Fine-Tuning LLMs on a Financial Instruction Dataset: A Comparative Study of LLMs for Financial Text Generation Tasks

By Jasmine Coloma
Professor Ying Mao, Fordham University
Fall 2023

ABSTRACT

This research explores the comparative analysis of large language models after fine-tuning on a finance instruction dataset. The study aims to assess their effectiveness in understanding natural language within financial contexts. Through a systematic comparison, the research provides detailed insights into the strengths and limitations of each language model, highlighting their adaptability and performance in processing finance-specific language patterns. These findings enhance our understanding of optimizing language models for specialized applications in the financial sector, aiding informed decisions for practitioners and developers. The central question guiding this research is: What is the comparative impact of fine-tuning Llama 2, GPT 2, Falcon, and Bloom language models on the performance of natural language understanding in the context of a finance instruction dataset? Specifically, how do these models differ in their ability to comprehend and generate accurate financial instructions?

1 INTRODUCTION

In the domain of natural language processing, the fine-tuning of language models for specific tasks has become crucial. This research focuses on financial text generation tasks, conducting a comparative assessment of the performance of four language models: Llama 2, GPT 2, Falcon, and Bloom. These models encounter a unique challenge due to the intricate nature of financial language, incorporating technical terms and subtle contextual details.

In the finance field, two widely recognized Large Language Models (LLMs), BloombergGPT [1] and FinGPT [2], have gained attention for their impact on financial analysis tasks. However, this research distinguishes itself by comparing more general language models—Llama 2, GPT 2, Falcon, and Bloom—instead of these specialized financial LLMs.

While pre-trained language models form a foundation for general language understanding, their adaptability to specialized domains like finance requires further investigation. This study critically assesses and compares the proficiency of Llama 2, GPT 2, Falcon, and Bloom in handling the linguistic challenges posed by finance instructions. The objective is to offer insights into the differences among these models.

This study seeks to enhance the overall comprehension of language models' adaptability and optimization capabilities, particularly when confronted with the complexities of financial language. By concentrating on general LLMs, the research establishes a fundamental benchmark for comparison, aiding a diverse audience, including those contemplating the use of non-specialized models for financial language tasks.

This comparative analysis aspires to contribute valuable knowledge to both academic and practical domains, offering a comprehensive understanding of the optimization potential of language models, particularly in the context of finance. The outcomes of this research are expected to guide decision-making processes for practitioners and developers engaged in deploying language models for finance-specific tasks. It is crucial to acknowledge certain constraints, and extending these findings to different language models or financial contexts might require more in-depth investigation.

2 LITERATURE REVIEW

In recent years, there has been a notable increase in the fine-tuning of language models for specific tasks in the field of natural language processing (NLP), with a pronounced emphasis on applications within the financial sector. Notably, two LLMs, BloombergGPT and

FinGPT, have garnered significant attention for their specialized impact in finance. These models, designed specifically for finance-related tasks, have played a crucial role in automating functions like financial analysis and sentiment analysis, addressing essential requirements within the financial industry.

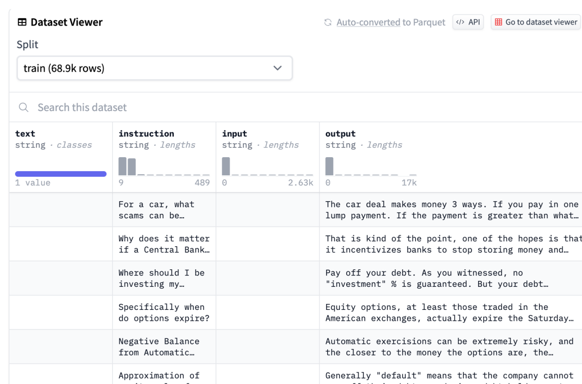
This literature review delves into pivotal studies that contribute to our understanding of fine-tuning large language models, and their application in the financial context. Yi Yang et al.'s work [3], conducted at the School of Business and Management at Hong Kong University of Science and Technology, introduces FinBERT—a pre-trained language model explicitly tailored for financial communications. FinBERT aims to enhance the comprehension and analysis of language used in financial contexts, thereby improving sentiment analysis and other tasks within the financial domain.

Similarly, Neng Wang et al. [4] present FinGPT in their paper titled "FinGPT: Instruction Tuning Benchmark for Open-Source Large Language Models in Financial Datasets." This benchmark is specifically crafted for fine-tuning LLMs on financial datasets. The authors underscore the importance of adapting language models for optimal performance in financial scenarios. FinGPT serves as a benchmark to assess the adaptability and efficiency of LLMs when applied to financial datasets.

These studies highlight the increasing trend of tailoring language models for finance-related purposes. FinBERT and FinGPT, designed for specific tasks in financial communication and benchmarking, contribute to a broader understanding of how language models can be customized for various applications within the financial domain. The comparative analysis of LLMs, including Llama 2, GPT 2, Falcon, and Bloom, further extends this exploration, providing insights into their strengths and limitations on causal language modeling and financial text generation when fine-tuned on finance instruction datasets. Together, these studies contribute valuable knowledge to both academic and practical domains, offering comprehensive insights into the optimization potential of language models in the context of finance.

3 DATASET

The Finance-Alpaca Dataset is an instruction dataset consisting of 4 total features: “text”, “instruction”, “input”, and “output”. The dataset comprises approximately 70,000 data points related to finance. The data primarily consists of finance-related questions accompanied by their respective answers. This dataset is created by merging Stanford’s Alpaca and FiQA. Additionally, about 1,300 additional pairs were generated specifically for this dataset using GPT3.5.



Dataset Viewer

Auto-converted to Parquet API Go to dataset viewer

Split

train (68.9k rows)

Search this dataset

text string · classes 1 value	instruction string · lengths 9 489	input string · lengths 0 2.63k	output string · lengths 0 17k
	For a car, what scams can be…		The car deal makes money 3 ways. If you pay in one lump payment. If the payment is greater than what…
	Why does it matter if a Central Bank…		That is kind of the point, one of the hopes is that it incentivizes banks to stop storing money and…
	Where should I be investing my…		Pay off your debt. As you witnessed, no "investment" % is guaranteed. But your debt…
	Specifically when do options expire?		Equity options, at least those traded in the American exchanges, actually expire the Saturday…
	Negative Balance from Automatic…		Automatic exercisions can be extremely risky, and the closer to the money the options are, the…
	Approximation of equity value for…		Generally "default" means that the company cannot pay off their debts, and since debt holders get…

Table 1: Dataset [5]

4 METHODOLOGY

The methodology employed for this research involved a systematic evaluation of various language models (LLMs) in the domain of financial instructions. The selected models, including Llama 2, GPT 2, Falcon, and Bloom, were chosen based on their relevance and established performance in natural language processing tasks. The initial phase encompassed loading each pre-trained LLM to 4-bit precision in order to facilitate the subsequent model evaluations and fine-tuning.

4.1 Experiment Setup Each model was executed on Google Colab Pro using the V100 GPU as the runtime type. Consistently employing this runtime type across all models, despite variations in size, facilitated efficient memory availability.

4.2 Model Selection The choice of LLMs for this research focused on choosing from a class of autoregressive / decoder-only models, each varying in different parameter size. The models were run on a single GPU on GoogleColab. Due to limited resources, opting for models with the lower parameters was deemed optimal. Therefore, the following versions of Llama 2, GPT 2, Falcon, and Bloom were selected to align with the specific project requirements and constraints. All models were directly downloaded from Hugging Face.

Llama 2 Developed by Meta in July 2023, the Llama 2 family comprises pre-trained generative text models ranging in size from 7 billion to 70 billion parameters [6]. This research utilized the 7 billion parameter chat version, directly downloaded from Hugging Face.

Falcon Developed by the Technology Innovation Institute in Abu Dhabi [8], the Falcon LLM, specifically the 7B version, was chosen for this research. It was trained on 1,500B tokens of RefinedWeb, a high-quality filtered and deduplicated web dataset enhanced with curated corpora, inspired by The Pile[9].

Bloom A sophisticated autoregressive LLM, Bloom operates by predicting and generating text based on extensive training data. It can produce coherent text in 46 languages and 13 programming languages, handling text tasks not explicitly part of its training [10].

GPT 2 Developed by OpenAI, GPT 2 is a large language model pre-trained on a vast corpus of raw English data in a self-supervised fashion, trained to predict the next word in sentences [11].

Model	Llama 2	Falcon	Bloom	GPT 2
Model size	7B	7B	3B	124M

Table 2: LLM Comparisons

4.3 Model Quantization Fine-tuning a LLM poses a challenge due to the unclear nature of the process, and the significant computational resources needed to train a billion parameter model without optimizations can be quite restrictive. To drastically reduce storage and memory usage, the models were configured to 4-bit precision using QLoRa, or “Quantized Low-Rank Adaptation” [12]. Using this approach combines quantization and low-rank adaptation to achieve efficient fine-tuning of LLMs. It helps reduce the memory required for fine-tuning LLM without sacrificing the performance.

4.4 Base Model Evaluation The initial phase involved evaluating the base LLMs, using perplexity as the chosen metric. A subset of 100 instances from the training dataset was employed to assess the perplexity of the models in understanding financial language patterns. Furthermore, the generation capabilities of each LLM was evaluated by asking the model a specific prompt: “Why should portfolios be diversified?”. The responses were analyzed to assess how proficient the models were in generating contextually relevant information.

4.5 Fine Tuned Strategy The fine-tuning task for the selected models was focused on text generation. Using the Supervised Fine-tuning Trainer (`SFTtrainer`) from the Transformer Reinforcement Learning (`trl`) library, the models underwent training for 100 steps and 1 epoch. The learning rate was set at 2e-4, and a batch size of 4 was employed (indicating the number of examples processed per optimization step). These parameters were carefully chosen to optimize the fine-tuning process for text generation, enhancing the models' domain-specific performance on financial instruction datasets. The fine-tuned models were subsequently evaluated on the same test dataset used for base model evaluations, with perplexity serving as the primary metric to quantify improvements in the comprehension of financial language.

4.6 Evaluation Metrics Incorporating the perplexity metric into each model evaluation involved a comprehensive approach to assessing the model's predictive performance on a given sequence of words. Perplexity is a measure of how well the language model can predict the next word in a sequence. It is calculated as the exponentiation of the average negative log-likelihood per word.

To implement this metric, a function was created to develop a sliding-window strategy that was utilized during model evaluation. The input sequence was divided into subsequences with a specified stride equal to the model's maximum input size. For each subsequence, the model predicted the next word based on the preceding context. The negative log-likelihood of each prediction was then computed. By accumulating these values and taking their average, the perplexity of the model on the entire sequence was obtained. Through the utilization of this approach, the performance of each model was assessed by evaluating its ability to predict each word within the context of preceding words. This method provided insights into the overall language understanding and generation capabilities of the models.

4.7 Model Evaluation The training loss of the fine-tuned model was evaluated and perplexities were compared between the base model and the fine-tuned model. Furthermore, the generation capabilities of the models were reassessed post fine-tuning by asking the same question, "Why should portfolios be diversified?"

5 Experiment Results

In this section, the results of the experiments are presented, comparing the training performance of all four language models. Additionally, the evaluation criteria, which include training loss and perplexity, provide insights into how well each model adapted to the finance instruction dataset. Finally, a comparison is made between the inferences of the base model and the fine-tuned model.

5.1 Training Loss Comparison Each model was trained on the dataset for 1 epoch and a total of 100 steps. The following diagram (figure 3) shows the net training loss from step 10 and step 100.

The results indicate that Llama 2 achieved the highest total training loss (0.6725), followed by Bloom (0.5951), Falcon (0.5334), and GPT 2 (0.2). These findings suggest that Llama 2 and Bloom performed relatively well in adapting to the finance instruction dataset, demonstrating lower discrepancies between predicted and actual values.

Steps	Llama 2	Falcon	Bloom	GPT 2
10	2.14	2.0287	2.5505	3.6686
20	2.326	1.9541	2.5177	4.2383
30	1.7484	1.7288	2.1463	3.561
40	1.9608	1.8625	2.2457	3.9185
50	1.6703	1.7278	2.0993	3.8756
60	1.8701	1.9993	2.4935	3.6021
70	1.5771	1.7149	2.1752	3.6971
80	1.64	1.7292	2.2324	3.3206
90	1.6459	1.6588	2.2505	3.7418
100	1.4675	1.4953	1.9554	3.4193
Total loss	**0.6725**	**0.5334**	**0.5951**	**0.2**

Table 3: Training Loss Comparisons

5.2 Perplexity Comparison

Due to simplicity and storage limitations, the perplexity was evaluated over a 100 samples of the dataset. The table below presents the perplexity comparisons between the original models and their fine-tuned counterparts based on the provided results.

The perplexity values reflect the model's predictive accuracy, with lower values indicating improved performance. Notably, Llama 2 experienced an increase in perplexity after fine-tuning (+1.04), suggesting that the model faced challenges in adapting to the specific nuances of the finance instruction dataset. Falcon demonstrated a slight decrease in perplexity (-0.04), showcasing a slight improvement. Both Bloom (-0.18) and GPT 2(-0.13) exhibited a decrease in perplexity, indicating enhanced performance after fine-tuning.

Model Name	Base Model	Fine Tuned Model	Performance Gain
Llama 2	2.58	3.62	+1.04
Falcon	1.73	1.69	-0.04
Bloom	3.14	2.96	-0.18
GPT 2	1.3125	1.016	-0.13

Table 4: Perplexity Comparisons

5.3 Model Inference

In assessing the model inferences of all four fine-tuned LLMs, the responses to the query, "Why should portfolios be diversified?" are examined. This analysis aims to identify unique communication styles and the depth of explanation exhibited by each model.

5.3.1 Llama 2 *(shown in Table 5)*

Base Model Response: Llama 2's base model showcases a meticulous and detailed response, characterizing the importance of portfolio diversification. The model articulates key principles with clarity, emphasizing risk minimization, consistent returns, and alignment with personal financial goals. The response reflects a balanced and refined understanding of the intricacies of diversification.

Fine-Tuned Model Response: Post fine-tuning, Llama 2 maintains its detailed approach while experiencing a slight increase in response length. The fine-tuned model successfully preserves clarity and coherence, delivering a concise yet thorough explanation of the significance of diversification.

5.3.2 Falcon *(shown in table 6)*

Base Model Response: Falcon's base model provides a succinct explanation, emphasizing risk reduction through diversification by spreading investments across different assets and industries. However, the brevity of the response may result in a less comprehensive understanding for users seeking a detailed explanation.

.

Fine-Tuned Model Response: The fine-tuned model enriches its response by offering a more extensive and detailed explanation. It emphasizes diversification's role in risk reduction, volatility management, and resilience to market downturns. The fine-tuned model's responses are characterized by a balance between depth and accessibility.

5.3.3 Bloom *(shown in table 7)*

Base Model Response: Bloom's base model focuses on diversification as a means to reduce the risk of losing money, with an emphasis on managing risks and hedging. While the response is brief, it falls short of providing a comprehensive exploration of the diverse advantages associated with diversification.

Fine-Tuned Model Response: The fine-tuned model refines its explanation, presenting diversification as a risk reduction strategy and highlighting its role in managing portfolios by spreading risks across different assets. The response maintains a concise yet focused style, catering to users seeking clarity in risk management strategies.

5.3.4 GPT 2 *(shown in table 8)*

Base Model Response: GPT 2 's base model delivers an abstract discussion, framing portfolios as collections of assets and touching on investment strategies without a distinct emphasis on diversification. The response lacks specificity and may leave users wanting a more targeted explanation.

Fine-Tuned Model Response: The fine-tuned model rectifies the abstract nature of the base response, providing a more coherent explanation. It describes portfolios as collections of assets used for buying and selling securities and underscores the importance of diversification in managing risks during market volatility. The fine-tuned model's responses exhibit a more concrete and focused manner.

6 Discussion

In this section, we break down and interpret the outcomes of the experiments, shedding light on how well Llama 2, Falcon, Bloom, and GPT 2 performed in processing finance instructions.

The examination of training loss reveals notable distinctions among the models. Llama 2 emerges as a frontrunner, achieving the lowest total training loss. Falcon and Bloom closely follow, demonstrating competitive performance in minimizing errors during training. GPT 2 exhibits commendable outcomes, further contributing to the diversity of model performances.

An improved understanding of model behavior emerges when assessing perplexity dynamics. While Llama 2 excels in training loss, it experiences an unexpected increase in perplexity after fine-tuning. This signifies potential challenges in the model's confidence when predicting finance sequences. In contrast, Falcon, Bloom, and GPT 2 showcase improvements in predictive accuracy, evident in reduced perplexity values post fine-tuning. Each model exhibits unique responses, highlighting the complexity of their interactions with the finance instruction dataset.

In examining the responses generated by language models—Llama 2, Falcon, Bloom, and GPT 2 —to the query "Why should portfolios be diversified?" several noteworthy patterns emerge. Firstly, the process of fine-tuning consistently yields responses that strike a balance between providing more in-depth insights and maintaining clarity across all models. This signifies an enhancement in the models' capability to offer detailed explanations while ensuring that responses remain accessible and clearly expressed. The adaptability of these language models stands out, demonstrating their capacity to adjust to response styles through fine-tuning, catering to specific user expectations and the demands of different applications. Additionally, a clear trend is observed in the specificity of communication: fine-tuned models excel in delivering precise information related to the query, whereas base models often produce more generalized responses. In conclusion, these overall observations underscore the dynamic nature of language models, showcasing their ability to refine and tailor response styles for effective communication.

7 Limitations and Future Work

This research reveals several limitations to improve in future research. Firstly, this evaluation is limited to a select set of notable models. There may be other LLMs that prove superior or varied performances compared to the models reported here. Furthermore, each model underwent training for only one epoch, potentially limiting their performance. Improvements might be observed by adjusting training parameters or increasing the number of dataset data points.

Lastly, it is important to note that this research exclusively concentrates on general financial text generation and the assessment of each LLM's adaptability to financial text data. Further research can extend to developing benchmark comparisons for other various tasks within the financial domain.

6 Conclusion

In conclusion, this research delves into a comprehensive comparative analysis of Llama 2, GPT 2 , Falcon, and Bloom language models when fine-tuned on a finance instruction dataset. The study is driven by the overarching goal of evaluating the effectiveness of these models in natural language understanding within the financial context. Through systematic comparisons, the research provides more insights into the adaptability and performance variations of each language model in processing finance-specific language patterns. Overall, this research significantly contributes to the broader understanding of language models' adaptability and optimization potential, particularly in dealing with the challenges of financial language. The outcomes aim to guide decision-making processes for practitioners and developers engaged in deploying language models for finance-specific tasks. While acknowledging certain constraints, the research underscores the importance of ongoing investigation in different language models and financial contexts to further enhance our understanding in this evolving field.

Works Cited

1. Wu, Shijie, et al. "Bloomberggpt: A large language model for finance." *arXiv preprint arXiv:2303.17564* (2023).
2. Yang, Hongyang, Xiao-Yang Liu, and Christina Dan Wang. "FinGPT: Open-Source Financial Large Language Models." *arXiv preprint arXiv:2306.06031* (2023)
3. Yang, Yi, Mark Christopher Siy Uy, and Allen Huang. "Finbert: A pretrained language model for financial communications." *arXiv preprint arXiv:2006.08097* (2020).
4. Wang, Neng, Hongyang Yang, and Christina Dan Wang. "FinGPT: Instruction Tuning Benchmark for Open-Source Large Language Models in Financial Datasets." arXiv preprint arXiv:2310.04793 (2023).
5. Bharti, Gaurang. "GBHARTI/Finance-Alpaca · Datasets at Hugging Face." Hugging Face , huggingface.co/datasets/gbharti/finance-alpaca. Accessed 11 Dec. 2023.
6. Touvron, Hugo, et al. "Llama 2: Open foundation and fine-tuned chat models." arXiv preprint arXiv:2307.09288 (2023).
7. "NousResearch/LLAMA-2-7B-CHAT-HF ·
8. Hugging Face." Hugging Face, huggingface.co/NousResearch/Llama-2-7b-chat-hf. Accessed 11 Dec. 2023.
9. "Falcon LLM." LLM, falconllm.tii.ae/falcon.html. Accessed 11 Dec. 2023.
10. Gao, Leo, et al. "The pile: An 800gb dataset of diverse text for language modeling." arXiv preprint arXiv:2101.00027 (2020).
11. "BLOOM." Bigscience.huggingface.co, bigscience.huggingface.co/blog/bloom.
12. Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.
13. Dettmers, Tim, et al. "Qlora: Efficient finetuning of quantized llms." arXiv preprint arXiv:2305.14314 (2023).

Table 5: Llama 2 Model Inference

Question: Why should portfolios be diversified?	
Base Model Response:	Fine Tuned Model Response:
why should portfolios be diversified? Diversification is a key principle of investment management that involves spreading investments across different asset classes, industries, and geographic regions to reduce risk. nobody likes to lose money, but diversification can help reduce the impact of market volatility on a portfolio. by investing in different types of assets, investors can reduce their exposure to any one particular asset class, industry, or geography. this can provide a more stable return over time, as gains in one area can offset losses in another. for example, if an investor has all of their money invested in a single stock, they are at risk of losing everything if that stock experiences a significant decline in value. but if they have a well-diversifiedportfolio that includes stocks, bonds, real estate, commodities, etc., they may only lose a small amount if the stock market declines. divers	Portfolio diversification is a crucial aspect of investing, as it helps to minimize risk and maximize returns. By spreading investments across different asset classes, sectors, and geographic regions, a diversified portfolio can provide a more consistent return over time, while also reducing the impact of any individual investment's performance on the overall portfolio. This is because different assets tend to perform differently in response to market conditions, so a diversified portfolio can help to smooth out volatility and reduce the likelihood of large losses. Additionally, diversification can help to ensure that an investor's investments align with their personal financial goals and risk tolerance, by providing a range of options for growth, income, and capital preservation. Overall, diversification is an essential component of any successful investment strategy, as it can help to protect against unexpected market events and ensure that an investor's money works hardest for them over the long term.</s>

Table 6: Falcon Model Inference

Question: Why should portfolios be diversified?	
Base Model Response:	Fine Tuned Model Response:
Portfolios should be diversified because it reduces the risk of the portfolio. Diversification reduces the risk of the portfolio by spreading the risk across different assets. Diversification also reduces the risk of the portfolio by spreading the risk across different industries.	Diversification is a key principle of investing. It helps to reduce the risk of a portfolio by spreading investments across different asset classes and sectors. This can help to reduce volatility and increase returns over time. Additionally, diversification can help to smooth out fluctuations in the market, making it easier to manage risk. Diversification can also help to reduce the impact of individual investment losses. By spreading investments across different asset classes and sectors, a portfolio can be more resilient to downturns in the market. Diversification can also help to reduce the risk of a portfolio by spreading investments across different asset classes and sectors. This can help to reduce volatility and increase returns over time. Additionally, diversification can help to smooth out fluctuations in the market, making it easier to manage risk. Diversification can also help to reduce the impact of individual investment losses. By spreading investments across different asset classes and sectors, a portfolio can be more resilient to downturns in the market. Diversification can also help to reduce the risk of a portfolio by spreading investments across different asset classes and sectors. This can help to

Table 7: Bloom Model Inference

Question: Why should portfolios be diversified?	
Base Model Response:	Fine Tuned Model Response:
`diversification is a way to reduce the risk of losing money in case you are wrong about your predictions.` `- portfolio diversification can also help with hedging, which will make it easier for you to manage risks.`	`diversification is a way to reduce risk. It helps you manage your portfolio by spreading out the risks across different assets, which can help protect against market downturns and volatility. Diversification also allows for better returns in case of good times or bad.</s>`

Table 8: GPT 2 Model Inference

Question: Why should portfolios be diversified?	
Base Model Response:	Fine Tuned Model Response:
`Diversification is about building a portfolio that's flexible enough to move into the future. It's a great way to start making money, or at least making money in the way we want to.`	`The portfolio is a collection of assets that are used to buy and sell securities. It can also include stocks, bonds or other investments in which the issuer has an interest (such as mutual funds). A fund's investment strategy may involve buying shares from investors who have invested their money into it for short-term gains; selling them at market prices when they become available later on during periods of low demand such like those seen with oil price appreciation but not so high volatility markets where there was no long term supply`