE effectively integrates domain-specific financial expertise with strong general knowledge, achieving state-of-the-art performance across both general and financial benchmarks at a comparable scale. Its innovative architecture, comprehensive training techniques enable it to deliver accurate and scalable solutions for complex financial tasks. Future work will focus on further enhancing FinMoE's adaptability to evolving financial contexts and expanding its applications to other real-world scenarios (Zhang and Yang, 2021b; Zhang et al., 2021, 2022b,a; Zhang and Yang, 2025), solidifying its role as a powerful tool in advancing AI research and practice.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2024. A survey on mixture of experts. *Authorea Preprints*.

Ronan Collobert, Samy Bengio, and Yoshua Bengio. 2001. A parallel mixture of svms for very large scale problems. *Advances in Neural Information Processing Systems*, 14.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Michael I Jordan and Robert A Jacobs. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Xiaonan Nie, Xupeng Miao, Shijie Cao, Lingxiao Ma, Qibin Liu, Jilong Xue, Youshan Miao, Yi Liu, Zhi Yang, and Bin Cui. 2021. Evomoe: An evolutional mixture-of-experts training framework via dense-to-sparse gate. *arXiv preprint arXiv:2112.14397*.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Xun Wu, Shaohan Huang, and Furu Wei. 2024. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

Xuanyu Zhang. 2019. MCˆ2: Multi-perspective convolutional cube for conversational machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6185–6190, Florence, Italy. Association for Computational Linguistics.

Xuanyu Zhang. 2020. Cfgnn: Cross flow graph neural networks for question answering on complex tables. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9596–9603.

Xuanyu Zhang, Bingbing Li, and Qing Yang. 2023a. Cgce: A chinese generative chat evaluation benchmark for general and financial domains. *arXiv preprint arXiv:2305.14471*.

Xuanyu Zhang, Zhepeng Lv, and Qing Yang. 2023b. Adaptive attention for sparse-based long-sequence transformer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8602–8610, Toronto, Canada. Association for Computational Linguistics.

Xuanyu Zhang and Zhichun Wang. 2020. Rception: Wide and deep interaction networks for machine reading comprehension (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(10):13987–13988.

Xuanyu Zhang and Qing Yang. 2021a. Dml: Dynamic multi-granularity learning for bert-based document reranking. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3642–3646, New York, NY, USA. Association for Computing Machinery.

Xuanyu Zhang and Qing Yang. 2021b. Position-augmented transformers with entity-aligned mesh for textvqa. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 2519–2528, New York, NY, USA. Association for Computing Machinery.

Xuanyu Zhang and Qing Yang. 2023a. Generating extractive answers: Gated recurrent memory reader for conversational question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7699–7704, Singapore. Association for Computational Linguistics.

Xuanyu Zhang and Qing Yang. 2023b. Self-qa: Unsupervised knowledge guided language model alignment. *arXiv preprint arXiv:2305.11952*.

Xuanyu Zhang and Qing Yang. 2023c. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 4435–4439, New York, NY, USA. Association for Computing Machinery.

Xuanyu Zhang and Qing Yang. 2025. Extracting the essence and discarding the dross: Enhancing code generation with contrastive execution feedback. In *Proceedings of the 31th International Conference on Computational Linguistics*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2021. Combining explicit entity graph with implicit text information for news recommendation. WWW '21, page 412–416, New York, NY, USA. Association for Computing Machinery.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2022a. Deepvt: Deep view-temporal interaction network for news recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 2640–2650, New York, NY, USA. Association for Computing Machinery.

Xuanyu Zhang, Qing Yang, and Dongliang Xu. 2022b. TranS: Transition-based knowledge graph embedding with synthetic relation representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1202–1208, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.