**P1 Preliminary Report**

### **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. Summary:**

The project addresses a critical gap in personal finance management, where traditional tools often fail to keep up with the rapidly changing and complex demands of modern financial planning. This challenge results in delayed, inaccurate, or generalized advice, which can hinder individuals from making optimal financial decisions. Leveraging advancements in Artificial Intelligence (AI), particularly Large Language Models (LLMs), this project aims to develop an AI-driven financial advisory tool capable of delivering real-time, personalized financial insights that cater to individual needs in areas like budgeting, investment analysis, and risk assessment.

### **4. Methods and Technical Implementation**

#### **a. Data Collection**

* **Sources**: Financial data is collected from two key sources—the Alpaca News API, which provides real-time financial news articles, and the Hugging Face finance-alpaca dataset, a question-answer dataset focused on finance. These sources supply diverse and current financial information essential for building a reliable, contextually aware financial advisory model.
* **Purpose**: By using up-to-date and comprehensive financial data, the model can better understand trends, terminology, and financial context, enhancing its ability to provide accurate, relevant advice.

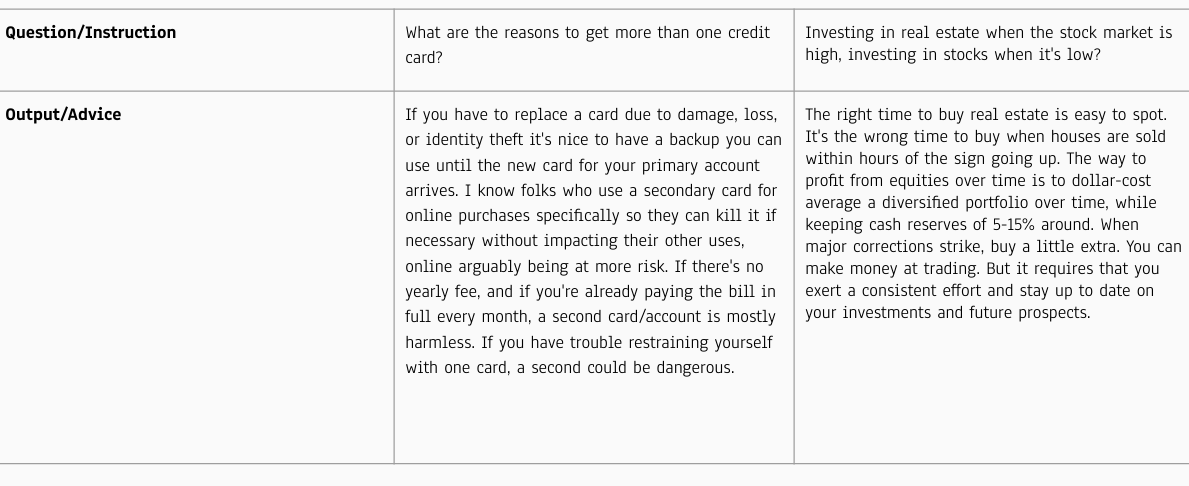


Fig 1: Sample Data

#### **b. Data Processing**

* **Cleaning and Preprocessing**: Raw financial data undergoes thorough cleaning and preprocessing to prepare it for model training.
  + **Text Normalization**: The data is standardized by removing special characters, converting text to lowercase, and handling punctuation for consistency.
  + **Tokenization**: Text data is broken down into tokens to facilitate embedding and storage in the vector database.
* **Embedding and Storage in Qdrant**: Processed data is transformed into vector embeddings and stored in Qdrant, a vector database optimized for fast and efficient retrieval. This enables quick access to relevant information during model inference, ensuring that data is organized and readily available when needed.

#### **c. Modeling**

* **Fine-Tuning with LoRA (Low-Rank Adaptation)**:
  + **LoRA Overview**: LoRA is used to fine-tune the language model, adjusting a minimal set of parameters that allow it to specialize in financial contexts without requiring complete retraining.
  + **Training**: Through LoRA adaptation, the model learns to interpret financial data, allowing it to deliver accurate and context-sensitive financial advice.
  + **Performance Tracking**: Model performance is continuously monitored using Comet, a machine learning experiment management tool that tracks metrics and validates fine-tuning accuracy.
* **Retrieval-Augmented Generation (RAG)**:
  + **Purpose of RAG**: RAG enhances the model's contextual accuracy by fetching relevant data from Qdrant during response generation.
  + **Implementation**: When the model receives a query, RAG retrieves related information from Qdrant, allowing the model to incorporate the most recent and pertinent data into its responses. This ensures that responses are both accurate and contextually relevant.

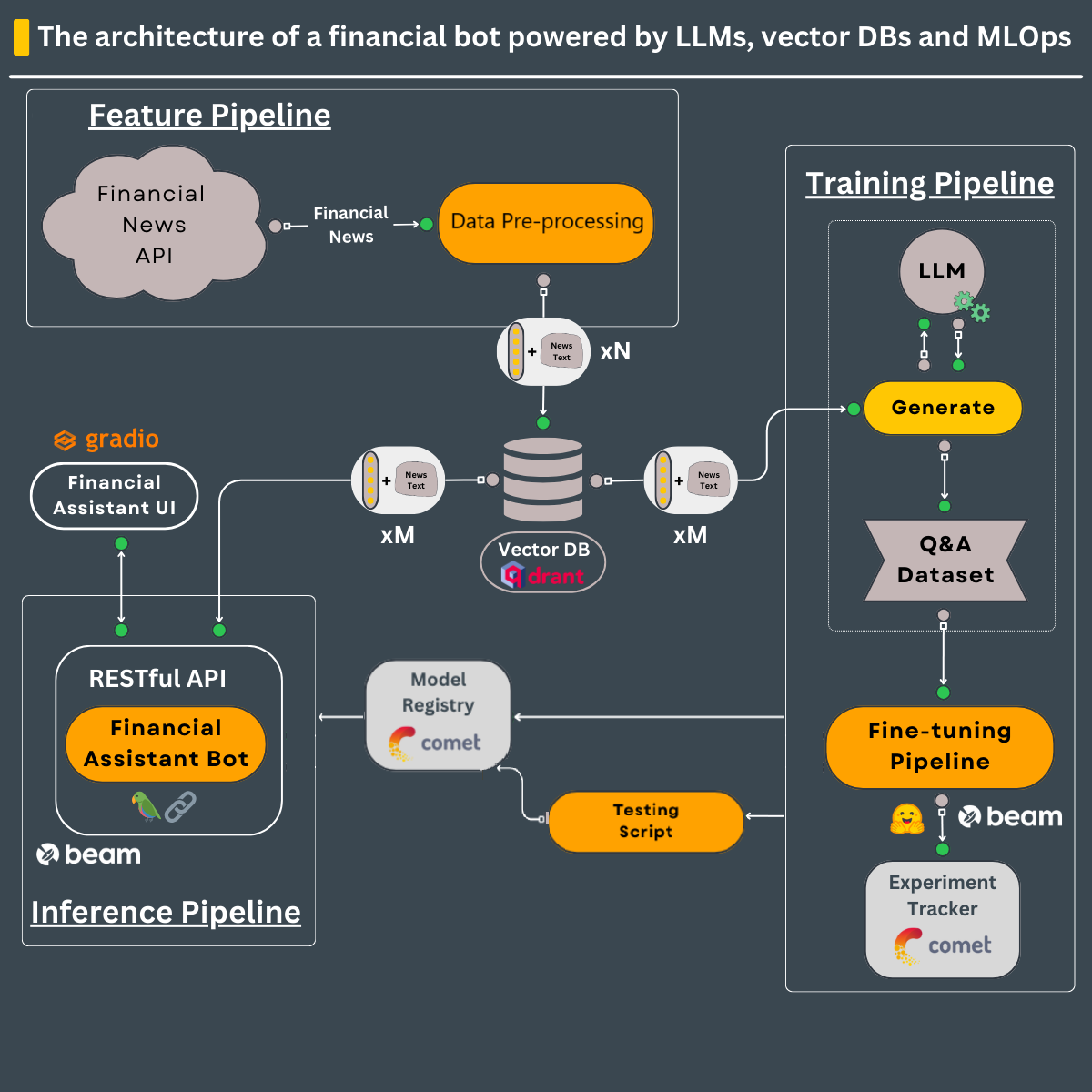


Fig 2: Methodology Pipeline

**d. Model Deployment**

* **RESTful API**: The fine-tuned model is deployed via a RESTful API, providing real-time accessibility for user interaction.
* **User Interaction with Gradio**: The API integrates with Gradio, an interactive interface where users can input financial questions and view responses. Gradio converts user queries into model-friendly formats, manages response displays, and supports real-time engagement.

#### **e. Inference Pipeline**

* The inference pipeline is optimized for handling real-time data retrieval and efficient model inference. By integrating Qdrant with RAG, the system ensures that each response generated by the model is up-to-date and contextually accurate.
* **Outcome**: This approach provides a fast, responsive, and user-friendly experience, allowing users to receive tailored financial advice on demand.

### **5. Results**

The following results highlight key insights and demonstrate the model's effectiveness in providing personalized financial advice based on real-life data.

* **Exploratory Data Analysis (EDA): Word Cloud of Financial Terms**: The word cloud generated from the financial news dataset reveals the most frequent terms, such as "investment," "risk," "credit," and "market." These terms indicate the primary focus areas the model must handle, ensuring it is attuned to core financial themes. This visual provides confidence that the data aligns with the intended use of the model, focusing on areas relevant to personal finance advice.

The word cloud provides a quick overview of these key focus areas in financial data, helping readers understand the primary domains that the model will be trained on. By illustrating which terms appear most often, this visualization highlights the financial themes that the model must interpret and respond to effectively. This overview also reinforces the model's primary training context, ensuring that it is well-equipped to provide relevant advice on these commonly discussed financial topics.



Fig 3: Word Cloud of Financial Terms

* **Prompt Length Distribution:** The prompt length distribution is displayed as a histogram, showing the range and frequency of input lengths within the dataset. This figure plays a critical role in determining the optimal context window size for Large Language Models (LLMs), as it helps identify typical input sizes. By aligning input lengths with the model’s capabilities, we can ensure that the model is both efficient and accurate in processing financial queries. In this histogram, any peaks or patterns—such as a concentration of short or long prompts—can indicate the typical range of user queries, which has implications for computational efficiency and performance optimization. This insight allows for fine-tuning of input parameters, which is essential for maximizing model performance without unnecessary computational load.

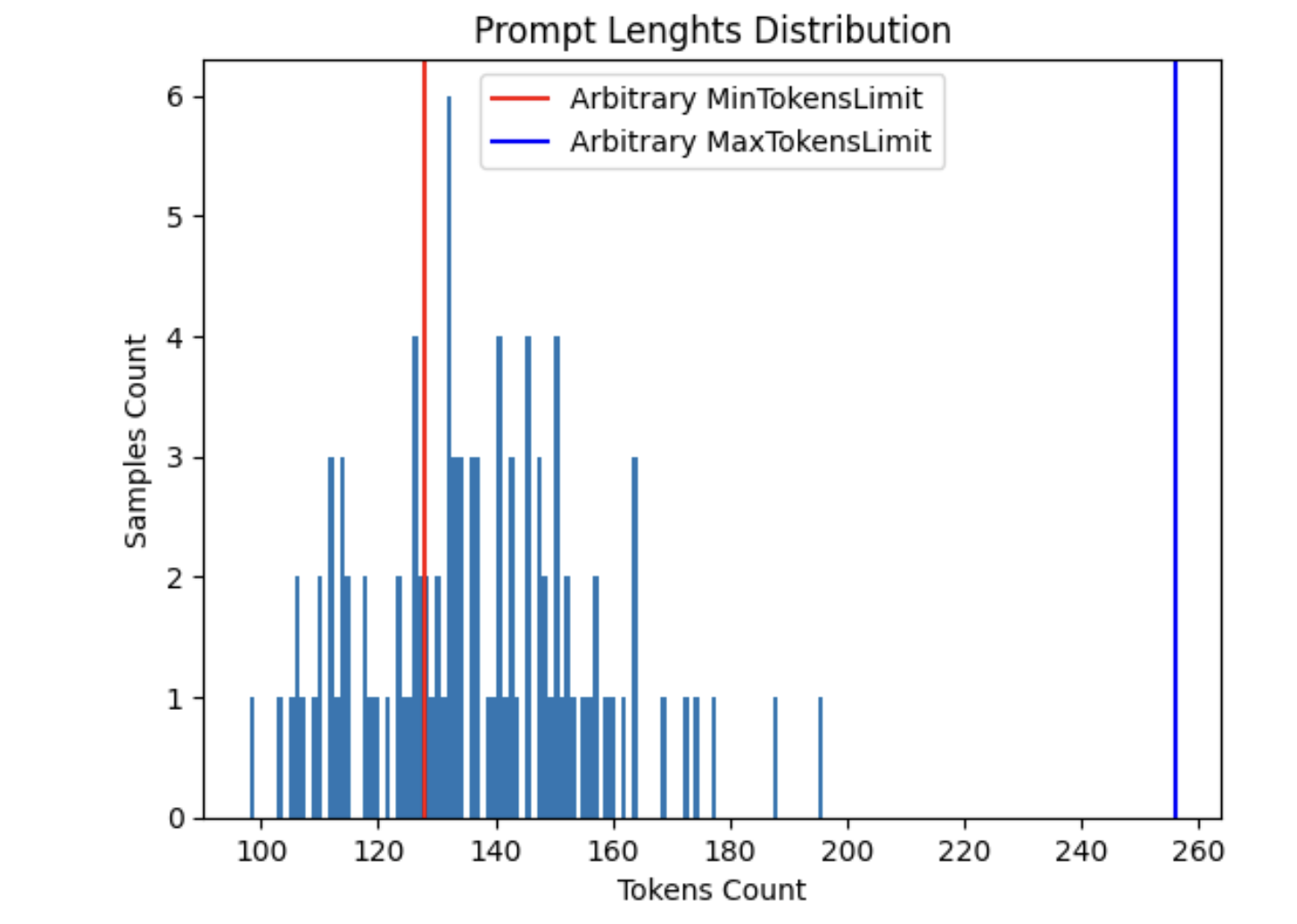


Fig 4: Prompt Length Distribution

### **7. Discussion**

The preliminary findings from FinAdvisor, an AI-driven financial advisory tool, offer several insights and implications for further development. The Exploratory Data Analysis (EDA) revealed a notable emphasis on terms such as "investment," "risk," and "credit," indicating alignment with essential topics in personal finance. This alignment suggests that the model is capable of addressing key areas of financial management, including budgeting, investment analysis, and risk assessment, which are crucial for providing accurate and tailored advice to users. Additionally, the analysis of prompt length distribution underscores the importance of optimizing the model’s context window size, ensuring that FinAdvisor operates with both computational efficiency and high levels of accuracy.

The results of this phase demonstrate the potential applicability of FinAdvisor in real-world financial decision-making contexts. By focusing on commonly referenced financial themes and using a contextually aware retrieval mechanism, FinAdvisor can deliver prompt, relevant, and context-sensitive responses that aid users in making informed financial choices. Furthermore, the integration of Qdrant and the Retrieval-Augmented Generation (RAG) approach enables real-time data retrieval, ensuring that each response generated reflects the most current financial information available, thus enhancing the model’s relevance and utility.

The primary goals for this project phase were achieved, as the team successfully implemented foundational aspects of data preprocessing, model fine-tuning, and real-time data integration. However, certain areas have been identified for improvement. For instance, expanding the dataset to cover a broader range of financial scenarios could increase the model's versatility, and refining the tokenization process may enhance FinAdvisor’s capacity to handle complex, multi-part questions. The exploration of alternative vector storage solutions may also offer opportunities to optimize retrieval efficiency.

Looking ahead, earlier incorporation of user feedback could further enhance model customization, allowing for adjustments that would align more closely with user expectations and preferences. Additionally, tracking a wider array of performance metrics could improve FinAdvisor’s responsiveness and accuracy in varied financial contexts. By addressing these areas, FinAdvisor could more effectively support diverse user needs, ultimately providing personalized and actionable financial advice that adapts to the complexities of modern financial landscapes.

### **8. Statement of Contributions**

The following outlines the contributions of each team member, who equally participated in brainstorming, researching, and implementing the project:

* **Krishna Venkatesh:**
  + Led data collection from the Alpaca News API and embedded selected financial data into the Qdrant vector database.
  + Managed text normalization by removing special characters, converting text to lowercase, and handling punctuation for consistency.
  + Contributed to exploratory data analysis (EDA) with visualizations like a word cloud and assisted in documenting data processing methods.
* **Laasya Anantha Prasad:**
  + Played a key role in tokenizing text data for embedding and model training during data preprocessing.
  + Managed embedding the processed data into Qdrant, ensuring efficient data storage and retrieval.
  + Equally contributed to EDA by identifying significant terms and drafting sections on modeling and technical implementation.
* **Rakshak Kunchum:**
  + Assisted in data collection from the Hugging Face finance-alpaca dataset, ensuring quality for model training.
  + Researched data storage techniques for Qdrant, developing an efficient retrieval plan.
  + Contributed to EDA by generating insights and writing about results interpretation in the report.

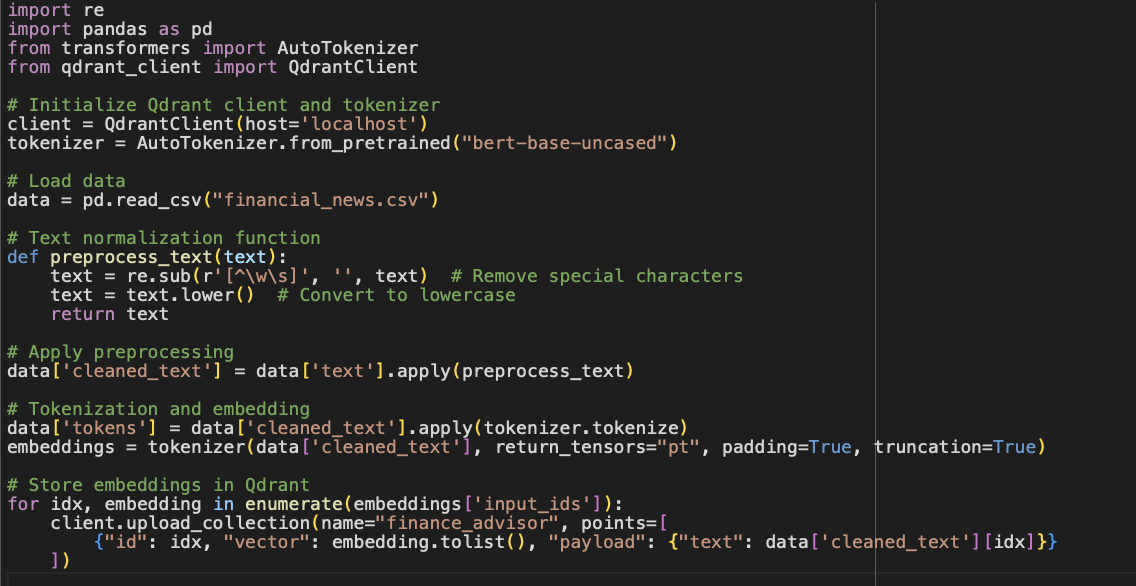
### **9. References**

1. Alpaca Real Time News API. *Alpaca Market News API Documentation*. Available at: <https://docs.alpaca.markets/docs/streaming-real-time-news>
2. Gaurav Bharti. "Finance Alpaca Dataset." *Hugging Face Datasets*. Available at: <https://huggingface.co/datasets/gbharti/finance-alpaca>
3. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. "LoRA: Low-Rank Adaptation of Large Language Models." *arXiv preprint arXiv:2106.09685*. Available at: <https://arxiv.org/abs/2106.09685>
4. Izacard, G., & Grave, E. "Leveraging Passage Retrieval with Generative Models for Qdrant. "Qdrant Vector Database Documentation." Available at: <https://qdrant.tech/>Open Domain Question Answering." *arXiv preprint arXiv:2007.01282*. Available at: <https://arxiv.org/abs/2007.01282>
5. Gradio. "Gradio Documentation." Available at: https://gradio.app/docs/
6. Comet. "Comet Machine Learning Experiment Management." Available at: <https://www.comet.ml/>

### **10. Appendix**

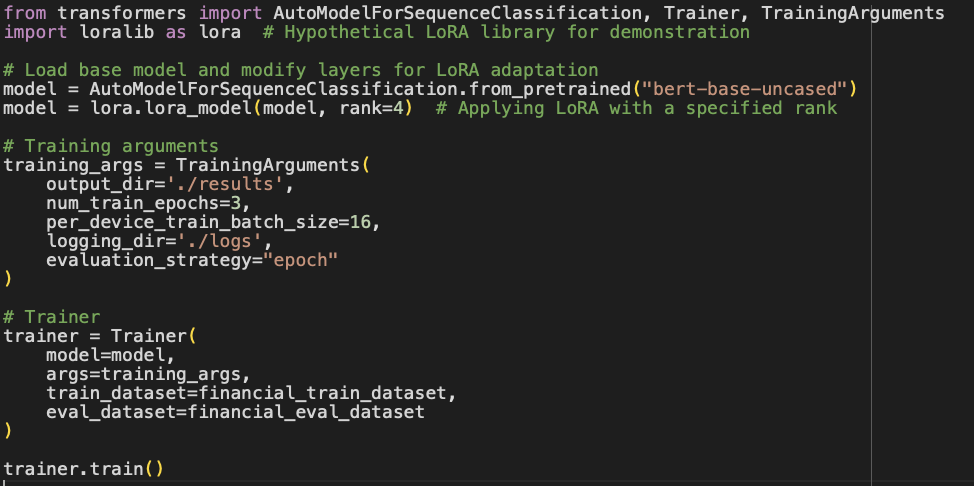
#### **Data Preprocessing**

The following code snippet demonstrates the data cleaning and preprocessing steps, including text normalization, tokenization, and embedding storage in Qdrant.



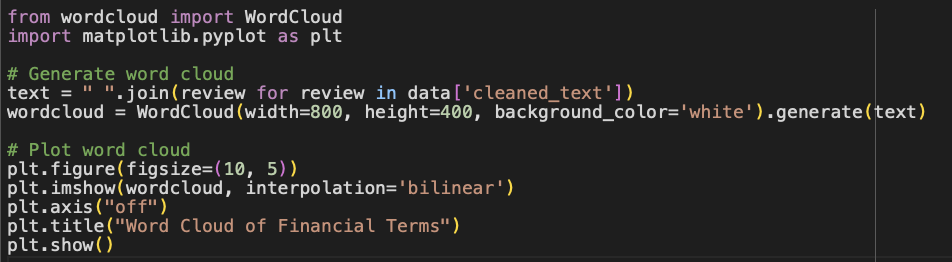
#### **Model Fine-Tuning with LoRA**

Basic outline of the code for fine-tuning a language model using LoRA (Low-Rank Adaptation) for financial context specialization.



#### **Exploratory Data Analysis (EDA) Plot - Word Cloud of Financial Terms**

The word cloud visualizes the most frequent terms in the financial dataset, giving a quick overview of the model's primary focus areas.



#### **Exploratory Data Analysis (EDA) Plot - Prompt Length Distribution**

