ASSIGNMENT 3: TRAIN YOUR OWN LLMS

Chua Qin Di, 125400010
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
125400010@link.cuhk.edu.cn

ABSTRACT

This project investigates the fine-tuning of a pre-trained large language model (LLM) for English financial question answering. Using a subset of the Finance-Alpaca dataset, a supervised instruction-tuning pipeline was implemented to adapt a general-purpose model toward domain-specific financial reasoning. To enable training under limited GPU resources, the QLoRA framework was used with 4-bit NF4 quantization. Evaluation was conducted using automatic metrics, including ROUGE scores and character-level semantic similarity. Results demonstrate that even small-scale fine-tuning on 100 examples leads to measurable improvements in alignment with reference answers, producing responses that are more grounded, structured, and financially relevant compared to the unfine-tuned baseline. The findings highlight the practicality of parameter-efficient fine-tuning for domain adaptation in specialized, knowledge-intensive settings.

1 INTRODUCTION

Research Topic This study focuses on adapting a Large Language Model (LLM) for English financial question answering. The goal is to enhance a general-purpose LLM's ability to interpret and respond to financial queries involving investing, personal finance, macroeconomic concepts, and corporate fundamentals.

Motivation General-purpose LLMs often produce vague answers when handling technical financial topics. Financial question answering requires numerical awareness, precise terminology, and the ability to reason about mechanisms such as valuation, interest rates, retirement planning, or portfolio allocation. Improving an LLM' s reliability in this domain is essential for educational tools, financial research assistants, and advisory-support systems.

Project Summary Pre-trained Qwen2.5-3B and Qwen2.5-7B models was fine-tuned using a small curated subset of the Finance-Alpaca dataset. The training pipeline incorporated data preprocessing, supervised instruction formatting, and QLoRA parameter-efficient fine-tuning with 4-bit quantization. The resulting model was evaluated against the original backbone model using ROUGE-based metrics and semantic similarity. Despite the small training set size, the fine-tuned model showed improved factual alignment and domain reasoning.

2 EXPERIMENT DESIGN

Definition of the task The task is formulated as a single-turn, open-ended financial question-answering problem. The model receives a natural-language financial prompt and generates a concise, accurate, and contextually relevant response. Evaluation emphasizes similarity to reference answers but also considers qualitative reasoning quality.

Design of the experiment The dataset was converted into instruction-response conversation pairs and split into 100 training samples, 20 validation samples, and an additional test set of 50 samples. Two models were evaluated: Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct. Both models were fine-tuned using QLoRA with 4-bit NF4 quantization, allowing training under limited GPU memory.

Evaluation on the test set used ROUGE metrics and character-level similarity between generated and reference answers.

3 EXPERIMENTS

All experiments were performed in Google Colab using an A100 GPU.

3.1 QUANTIZATION AND PARAMETER-EFFICIENT FINE-TUNING

Given hardware limitations, the QLoRA method was used to reduce the memory footprint during training. The pretrained backbone weights were quantized using the 4-bit NF4 format, which preserves statistical structure while reducing memory usage by approximately 75%. Only LoRA adapter parameters were trained in `bfloat16` precision, while the quantized backbone remained frozen.

This approach provides:

- **Feasibility under limited GPU memory:** Allows fine-tuning of multi-billion parameter models on 12–16 GB GPUs.
- **Low training overhead:** Reduced data movement leads to faster iterations.
- **Stable optimization:** Trainable adapter weights avoid numerical degradation from quantization.

3.2 TRAINING CONFIGURATION

The instruction-tuning process followed the configuration implemented in the training code. The main hyperparameters are:

- **Model:** Qwen2.5-3B-Instruct
- **Quantization:** 4-bit NF4 with double quantization
- **Precision:** `bfloat16` for LoRA adapters
- **Epochs:** 3
- **Batch Size:** 2
- **Gradient Accumulation:** 4 (effective batch size = 8)
- **Learning Rate:** 2×10^{-5}
- **Scheduler:** Linear learning rate decay
- **Weight Decay:** 0.0
- **Optimizer:** Paged AdamW (32-bit)

These settings ensured stable training within Colab's memory limits.

3.3 Dataset and Evaluation Setup

Two datasets were used in this project. The primary dataset was a subset of the Finance-Alpaca dataset, which provided supervised instruction–response pairs for fine-tuning. The data was partitioned as follows:

- 100 samples for training,
- 20 samples for validation,
- 50 samples for testing.

In addition to the Finance-Alpaca evaluation set, an external dataset—the Financial Advisor 100 dataset—was used for further testing. This dataset consists of real-world, long-form financial advice questions. A sample of 50 examples from this dataset was used to assess the generalization ability of both baseline and fine-tuned models on queries not seen during training.

Across both test sets, evaluation relied on the same automatic metrics: ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum, and character-level semantic similarity. These metrics measure lexical and structural alignment between model-generated answers and human-written reference responses.

Evaluation relied entirely on automatic metrics:

- **ROUGE-1** measures unigram (single-word) overlap between the model output and the reference answer. It captures whether the model uses the same key terms as the human-written response.
- **ROUGE-2** measures bigram (two-word sequence) overlap. This is a stricter metric that reflects whether the model reproduces short phrases or financial expressions in a manner similar to the reference.
- **ROUGE-L** measures the longest common subsequence (LCS) between the generated and reference answers. It reflects how well the model follows the overall structure and ordering of the reference explanation.
- **ROUGE-Lsum** is a sentence-level extension of ROUGE-L that aggregates LCS-based similarity across multiple sentences. This metric is useful for long-form financial answers where structure and flow matter.
- **Character-level similarity** using Python's `difflib.SequenceMatcher` provides a matching score based on character alignment and is used to capture broader semantic similarity beyond exact word choices.

These metrics quantify textual alignment between the model's answer and the dataset's reference answer.

3.4 RESULTS

The evaluation compares three distinct models: (1) Qwen2.5–3B backbone, (2) fine-tuned Qwen2.5–3B model, and (3) fine-tuned Qwen2.5–7B model. Two different test sets were used (Finance-Alpaca and Financial Advisor).

Table 1 reports ROUGE-based metrics and character-level similarity over 50 samples for each model-dataset configuration. Both fine-tuned models outperform the backbone model across all ROUGE metrics, showing improved lexical overlap (ROUGE-1/2) and structural correspondence (ROUGE-L/Lsum). Fine-tuning greatly increases alignment between the model's output and the reference answers, despite the small 100-sample training set.

The Qwen2.5–7B model consistently achieves higher scores than the 3B model when evaluated on the same dataset, reflecting the benefits of increased capacity. Performance is higher on the Alpaca test set than on the Financial Advisor test set because the Alpaca references are shorter, more templated, and stylistically similar to the fine-tuning data.

Qualitative inspection further shows:

- responses became more structured and concise,
- financial terminology was used more appropriately,
- reasoning errors and irrelevant content decreased,
- answers were more consistent across similar question types.

Overall, even with limited domain-specific supervision, the fine-tuned models exhibit substantial improvements in both semantic alignment and financial reasoning, with Qwen2.5–7B achieving the strongest performance.

Table 1: ROUGE and similarity scores for the unfine-tuned backbone and the two fine-tuned Qwen2.5 models, each evaluated on two different test sets.

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum	Similarity
Qwen2.5–3B (Backbone)	0.2592	0.0822	0.1686	0.1969	0.0897
Qwen2.5–3B (Alpaca Test Set)	0.2658	0.0820	0.1856	0.2075	0.1531
Qwen2.5–7B (Alpaca Test Set)	**0.3092**	**0.1332**	**0.2236**	**0.2480**	**0.1765**
Qwen2.5–3B (Advisor Test Set)	0.2772	0.0956	0.1530	0.2287	0.0652
Qwen2.5–7B (Advisor Test Set)	0.2791	0.0970	0.1595	0.2337	0.0680

4 CONCLUSION

This project shows that domain-specific instruction tuning can meaningfully improve the financial question-answering abilities of large language models. Using QLoRA with 4-bit quantization enabled efficient fine-tuning of 3B and 7B models on limited hardware, while preserving training stability.

Both fine-tuned models outperformed the backbone model across all automatic metrics on two distinct evaluation sets. The gains reflect stronger lexical alignment, clearer structure, and more accurate use of financial terminology. Qualitative analysis also showed fewer reasoning errors and more coherent explanations.

These findings demonstrate that even small, targeted datasets can effectively adapt general-purpose LLMs to specialized financial tasks. Future extensions may incorporate larger corpora, retrieval-augmented methods, or broader financial reasoning benchmarks to further improve robustness.

ACKNOWLEDGMENT

This is the Assignment3 for CSC4100, see details in https://nlp-course-cuhksz.github.io/.

REFERENCES