FinLoRA: Benchmarking LoRA Methods for Fine-Tuning LLMs on Financial Datasets

Dannong Wang[1] **Jaisal Patel**[1] **Daochen Zha**[2] **Steve Y. Yang**[3] **Xiao-Yang Liu**[2]
[1]Rensselaer Polytechnic Institute [2]Columbia University [3]Stevens Institute of Technology
Emails: {wangd12, patelj8}@rpi.edu, syang14@stevens.edu, xl2427@columbia.edu

Abstract

Low-rank adaptation (LoRA) methods show great potential for scaling pre-trained general-purpose Large Language Models (LLMs) to hundreds or thousands of use scenarios. However, their efficacy in high-stakes domains like finance is rarely explored, e.g., passing CFA exams and analyzing SEC filings. In this paper, we present the open-source FinLoRA project that benchmarks LoRA methods on both general and highly professional financial tasks. First, we curated 19 datasets covering diverse financial applications; in particular, we created four novel XBRL analysis datasets based on 150 SEC filings. Second, we evaluated five LoRA methods and five base LLMs. Finally, we provide extensive experimental results in terms of accuracy, F1, and BERTScore and report computational cost in terms of time and GPU memory during fine-tuning and inference stages. We find that LoRA methods achieved substantial performance gains of 36% on average over base models. Our FinLoRA project provides an affordable and scalable approach to democratize financial intelligence to the general public. Datasets, LoRA adapters, code, and documentation are available at https://github.com/Open-Finance-Lab/FinLoRA

1 Introduction

Large language models (LLMs) [49, 50] have demonstrated impressive general capabilities in various vertical domains, such as finance [42, 22, 17, 1], healthcare [41, 5, 51], law [43], education [20], and scientific discovery [26, 3]. In the finance sector, LLMs have been applied to tasks such as sentiment analysis [47], question-response, and stock market prediction [17].

Cost-effective adaptation is critical for applying LLMs to vertical domains like finance, since general-purpose LLMs lack the specialized knowledge to excel in professional-level tasks. Full fine-tuning can close such performance gaps but is prohibitive for most organizations due to its computationally demanding nature. As such, parameter-efficient fine-tuning (PEFT), particularly Low-Rank Adaptation (LoRA) [12] and its variants [7, 29, 46, 13, 19, 4, 31], has emerged as an affordable and scalable solution. LoRA methods can enhance pre-trained general-purpose LLMs with domain-specific knowledge and improve performance on downstream tasks [29].

Recent research like FinGPT [22, 23] has applied a quantized LoRA method [7] to general financial tasks; however, the comparative performance of various LoRA variants in complex, professional-level financial tasks remains rarely explored. Previous research shows that LLMs are struggling with professional-level financial tasks, such as analyzing SEC filings [15] and passing financial certificate exams [2]. A critical area within professional finance involves eXtensible Business Reporting Language (XBRL) data [34], the de facto global standard for business reporting. Despite XBRL's importance, dedicated datasets for related analytical tasks are scarce. This deficiency, coupled with the need to evaluate different LoRA methods on highly specialized financial tasks, motivates our

Preprint. Under review.

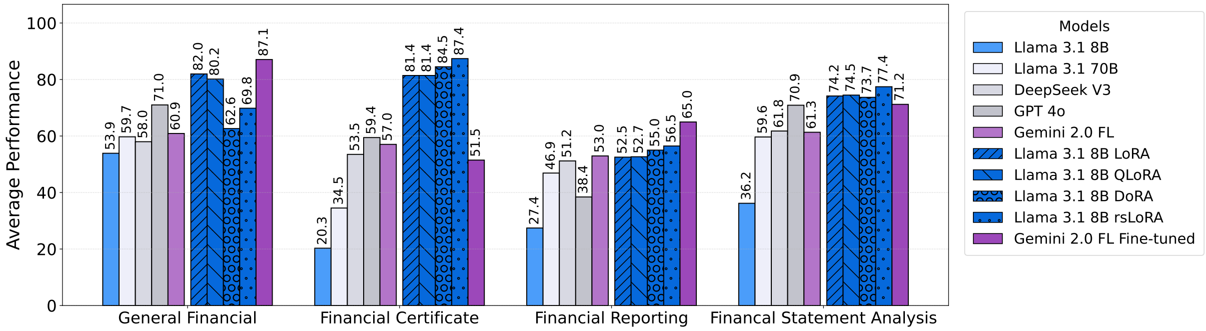

Figure 1: Average performance of base models and LoRA models.

introduction of FinLoRA: a comprehensive benchmark designed to assess LoRA variants across diverse financial scenarios, with an emphasis on professional XBRL applications.

This paper demonstrates that fine-tuning state-of-the-art LLMs can significantly improve performance across a range of financial tasks, including specialized XBRL analysis, locally and cost-effectively using widely accessible GPUs. As illustrated in Fig. 1, our LoRA-adapted models achieve notable performance improvements over baseline models across four categories of financial tasks. Our main contributions are summarized as follows:

- We curated 19 financial datasets, including general financial tasks, financial analysis, and professional-level XBRL tasks. In particular, we created four novel XBRL analysis datasets. This enables future research to perform rigorous evaluation of LoRA methods in financial tasks.
- We implemented and fairly compared five LoRA methods—including LoRA [12], QLoRA [7], DoRA [21], rsLoRA [16], and Federated LoRA—by fine-tuning models on financial datasets. LoRA methods achieved an average increase of 36% in accuracy over baseline models, which validates the effectiveness of low-rank adaptation and quantization for fine-tuning LLMs.
- We conducted an extensive analysis with 46 rounds of fine-tuning and 194 rounds of evaluations for LoRA methods from four angles: (i) a comprehensive comparison across different base models and datasets, (ii) performance on various types of financial tasks, (iii) resource requirements for fine-tuning and inference, and (iv) practical considerations for LoRA deployment in finance.

Our benchmark is open-sourced at `https://github.com/Open-Finance-Lab/FinLoRA`.

2 Is Fine-tuning of LLMs Needed on Financial Tasks?

While general-purpose LLMs—such as GPT-4o [14], Llama 3.1 [8], and DeepSeek-V3 [18]—demonstrate broad NLP competence, their performance often falls short on nuanced financial tasks [15, 2]. This section discusses three key reasons that underscore the necessity of fine-tuning, particularly with methods like LoRA, for developing effective financial LLMs: ***(i)* Lack of High-Quality Financial Data in Pre-Training Datasets.** Many pre-training datasets, such as The Pile [9], primarily draw from general web crawls (e.g., GitHub, arXiv). These sources often under-represent high-quality, specialized financial data, which may be private and exist in complex formats (like XBRL). Consequently, to equip LLMs with the understanding required for complex financial analysis, targeted fine-tuning on curated, domain-specific datasets becomes essential. ***(ii)* General LLMs' Failure in Specialized Financial Tasks.** General LLMs often struggle with specialized tasks that demand deep domain-specific knowledge. XBRL analysis provides a clear illustration of these difficulties: Table 1 details the typical Llama base model's errors on XBRL questions and its improved outputs after LoRA fine-tuning. In Question 1, the base model mistags a $2.0 billion value because it relies on superficial keyword matches (e.g., "equity," "carrying value,") and applies the generic tag *us-gaap:MajorityEquityInterest*, ignoring the context "under the equity method." In Question 2, the base model incorrectly selects a tag referencing the value 1,209,000,000 by matching the keyword "Equity" and also ignores the decimals="-6" attribute. ***(iii)* Cost and Time.** As shown in Table 5, the training-from-scratch approach of BloombergGPT [42], which reportedly cost $2.7 million and took

Table 1: Case study—XBRL tagging (Google 10-Q 2025-Q1) and XBRL formula calculation (Travelers 10-K FY-2023)—from base Llama 3.1 8B Instruct and our LoRA fine-tuned version.

Question 1	What is the appropriate XBRL US-GAAP tag for "2.0" in **"...equity securities accounted for under the equity method had a carrying value of approximately $2.0 billion"** ?
Llama 3.1 8B	`us-gaap:MajorityEquityInterest`
Llama 3.1 8B LoRA (8bit r8)	`us-gaap:EquityMethodInvestments`
Ground truth	`us-gaap:EquityMethodInvestments`
Question 2	What is Travelers Companies Inc's **Equity Multiplier** for FY 2023? (Answer with a formula substituted with values.) {XBRL Context}
Llama 3.1 8B	(1,209,000,000 / 249,210,000,000)
Llama 3.1 8B LoRA (8bit r8)	$125,978,000,000 \ / \ 249,210,000,000$
Ground truth	$125,978,000,000 \ / \ 249,210,000,000$

53 days to train (Table 5), is economically nonviable for most organizations. In contrast, fine-tuning existing foundational models using LoRA methods is significantly more accessible and time-efficient.

3 FinLoRA Benchmark

3.1 Benchmark Tasks, Datasets, and Metrics

As displayed in Table 2, we consider four types of tasks: general financial tasks, financial certificate, financial reporting, and financial statement analysis.

Public Financial Datasets FinLoRA includes 15 public financial datasets. *(i)* Sentiment analysis (SA): Financial Phrase Bank (FPB) [28], Financial QA Sentiment Analysis (FiQA SA) [27], Twitter Financial News Sentiment (TFNS) [33], and News with GPT Instruction (NWGI) [23], each with financial text from news or tweets and sentiment labels. *(ii)* Headline analysis: The Headline dataset [37] classifies financial headlines based on various questions into two classes: "yes" and "no". *(iii)* Named-entity recognition (NER): NER dataset [35] annotates one entity per sentence, categorized into one of three classes: "location", "person", and "organization". *(iv)* Financial certificate: CFA Level I, II, and III, and CPA Regulation. *(v)* Financial reporting: XBRL Terminology [10], Financial Numeric Entity Recognition (FiNER) [25], and Financial Numeric Extreme Labeling (FNXL) [36]. *(vi)*: Financial statement analysis: Financial Math [10] and FinanceBench [15, 10].

Newly-added XBRL Analysis Datasets We introduce 4 novel XBRL analysis datasets, i.e., extracting and analyzing SEC financial reports in XBRL format. These question-answering datasets, derived from the 2019-2023 annual reports of Dow Jones 30 companies, provide each example with a question, a relevant filtered XBRL text segment as source material, and a ground truth answer. The datasets cover four distinct task types: *(i)* **XBRL tag extraction** involves extracting a specific XBRL tag from a raw XBRL text segment given a natural language description of the tag. *(ii)* **XBRL value extraction** focuses on extracting a numeric value from the raw XBRL text segment given a natural language description of the value. *(iii)* **XBRL formula construction** tasks the LLM to first identify and select multiple relevant facts (and their corresponding XBRL tags) from the XBRL data, and then construct a standard financial formula (e.g., Net Profit Margin, Quick Ratio) using these selected tags as components. *(iv)* **XBRL formula calculation** builds on the previous task and requires the LLM to substitute the actual numeric values into the formula and compute the final result.

Dataset Construction Pipeline Initially, we classified financial tasks into nine categories, creating a training set for each to develop category-specific LoRA adapters per configuration. The four novel XBRL analysis datasets were constructed using XBRL-formatted 10-K annual reports from Dow Jones 30 companies (2019-2023). For these, we generated the four aforementioned types of questions by applying five distinct templates to consolidated, company-specific facts. To ensure contextual

Table 2: Benchmark tasks and datasets.

Datasets	Types	#Train/#Test	Average Prompt Length	Metrics	Sources & License
General Financial Tasks (Total: 122.9k/31.7k)					
FPB [28]	Sentiment Analysis	3.1k/970	56	Accuracy, F1	HF, CC BY-SA 3.0
FiQA SA [27]	Sentiment Analysis	822/234	48	Accuracy, F1	HF MIT
TFNS [33]	Sentiment Analysis	9.5k/2.4k	52	Accuracy, F1	HF MIT
NWGI [22]	Sentiment Analysis	12.9k/4.1k	81	Accuracy, F1	HF MIT
Headline [37]	Headline Analysis	82.2k/20.5k	43	Accuracy, F1	HF CC BY-SA 3.0
NER [35]	NER	13.5k/3.5k	138	Accuracy, F1	HF CC BY-SA 3.0
Financial Certificate Tasks (Total: 472/346)					
CFA Level I	Analyst Exam	180/90	181	Accuracy, F1	Internet (Public; Not Released Due to Copyright)
CFA Level II	Analyst Exam	88/77	1.0k	Accuracy, F1	
CFA Level III	Analyst Exam	80/78	961	Accuracy, F1	
CPA REG	Accountant Exam	124/101	147	Accuracy, F1	
Financial Reporting Tasks (Total: 15.9k/8.3k)					
FiNER-139 [25]	XBRL Tagging	10.0k/7.4k	1.8k	Accuracy, F1	HF CC BY-SA 4.0
FNXL [36]	XBRL Tagging	-/247	7.1k	Accuracy, F1	GitHub Public
XBRL Term [10]	Terminology	5.9k/651	25	BERTScore	GitHub MIT
Financial Statement Analysis Tasks (Total: 27.9k/7.3k)					
Financial Math [10]	Math	800/200	116	Accuracy	GitHub MIT
FinanceBench [15, 10]	Math	86/43	983	BERTScore	GitHub CC BY-NC 4.0
Tags Extraction	XBRL Analysis	10.1K/2.9k	3.8k	Accuracy, F1	HF MIT
Values Extraction	XBRL Analysis	10.1k/2.5k	3.8k	Accuracy, F1	HF MIT
Formula Construction	XBRL Analysis	3.4K/835	3.8k	Accuracy, F1	HF MIT
Formula Calculation	XBRL Analysis	3.4K/835	3.8k	Accuracy, F1	HF MIT

relevance, XBRL file segments were automatically filtered based on pertinent factors like year and reporting axes. Further details on the XBRL dataset creation and the processing of other public datasets are available in Appendix A.

Metrics For all general financial tasks, financial analysis, XBRL tagging, financial math, and XBRL analysis tasks, we use Exact Match (EM) to evaluate the LLMs' output and report both the accuracy and weighted F1 score (in the supplementary materials). For XBRL Term and FinanceBench, we report BERTScore F1 [48] instead. We also report the average of scores across the tasks with BERTScore F1 multiplied by 100.

3.2 Base Models and LoRA Methods

Base Models We benchmark two models for both base model and LoRA fine-tuning performance—Llama 3.1 8B Instruct [8] and Gemini 2.0 Flash Lite [39]. We also evaluated three additional models—Llama 3.1 70B Instruct [8], DeepSeek V3 [18], and GPT-4o [14]—as base models only.

LoRA Methods We considered the following five popular LoRA methods.

- (Vanilla) **LoRA**: Low-rank adaptation (LoRA) [12] is a parameter-efficient fine-tuning method that preserves the weights of the pre-trained model and introduces a smaller set of trainable weights. The updated weights follow the low-rank decompositions $\Delta \boldsymbol{W} = \gamma_r \boldsymbol{B}\boldsymbol{A}$, where γ_r is a scaling factor ($\gamma_r = \frac{\alpha}{r}$ with $\alpha > 0$ and rank $r > 0$), $\boldsymbol{A} \in \mathbb{R}^{r \times k}$ and $\boldsymbol{B} \in \mathbb{R}^{d \times r}$ are trainable parameters, and $\boldsymbol{W}_0 \in \mathbb{R}^{d \times k}$ denote the pre-trained weights. During the fine-tuning stage, the forward pass is $\boldsymbol{y} = \big(\boldsymbol{W}_0 \ + \ \gamma_r \boldsymbol{B}\boldsymbol{A}\big)\boldsymbol{x} = \boldsymbol{W}_0\boldsymbol{x} \ + \ \gamma_r \boldsymbol{B}\boldsymbol{A}\boldsymbol{x}$.
- **QLoRA**. Quantized LoRA (QLoRA) [7] further reduces memory usage by using 4-bit quantization. During fine-tuning, all weights of the pre-trained model are quantized to 4 bits. Weights will be dynamically dequantized back to 16 bits when performing computation with the input sequence $\boldsymbol{x}$ and the adapter matrix $\boldsymbol{A}$ and $\boldsymbol{B}$, which remain in 16-bit precision throughout the process, where

$\boldsymbol{y} = p_{16}(\boldsymbol{W}_0^{\text{NF4}})\boldsymbol{x} + \gamma_r \boldsymbol{BAx}$. The process is similar in the inference stage, where the merged weights $\boldsymbol{W}$ are loaded in 4-bit precision.

- **DoRA**. Weight-Decomposed Low-Rank Adaptation (DoRA) [21] decomposes $\boldsymbol{W}_0 \in \mathbb{R}^{d\times k}$ into a column-wise magnitude vector $\boldsymbol{m} \in \mathbb{R}^{1\times k}$ and a direction matrix $\boldsymbol{V} \in \mathbb{R}^{d\times k}$, where $\boldsymbol{m} = \|\boldsymbol{W}_0\|_c$ (with $\|\cdot\|$ being column-wise norm) and $\boldsymbol{V} = \boldsymbol{W}_0$. Only the direction matrix receives updates through LoRA. The magnitude vector is updated separately. DoRA can achieve accuracy close to that from full fine-tuning while keeping the same parameter count as LoRA.
- **rsLoRA**. Vanilla LoRA uses a scaling factor α/r, which may cause gradients to explode or diminish as the rank r increases. Rank-Stabilized LoRA (rsLoRA) [16] uses a scaling factor $\alpha/\sqrt{r}$: $\boldsymbol{W}' = \boldsymbol{W}_0 + \frac{\alpha}{\sqrt{r}}\boldsymbol{BA}$. This scaling results in gradient-scale stability at higher ranks, enabling the rank to be higher for long-context tasks like XBRL analysis.
- **LoRA with Federated Learning**. In the finance sector, multiple institutions may want to collaborate using their own proprietary datasets, but they cannot share their data due to compliance reasons and privacy concerns. Federated learning solves this issue by fine-tuning a model on local data and aggregating LoRA updates to a central node.

3.3 Benchmark Angles

Angle I: LoRA Methods' Performance on Financial Datasets We seek to learn which LoRA method is most effective in financial tasks, in terms of both category-specific and overall performance, and how these LoRA fine-tuned models perform compared to existing state-of-the-art (SOTA) models. We fine-tuned Llama 3.1 8B Instruct using LoRA, QLoRA, rsLoRA, and DoRA, representing open-source models and fine-tuning approaches, and fine-tuned Gemini 2.0 Flash Lite using Google's proprietary fine-tuning methods as a baseline representing closed-source counterparts.

Angle II: LoRA Suitability for Financial Tasks We wish to investigate how the benefits of LoRA fine-tuning vary across different financial tasks. This angle is motivated by the need to identify which specific applications (e.g., sentiment analysis, XBRL tagging, XBRL analysis) are most responsive to fine-tuning, and what properties of the datasets cause this.

Angle III: Resources of LoRA Fine-tuning and Inference We aim to compare which LoRA methods, out of the tested methods, are the most cost-effective in fine-tuning and compare the fine-tuning cost to closed-source fine-tuning services. We are also motivated to measure and compare the inference speeds of LoRA-fine-tuned models against their larger base model counterparts. The goal is to quantify the potential for reduced latency and increased throughput, which are critical for real-time financial applications and operational efficiency.

Angle IV: Practical Considerations for LoRA Deployment in Finance To assess the viability of deploying LoRA-fine-tuned models in real-world financial scenarios, we investigate two key concerns: *(i)* Data Privacy in Collaborative Training: While local LoRA fine-tuning enhances data protection, collaborative model training across multiple institutions often requires approaches like Federated Learning to preserve the privacy of proprietary training data. We investigate this by simulating data distribution across several nodes and evaluating LoRA fine-tuning performance against centralized training. *(ii)* Catastrophic Forgetting: Fine-tuning can risk degrading a model's pre-existing general knowledge and capabilities. To quantify this, we evaluate our LoRA-fine-tuned models on established general-domain benchmarks, such as MMLU [11], measuring any performance changes on tasks outside their financial fine-tuning scope.

4 Benchmark Results

Setup Our experiments were conducted on four NVIDIA A5000 GPUs. For closed-source models, we employed various inference and fine-tuning APIs. For each LoRA method, we fine-tuned 9 LoRA adapters based on their respective training sets merged by task categories. We used a learning rate of 1e-4 and a batch size of 2–8 based on prompt length (Refer to Appendix C for details). For inference, we used a temperature of 0.0. Overall, we conducted 46 rounds of fine-tuning and 194 rounds of evaluations to benchmark these LoRA methods from different angles.

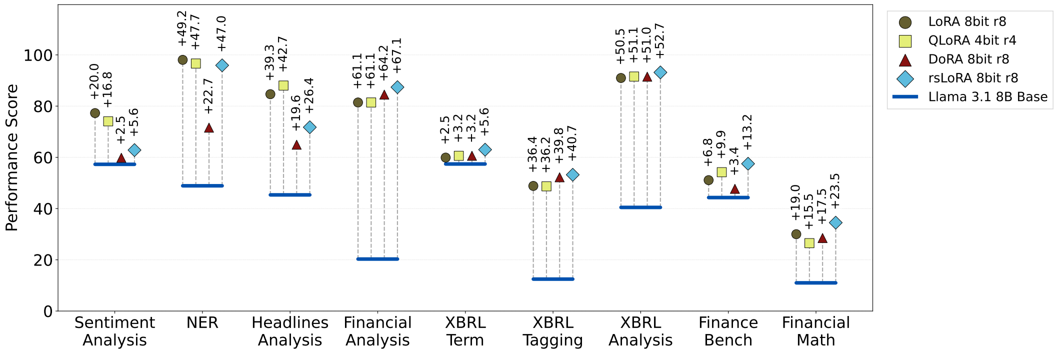

Figure 2: Task suitability.

4.1 Angle I: LoRA Methods Performance on Financial Datasets

Comparative Performance of LoRA Variants Table 4 shows the performance of base models and different LoRA fine-tuned models. Vanilla LoRA (8-bit, rank 8) achieves the highest overall average score (74.74), a 37.69% increase over the Llama 3.1 8B base model's 37.05. Fig. 1 shows the performance by category. Vanilla LoRA outperforms other LoRA variants in general financial tasks, while rsLoRA leads in financial analysis, financial reporting, and financial statement analysis.

rsLoRA Performs Better at High Ranks rsLoRA scales with $\alpha/\sqrt{r}$ instead of α/r to prevent gradient exploding or vanishing at large ranks. We set $r = 8$ for memory efficiency. rsLoRA just slightly underperforms against LoRA and QLoRA. The rsLoRA paper's experiments [16] led to lower perplexity at higher ranks (e.g., r = 64). This lower perplexity and the fact that higher rank LoRA captures more details suggest rsLoRA's benefits are primarily exploited at high ranks.

DoRA Benefits from Two Learning Rates DoRA performed worse than the other three LoRA methods. We used the same learning rate for updating the magnitude vector and direction matrix. However, as shown in Table 4, this can lead to sub-optimal performance in some cases due to the gradient scales being different between the two types of updates in DoRA. This leads to DoRA sometimes under-training the magnitude vector in our experiments, which uses the same low learning rate. Thus, DoRA may achieve higher performance if the magnitude vector has its own learning rate that is higher than the low-rank update's learning rate.

LoRA-Tuned Llama 3.1 8B vs. Baseline Models and Gemini Fine-Tuned Compared to SOTA base LLMs, the LoRA-tuned Llama 3.1 8B Instruct models generally show superior performance across most datasets, with NWGI and FNXL being the exceptions. Against another fine-tuned baseline, the Gemini 2.0 FL fine-tuned model, this Gemini model excels in general financial tasks and XBRL data reporting. However, our Llama 3.1 8B Instruct LoRA variants demonstrate stronger average performance in financial analysis and XBRL data analysis tasks.

4.2 Angle II: Financial Task LoRA Suitability

Fig. 2 highlights LoRA's varying effectiveness across different financial tasks. A key observation is the contrast in LoRA method improvements between XBRL Analysis tasks and FinanceBench. Although both aim to analyze financial statements, tasks based on XBRL data demonstrate substantial LoRA-induced performance improvements, whereas FinanceBench exhibits minimal gains. This disparity underscores XBRL's superior suitability for financial statement analysis. The standardized semantics and taxonomy inherent in XBRL likely provide a more structured and consistent learning environment for LLMs, facilitating more effective adaptation compared to FinanceBench, which relies on OCR-processed PDF data lacking such rich, standardized metadata. These findings emphasize the crucial role of XBRL in enabling effective LLM integration for financial report analysis.

Table 4: Performance on financial tasks: accuracy in blue, F1 in gray, and BERTScore F1 in green.

Datasets	**Base Models**					**Fine-tuned Models**				
	Llama 3.1 8B [8]	Llama 3.1 70B [8]	DeepSeek V3 [18]	GPT-4o [14]	Gemini 2.0 FL [39]	Llama 3.1 8B LoRA 8bit-r8	Llama 3.1 8B QLoRA 4bit-r4	Llama 3.1 8B DoRA 8bit-r8	Llama 3.1 8B rsLoRA 8bit-r8	Gemini 2.0 FL N/A N/A
General Financial Tasks										
FPB	68.73 0.677	74.50 0.736	78.76 0.764	81.13 0.818	81.02 0.894	85.64 **0.922**	84.16 0.909	81.93 0.901	82.84 0.853	**87.62** 0.878
FiQA SA	46.55 0.557	47.27 0.565	60.43 0.686	72.34 0.773	68.09 0.810	81.28 **0.884**	78.30 0.874	78.72 0.874	73.19 0.806	**88.09** 0.879
TFNS	69.97 0.683	68.42 0.686	84.38 0.846	73.32 0.740	26.38 0.385	88.02 **0.932**	83.84 0.910	59.09 0.702	59.51 0.655	**89.49** 0.896
NWGI	43.86 0.583	50.14 0.596	7.44 0.097	66.61 0.656	48.16 0.614	54.16 **0.690**	49.96 0.645	19.57 0.281	35.80 0.464	**62.59** 0.581
NER	48.89 0.569	46.28 0.454	40.82 0.360	52.11 0.523	65.13 0.769	**98.05** **0.981**	96.63 0.966	71.59 0.834	95.92 0.963	97.29 0.973
Headline	45.34 0.558	71.68 0.729	76.06 0.779	80.53 0.814	76.60 0.847	84.66 0.852	88.03 0.886	64.93 0.781	71.75 0.828	**97.32** **0.973**
Financial Certificate Tasks										
CFA Level 1	13.33 0.133	42.22 0.418	54.44 0.556	63.33 0.631	55.56 0.556	86.67 0.867	**87.78** **0.878**	**87.78** **0.878**	**87.78** **0.878**	52.22 0.530
CFA Level 2	19.48 0.199	29.87 0.303	46.75 0.485	55.84 0.563	56.67 0.567	88.31 0.883	83.12 0.835	90.91 0.909	**92.21** **0.922**	51.11 0.519
CFA Level 3	16.67 0.179	24.36 0.271	47.44 0.496	51.28 0.517	52.56 0.538	70.51 0.705	66.67 0.675	69.23 0.697	**79.49** **0.795**	51.28 0.557
CPA REG	31.68 0.317	41.58 0.426	65.35 0.654	67.33 0.667	63.37 0.638	80.20 0.802	88.12 0.885	**90.10** **0.901**	**90.10** **0.901**	51.28 0.557
Financial Reporting Tasks										
FiNER	21.28 0.232	61.82 0.606	68.92 0.699	72.29 0.725	63.91 0.638	74.10 0.759	74.32 0.760	70.92 0.732	70.72 0.724	**80.32** **0.802**
FNXL	3.64 0.045	20.14 0.210	27.33 0.288	42.41 0.398	37.75 0.356	23.57 0.250	23.05 0.253	33.50 0.311	35.68 0.348	**47.98** **0.438**
XBRL Term	0.574	0.587	0.573	0.584	0.572	0.599	0.606	0.606	0.630	**0.666**
Financial Statement Analysis Tasks										
Tag Extraction	69.16 0.739	69.64 0.782	85.03 0.849	81.60 0.864	80.27 0.811	**89.13** 0.886	86.89 0.872	80.44 0.896	85.26 0.879	85.03 **0.907**
Value Extraction	52.46 0.565	88.19 0.904	98.01 0.982	97.01 0.974	98.02 0.980	98.49 0.986	97.14 0.974	98.57 0.988	99.13 **0.992**	**99.20** **0.992**
Formula Construction	12.92 0.201	59.28 0.665	22.75 0.315	79.76 0.820	61.90 0.644	77.61 0.876	89.34 **0.898**	88.02 0.882	**89.46** 0.893	67.85 0.786
Formula Calculation	27.27 0.317	77.49 0.783	85.99 0.868	83.59 0.857	53.57 0.536	98.68 0.990	92.81 0.947	**98.92** **0.993**	98.80 0.988	54.76 0.548
Finance Bench	0.443	0.528	0.573	0.564	0.552	0.511	0.542	0.477	**0.575**	0.544
Financial Math	11.00 0.136	10.50 0.134	21.50 0.255	27.00 0.296	19.00 0.204	30.00 0.332	26.50 0.307	28.50 0.317	34.50 0.370	**66.00** **0.785**
Overall Average (Using BERTScore F1 × 100)										
Aggregated	37.05	52.36	57.16	63.39	58.97	**74.74**	74.29	69.53	73.82	71.08

Table 5: Comparison of fine-tuning cost. GPT-4o cost is estimated based on 4 epochs of fine-tuning at OpenAI fine-tuning pricing [32].

Models	Time	GPUs	Est. Cost (USD)
BloombergGPT [42]	53 days	512×A100	$2.7 M
LoRA	14.9h	4 × A5000	$15.50
QLoRA	14.1h	4 × A5000	$14.66
DoRA	15.9h	4 × A5000	$16.54
rsLoRA	14.5h	4 × A5000	$15.11
Gemini 2.0 FL	8.8h	-	$162.02
GPT-4o-mini	-	-	$312.00

Figure 3: Average inference time of LoRA fine-tuned Llama 3.1 8B and LoRA fine-tuned Gemini 2.0 FL across tasks

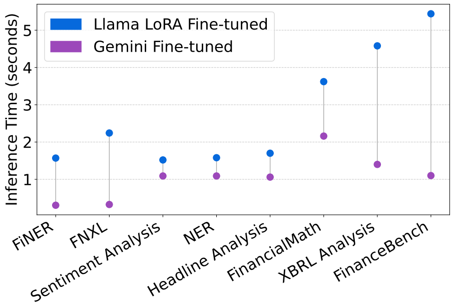

Table 6: Accuracy on MMLU & GSM8K benchmarks for Llama 3.1 8B base and eight LoRA adapters. Scores are colored relative to base: gray (same), green (higher), red (lower)

Dataset	Llama 3.1 8B (base)	Llama 3.1 8B Adapters			
		LoRA 8bit-r8	QLoRA 4bit-r4	DoRA 8bit-r8	rsLoRA 8bit-r8
MMLU [11] (*Sentiment*)	0.229	0.229	0.229	0.229	0.229
MMLU (*FiNER*)	0.229	0.229	0.229	0.229	0.229
GSM8K [6] (*Sentiment*)	0.011	0.014	0.014	0.011	0.011
GSM8K (*FiNER*)	0.011	0.011	0.014	0.011	0.016

Table 7: Performance comparison of central LoRA and LoRA federated learning using four nodes on sentiment analysis tasks: accuracy (blue) and F1 score (gray).

Llama 3.1 8B 8bit-r8	FPB	FiQA SA	TFNS	NWGI
Base	68.73 0.677	46.55 0.557	69.97 0.683	46.58 0.412
Central (LoRA)	**89.11** **0.941**	**88.09** **0.923**	**91.96** **0.955**	**61.92** **0.748**
FedAvg [30]	82.43 0.902	76.17 0.860	73.41 0.842	56.02 0.698

4.3 Angle III: Resource Usage and Performance Trade-offs of LoRA methods

Table 5 details the computational costs of LoRA fine-tuned models. Using four NVIDIA A5000 GPUs, the wall-clock time for fine-tuning ranged from 14.1 hours (QLoRA) to 15.9 hours (DoRA), corresponding to a total of approximately 56.4 to 63.6 GPU hours. At an estimated rate of $0.26 per GPU hour, this translates to a cost of roughly $14.66 to $16.54. This is substantially more cost-effective than fine-tuning services from providers like Google or OpenAI. Figure 3 illustrates the inference time of fine-tuned models on various datasets. Gemini API generally exhibits lower inference latency and is less sensitive to increasing prompt lengths than local Llama 3.1 8B Instruct inference, even when accounting for network overhead for the API. However, the inference speed of locally deployed Llama models can be significantly enhanced through the use of larger batch sizes.

4.4 Angle IV: Practicability of Applying LoRA in Real-world Financial Scenarios

Federated LoRA The sensitive nature of financial data necessitates privacy-preserving techniques like Federated Learning for collaborative training. To explore this, we evaluated Federated LoRA [38], with results presented in Table 7. Our experimental setup simulated a four-node environment employing the FedAvg algorithm [30], where the sentiment analysis dataset was partitioned across these nodes. The performance of this approach was benchmarked against both the base Llama model and standard centralized LoRA fine-tuning. While Federated LoRA did not match the performance levels of centralized LoRA, the results demonstrate a notable improvement compared to the base Llama model.

Catastrophic Forgetting A major concern with PEFT is that fine-tuning on domain-specific tasks leads to the model forgetting pre-training knowledge. To investigate this, we evaluated eight adapters—covering both sentiment and FiNER tasks and all four LoRA variants—as well as the Llama 3.1 8B Instruct base model on two out-of-domain benchmarks, MMLU [11] and GSM8K [6]. We used a zero-shot, no chain-of-thought setting to isolate stored knowledge. Table 6 shows identical MMLU accuracy across all adapters and the base model, and equal or higher scores on GSM8K. Hence, at the ranks r we tested (4 and 8) with $\alpha{:}r$ equal to 8:1 or 4:1, we observe that LoRA does not exhibit catastrophic forgetting. In fact, the slight GSM8K performance improvements hint at cross-domain knowledge transfer—fine-tuning on financial data may improve the model's numerical reasoning skills.

5 Related Works

5.1 Financial LLMs and Benchmarks

BloombergGPT [42] is the first LLM specialized for the financial domain. The-50-billion parameter model was trained from scratch using a mix of financial and general datasets. The evaluation was conducted on a series of financial tasks including sentiment analysis, named entity recognition (NER), and question answering (QA) as well as general benchmarks, showing performance exceeding comparable models on financial tasks and strong performance on general tasks.

FinGPT [22, 24, 23] aims to provide a customized and personalized financial LLM. Instead of training from the ground up, FinGPT applied LoRA fine-tuning on open-source LLMs using general financial training sets. Performance evaluation displayed noticeable improvement over the base model, even surpassing that of BloombergGPT, while having substantial memory reduction and training speedup compared to training-from-scratch.

FinBen [44] and PIXIU [45] are financial benchmarks that offer a broad array of curated datasets. Benchmarking various general LLMs, they conclude that while LLMs demonstrate strong capabilities in textual analysis, they face challenges with advanced reasoning and complex financial problem-solving.

5.2 Parameter-Efficient Fine-Tuning (PEFT) with Low-rank Adaptation (LoRA) Methods

Full fine-tuning, which fine-tunes the full parameters of an LLM, is extremely computationally expensive. Parameter-efficient fine-tuning (PEFT) was proposed to reduce the number of trainable parameters by only fine-tuning a small number of model parameters [29]. Low-rank adaptation (LoRA) [12] is a widely used PEFT method that inserts a smaller set of pluggable low-rank trainable weights. The performance of downstream tasks after LoRA fine-tuning is comparable to that of full fine-tuning. Quantized LoRA (QLoRA) [7] quantizes the LLM to 4 bits and applies LoRA fine-tuning on such a model. QLoRA can significantly reduce GPU memory usage.

5.3 LoRA Methods with Federated Learning

In the financial domain, private training data might be spread across multiple institutions. To fine-tune LLMs with non-centralized data, federated learning is needed. Several research papers have applied LoRA on federated learning, such as PrivateLoRA [40] and Federated Freeze A LoRA (FFA-LoRA) [38].

6 Conclusion and Future Work

In this paper, we present FinLoRA, a benchmark that evaluates LoRA methods on both general and highly specialized financial tasks. We curated 19 diverse datasets covering a wide range of financial applications. Our study includes 46 rounds of fine-tuning and 194 rounds of evaluation to thoroughly assess and analyze commonly used LoRA methods. FinLoRA offers insights into overall performance, task-specific results, resource requirements for fine-tuning and inference, and practical considerations for real-world deployment—including data privacy in collaborative training and catastrophic forgetting. Our results demonstrate that fine-tuning can significantly enhance the effectiveness of LLMs on financial tasks. Additionally, FinLoRA provides a comprehensive collection of datasets with baseline results, laying a solid foundation for future research in this field. Moving forward, we plan to expand FinLoRA by incorporating additional LoRA methods into the project.

References

[1] Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. FinTral: A family of GPT-4 level multimodal financial large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13064–13087, August 2024.

[2] Ethan Callanan, Amarachi Mbakwe, Antony Papadimitriou, Yulong Pei, Mathieu Sibue, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. Can GPT models be financial analysts? an evaluation of ChatGPT and GPT-4 on mock CFA exams. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 23–32, Jeju, South Korea, 3 August 2024. -.

[3] Lulu Chen, Yingzhou Lu, Chiung-Ting Wu, Robert Clarke, Guoqiang Yu, Jennifer E Van Eyk, David M Herrington, and Yue Wang. Data-driven detection of subtype-specific differentially expressed genes. *Scientific Reports*, 11(1):332, 2021.

[4] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

[5] Tianyi Chen, Nan Hao, Capucine Van Rechem, Jintai Chen, and Tianfan Fu. Uncertainty quantification and interpretability for clinical trial approval prediction. *Health Data Science*, 4:0126, 2024.

[6] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, and et al. The Llama 3 herd of models, 2024.

[9] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[10] Shijie Han, Haoqiang Kang, Bo Jin, Xiao-Yang Liu, and Steve Y Yang. XBRL Agent: Leveraging large language models for financial report analysis. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 856–864, 2024.

[11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[12] Edward J Hu, Yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[13] Muchen Huan and Jianhong Shun. Fine-tuning transformers efficiently: A survey on LoRA and its impact. *Preprints*, February 2025.

[14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[15] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering, 2023.

[16] Damjan Kalajdzievski. A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA, 2023.

[17] Jean Lee, Nicholas Stevens, and Soyeon Caren Han. Large language models in finance (finllms). *Neural Computing and Applications*, January 2025.

[18] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[19] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

[20] Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. SocraticLM: Exploring socratic personalized teaching with large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[21] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

[22] Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. Data-centric FinGPT: Democratizing internet-scale data for financial large language models. In *Workshop on Instruction Tuning and Instruction Following, NeurIPS*, 2023.

[23] Xiao-Yang Liu, Jie Zhang, Guoxuan Wang, Weiqin Tong, and Anwar Walid. Efficient Pretraining and Finetuning of Quantized LLMs with Low-Rank Structure . In *IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*, pages 300–311, July 2024.

[24] Xiao-Yang Liu, R. Zhu, Daochen Zha, J. Gao, S. Zhong, Matt White, and Meikang Qiu. Differentially private low-rank adaptation of large language model using federated learning. *ACM Transactions on Management Information Systems*, 2024.

[25] Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, May 2022.

[26] Yingzhou Lu, Chiung-Ting Wu, Sarah J Parker, Zuolin Cheng, Georgia Saylor, Jennifer E Van Eyk, Guoqiang Yu, Robert Clarke, David M Herrington, and Yue Wang. COT: an efficient and accurate method for detecting marker genes among many subtypes. *Bioinformatics Advances*, 2(1), 2022.

[27] Macedo Maia, Siegfried Handschuh, Andre Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: Financial opinion mining and question answering. pages 1941–1942, 04 2018.

[28] Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. Good debt or bad debt: Detecting semantic orientations in economic texts, 2013.

[29] Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. A survey on lora of large language models. *Frontiers of Computer Science*, 19(7), December 2024.

[30] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

[31] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024.

[32] OpenAI. OpenAI API pricing. `https://platform.openai.com/docs/pricing`, May 2025. Accessed: 2025-05-14.

[33] Md. Abdur Rahman. Twitter financial news sentiment. `http://precog.iiitd.edu.in/people/anupama`, 2022.

[34] Ali Saeedi, Jim Richards, and Barry Smith. An introduction to XBRL. In *British Accounting Association's Annual Conference*, 2007.

[35] Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. Domain adaption of named entity recognition to support credit risk assessment. In Ben Hachey and Kellie Webster, editors, *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia, December 2015.

[36] Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. Financial numeric extreme labelling: A dataset and benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3550–3561, July 2023.

[37] Ankur Sinha and Tanmay Khandait. Impact of news on the commodity market: Dataset and results, 2020.

[38] Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving loRA in privacy-preserving federated learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[39] Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. Gemini: A family of highly capable multimodal models, 2024.

[40] Yiming Wang, Yu Lin, Xiaodong Zeng, and Guannan Zhang. PrivateLoRA for efficient privacy preserving LLM, 2023.

[41] Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. TWIN-GPT: Digital twins for clinical trials via large language model. *ACM Trans. Multimedia Comput. Commun. Appl.*, July 2024. Just Accepted.

[42] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[43] Yang Wu, Chenghao Wang, Ece Gumusel, and Xiaozhong Liu. Knowledge-infused legal wisdom: Navigating llm consultation through the lens of diagnostics and positive-unlabeled reinforcement learning. In *ACL (Findings)*, pages 15542–15555, 2024.

[44] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Huang Jiajia, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. FinBen: An holistic financial benchmark for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[45] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. PIXIU: A comprehensive benchmark, instruction dataset and large language model for finance. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[46] Menglin Yang, Jialin Chen, Yifei Zhang, Jiahong Liu, Jiasheng Zhang, Qiyao Ma, Harshit Verma, Qianru Zhang, Min Zhou, Irwin King, and Rex Ying. Low-rank adaptation for foundation models: A comprehensive review, 2024.

[47] Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models. In *ACM International Conference on AI in Finance*, pages 349–356, 2023.

[48] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[49] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

[50] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[51] Shuang Zhou, Zidu Xu, Mian Zhang, Chunpu Xu, Yawen Guo, Zaifu Zhan, Sirui Ding, Jiashuo Wang, Kaishuai Xu, Yi Fang, et al. Large language models for disease diagnosis: A scoping review. *arXiv preprint arXiv:2409.00097*, 2024.