**P2 Proposal**

### **DS5500: Data Science Capstone**

Prof. Kylie A. Bemis

### **1. Title:** FinAdvisor - Financial Advisory Tool for contextually-aware personal finance assistance.

### **2. Authors:**

* Laasya Anantha Prasad
* Krishna Venkatesh
* Rakshak Kunchum

### **3. Github:** https://github.com/sriksven/FinAdvisor\_LLM

### **4. Summary of Previous Work:**

1. The FinAdvisor project is an AI-driven tool designed to provide personalized financial advice in real-time. Traditional financial advisory services are often hindered by limitations such as restricted accessibility, reliance on static knowledge, and dependence on human expertise, which can lead to delayed and generalized recommendations. FinAdvisor seeks to address these constraints by leveraging Large Language Models (LLMs) to provide tailored, up-to-date financial insights on topics like budgeting, investment, and risk management.
2. To train the models, we have utilized the **Financial Alpaca** and **FinTalk-19k** datasets, which contain a diverse range of real-world financial scenarios and user-generated content. This data equips the model to offer relevant and contextually appropriate advice. Through the integration of advanced LLMs and real-time data processing, the project aims to democratize access to sophisticated financial guidance, empowering users to make informed financial decisions.

### **5. Proposed Plan:**

Phase 2 will focus on enhancing the functionality and scalability of FinAdvisor. This includes further fine-tuning of the models, implementing a real-time user interface, and optimizing the model’s performance. The proposed plan includes the following components:

1. **Data Processing Pipeline**We successfully built a data processing pipeline to clean and merge datasets by cleaning whitespaces, dashes, unicode characters, non-ascii characters. We performed embedding analysis using **Sentence Transformer model** (**all-MiniLM-L6-v2**) to extract relevant financial information for the dataset. Additionally, we generated multiple exploratory data analysis (EDA) plots to visualize several hypothesis and insights regarding the relationships between the feature variables and if longer instructions would lead to longer resposonses. All code and documentation are available on GitHub.
2. **Data Streaming Pipeline Completion**The team will finalize the vector data pipeline using **Qdrant**, a high-dimensional vector database designed to support efficient similarity searches. This component will enable quick retrieval of relevant financial context, enhancing the accuracy and relevance of responses generated by the model.
3. **Training Pipeline and Fine-Tuning**
   * **Primary Models:** The primary models fine-tuned in this phase are **LLaMA 7B and Mistral 7B**. Both are transformer-based, auto-regressive language models with **7 billion parameters**, making them well-suited for interpreting complex financial language. **LLaMA 7B**, developed by **Meta AI’s FAIR team**, is known for its versatility and computational efficiency, and its fine-tuning has been **completed**. **Mistral 7B**, leveraging sparse attention mechanisms for high performance with reduced computational demand, is **nearly fine-tuned**. We are currently building the inference pipeline.
   * **Additional Models:** Depending on available resources, the team may explore additional models such as **Falcon 7B, LLaMA 3B, and Distill-GPT** to increase flexibility and computational efficiency in handling financial queries.
   * **Fine-Tuning Method:** To adapt the models effectively to financial contexts, we have employed **Parameter-Efficient Fine-Tuning (PEFT) with QLoRA** in the **training pipeline**. This technique adjusts only a minimal set of model parameters—specifically the **‘q\_proj’** and **‘v\_proj’** modules—while keeping the rest frozen. By reducing the number of trainable parameters, PEFT with QLoRA allows us to fine-tune the models on financial data without excessive VRAM usage, thereby maintaining both **accuracy and computational efficiency**.
4. **Inference Pipeline and User Interface Development**
   * **RESTful API Development:**A RESTful API is being developed to act as the bridge between the fine-tuned models and the user interface, enabling real-time, seamless interactions. The API processes user queries, forwards them to the fine-tuned models for inference, and returns tailored financial advice. Designed with scalability and efficiency in mind, it ensures the system can handle multiple concurrent requests without compromising performance. Security features such as authentication and encryption are incorporated to protect sensitive financial data, ensuring user trust and privacy.
   * **User Interface Development:**The user interface, built using Gradio, provides an intuitive platform for users to interact with the model. Designed for simplicity and ease of use, the interface allows users to input financial questions and receive personalized, real-time advice instantly. The Gradio-based UI ensures accessibility across devices and includes user-friendly features such as customizable inputs, clear error handling, and responsive design. By combining this interactive front end with the API, the system offers a seamless and efficient experience for users seeking AI-driven financial guidance.
5. **Performance Monitoring and Evaluation**The project is utilizing the **Weights & Biases** platform to monitor key performance metrics, such as training loss, accuracy, and perplexity, to ensure continuous improvement. Additionally, the **ROUGE** metric is being integrated to assess linguistic relevance and evaluate the model’s ability to simulate human-like responses. These metrics will allow the team to refine model performance, enhancing the tool’s reliability and contextual accuracy.
6. **Defining Success for FinAdvisor**

Success for FinAdvisor is defined in terms of both quantitative metrics and qualitative user experience goals:

**Quantitative Goals**:

* Achieving a perplexity score within the range of 4 to 6, signifying strong contextual understanding and coherence in responses.
* Attaining high ROUGE scores that demonstrate the model’s ability to produce linguistically accurate and relevant responses, effectively simulating human-like financial advice.
* A steady decrease in training loss and improved accuracy, indicating robust model learning and adaptation to financial data.

**Qualitative Goals**:

* **User Satisfaction**: Positive feedback from users regarding the relevance, clarity, and usability of FinAdvisor's recommendations.
* **Reliability**: The model consistently provides accurate and contextually appropriate advice across various financial topics, such as budgeting, investment, and risk assessment.
* **Accessibility**: Users find the Gradio-based interface intuitive and efficient for obtaining timely financial insights.

### **6. Phase 1 Results :**

The model training is closely monitored using key performance metrics, particularly focusing on the LLaMA 7B model's adaptation process. Training loss and perplexity scores, crucial indicators of model performance, have shown considerable improvement. Specifically, the training loss for the LLaMA 7B model has steadily decreased from 2.2643 at step 50 to 1.527531 at step 250, demonstrating the model's effective learning and its enhanced ability to generate accurate financial advice. This reduction in loss is critical as it shows the model's growing proficiency in handling complex financial queries.

In terms of assessing the model’s capability to understand context and predict subsequent tokens, the current perplexity score is **7.12**. Although state-of-the-art models typically achieve scores between 4 and 6, this score is commendable given our limited computational resources and suggests that our model is nearing industry-leading standards. Going forward, we plan to integrate additional metrics such as the ROUGE Score to evaluate the overlap between model-generated responses and reference texts, further enhancing our insights into the model's linguistic accuracy and its effectiveness in providing contextually relevant financial advice.

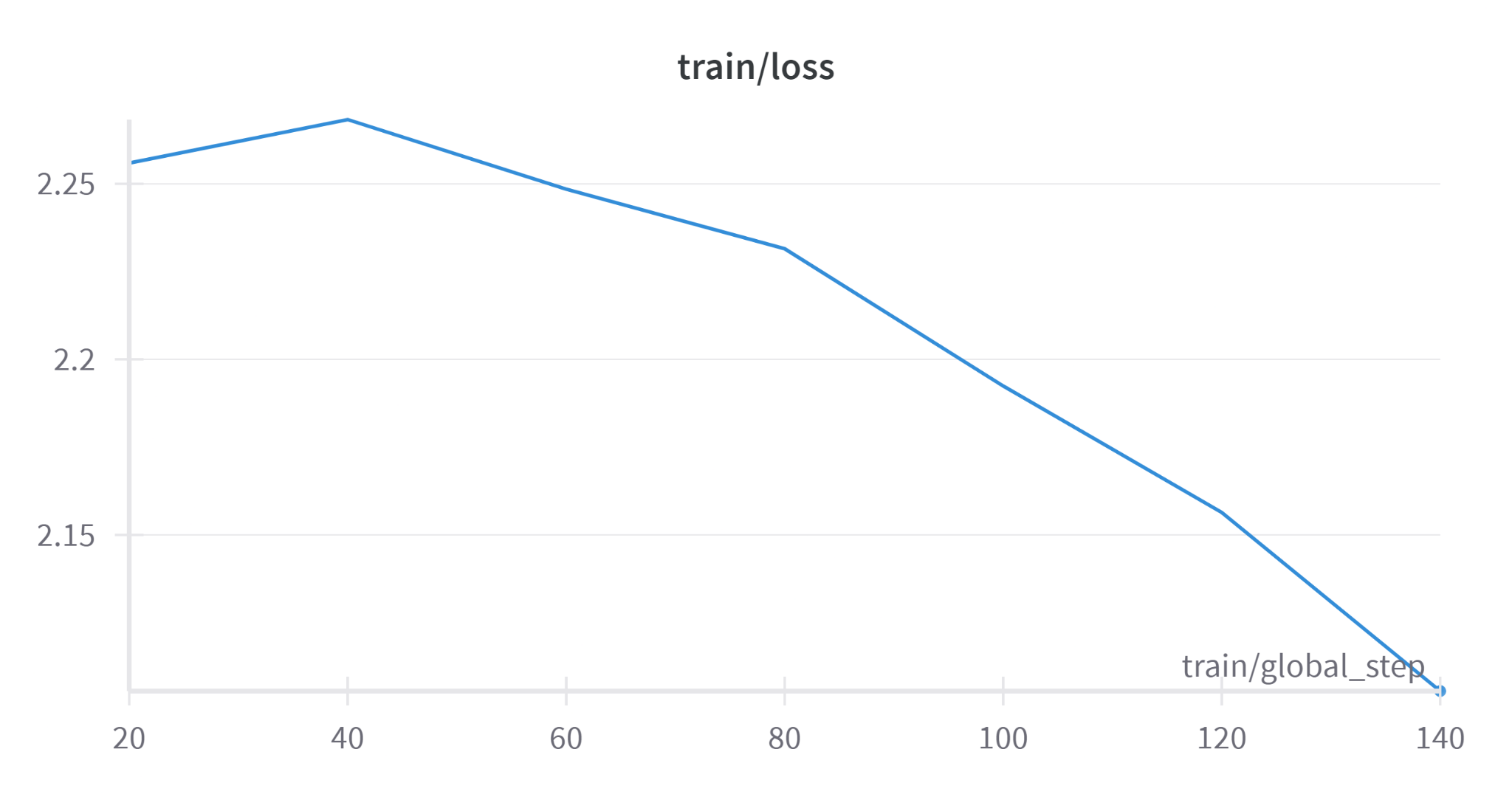


Figure 1:. Fine Tuning Training Loss

### **7. Project Milestones and Timeline**

The FinAdvisor project is organized into two primary phases over the next several weeks to ensure structured development, testing, and refinement of the tool. We have divided the project into Mid-Phase and End-of-Phase Milestones, each addressing critical components that will bring FinAdvisor closer to delivering real-time, reliable, and accessible financial advice. This timeline ensures that the core features are completed sequentially, with ample time for testing, user feedback integration, and performance evaluation.

Below is a detailed breakdown of the project’s timeline:

**Mid-Phase (Next 4 Weeks)**:

* Complete the Qdrant-based vector data pipeline for similarity searches.
* Finalize fine-tuning of the Mistral 7B model.
* Fully integrate performance monitoring metrics, including ROUGE and training loss, for ongoing evaluation.
* Develop a prototype of the Gradio-based user interface for initial testing.
* Complete the RESTful API to support real-time interactions with the model.

**End-of-Phase (Completion)**:

* Conduct user testing of the Gradio interface and refine the interface based on user feedback.
* Finalize documentation and prepare a comprehensive project report summarizing model performance, user feedback, and recommendations for future development.

### **8. References**

1. **Financial News Articles and Blogs**. A collection of current and historical articles sourced from financial media outlets such as Reuters, Bloomberg, and The Financial Times.
2. **Hugging Face**. Pre-trained Large Language Models (LLMs). Available at: [https://huggingface.co](https://huggingface.co/)
3. **Qdrant Vector Database**. High-performance vector database for storing embeddings. Available at: [https://qdrant.tech](https://qdrant.tech/)
4. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). *LoRA: Low-Rank Adaptation of Large Language Models*. Available at: <https://arxiv.org/abs/2106.09685>
5. Biewald, L. (2020). Experiment Tracking with **Weights and Biases**. Available at:<https://www.wandb.com/>
6. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. Available at: https://arxiv.org/abs/2302.13971
7. **Flask/ FastAPI**. RESTful API frameworks for Python. Available at: https://flask.palletsprojects.com/ and <https://fastapi.tiangolo.com/>