Baseline Model Report

1. Analytic Approach

Target Definition

Our main goal is to build a **financial chatbot** that provides accurate and relevant answers to user queries about financial topics (e.g., stock prices, investment strategies, financial terminology, economic indicators).

The **target** is:

- Generate coherent, contextually correct answers to financial queries
- Provide factually accurate information based on both the model's knowledge and retrieved references

Inputs (Description)

- **Training Text Data**: A curated collection of financial texts such as articles, reports, and FAQs.
- **Validation and Testing Data**: Representative financial questions with known correct answers.
- **User Prompt** : A prompt provided by everyday investors, advisors, or other participants. Due to their limited financial background, it may be incomplete or lacking clarity.
- **Qdrant Vector Database (RAG)**: A set of financial documents stored in Qdrant, which are retrieved to aid in grounding the model's responses.

What Kind of Model Was Built?

A **Large Language Model (LLM)** approach based on **Falcon 7B** is used:

1. **Preprocessed Data**: Cleaned and formatted text.
2. **Fine-tuning**: Trained the model on financial domain data.
3. **Retrieval-Augmented Generation (RAG)**: Relevant documents fetched from Qdrant to support context for the user query.

2. Model Description

Models and Parameters

- **Falcon 7B**: A 7-billion-parameter open-source model.
- **Hyper-Parameters**:

- Learning Rate: ~2e-5
- Batch Size: 4–8
- Number of Epochs: 3–5
- Max Sequence Length: 512

Data Flow Graph (Conceptual)

1. **User Query →**
2. **Retrieve relevant docs from Qdrant →**
3. **Combine query + retrieved docs →**
4. **Falcon 7B** generates answer →
5. **Answer returned**.

3. Results (Model Performance)

Because this is a **generative model**, we look at both **perplexity** and **qualitative QA metrics**. In addition, we measured **faithfulness** and how well the retrieved context is used:

- **Answer Similarity**: ~0.1343
  - Measures how semantically similar the retrieved response is to the ground-truth answer, assessing the relevance and correctness of the generated answer.
- **Faithfulness**: ~0.293
  - Measures alignment of the answer with the given query and any retrieved context (lower indicates more "hallucination" or off-topic content).

These results suggest that the system is somewhat knowledgeable but **sometimes drifts** from the context.

4. Model Understanding

Variable Importance

With large language models, there is no traditional "feature importance." Instead, attention and learned embeddings guide the responses.

Insights Derived

1. **Basic Financial Queries**: The model can explain common finance concepts.
2. **Dependency on RAG**: Performance depends on retrieving relevant documents from Qdrant.

3. **Hallucinations**: The model sometimes invents or misattributes facts, highlighted by the relatively low faithfulness metric.
4. **Technical problems**: The model sometimes gets stuck or repeats the same sentence many times.

---

5. Conclusion and Discussions for Next Steps

Conclusion on Feasibility

- **Feasibility**: We can build a financial chatbot using Falcon 7B, leveraging a RAG approach.
- **Challenges**: Te calculated ragas scores show that the system needs more robust retrieval and fine-tuning methods to ensure accurate, on-topic answers.

Discussion on Overfitting

- **Low Overfitting Risk**: Few training epochs and a moderate learning rate reduce overfitting risk.
- **Potential Underfitting**: The model may not be fully capturing nuanced financial knowledge.

What Other Features Can Be Generated from the Current Data?

1. **Named Entity and Sector Tags**: Mark up companies, symbols, or industries to improve retrieval.
2. **Structured Prompting**: Guide the model with clearer instructions (e.g., "Focus on current market data from Qdrant retrieval").

What Other Relevant Data Sources Are Available?

- **Regulatory Filings (10-K, 10-Q)** for deeper company insight.
- **Real-time Market Data** feeds to handle up-to-date queries.
- **User-Generated Q&A** for real-world patterns in financial questions.

Future improvement approaches

- **Improved Prompt Engineering**: Better structured prompts or "chain-of-thought" to reduce hallucinations.
- **Better context Retrieval**: Improve the RAG-based approach by enhancing the relevance of the context provided to the model, ensuring it is highly aligned with the query for more accurate and concise responses.