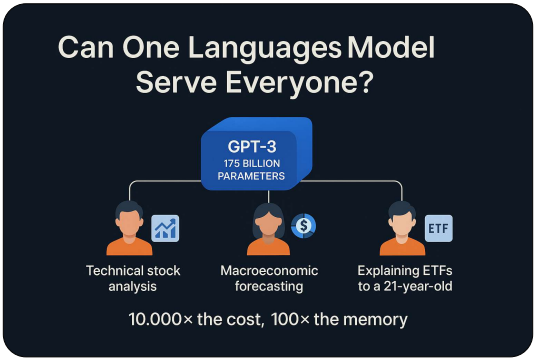
Priyanshu Gupta (19750534)

LoRA and Efficient LLM Serving for Financial Domain Expert Agents

An Efficient Approach for Specialized Financial Intelligence

Table of contents

Introduction

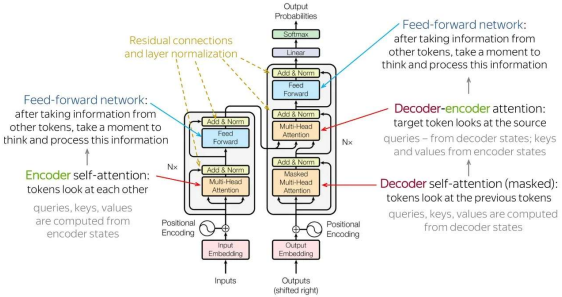


This is where LoRA and serving frameworks like S-LoRA come in, enabling large-scale personalization at minimal cost, without fine-tuning full models.

Image source: AI Generated
Hu et al., *LoRA*

Background Theory

How Transformers Work: A Step-by-Step Breakdown



Input Sequence

A sequence of words or tokens that the model will process.

Image source:
https://medium.com/@asadali.syne/understanding-the-transformer-architecture-in-llm-e475453879fe

Challenges

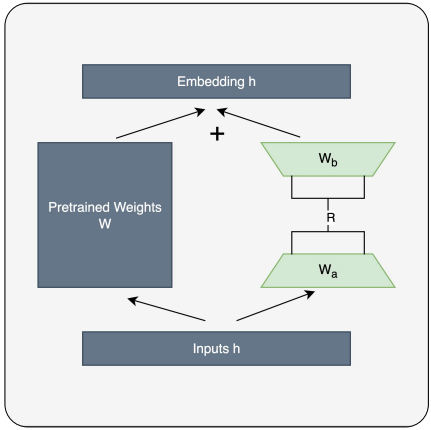
- Specialized Knowledge: Niche, complex, time-intensive
- LLM Limitations: Hallucinations, limited memory, complex instructions
- Traditional Fine-tuning: Prohibitively expensive for large models

Kaler et al., "A Natural Way of Building Financial Domain Expert Agents"
Hu et al., *LoRA*

Outline

Introducing LoRA: Low-Rank Adaptation

- Efficient Adaptation: Freezes base weights, injects small trainable matrices (A, B)
- Low Parameters: Reduces trainable parameters by 10,000x for GPT-3 175B
- GPU Memory: Cuts GPU memory by 3x
- No Inference Latency: Merges with base model for deployment



Hu et al., *LoRA*

The Need for Specialized LLM Serving

- Proliferation of LoRAs: Thousands of specialized models
- Serving Challenge: Managing many LoRA adapters concurrently
- Memory Bottleneck: KV Cache and Adapter Weights
- Throughput Constraint: Inefficient batching, dynamic request lengths

Chen et al., *Punica*
Sheng et al., *S-LoRA*
Kwon et al., *Efficient Memory Management for Large Language Model Serving with PagedAttention*

Different Approaches for Training LoRA

- Single LoRA on Combined Datasets
  - Combine all datasets $\rightarrow$ train one unified LoRA
  - Captures patterns across domains
  - Risk: Domain imbalance. Mitigate with data balancing or curriculum learning
- Separate LoRAs per Dataset
  - Train one LoRA per dataset/domain
  - LoRAs can be: Swapped, merged, blended, or used independently
  - Used in PEFT, diffusion models, LoRA merging tools
- Sequential (Chained) LoRA Training
  - Train on Dataset A $\rightarrow$ continue training on Dataset B
  - Ideal for domain adaptation
  - Risk: Catastrophic forgetting

Bafghi et al., *Fine Tuning without Catastrophic Forgetting via Selective Low Rank Adaptation*
Kalajdzievski, *Scaling Laws for Forgetting When Fine-Tuning Large Language Models*

LoRA Variants

Variant	Key Feature	Use Case
LoRA (Original)	Adds low-rank layers to frozen base	Fine-tune small/mid models per domain (e.g., sentiment)
QLoRA	4-bit quantized base model	Finetune large model (13B+) on low-end GPUs
S-LoRA	Runtime adapter switching	Multi-skill LLM/chatbot switching at runtime
Multi-Tenant LoRA	Hosts multiple adapters concurrently	Serve many users/tasks with isolated adapters

Table: LoRA Variants and Their Applications

Hu et al., *LoRA*
Sheng et al., *S-LoRA*
Wu et al., "dLoRA: Dynamically Orchestrating Requests and Adapters for LoRA LLM Serving"
Chen et al., *Punica*

Evaluation Metrics

- Language Model Performance
  - Accuracy, F1-Score
  - BLEU, ROUGE-L, METEOR (text generation)
  - BERTScore (semantic similarity)
  - Binary Accuracy (binacc) (e.g., market movement classification)
  - Mean Squared Error (MSE) (numerical predictions)
  - Matthew's / Pearson Correlation (CoLA, STS-B)

D. Wang et al., *FinLoRA*

Evaluation Metrics

- Serving System Metrics
  - Throughput Latency: Requests/sec, Tokens/sec. First Token Latency, P90, Decode, Prefill, Kernel Latency
  - Resource Efficiency: GPU Utilization, KV Cache Memory Usage, Memory Fragmentation, Unified Memory Pool Utilization
  - Scalability Overhead: Switching Overhead (LoRA adapters), I/O Computation Overhead, Multi-GPU Communication Cost
  - User-Centric Metrics: SLO Attainment (% within latency target), User Satisfaction Score

Evaluation Metrics

- LoRA-Specific Metrics
  - Overfitting / Underfitting
  - Dataset Confusion
  - Performance vs Efficiency Tradeoff

Outline

Building Financial Domain Expert Agents (Architecture)

- Iterative Approach: LLM + Layers
- Data Extraction Layer: SQL/API access
- Scripting Layer: Python for complex analysis, error handling
- Memory Layer: Stores processes, domain knowledge, keyword meanings

Kaler et al., “A Natural Way of Building Financial Domain Expert Agents”

Financial LLMs: Key Approaches

- Prompt Engineering: Role-playing, few-shot examples
  - Cons: High token usage if the task is complex
- Retrieval-Augmented Generation (RAG): Finance-specific docs (10-Ks)
  - Cons: Initial setup cost, model migration is difficult, not for real-time high frequency data
- Tool-Augmented Agents: Python, Web search, PDF parsers

Financial LLMs: Key Approaches

- LoRA/PEFT: Internalize financial tone, structure
- Finetuned GPT Models: Report summarization, KPI extraction
- Instruction Tuning: Financial reasoning tasks
- Domain-Adaptive Pretraining (DAPT): Custom base model

Data Sources for Finance LLMs

Source Type	Examples	Data Formats
Regulatory Filings	10-K, 10-Q, 8-K, S-1	XBRL, PDF
Earnings Calls	Transcripts, Analyst Q&A	PDF
Research Reports	Equity Research, DCF Models, Market Outlooks	PDF, DOCX
News Articles	Reuters, Bloomberg, Yahoo Finance	Raw Text, JSON
Market Data	Technical Indicators, OHLCV Prices	JSON

Table: Representative Data Sources and Formats for Finance LLMs

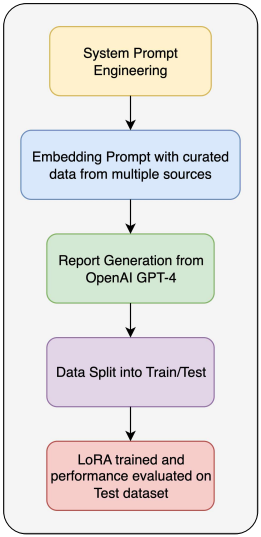
FinGPT

- A large language model tailored for financial tasks
- Learns user preferences via Reinforcement Learning from Human Feedback (RLHF)
- Inspired by techniques used in ChatGPT/GPT-4

Yang, Liu, and C. D. Wang, “FinGPT: Open-Source Financial Large Language Models”

FinGPT - Architecture Overview

- Analyzes News (from Finnhub) and Stock data (from yfinance)
- Workflow: Structured prompts $\rightarrow$ GPT-4 $\rightarrow$ store prediction
- Trained with LoRA
- Models: chatglm2-6b, LLaMA-2-7b-chat-hf



Yang, Liu, and C. D. Wang, "FinGPT: Open-Source Financial Large Language Models"

FinLoRA

- A parameter-efficient fine-tuning framework for financial LLMs
- Targets the high cost of models like BloombergGPT
- Cuts fine-tuning cost to <$100 using LoRA methods

D. Wang et al., *FinLoRA*

FinLoRA

- Works with 150 SEC filings to handle complex XBRL formats
- Two primary applications: Financial Reporting for SMBs, Statement Analysis from structured reports
- LLaMA 3.1 + LoRA variants (LoRA, QLoRA, DoRA, rsLoRA)

FinLoRA achieved +36% performance over base models using lightweight, low-cost LoRA tuning — making financial LLMs more accessible.

D. Wang et al., *FinLoRA*

Outline

Conclusion: The Future of Financial AI

- LoRA's Impact: Efficient specialization, lower barriers
- Serving Systems: Scalable, efficient, multi-tenant
- Expert Agents: Mimic human expertise, automate tasks
- Enhanced Capabilities: Precision, multi-step reasoning, memory

Hu et al., *LoRA*
Sheng et al., *S-LoRA*
Kaler et al., "A Natural Way of Building Financial Domain Expert Agents"

Future Work

- Curate high-quality datasets from reliable financial sources
- Experiment with different LoRA training strategies
- Mix and match LoRA variants and base models
- Explore adapter fusion for better generalization
- Study weight selection logic and rank-deficiency

References I

Bafghi, Reza Akbarian et al. *Fine Tuning without Catastrophic Forgetting via Selective Low Rank Adaptation*. Jan. 26, 2025. DOI: 10.48550/arXiv.2501.15377. arXiv: 2501.15377 [cs]. URL: http://arxiv.org/abs/2501.15377 (visited on 07/23/2025). Pre-published.

Chen, Lequn et al. *Punica: Multi-Tenant LoRA Serving*. Oct. 28, 2023. DOI: 10.48550/arXiv.2310.18547. arXiv: 2310.18547 [cs]. URL: http://arxiv.org/abs/2310.18547 (visited on 06/02/2025). Pre-published.

Hu, Edward J. et al. *LoRA: Low-Rank Adaptation of Large Language Models*. Oct. 16, 2021. DOI: 10.48550/arXiv.2106.09685. arXiv: 2106.09685 [cs]. URL: http://arxiv.org/abs/2106.09685 (visited on 06/02/2025). Pre-published.

Kalajdzievski, Damjan. *Scaling Laws for Forgetting When Fine-Tuning Large Language Models*. Jan. 11, 2024. DOI: 10.48550/arXiv.2401.05605. arXiv: 2401.05605 [cs]. URL: http://arxiv.org/abs/2401.05605 (visited on 07/23/2025). Pre-published.

Kaler, Gagandeep Singh et al. "A Natural Way of Building Financial Domain Expert Agents". In: (2024).

Kwon, Woosuk et al. *Efficient Memory Management for Large Language Model Serving with PagedAttention*. Sept. 12, 2023. DOI: 10.48550/arXiv.2309.06180. arXiv: 2309.06180 [cs]. URL: http://arxiv.org/abs/2309.06180 (visited on 06/02/2025). Pre-published.

Sheng, Ying et al. *S-LoRA: Serving Thousands of Concurrent LoRA Adapters*. June 5, 2024. DOI: 10.48550/arXiv.2311.03285. arXiv: 2311.03285 [cs]. URL: http://arxiv.org/abs/2311.03285 (visited on 06/02/2025). Pre-published.

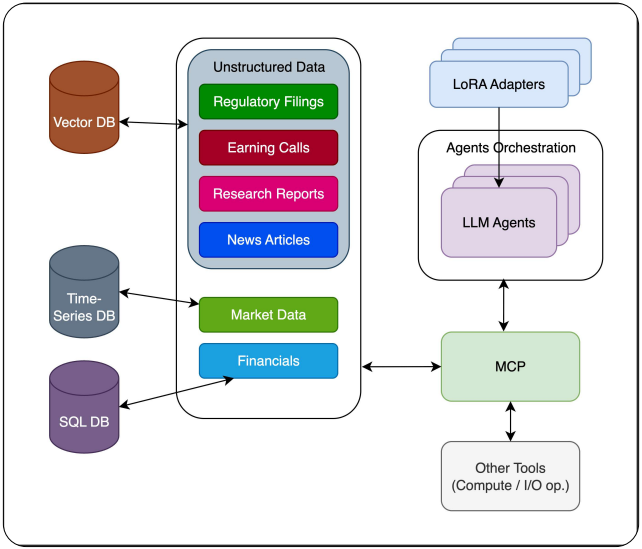
References II

Wang, Dannong et al. *FinLoRA: Benchmarking LoRA Methods for Fine-Tuning LLMs on Financial Datasets*. May 26, 2025. DOI: 10.48550/arXiv.2505.19819. arXiv: 2505.19819 [cs]. URL: http://arxiv.org/abs/2505.19819 (visited on 07/23/2025). Pre-published.

Wu, Bingyang et al. "dLoRA: Dynamically Orchestrating Requests and Adapters for LoRA LLM Serving". In: ().

Yang, Hongyang, Xiao-Yang Liu, and Christina Dan Wang. "FinGPT: Open-Source Financial Large Language Models". In: *FinLLM Symposium at IJCAI 2023* (2023).

Overview


Vector DB
Unstructured Data
Regulatory Filings
Earning Calls
Research Reports
News Articles
LoRA Adapters
Agents Orchestration
LLM Agents
Time-Series DB
Market Data
Financials
MCP
SQL DB
Other Tools (Compute / I/O op.)