FinBench Arena: An Independent Benchmarking Framework for Applied Financial Intelligence

Abrar Rahman, Anish Sundar

Abstract

The adoption of Artificial Intelligence (AI) in financial services has surged, yet the absence of rigorous, domain-specific benchmarking for model evaluation presents critical risks. Without standardized methods to assess alignment, reliability, and performance, institutions face potential vulnerabilities in compliance and user trust.

We introduce FinBench Arena, an open-source benchmarking platform designed to evaluate and compare Large Language Models (LLMs) within financial contexts. Built as a fork of Chatbot Arena, FinBench employs a pairwise comparison methodology, crowdsourced human feedback, and advanced statistical evaluation to measure model performance across financial tasks. The platform supports both general-purpose and proprietary models, allowing for nuanced assessment tailored to the unique demands of financial verticals. FinBench offers an intuitive interface and focuses on tasks critical to financial institutions, such as legal document summarization, compliance validation, and risk analysis.

This paper details the design, methodology, and preliminary results of FinBench Arena, demonstrating its utility in identifying strengths and limitations of LLMs in financial applications. We also discuss its potential to serve as a foundation for AI alignment and safety within tightly regulated industries.

I. Introduction

The rise of Large Language Models (LLMs) has transformed various industries. However, the unique demands of the financial world—such as complex regulatory language, interpreting vast amounts of market data, and numerical reasoning—have led to the development of **Financial Large Language Models (FinLLMs)**. These specialized models are trained to perform tasks specific to finance, including investment analysis, regulatory compliance, fraud detection, and/or real-time decision-making in trading environments.

Evolution of FinLLMs

The evolution of FinLLMs begins with the general-purpose language models that laid the groundwork for modern NLP techniques. Early models like **GPT-1** (Radford et al., 2018) [1], **GPT-2** (Radford et al., 2019) [2], and **BERT** (Devlin et al., 2018) [3] revolutionized language modeling through transfer learning, enabling models to handle diverse tasks with minimal

task-specific data. These models evolved into increasingly powerful architectures, such as **GPT-3** (Brown et al., 2020) [4], which introduced few-shot learning and was quickly followed by **GPT-4** (2023) [5], pushing the limits of language understanding and generation.

However, general-domain models, while powerful, were not fully equipped to handle the complexities of financial data. This led to the emergence of **financial-domain language models**, with **FinBERT-19** (Araci, 2019) [6] being one of the first domain-specific adaptations. Trained on financial texts, FinBERT set the stage for subsequent models that are now a cornerstone in financial NLP tasks, such as sentiment analysis, risk assessment, and financial question answering.

Techniques: From Pre-training to Instruction Tuning

The progression from general LLMs to FinLLMs involved various pre-training techniques tailored for the financial domain. **FinBERT-20** (Yang et al., 2020) [7] introduced **domain-specific pre-training from scratch**, where models were trained exclusively on financial corpora, resulting in improved performance on financial sentiment and text classification tasks. Subsequent models like **FinBERT-21** [8] and **FLANG** (Luo et al., 2022) [9] employed **mixed-domain pre-training**, combining general and financial data to balance performance across a wider range of tasks.

The most recent advancements, such as **BloombergGPT** (Wu et al., 2023) [10], represent **mixed-domain large language models** with prompt engineering specifically tailored for financial applications. These models are fine-tuned on diverse financial datasets and leverage instruction-based tuning to improve their reasoning abilities across complex financial tasks. New models like **FinMA** [11], **InvestLM** [12], and **FinGPT** [13] further enhance this process through **instruction tuning**, allowing LLMs to follow task-specific instructions for superior performance in tasks like market forecasting and regulatory compliance.

Existing Benchmarks and Datasets

A key challenge in evaluating FinLLMs is the development of suitable benchmark tasks that reflect the complexity of financial reasoning. Traditional NLP benchmarks, such as question answering or named entity recognition, need to be adapted to account for the domain-specific requirements of financial language, such as handling complex regulations or predicting stock market movements based on both textual and numerical data.

Some notable datasets used for benchmarking FinLLMs include:

- **Financial PhraseBank** [14] for sentiment analysis, which evaluates models on their ability to classify the sentiment of financial news and reports.
- **FinQA** [15] and **ConvFinQA** [16], datasets designed for numerical reasoning and question answering, where models must extract and compute information from both tabular and textual financial data.

- **StockNet** [17] and **StockEmotions** [18], which assess FinLLMs' ability to predict stock movements based on news articles and social media sentiment.

In addition to these standard tasks, **advanced financial tasks** such as **event detection** (e.g., identifying corporate actions that impact market movements), **relation extraction** (e.g., mapping connections between companies, transactions, and entities), and **causality detection** are essential for evaluating FinLLMs' understanding of more nuanced financial interactions.

Challenges and Opportunities in FinLLM Development

Despite significant progress, FinLLMs still face substantial challenges. One of the most pressing issues is the need for more advanced **numerical reasoning** capabilities. While general LLMs excel at textual understanding, financial decision-making often requires precise calculations, handling large datasets, and interpreting complex, domain-specific concepts like derivative pricing or risk modeling.

Another challenge lies in ensuring **fairness and transparency** in financial decision-making, especially as FinLLMs are increasingly applied to high-stakes areas like loan approvals, fraud detection, and compliance with anti-money laundering (AML) regulations. As the use of FinLLMs expands, there is a growing need for evaluation frameworks that assess not just accuracy, but also bias, ethical implications, and regulatory alignment.

On the opportunity side, FinLLMs hold the potential to transform financial services by enabling more accurate risk assessments, improving fraud detection mechanisms, and enhancing customer interactions through intelligent, real-time analysis of financial data. The use of FinLLMs in **market forecasting**, **algorithmic trading**, and **automated reporting** is expected to grow as models become more capable of handling multi-modal inputs (e.g., integrating textual, numerical, and visual data).

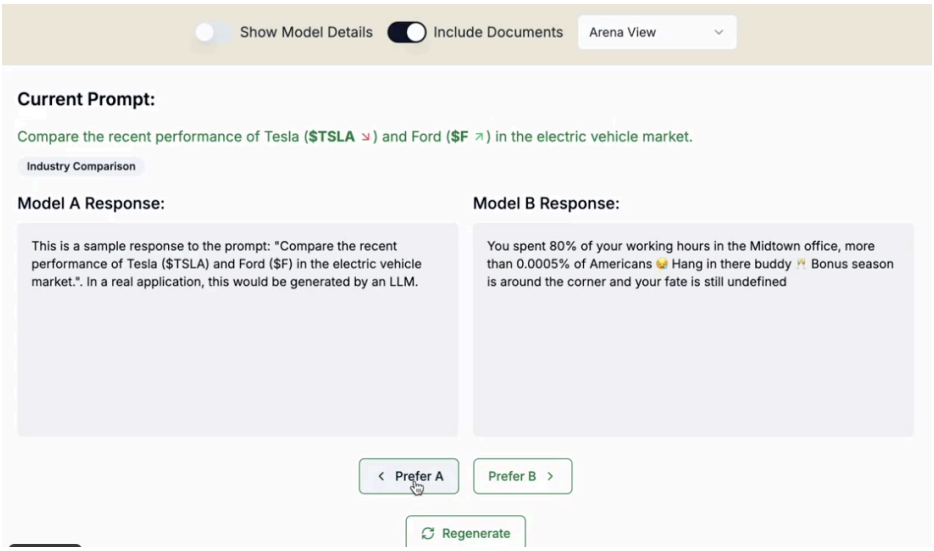
Financial Chatbot Arena: A Benchmarking Platform for FinLLMs

To address the need for comprehensive evaluation, we introduce **FinBench Arena**, a benchmarking platform designed specifically for FinLLMs. This platform facilitates real-world testing through crowdsourced financial queries and dynamic multi-step reasoning tasks. FinLLMs are evaluated on their ability to handle diverse financial challenges, including sentiment analysis, market predictions, and compliance verification.

The platform builds on previous work such as the original **Chatbot Arena** [19] and incorporates financial datasets like **FinPile** [20], **FinancialPhraseBank**, and **ConvFinQA**. By focusing on live interactions and explainability, the Financial Chatbot Arena offers a more holistic assessment of FinLLMs' capabilities in real-world financial settings.

II. Live Preference Tuning Framework

A key feature is its live preference tuning system, where user feedback directly informs alignment. However, given the scale of models (up to 70 billion parameters) and the nature of real-time interactions, this tuning process faces challenges. User feedback is sparse (one bit per generation), highly variable in quality, and subject to noisy, even adversarial input. These factors make extracting reliable signals a significant challenge, particularly in environments where accuracy and confidence are critical.



To address these challenges, the Arena leverages conformal prediction methods to manage uncertainty and ensure reliability. These techniques, used without modification, provide probabilistic guarantees about the confidence and validity of model outputs even under messy, dynamic conditions. By calibrating predictions and adapting to user preferences in real time, conformal methods allow the Arena to maintain robustness despite sparse and noisy feedback.

III. Holistic Benchmarking Framework

Due to space considerations, the list of question answer categories represented in our growing dataset will be relegated to the appendix.

General LLM Evaluation Metrics

Holistic Evaluation of Language Models (HELM) framework [22] provides a comprehensive approach to evaluating LLMs, including FinLLMs, by considering diverse modalities such as adversarial robustness, fairness, and calibration. HELM allows for a more extensive assessment of FinLLMs by capturing both standard and edge-case financial scenarios, ensuring they perform reliably across different financial contexts.

Metric	Description	Relevant Client Types	Reasoning Type
Heldout Loss	Measures model accuracy on unseen financial data sets	All client types	General Reasoning
Reading Comprehension	Evaluates the model's ability to generate correct answers from input texts	Verified Academics, Financial Research Analysts	Textual Understanding
Few-shot Learning	Evaluates model adaptability to new prompts without retraining	Venture Capital, Corporate M&A	Adaptive Reasoning

Financial-Specific Evaluation Metrics [23]

Metric	Description	Relevant Client Types	Reasoning Type
Financial Sentiment Analysis	Detects sentiment polarity in financial news, social media, earnings reports	Investment Banking, HFT, Equities	Textual Analysis
Numerical Reasoning	Solves financial calculations and interprets earnings reports	High-Frequency Traders, Research Analysts	Numerical Reasoning
Multi-Turn Financial QA	Handles multi-step financial inquiries, maintaining long-term context	Investment Banking, Corporate M&A, Customer Service	Complex Dialogue
Difficulty-Level Filters	Allows sorting based on task complexity (easy, medium, hard)	All clients	Task Adaptability

Synthetic data pipelines, powered by tools like NVIDIA Nemotron [24], will be used alongside real-world financial datasets (ex: Finance Alpaca [25], a combination of Stanford's Alpaca [26] and Google's FiQA [27] with synthetic data added via GPT3.5). These datasets provide a broad spectrum of cases for fairness evaluation and ensure that edge cases, which often reveal hidden biases, are adequately tested.

IV. Cohort & Fairness Evaluation Framework

To evaluate fairness in GPT-based financial decision systems, we adapted **Seismometer [28]**, an open-source statistical fairness evaluation library originally designed for binary classification

tasks. While Seismometer's core functionality focuses on cohort-based fairness in static decision models, we extend its capabilities to assess fairness in dynamic, multi-turn interactions characteristic of GPT-driven systems. This involves modifying its evaluation framework to accommodate the nuanced and probabilistic outputs of large language models (LLMs).

The fairness evaluations will be conducted retrospectively on logged interaction data. As the system operates and collects user interactions over time, we will apply fairness analyses to this historical data. This approach ensures that fairness assessments are grounded in real-world usage patterns rather than hypothetical scenarios, capturing the complexities of financial interactions at scale. Logged data will be grouped into cohorts (e.g., demographic or socioeconomic groups), and outcomes will be analyzed for statistical disparities using fairness metrics such as disparate impact, equal opportunity, and demographic parity.

Query Type	**Cohort Focus**	**Method to Detects Bias**	**Example Use Case**
Lending Decisions	Race, Gender, Income Levels	Identifies disproportionate rejection rates for minority or low-income groups.	Jamal vs. Jack receive different loan approvals for similar profiles.
Credit Scoring	Minority Groups, Geographic Region	Detects lower credit scores being disproportionately assigned to minority or rural populations.	Ensures that rural regions are not penalized in credit scoring algorithms.
Insurance Risk/Price	Health Disparities, Age, Gender	Finds discrepancies in insurance risk models that penalize older individuals or minority groups.	Detects higher premiums charged for minority or elderly populations.
Fraud Detection	Income Brackets, Transaction Types	Ensures that low-income individuals aren't disproportionately flagged as fraud risks.	Low-income earners aren't wrongly penalized by false positives.
Market Analysis	International Localization, Firm Region	Detects inconsistencies in market analysis outcomes across different regions or countries.	Cross-compare market predictions between international branches of a firm.
M&A Valuation	International Firms, Emerging Markets vs. US	Detects bias in valuing companies based on region or size.	Prevents emerging market companies from being undervalued.

IV. Conclusion

As generative AI advances, the surface area for negative consumer interactions in financial services expands significantly. Consider a voice-based financial assistant providing automated loan consultations. Unlike earlier systems that performed narrow, contained functions (e.g., credit scoring), generative AI's real-time, humanlike interactions bring new challenges to light. A customer could unknowingly receive advice influenced by biased data, believing it to be objective and human-generated. This raises ethical concerns about transparency, accountability, and trust in financial decision-making.

Generative AI's scale compounds these risks. While traditional systems might impact a subset of users—such as loan applicants—LLMs can engage millions simultaneously, amplifying systemic biases embedded in historical datasets. These biases, when presented through hyper-realistic interfaces, may mislead consumers into believing the advice they receive is accurate and impartial, increasing the potential for harm.

For example, as financial firms adopt AI-powered voices and digital avatars to explain products or conduct meetings, the line between human and machine continues to blur. Without clear disclosure, both consumers and professionals might unknowingly collaborate with AI systems in decisions that affect financial futures. The ethical implications are profound: these systems often lack the transparency and accountability inherent to human decision-makers.

FinBench Arena provides a critical safeguard against these risks. It offers a transparent framework for evaluating financial AI systems, benchmarking not just their performance but also their fairness, explainability, and compliance. Unlike static evaluations, FinBench focuses on dynamic, multi-turn scenarios, such as voice-to-voice financial consultations, where maintaining transparency and sound decision-making across extended dialogues is essential.

In a financial landscape increasingly dominated by digital agents, the stakes for ethical and accountable AI systems are immense. By proactively identifying weaknesses in transparency and fairness, FinBench equips financial institutions to deploy generative AI responsibly, ensuring these technologies meet the highest standards before they scale.

Appendices

A. Modality-Specific Metrics

Metric	Description	Relevant Client Types	Modality
Text Modality Evaluation	Tests how well the model processes pure text inputs	Verified Academics, Financial Research Analysts	Text
Voice-to-Text Evaluation	Evaluates models converting voice into text (e.g., financial advisory services)	Customer Service, Corporate M&A	Voice-to-Text
Graph Understanding	Assesses the model's ability to interpret and analyze graph-structured data (e.g., financial networks, social graphs)	Risk Analysts, Fraud Detection Teams	Graph
Time Series Data Evaluation	Measures model performance on time-dependent datasets (e.g., stock prices, economic indicators)	Investment Analysts, Quantitative Traders	Time Series
Image-to-Text Evaluation	Evaluates how the model translates images (e.g., documents, charts) into text summaries or insights	Insurance Underwriters, Legal Teams	Image-to-Text
Multimodal Synthesis	Assesses how the model integrates data across multiple formats (e.g., combining text with financial graphs or time series)	Portfolio Managers, Market Strategists	Multimodal
Selective Information Surfacing	Tests the model's ability to filter and highlight relevant data from large datasets (e.g., surfacing critical news events)	Hedge Fund Managers, Business Intelligence Teams	Text, Multimodal
Document Structure Recognition	Evaluates the model's capability to recognize and interpret structured documents (e.g., financial statements, contracts)	Legal Teams, Auditors	Text, Image-to-Text

B. Evaluation Metrics and Perturbation Robustness Metrics

Metric	Description	Relevant Client Types	Reasoning Type / Modality
Prompt Types	Zero-shot, few-shot, role-oriented, task-oriented prompts	Developers, LLM users, Financial Analysts	LLM-based natural language processing (NLP)
Attack Types	Character-level, word-level, sentence-level, semantic-level attacks	Security professionals, AI researchers, NLP engineers	Adversarial testing and robustness evaluation in LLMs
Tasks Evaluated	Sentiment analysis, grammar correctness, duplicate sentence detection, natural language inference, multi-task knowledge, reading comprehension, translation, math problem-solving	Verified academics, Financial research analysts, Developers	Task-based performance testing, LLM task adaptability
Adversarial Prompt Testing	Perturbations like typos, synonyms, and strategically crafted adversarial attacks	Security professionals, NLP engineers, AI developers	Adversarial prompt-based reasoning
LLMs Evaluated	Flan-T5-large, ChatGPT, GPT-4, Vicuna, Llama2, UL2	Developers, AI researchers, Financial institutions	Language models for robustness testing and adversarial attacks
Word-level Attack Impact	Highest performance drop (39%) observed in LLMs when faced with word-level attacks	AI security experts, NLP engineers, Financial institutions	Performance vulnerability due to word-level perturbations
APDR (Average Performance Drop Rate)	Quantitative measurement of the robustness across various models and datasets	AI researchers, NLP developers, Financial research analysts	Drop rate analysis across models, focusing on robustness