FinTech Frankenstein: Utilizing Multiple LLMs and Routing to Produce the Best Response

Akaash Dash
GA Institute of Technology
adash37@gatech.edu

Erik Larson
GA Institute of Technology
elarson43@gatech.edu

Aditya Natham
GA Institute of Technology
anatham3@gatech.edu

Sapan Patel
GA Institute of Technology
spatel794@gatech.edu

Max Zhao
GA Institute of Technology
qzhao305@gatech.edu

Abstract

This study introduces FinTech Frankenstein, an innovative approach designed to optimize the use of Large Language Models (LLMs) in the financial domain. With the proliferation of LLMs, selecting the most suitable model for specific tasks has become increasingly challenging, especially in finance where small variations in content can lead to significantly different responses. FinTech Frankenstein addresses this challenge by integrating a fine-tuned classification model, routing logic, and three leading financial LLMs. This setup categorizes user prompts and directs them to the most appropriate LLM, based on performance benchmarks across various financial tasks. The paper discusses the development and testing of this system, emphasizing its ability to outperform single-model approaches. Despite its success, the study acknowledges limitations such as token constraints in BERT and the need for further hyperparameter optimization. The paper concludes with future directions, highlighting the potential integration of additional LLMs and the continual adaptation to evolving benchmarks in the financial LLM landscape.

1 Introduction

As new Large Language Models (LLMs) are being rapidly published, it is becoming harder to decipher which model is best for specific prompting tasks and use cases. Small variations in content and phrasing can result in vastly different responses across models, creating strong discrepancies. This is especially true in the financial domain, where the use of LLMs is becoming widespread. All of the existing complexities, along with the ever-changing nature of the LLM space, make it difficult for end users in the financial space such as analysts, advisors, and educators to utilize the wide selection of LLMs to their best capabilities. As a result, the end user's current typical pattern of usage is to rely on one model, such as GPT-4 with ChatGPT, and use that model for every query, which isn't optimal since there are queries that other LLMs like FinMA produce better responses for.

To solve this non-optimal LLM usage issue, we propose FinTech Frankenstein, a new approach that combines a fine-tuned classification model, routing logic, and three leading LLMs in the financial space to categorize the user's prompt and send the query to the LLM that is best suited for the prompt. To determine which LLM is best for which prompt, we followed existing benchmarks on the performance of a wide variety of LLMs on certain categories of financial tasks.

2 Problem Statement

The wide range of LLMs in the field of finance has given rise to a pressing challenge for end users: figuring out the most appropriate model for any given use case. Of the extensive variety of available models, each is designed with specific strengths and limitations. This creates a broad assortment of potential discrepancies, resulting in users faced with the difficult task of determining which model is most suitable for their particular queries and tasks. In the financial domain, where the nuances of tasks can vary significantly, this is especially true since selecting the right LLM becomes crucial for accurate and reliable results. Compounding this issue is the fact that many individuals who utilize LLMs in the financial industry lack the required technical expertise to effectively navigate the intricacies of these models. The end users, which may include financial analysts, traders, or other professionals, often find themselves in a situation where they must choose between models without the requisite knowledge or understanding. Unable to make an informed decision, the inclination of users is to opt for a one-size-fits-all approach, as the complexities of different LLMs and their optimal applications remain unknown to the majority of non-technical users. Even if users were to pos-

sess a general understanding of which model suits specific prompts, it is inefficient to manually select the most suitable model for every query, especially if they are querying a model many times a day. Furthermore, it increases the potential for human error, emphasizing the need for a streamlined solution in model selection for the finance domain.

3 No One-Model-Fits-All

In the realm of Large Language Models (LLMs), the concept of a singular, superior model that excels in all aspects remains an elusive goal. The primary reason for this lies in the inherent "seesaw effect" encountered in model development. This effect illustrates the trade-offs that typically arise during the creation of LLMs. For instance, enhancing a model's performance in one specific area, such as financial analysis, might inadvertently reduce its effectiveness in another, like general conversational abilities. These compromises are a result of the intricate balance between specialized knowledge and broad adaptability, which are often at odds in AI model development. Furthermore, the complexity and resource intensity required to engineer an all-encompassing model make it a daunting, if not impossible, task. Current technological and computational limitations, alongside the vast and diverse nature of language and context-specific requirements, contribute to the unlikelihood of developing a single model that can master every domain effectively. Consequently, this leads to a preference for multiple specialized models over a hypothetical, universally proficient one.

4 Evolution of Financial LLMs

In the realm of financial technology, the evolution of Large Language Models (LLMs) has been marked by significant advancements, highlighted by models such as FinBERT, FLANG, BloombergGPT, and others. FinBERT emerged as one of the earliest financial-specific LLMs (Araci, 2019). It leveraged the BERT framework, pretraining it with a substantial corpus of financial texts, including open-source financial databases like TRC2-financial and Financial Phrase Bank. This model showed a notable improvement over traditional neural network methods, particularly in tasks related to financial sentiment classification. It underscored the potential of domain-specific pretraining in enhancing the LLM's performance for finance-related tasks.

Further evolution in this domain was marked by the development of FLANG (Shah et al., 2022). FLANG distinguished itself by integrating both BERT and ELECTRA architectures, focusing on financial texts primarily in English. This integration showcased the potential of combining different model architectures to enhance understanding and processing of financial language, aiming to provide more accurate and contextually relevant insights.

The introduction of BloombergGPT represented a significant leap in the scale and capability of financial LLMs (Wu et al., 2023). With a colossal 50 billion parameters, BloombergGPT was pre-trained on a mix of datasets from both general and financial domains, aiming to bridge the gap between broad contextual understanding and domain-specific expertise. However, unlike other contemporary LLMs like ChatGPT and GPT-4, BloombergGPT does not conform to the instruction-following paradigm and is not open-sourced, limiting its accessibility and application in broader research and development contexts.

Each of these models has contributed to the evolving landscape of LLMs in the financial sector, highlighting the trend towards more specialized, domain-focused models. They collectively demonstrate the progression towards models that not only understand the complex language of finance but also provide tailored, context-aware insights specific to the financial industry's unique needs.

5 Evolution of Financial LLM Benchmarks

The progression of benchmarks in the financial sector for Large Language Models (LLMs) reflects a dynamic interplay between the advancement of these models and the evolving criteria used to evaluate them. Notable among these benchmarks is the Financial Phrase Bank (FPB), which serves as a foundational tool for assessing sentiment analysis in the financial domain (Malo et al., 2013). It provides a collection of English sentences from financial news, each meticulously labeled with sentiment annotations by domain experts. This benchmark plays a crucial role in evaluating the capability of LLMs to accurately interpret and classify financial sentiments.

Another significant benchmark is FiQA, which extends the scope of sentiment analysis by including not only financial news but also microblog posts (Maia et al., 2018). The unique aspect of FiQA is

its quantification of sentiment on a numerical scale, offering a more nuanced understanding of sentiment intensity. This benchmark challenges LLMs to discern and quantify sentiment, a skill essential in the volatile and sentiment-driven world of finance.

Further enriching the landscape of financial benchmarks, FLUE, offers a heterogeneous suite of tasks including financial sentiment analysis, news headline classification, and named entity recognition (Shah et al., 2022). This comprehensive benchmark not only evaluates the LLMs' understanding of financial language but also their ability to perform more complex tasks like identifying key financial entities and classifying news headlines, thereby testing the models' depth and breadth of financial comprehension.

The introduction of FLARE represents a significant leap in benchmarking, as it not only covers a range of financial NLP tasks but also ventures into the realm of financial prediction (Xie et al., 2023). This addition marks an important evolution, acknowledging the need for LLMs to be adept not just in language processing but also in making predictions based on financial data. By incorporating tasks like stock movement prediction, FLARE provides a more robust and practical framework for assessing LLMs, aligning more closely with real-world financial applications.

6 Evaluating Financial LLMs

The basis of this problem is determining what makes a model 'better' than another model, and determining which model is more suitable for different use cases. This question is at the crux of our project since understanding the trade-offs between different models and how they relate to the category of questions allows us to better determine what categories prompts belong to, and which LLMs are best suited to respond to them. Luckily, FLARE is a set of benchmarks that range across several tasks and inputs that we can use to answer our questions. Flare covers 4 financial NLP tasks with 6 datasets, and 1 financial prediction task with 3 dataset and is the latest and most comprehensive LLM benchmark at the time of writing (Xie et al., 2023).

Advanced LLMs, financial specific or general, such as GPT-4, FinMA, BloombergGPT were evaluated with most tasks in FLARE, and was found that FinMA significantly outperforms many popular LLMs including BloombergGPT, ChatGPT and GPT-4 on most FLARE tasks, such as financial sentiment analysis, news headline classification, NER, and stock movement prediction (Xie et al., 2023). This could be attributed to the fine-tuning of FinMA specifically for the financial domain. However, FinMA falls to BloombergGPT, ChatGPT and GPT-4 on the question-answering category of FLARE, which is related to the quantitative reasoning ability of the model. This is attributed to the limitations of LLaMA concerning quantitative reasoning since LLaMA is the base model that FinMA is fine-tuned from (Touvron et al., 2023). Lastly, all of the models showed limited performance on tasks related to predicting stock movements, which shows there's room for improvement. Overall, these results show the limitations of using a single model for all of a user's queries, as there isn't a model that performs the best for all categories of prompts.

The top performing models of each category in FLARE alternate between GPT-4, FinMA 7B, FinMA 7B-full, and FinMA 30B, which are our top candidates for the end models our router routes to (https://huggingface.co/spaces/ChanceFocus/FLARE). We ultimately decided to remove FinMA 30B due to the size of the model and compute constraints, leaving GPT-4 FinMA 7B and FinMA7B-full as our end models.

7 Data Collection

In our research, a critical phase was the data collection and model selection process, focused on identifying the most efficient Large Language Models (LLMs) for financial tasks. We utilized the FLARE leaderboard hosted on Hugging Face as our primary data source. This leaderboard provides comprehensive insights into various benchmarks included in the FLARE suite and the performance of different models on these benchmarks.

From this rich dataset, we meticulously selected models that demonstrated superior performance across a range of tasks. Our choices included FinMA 7B, FinMA 7B-Full, FinMA 30B, and GPT-4, all of which stood out as top performers in the FLARE evaluations. This selection process was guided by the objective to integrate models that not only showed excellence in specific tasks but also offered a broad range of capabilities, essential for addressing the diverse needs in financial applications.

Furthermore, we collected detailed benchmark prompts from Hugging Face for all benchmarks

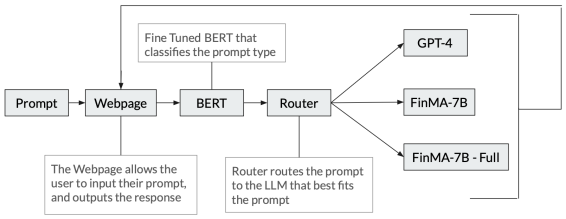

Figure 1: Project architecture

included in FLARE. This data was essential in constructing a comprehensive router, designed to navigate the intricacies of financial queries efficiently and direct them to the most suitable model. The collected prompts served as a foundational dataset, enabling us to fine-tune the router for optimal performance in real-world scenarios.

8 Methodology

Our desired approach can be seen in Figure 1. Prompts will be input by users through a webpage, which will be forwarded to a fine-tuned BERT classifier to classify the task in the prompt. The router will then select the strongest model on the task and query it with the prompt. The output will finally be returned back to the user.

8.1 Text Classification with BERT

BERT (Devlin et al., 2019), is a transformer-based model designed for bidirectional pre-training of language representations. Unlike traditional models that process text in a left-to-right or right-to-left manner, BERT captures contextual information from both directions, allowing it to understand the full context of a word within a sentence. This bidirectional context encoding results in highly contextualized embeddings, making BERT particularly effective for understanding the nuances of natural language.

To adapt BERT for our specific task of prompt categorization, we employed a fine-tuning process, which involves training the pre-trained BERT model on a labeled dataset that corresponds to our target prompt categories. The fine-tuning process allows the model to learn task-specific patterns and nuances, improving the baseline model. During fine-tuning, we feed the labeled FLARE benchmark suite into the pre-trained BERT model and adjust the model's parameters to minimize the classification error, using an 80/20 train/test split. The process tunes the baseline model to recognize the unique characteristics of our prompt categories,

Hyperparameter	Value
Maximum Tokens Length	256
Batch Size	32
Epochs	1
Learning Rate	1e-7

Table 1: Hyperparameters used when fine-tuning BERT

Metric	Value
Accuracy	0.8598
Precision	0.8598
Recall	0.8598

Table 2: BERT fine tuning results

capturing the more subtle semantic nuances that distinguish one category from another. The hyperparameters were chosen as seen in Table 1. The maximum tokens length and batch size were set to the highest value we could choose without running out of memory. We trained using a high-ram V100 GPU supported environment in google colab pro, but were still restricted by memory and compute constraints. Due to the large amount of training data, with 3907 batches, more than one epoch seemed unnecessary and was infeasible with computing constraints. Additionally, the model converged quickly, so a low learning rate was chosen to combat this

8.1.1 Fine Tuning Results

Fine tuning results can be seen in Table 2. Due to the highly structured nature of the data, the model converged very quickly and tended to overfit. Hyperparameters had to be constantly adjusted to avoid getting a 'perfect' result, which would be indicative of overfitting.

8.2 Routing

Having fine-tuned BERT for prompt categorization based on the FLARE benchmark suite and knowledge of which LLMs perform the best for each category, we can integrate this text classification model into the FinTech Frankenstein architecture to dynamically route user prompts to the most suitable LLM. We did this by using the categorization output from the BERT-based text classification model as a key input for the routing logic within FinTech Frankenstein, routing the prompt to the LLM that handles the prompt's category the best. The selection of the model was done by choosing the model with the highest

performance on the class that pertained to the classified user query. This was obtained from the FLARE leaderboard dataset.

GPT-4: Prompts that involve more general inquiries and tasks related to quantitative reasoning are directed to GPT-4 since the model performs better than the FinMA models in those aspects. Overall, it excels in providing diverse and contextually rich responses (OpenAI, 2023).

FinMA 7B: FinMA 7B is directed prompts that require a deeper understanding of financial context including analyzing financial sentiment, news analysis, and finance-specific named entity recognition. Its fine-tuning for the financial domain makes it well-suited for these specific tasks.

FinMA 7B-full: Like FinMA 7B, FinMA 7B-full is reserved for financial queries that demand more financial context. It outperforms FinMA 7B slightly in financial prompts related to number understanding and named entity recognition, so finance-specific prompts in those categories will be mostly directed to it.

9 Results and Testing

Our research culminated in the development and testing of FinTech Frankenstein — a solution designed to address the challenges associated with selecting the most model for various financial tasks. The system combines a fine-tuned classification model, routing logic, and three prominent LLMs, aiming to optimize user prompt categorization and LLM selection.

When separating the data, 10 prompts from each benchmark were set aside. We then queried FinTech Frankenstein with the prompts to evaluate performance. These can be seen in Table 3.

Overall, we found systematic testing on a large scale difficult to implement due to limited compute resources, and mainly relied on manually categorizing prompts and querying our system. Despite this, we found that prompting FinTech Frankenstein with queries ranging over multiple categories edges using any one model, which can be mainly attributed to the success of the routing logic combined with the fine-tuned classification BERT model in routing prompts to the most suited model.

Benchmark	Accuracy
FPB	0.7
FiQA-SA	0.8
Headlines	1.0
FinQA	0.4
ConvFinQA	0.7
Finer-ORD	0.7
SM-BigData	0.2
SM-ACL	0.4
SM-CIKM	0.1
German	0.3
Australian	0.3
EDTSum	0.4
FOMC	0.1
NER	0.8

Table 3: FinTech Frankenstein results against FLARE benchmarks

10 Limitations

Despite the positive outcomes from this project, there are still several limitations:

10.1 Token Limitation in BERT

The use of BERT for fine-tuning in our classification model introduces a constraint on the maximum number of tokens it can handle, capped at 256 tokens due to Colab Pro limitations. This restriction may impact the model's ability to capture extensive contextual information in longer prompts, potentially leading to suboptimal performance for lengthy queries.

10.2 Hyperparameter Exploration

Although the project extensively explored various models and combinations during benchmarking, the hyperparameter space for fine-tuning BERT and other components could have been further explored to further improve model performance.

10.3 Limited Control in Testing

The testing phase faced constraints imposed by computing limitations, which makes it difficult to verify the consistency of the results.

10.4 English-Only Support

FinTech Frankenstein is currently designed to support prompts in English only. While the underlying models, such as GPT-4 and FinMA, have the capability to handle multiple languages.

11 Conclusion and Looking Forward

We addressed the challenges posed by the continuous evolution of the LLM space within the financial domain with our FinTech Frankenstein solution, setting the basis to continue improving and refining our framework. As we reflect on our findings from this semester, we outline potential key directions for further improvements to our system.

The first idea we're the most excited about is to integrate additional LLMs into the FinTech Frankenstein framework. This expansion should focus on finding models that have a more 'spiked' performance chart, meaning models that excel in executing a couple of tasks exceptionally over models that perform every task pretty well. This would allow us to harness the routing idea to the fullest as we seek to receive the best output for each task. Each additional model would directly improve the response quality of the system, dramatically improving the response quality over time.

Another idea that builds upon the success of wielding three LLMs separately is to explore the potential benefits of integrating responses from multiple LLMs for a single user query. This approach could leverage the strengths of each model, providing users with responses that would be of higher quality than any one model could achieve.

Lastly, as the landscape of LLM benchmarks continues to evolve, we plan to continue running our models on newly developed benchmarks to ensure the FinTech Frankenstein system remains up-to-date and aligned with the latest standards. Resources like FinanceBench are good places to start here (Islam et al., 2023).

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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

Claudia Maia, Guilherme Lunardi, Andre Longaray, and Paulo Munhoz. 2018. Factors and characteristics that influence consumers' participation in social commerce. *Revista de Gestão*, 25(2):194–211.

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

OpenAI. 2023. Gpt-4 technical report.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance.