FIN-QA CHATBOT

A finance domain-specific chatbot fine-tuned on real-world Q&A pairs to provide accurate, context-aware answers to user questions

Nina Mwangi

15.06.2025
ML Techniques I Summative

INTRODUCTION

Dataset Name: `financial-qa-10K`

Source: [Hugging Face – virattt/financial-qa-10K]

Size: 10,000 examples

Language: English

Domain: Finance

Purpose of the Chatbot and Justification.

This project presents a domain-specific chatbot designed for the financial sector to provide accurate, context-aware answers to finance-related queries. The chatbot is fine-tuned on a curated dataset of question-answer pairs grounded in real-world financial documents, such as company 10-K filings. Its purpose is to assist users, from investors and analysts to students and the general public, in easily navigating complex financial information. The relevance of such a tool is underscored by the growing need for accessible financial literacy, the complexity of regulatory filings, and the time constraints faced by professionals who require fast and accurate insights. By focusing on a specialized domain, the chatbot avoids generic responses and delivers precise, reliable answers grounded in factual context.

Data structure

Field	**Description**
Question	A natural language financial question (e.g., *"What is a 401(k)?"*)
Answer	A corresponding answer in plain text (e.g., *"A 401(k) is a retirement savings plan sponsored by an employer."*)
Context	A paragraph or sentence containing the supporting evidence or background from a source document (e.g., SEC filings or 10-Ks)
Ticker	The stock ticker symbol for the company associated with the context
Filing	The name of the financial report file, such as “2023_10K”

Key Characteristics of the Dataset

Context-Aware: The dataset includes a `context` field, enabling models to learn grounded QA behavior by referencing actual financial documents.

Fact-Based: Answers are generally extractive or paraphrased from the context, reducing hallucination risks.

Diversity: Covers a wide range of financial topics, including retirement plans, earnings, IPOs, company performance, and regulations.

Preprocessing

The chatbot was trained using the **`financial-qa-10K`** dataset, a high-quality resource containing 10,000 question-answer pairs grounded in financial contexts such as SEC filings and annual reports. Each sample includes a user query, its corresponding answer, and the supporting context drawn from real-world documents. Comprehensive preprocessing was conducted to prepare the data for training. Questions and contexts were combined into structured prompts, while answers were used as target outputs. Text was cleaned by stripping unnecessary whitespace and punctuation where appropriate. Tokenization was performed using the **T5 tokenizer**, which applies a subword-level approach ensuring robust handling of rare financial terms and acronyms. Input and output sequences were truncated and padded to uniform lengths to maintain compatibility with the Transformer architecture. No missing values were encountered in the dataset, allowing for a smooth and complete preprocessing pipeline. All steps were documented and implemented using Hugging Face's **`datasets`** library for reproducibility.

Hyperparameter Tuning Summary Table

To optimize the performance of my Finance QA chatbot, I conducted a series of hyperparameter tuning experiments and I focused on improving generation quality based on **ROUGE-1**, **BLEU**, and **Exact Match (EM)** scores. The goal was to outperform the baseline configuration by systematically adjusting key training variables and prompt formatting strategies.

Exp	LR	Batch Size	Epoch	Input	ROUGE-1	BLEU	EM
1	5e-5	8	8	128	29.28%	14.29%	0.00
2	3e-5	8	10	256	30.73%	15.96%	0.00
3	5e-5	6	15	512	58.93%	35.23%	14.43%
4	5e-5	8	10	256	62.09%	35.47%	22.71%
5	3e-5	8	10	256	69.61%	43.75%	25.43%

Prompts

1. prompt = f"Answer the question based on the following context.\nContext: {context}\nQuestion: {question}"

2. prompt = f"Q: {question} A:"

Findings

Learning Rate: Lowering the learning rate from `5e-5` to `3e-5` resulted in more stable convergence and better generalization, as shown in experiments 2 and 5.

Batch Size: A batch size of `8` balanced memory efficiency and stability better than `6`.

Input Length: Increasing the max input length to `256` allowed more context and longer questions to be captured, improving generation fidelity.

Prompt Format: Prompt Format 1 (context-aware) performed significantly better than Format 2, confirming that providing explicit context is essential for accurate QA generation.

Conclusion: Experiment 5 produced the best results with a ROUGE-1 score of 69.61%, BLEU of 43.75%, and Exact Match of 25.43%, representing a >40% improvement over the baseline. These results confirm that thoughtful hyperparameter tuning and context inclusion meaningfully enhanced the model's performance.

ChatBot UI Design and Deployment.

To ensure an accessible and user-friendly experience, the Chatbot was deployed using Gradio. The interface is hosted on Hugging Face Spaces, enabling public interaction with the chatbot through a clean and intuitive web interface.

Key Features

- **Simple and Clean Design**: The chatbot UI consists of a two-part layout, a text input box for user queries and a dynamic chat window that displays conversational exchanges in styled bubbles.
- **Instructional Prompts**: Clear usage instructions are embedded within the interface to guide users in formulating finance-related questions.
- **Multi-turn Interaction**: Users can ask multiple questions in sequence without refreshing the page, creating a seamless chat-like experience.
- **Responsive Feedback**: The chatbot responds almost instantly, mimicking the natural flow of a dialogue system and enhancing user satisfaction.

Deployment

The fine-tuned T5 model was uploaded to Hugging Face Models Hub, ensuring reproducibility and public accessibility.The interactive Gradio interface was deployed to Hugging Face Spaces, allowing real-time web-based inference without requiring local installations. This approach ensures that the entire solution, from model inference to user interaction is accessible online with a single click.

CONCLUSION

This project successfully demonstrates the development and deployment of a domain-specific Transformer-based chatbot tailored for the financial sector. By leveraging the FLAN-T5 architecture, fine-tuned on a high-quality financial Q&A dataset, the model was able to generate context-aware and relevant answers to real-world finance questions.

Through iterative hyperparameter tuning, prompt engineering, and validation using ROUGE, BLEU, and Exact Match scores, the chatbot achieved significant performance improvements, with a final ROUGE-1 score of 69.61%, BLEU score of 43.75%, and Exact Match of 25.43%.

The user interface, built with Gradio and deployed on Hugging Face Spaces, offers an intuitive and seamless experience for end users, allowing them to interact with the model without requiring any technical background.

Overall, this project highlights the potential of combining pre-trained language models, domain-specific data, and accessible deployment tools to build intelligent and user-friendly financial assistants. Future work could explore expanding the chatbot's capabilities to support multilingual support, or integration with live financial APIs for real-time insights.