Advancing Financial Standards Comprehension

2025 IEEE 11th International Conference on Intelligent Data and Security (IDS)

Pavel Voropaev
Independent Researcher
Liverpool, UK
voropaev.p@gmail.com

Anna Detkina
School of Engineering
University of Liverpool
Liverpool, UK
anna.detkina@liverpool.ac.uk

Abstract— **This paper presents a novel Mixture of Experts (MoE) approach for optimizing Large Language Models in the interpretation of financial and accounting standards within Task 4 of the FinRL Contest 2025, focusing on the Common Domain Model, Markets in Financial Instruments Directive Operational Framework, and eXtensible Business Reporting Language. Operating under a 45GB VRAM hardware constraint, domain-specific data preparation was implemented using DeepSeek V3, and a Qwen 0.5B model was converted to MoE architecture with 16 experts, totaling 5.51B parameters. The router was initialized with random weights while experts retained pretrained knowledge of the instruct model. The research demonstrates that targeted fine-tuning with MoE architecture produced mixed results, significantly improving performance in recall related tasks, while reducing performance of more complex ones compared to a baseline model Qwen7B.**

Keywords— FinRL, MoE, Qwen, LLM, Financial Standards

INTRODUCTION

Large Language Models (LLMs) have become powerful tools for interpreting and managing complex financial tasks. Task 4 of the FinRL Contest 2025 aims to test and improve LLMs in financial and accounting standards, specifically focusing on the Common Domain Model (CDM), the Markets in Financial Instruments Directive Operational Framework (MOF), and the eXtensible Business Reporting Language (XBRL).

The LLMFinLegal challenge, organized by the same group, serves as a valuable precedent for the current work [1]. One significant insight from that challenge was the absence of mixture-of-experts approaches, which present an opportunity for innovation. Another notable finding was the relatively poor performance of contestant models on abbreviation and definition tasks, which we hypothesized could be improved through dedicated expert models and augmented custom datasets.

Our approach to the FinRL Contest was guided by the following objectives: creating datasets aligned with benchmark requirements and developing robust model architecture to mitigate domain-specific knowledge degradation [2, 3]. The project was conducted under two major constraints set by organizers: Retrieval-Augmented Generation (RAG) was prohibited despite its potential advantages in precision-driven regulatory domains [4], and the model size was restricted to a maximum of 8 billion parameters.

The study chose Qwen as a framework, starting with a 0.5B base model and converting it to a Mixture of Experts (MoE) architecture with 16 experts, creating domain-specific experts that improved financial standards interpretation. The Qwen 7B model served as the baseline for comparison. High-quality datasets were created using the DeepSeek API with tailored prompts.

Each domain presented distinct technical challenges for model optimization:

- The CDM required summarization as well as question answering.
- The MOF suffered from limited dataset size, creating risks of catastrophic forgetting during multi-domain training.
- The XBRL presented exceptional diversity in task types, spanning from token classification to complex financial mathematics computations.

While this study currently focuses on the LLM architecture and fine-tuning, RAG techniques offer potential to further enhance the precision and reliability of generative AI responses. Recent research has shown RAG's effectiveness in accurately handling qualitative financial data, thereby supporting better-informed decision-making and increasing overall response reliability of LLM [4]. This makes RAG a promising direction for future research.

METHODOLOGY

A. Data Preparation

Existing public datasets were utilized where available, supplemented by custom training sets generated from authoritative sources using DeepSeek V3 to produce domain-specific question-answer pairs.

For the CDM, documentation was obtained in markdown format from FinOS GitHub, eliminating the need for web scraping. The documents were processed in overlapping chunks using DeepSeek V3 to create the training dataset.

The MOF documentation was accessed directly from the OpenSource Initiative GitHub. The dataset was enhanced with diverse question formulations, including reverse questions for abbreviations and varied phrasing for license approval status [8]. General knowledge data was generated using DeepSeek based on MOF documentation and selected web content.

For XBRL, multiple datasets were utilized: redacted versions of FiNER [9] and FNXL [10] with broader taxonomy coverage, ConvFinQA [11] for numerical financial reasoning, and supplementary financial data for DOW30 companies from Yahoo Finance covering five years of key ratios and financial positions. Terminology-based question-answer pairs were generated using the XBRL glossary and specification.

To address the imbalanced dataset sizes, equal sampling from each dataset was implemented [7], resulting in more epochs for smaller datasets like CDM and MOF.

979-8-3315-9661-3/25/$31.00 ©2025 IEEE
DOI 10.1109/IDS66066.2025.00027

B. Model Architecture

A Mixture of Experts (MoE) architecture was implemented using the Qwen framework, converting the Qwen 0.5B base model to an MoE configuration with 16 experts of 0.5B parameters each. Experts retained the original pretrained weights and dimensions, while the router was initialized with random weights. The total parameter count was 5.51B with a top-k value of 2. The choice of 16 was dictated by the hardware, theoretically 24 experts would've been able to fit into the 8B limit.

Initial question stuffing proved ineffective with 0.5B expert models, so individual question formatting was used instead.

C. Training Configuration

The model was trained using an NVIDIA L40S GPU. Key training settings included a batch size of 2 per device, a learning rate of 5e-5, and 4 training epochs. The training also used 3000 warm-up steps, bfloat16 precision (bf16), gradient accumulation over 4 steps, and gradient checkpointing to optimize memory usage. Standard training was used with the PyTorch Trainer from the transformers library without any adapter for the MoE model. While the original plan considered memory optimization techniques, the final implementation pivoted to the Qwen framework as most optimization methods don't properly support MoE layers. MoE models generally lack mature library support for parameter-efficient fine-tuning methods [12].

D. Evaluation Methodology

The evaluation datasets were provided by the contest organizers on GitHub [13]. To evaluate models' responses, a GPT Judge approach was implemented, with GPT-4o being used as the evaluation model. The basis for the model evaluation was a detailed prompt that included specific evaluation criteria such as:

- accuracy (how well the generated response aligns with the reference answer in terms of factual correctness and completeness),
- relevance (how relevant the generated answer is to the prompt or question),
- clarity (how clear and easy to understand the generated response compared to the reference),
- concision (whether the generated answer provides necessary information without excessive wordiness),
- coherence (whether the response is logically organized),
- and style (how well the tone and writing style of the generated answer match that of the reference).

A structured scoring guide was used, with scores ranging from 10 (indicating strong similarity to the reference) to 0 (indicating major discrepancies).

RESULTS

The implementation of the MoE approach for financial standards interpretation has demonstrated mixed results across three domains compared to the baseline Qwen7B model. Qwen7B was chosen as baseline as it was the closest one by the size to the resulting MoE model.

Table 1 presents the performance comparison between Qwen-7B and Qwen-0.5B×16 MoE.

TABLE I. COMPARISON OF QWEN-7B AND QWEN-0.5B × 16 MOE

Model	MoF	CDM	XBRL	Average
Qwen7B	5.30	4.46	3.53	4.43
Qwen0.5B x16 (MoE)	6.96	2.64	1.81	3.8

The domain separation provided by the expert specialization proved particularly effective for handling the distinct terminology and reasoning patterns required by each standard. Specifically, we observed a regression in accuracy on CDM-related queries, which was unexpected and might be related to evaluation methodology, which favored summarized responses with high text cohesion over the precise documentation extracts. We expect to see different results, when organizers publish their evals. A significant improvement for MOF tasks was observed as anticipated, showing superior data retention capabilities of MoE architecture. A considerable regression on XBRL task against the baseline model was also seen. It was expected as XBRL source data for the queries was not provided and since we did not train on the raw XBRL documents, the final MoE model was unable to recall anything. Financial math results are worse than baseline, which is somewhat expected given small size of the base Qwen0.5B model and relatively small amount of time put into building good datasets for this task.

Table 2 represents the summary over the costs of the project implementation.

TABLE I I. COSTS OF THE PROJECT

Category	Cost, USD
API (deepseek/openai)	40
Compute (huggingface)	83.35
Total	**123.35**

The total project cost was 123.35 USD. Consumer grade NVIDIA GeForce RTX4080 GPU with 16GB VRAM was available to the team and is not included into the cost.

CONCLUSION

The implementation of the MoE approach for financial standards interpretation demonstrated promising results across MOF domain. The experimental evaluation showed that the MoE architecture significantly improved performance on data recall tasks, even with small fine-tuning datasets. However, its effectiveness on other tasks was limited by the use of a 0.5B base model, which led to performance regressions when the fine-tuning dataset didn't align well with the contest evaluation criteria.

The modest budget of this research suggests this approach is suitable for small research teams and organizations. Architectural modifications focusing on domain specialization can yield substantial performance gains on some of the tasks even within strict parameter constraints.

FUTURE WORK

Several promising directions emerge from this research. First, Retrieval-Augmented Generation (RAG) represents a

natural extension that would allow the model to reference authoritative documentation with a high degree of precision. While RAG was prohibited in the current challenge, incorporating it could significantly enhance the model's ability to provide accurate and up-to-date interpretations of financial standards.

Second, applying parameter-efficient fine-tuning methods such as LoRA specifically tailored for MoE architectures could further improve training efficiency and performance. Current limitations in library support for MoE-compatible parameter-efficient methods present an opportunity for innovation in this area. Developing specialized LoRA adapters for MoE experts could enable more targeted training while maintaining the benefits of parameter efficiency.

Third, architecture extensions should explore more sophisticated routing mechanisms and expert configurations, specifically implementing variable expert size, as XBRL would benefit from more parameters due to extensive taxonomies. Additionally, Chain-of-Thought fine-tuning could have substantial impact on performance and should be explored, as demonstrated in previous financial LLM competitions.

These future directions aim to build upon the foundation established in this work, further enhancing the capabilities of LLMs in specialized financial domains.

REFERENCES

[1] K. Wang, J. Patel, C. Shen, D. Kim, A. Zhu, A. Lin, L. Borella, C. Osborne, M. White, S. Yang, K. Xiao, and X.-Y. Liu, "FinNLP-FNP-LLMFinLegal-2025 Shared Task: Regulations Challenge," Proc. Joint Workshop on FinNLP, FNP, and LLMFinLegal, pp. 363–370, 2025.

[2] A. Rath, Y. Tang, P. Li, M. Tan, and H. Chen, "Large Language Model in Financial Regulatory Interpretation," arXiv:2405.06808, 2024. [Online]. Available: https://arxiv.org/abs/2405.06808

[3] "Implementing Financial Regulations Using Large Language Models," SSRN, 2024. [Online]. Available: https://papers.ssrn.com/abstract_id=5010694

[4] T. Zhu, J. Liu, and Y. Yang, "Financial Question Answering via Retrieval-Augmented Generation," Applied Sciences, vol. 14, no. 20, p. 9318, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/20/9318

[5] P. Chantangphol, P. Balee, K. Sucharitpongpan, C. Saetia, and T. Chalothorn, "FinMind-Y-Me at the Regulations Challenge Task: Financial Mind Your Meaning based on THaLLE," Proc. Joint Workshop on FinNLP, FNP, and LLMFinLegal, pp. 349–362, 2025.

[6] J. Huang, M. Jiang, and H. Zhu, "Audit-FT at the Regulations Challenge Task: An Open-Source Large Language Model for Audit," Proc. Joint Workshop on FinNLP, FNP, and LLMFinLegal, pp. 335–348, 2025.

[7] X. Cai, M. Xiao, Z. Ning, and Y. Zhou, "Resolving the imbalance issue in hierarchical disciplinary topic inference via LLM-based data augmentation," in Proc. IEEE ICDMW, 2023, pp. 1424–1429.

[8] A. Kumar, A. Bhattamishra, M. Bhandari, and P. Talukdar, "Submodular Optimization-based Diverse Paraphrasing for Improving Similarity Detection," in Proc. EMNLP-IJCNLP, 2019, pp. 4612–4623.

[9] L. Loukas, P. Lendvai, P. Ristoski, A. Menaj, and E. Milios, "FiNER: Financial Numeric Entity Recognition for XBRL Tagging," in Proc. ACL, 2022, pp. 4419–4435.

[10] A. Sharma, S. Singh, A. Mishra, and S. Garg, "Financial Numeric Extreme Labelling: A dataset and benchmarking," in Findings of ACL, 2023, pp. 3681–3693.

[11] Z. Chen, S. Li, C. Smiley, Z. Ma, S. Shah, and W. Y. Wang, "ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering," in Proc. EMNLP, 2022, pp. 6290–6306.

[12] G. Khatuya, S. Tripathi, V. Dahiya, M. Jain, and A. Sengupta, "Parameter-Efficient Instruction Tuning of Large Language Models For Extreme Financial Numeral Labelling," in Proc. NAACL-HLT, 2024, pp. 7258–7270.

[13] FinRL Contest 2025 testing dataset https://github.com/Open-Finance-Lab/FinRL_Contest_2025/tree/main/Task_4_FinLLM_Leaderboard_DRR/testing_dataset