FinBOT

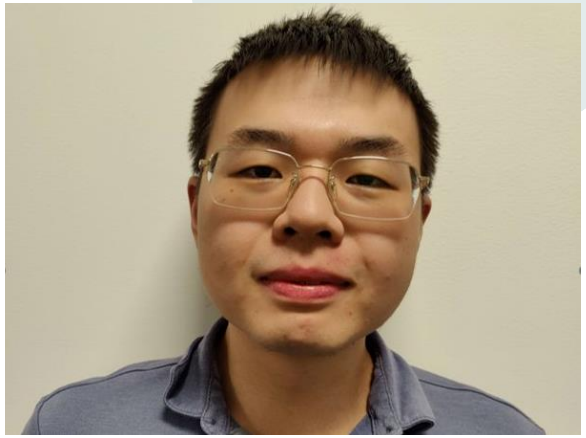
Retail Investments powered by Conversational AI

Conversational AI Research Project
Autumn 2023
November 13, 2023

Agenda

- Team Introduction
- Product Introduction
- Demo
- Data Infrastructure
- Model Iteration
- Next Steps

The Team



Shilong Dai

Snigda Gedela

Shefali Gupta

Takuma Koide

Yif Wang

Product Introduction

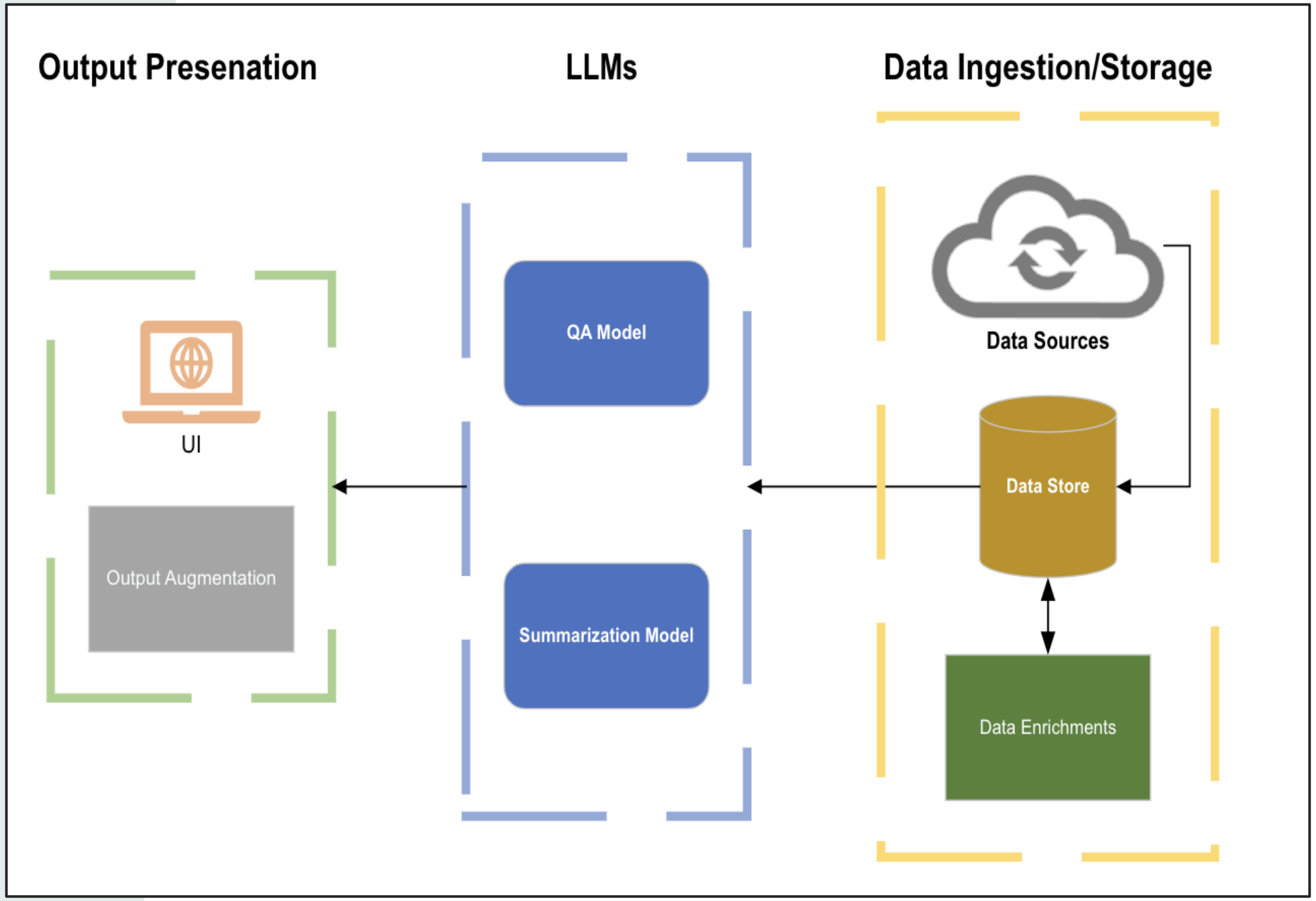
Executive Summary

FinBot takes in user queries and generate answers and key points regarding a company/industry to help user make investment decisions.

User's Query
- User specifies the format of response and asks a question

AI's response
- User receives report on the specified format

User's follow up
- User asks follow up question on the response received



Why FinBot? : Retail investors face major challenges that can be alleviated by Generative AI

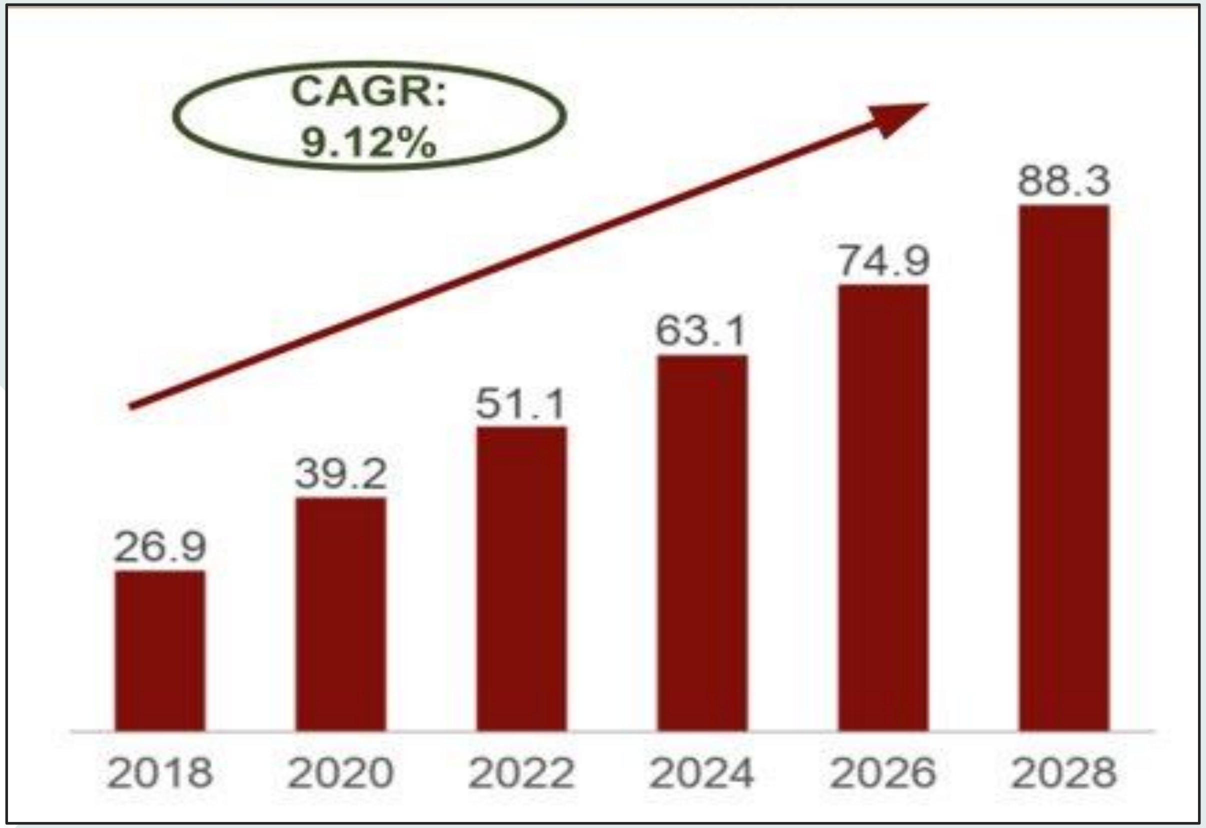
As is Scenario of Retail Investors

- **Limited time** to conduct in depth research and keep up with latest opportunities
- **High cost** associated with financial advisors of few hundred dollars per hour or more

Gen AI with LLM

- Gen AI rapidly analyzes vast data sets, offering insights, reducing the need to manually conduct research and seek opportunities
- Gen AI-powered tools, provide financial advice at a fraction of traditional costs

Market Size Asset Under Management (Trillion $)



An AI driven personal investment assistant, FinBot, can empower retail investors

Functionality & Expected Output

1. Direct QA over News

- **What? :** Generate direct and concise answers to user queries on business and financial news
- **Why? :** Up-to-date information and saves time

***Query:**

Which companies are working on …

***Time Period:**

1 month | 3 months | 6 months

Direct Answer:

The articles mentions several companies that are working on …

2. Key-points Summaries

- **What? :** Extract relevant key-points over multiple topics
- **Why? :** Broader awareness outside user's expertise Abstractive Summarization

Related to Your Query:

Meta, OpenAI and Google are …

Bullet Point Summary:

- On 2023-08-22, it was reported …
- Meta researchers gathered audio...

Source: CNBC.com

3. Intelligent Stock Screener

- **What? :** Intelligently provide key metrics commonly used to identify opportunities
- **Why? :** Identify potential investment opportunities

Public Stocks:

MSFT | AMZN | NVDA

Stock Price Over Time:

Financial KPIs:

Mkt Cap | PE% | Beta

Powered By

Technology

1. Web Crawler

- Crawls and augment thousands of news articles

2. Retrieval Augmented Generation (RAG)

- Technique for enhancing accuracy and reliability of Gen AI output

3. Abstractive Summarization

- Generates concise summary that captures main ideas of articles

***'Query' and 'Time Period' are user inputs. Other highlighted contents are model outputs**

Demo

Data Infrastructure

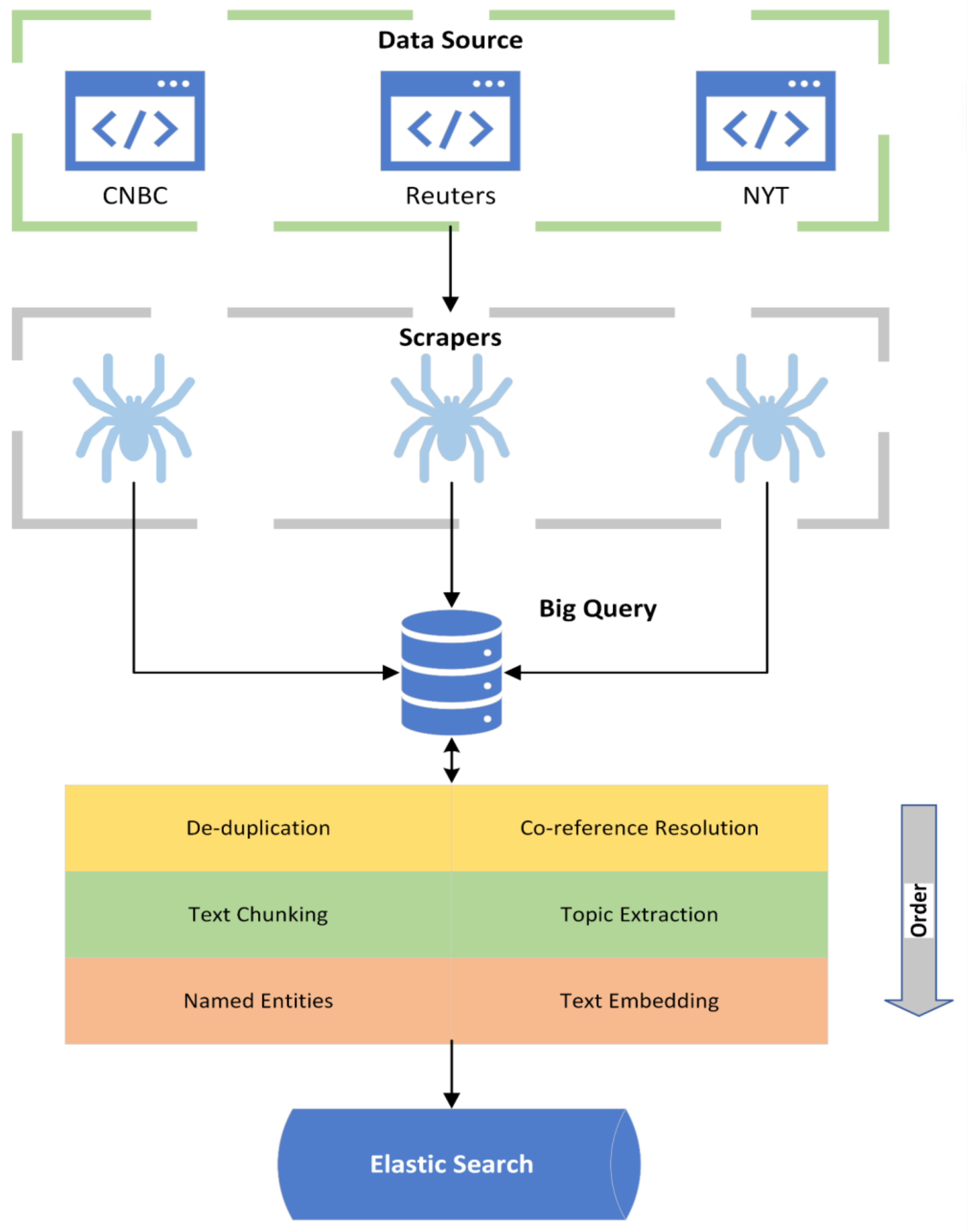
FinBot crawls and augments hundreds and thousands of articles to generate insights

Data Collection

- Web crawled articles from prominent news sources, including CNBC, The New York Times, and Reuters
- Articles are collected from Mar. 2008 to Oct. 2023
- Crawled data is stored in GCS

Data Augmentation

- Augmented the articles by applying a series of NLP techniques
  - Co-reference Resolution,
  - Topic Modeling
  - Text Chunking & NER
  - Embedding
  - Semantic Index



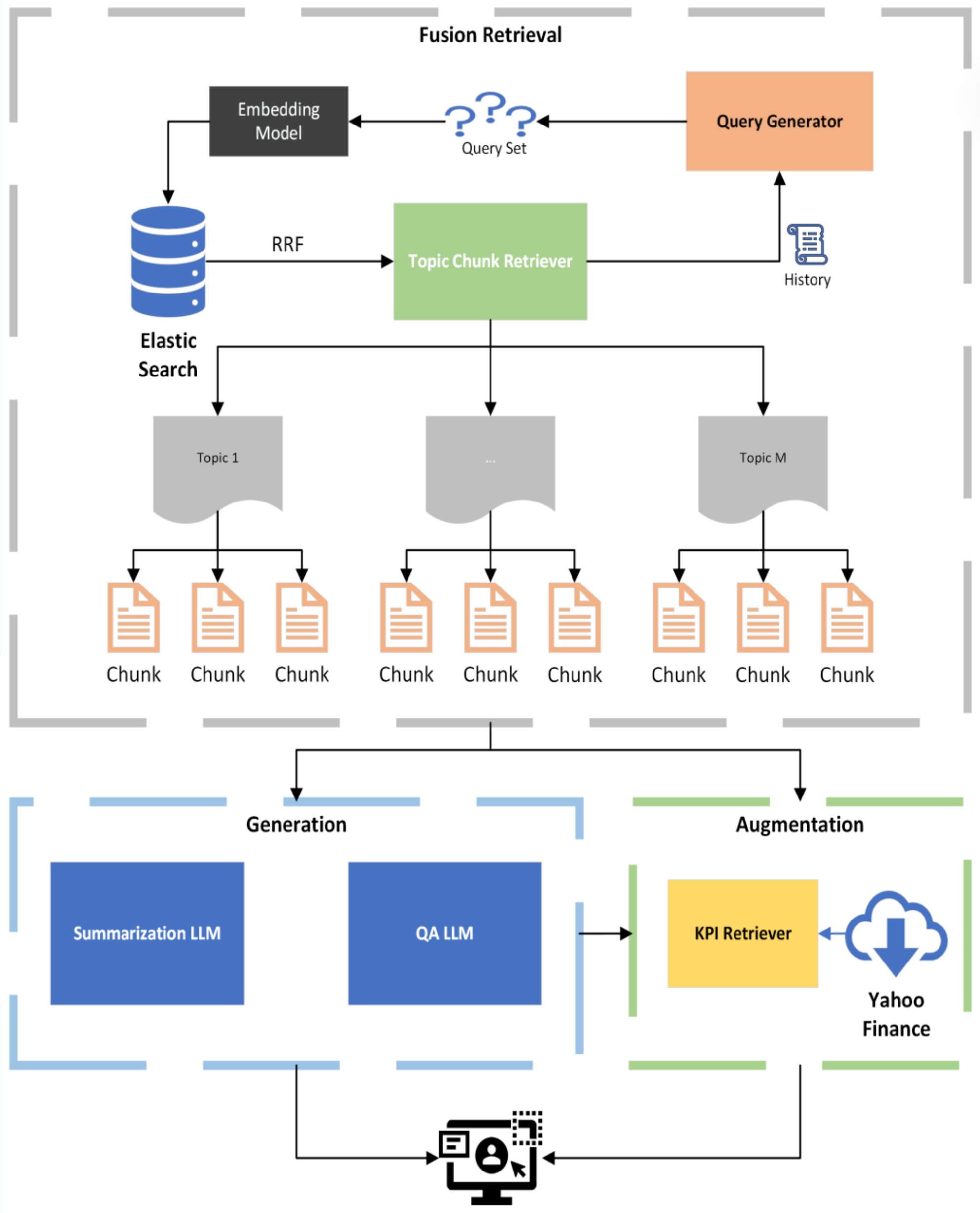
FinBot performs Fusion RAG to produce final insights

Retrieval

- A RAG (Retrieval Augmented Generation) is utilized for chunk retrieval
- It augments the user query based on interaction history and adds similar queries
- Then, for each query, the chunks are retrieved from elastic search using a hybrid RRF search

Generation

- Utilized custom-tuned Summarization LLM and
- Fine-tuned QA Model is used to directly answer the user query
- Enriched text generation by referring to KPI retrieved from Yahoo Finance



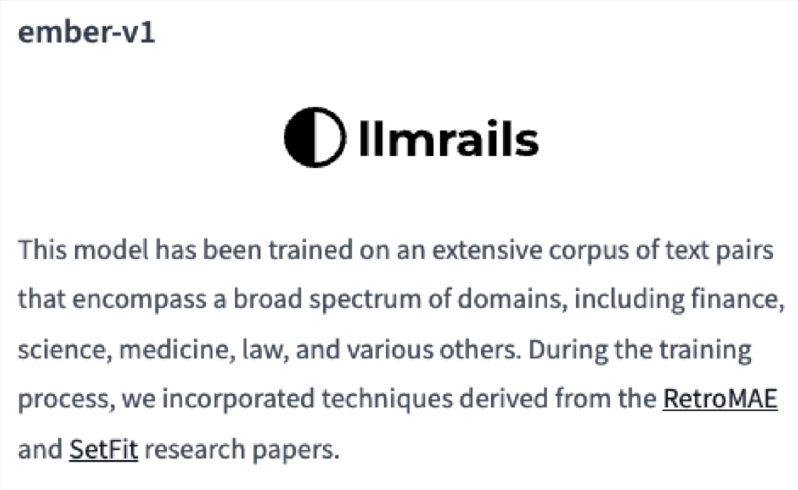
FinBot utilizes the state-of-the-art models as a foundation for text generation

Ember-v1

- The Ember-v1 model is used for creating the embeddings for retrieval
- It is known as the best non-instructed tuned embedding model for retrieval task on the HuggingFace MTEB Leaderboard
- Covers various domains including Finance

Llama-2 13B and Open-Orca Mistral-7B

- Open-Orca Mistral-7B is used as a foundation for key-points summarization
  - Achieves top performance among models ~7B for LLM tasks
  - Generates more similar summaries compared to human written summaries than Llama-Chat
- Llama-Chat-13B is used as the basis for generating concise QA answers
  - One of the most popular model family



ember-v1

This model has been trained on an extensive corpus of text pairs that encompass a broad spectrum of domains, including finance, science, medicine, law, and various others. During the training process, we incorporated techniques derived from the RetroMAE and SetFit research papers.

🐋 Mistral-7B-OpenOrca 🐋

Model Iteration

The embedding model was aligned via contrastive learning to improve retrieval performance

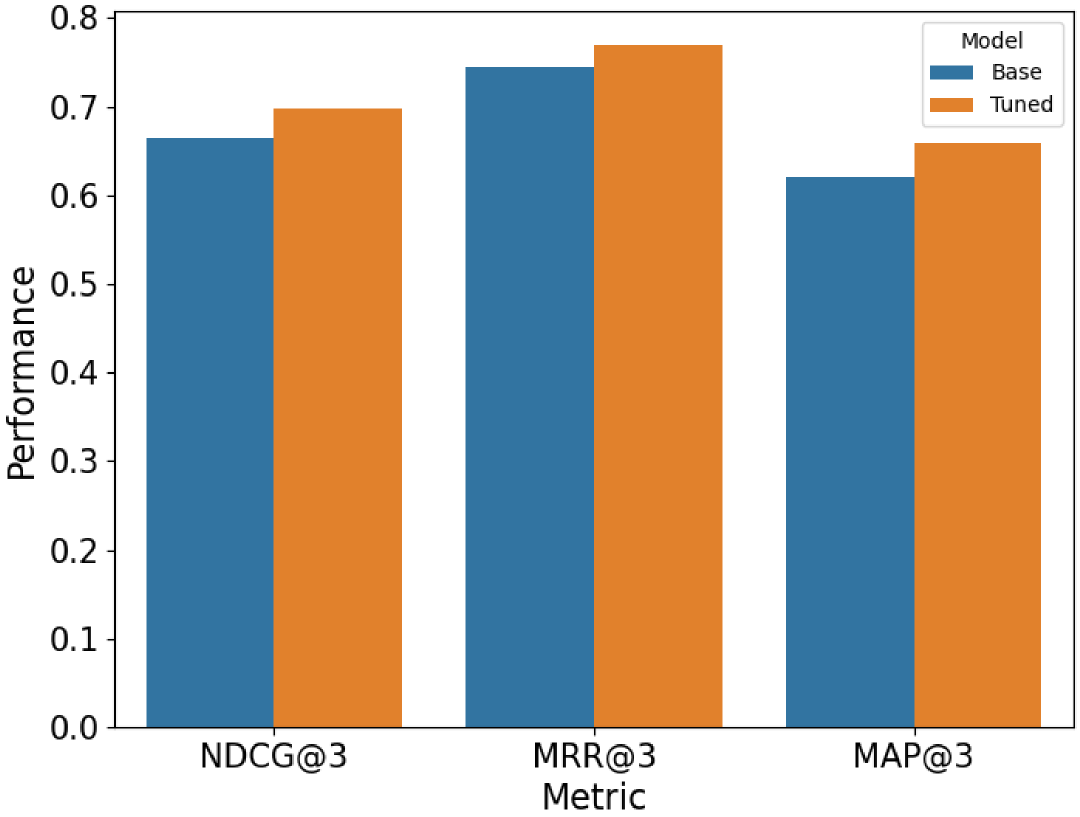
Data Selection

- The FIQA dataset
  - 30K finance related questions and answers scraped from QA forums
- Data augmented with PaLM2 to make the tone formal

Tuning Process

- Contrastive learning with Multiple Negative Ranking Loss
- Minimizing dissimilarity between question/answer vectors

Performance



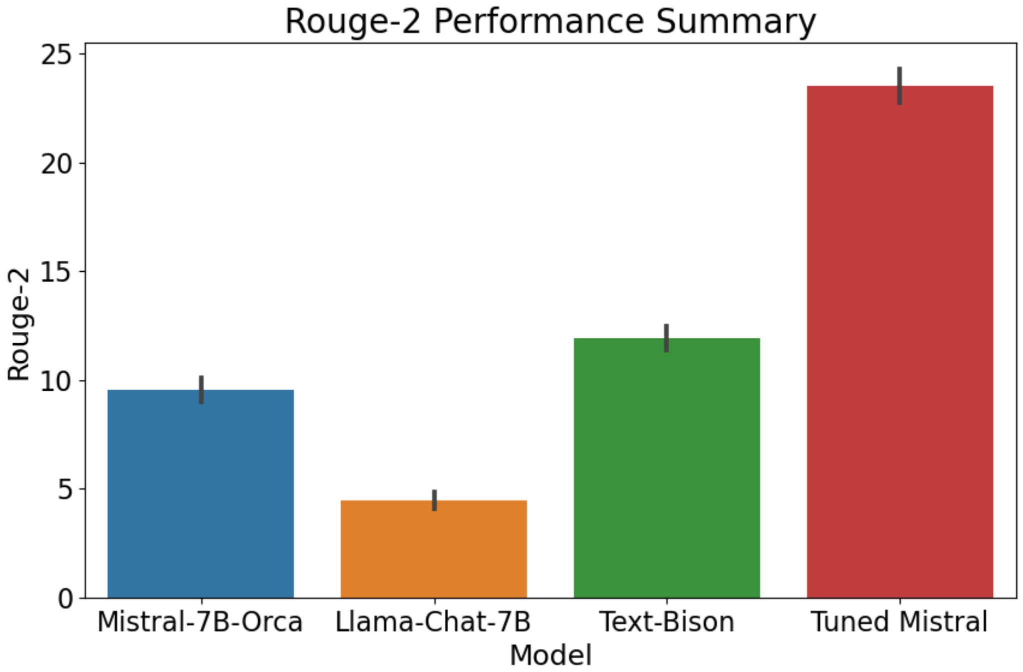
Human written key-points from scraped articles were used to improve the target summary model

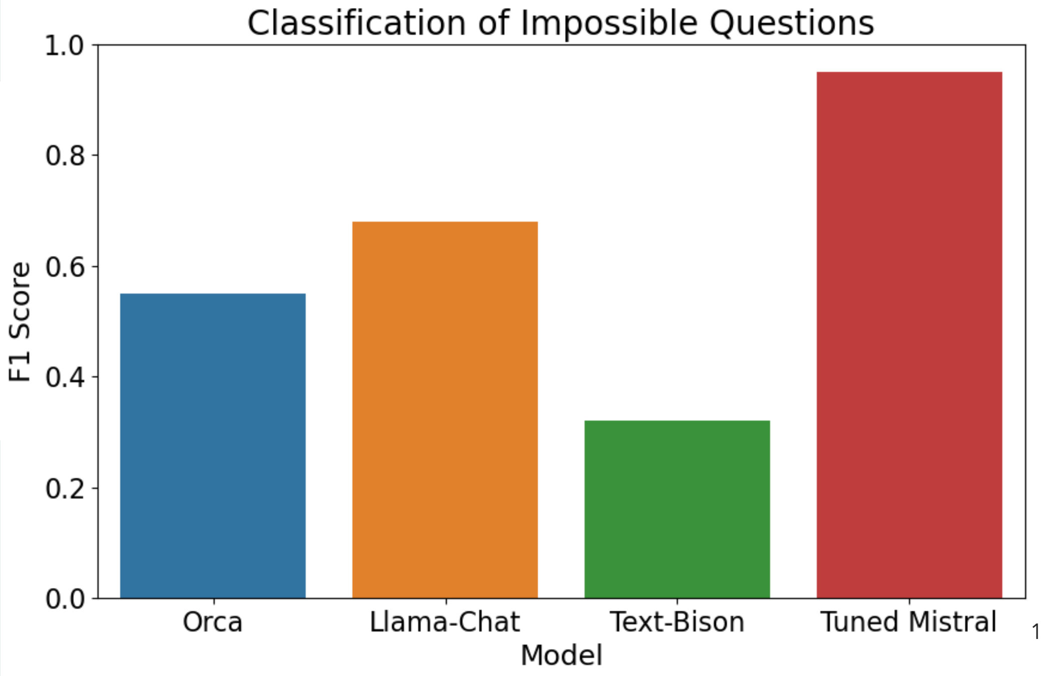
Mimicking Human Key-points

- Mistral-7B tuned on human-written key-points
  - Roughly 35K articles with appropriate key-points
  - Mostly sourced from Reuters and CNBC
- Data adjusted to improve readability
  - Injected published date information
  - Generated tagline over key-points

Targeted Summarization

- Data augmented with potential questions
  - A pair of answerable and unanswerable question
  - Classification of unanswerable question
- Noise were injected by adding random chunks and shuffling chunked context
- Achieved F1 score of 0.95





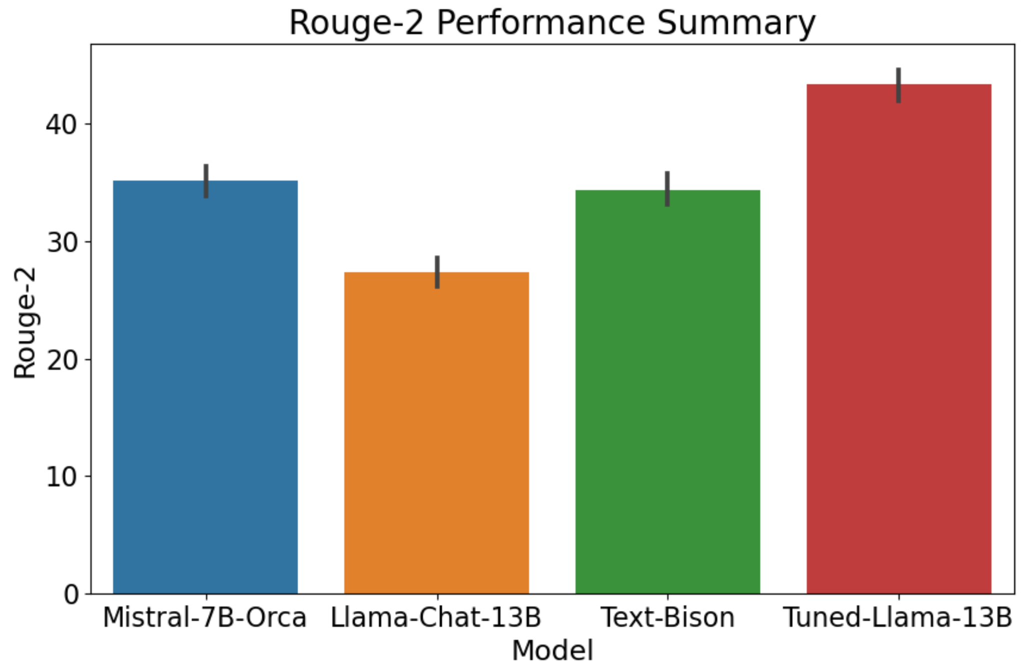
QA model was lightly fine-tuned using RAG specific datasets

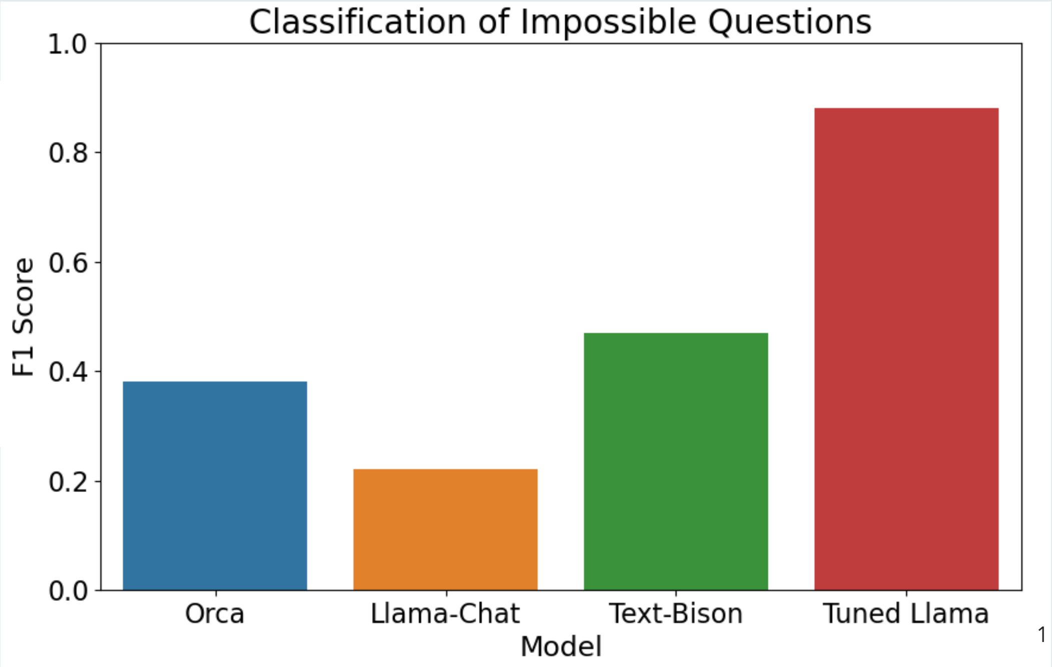
Generating Consistent Outputs

- Llama-Chat-13B tuned on 20K QA pairs with contexts
  - Finance related texts from TAT-QA/FINQA
  - General texts from WebGLM QA
  - Concise responses in under one paragraph
- Data adjusted to match RAG settings
  - Shuffled and mixed context chunks

Noise Injection

- Source datasets did not contain unanswerable questions
- Added 10% randomly chosen pure noise context with sampled questions
- Achieved F1 score of 0.9 with 1 epoch of tuning





Final Thoughts

Conclusions

Product

FinBot is a state-of-the-art AI driven financial assistant that empowers the retail investors by allowing them to make more informed decision on their own assets

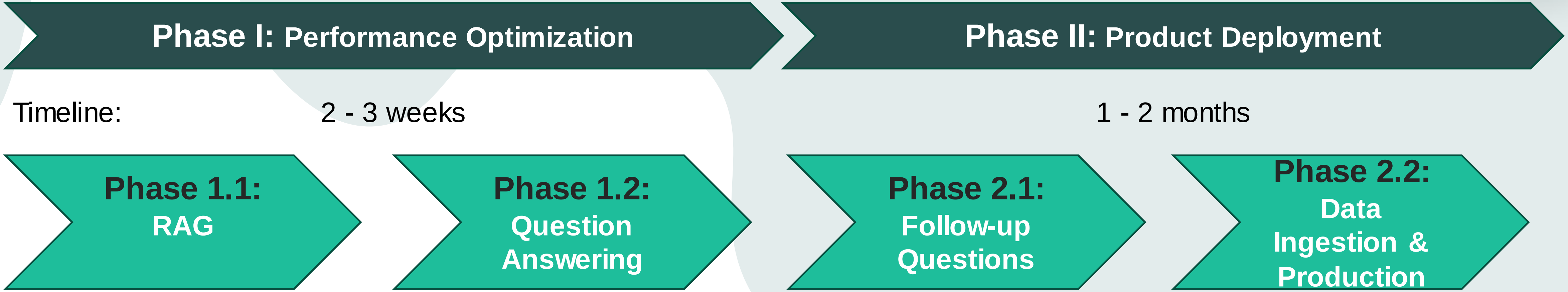
Solution

Our solution provides comprehensive market research to the users by utilizing retrieval augmented generation(RAG) over financial news articles

Methodology

Each component of RAG was improved and aligned by fine-tuning over financially related datasets

Next Steps



Phase I: Performance Optimization

Phase II: Product Deployment

Timeline: 2 - 3 weeks 1 - 2 months

Phase 1.1: **RAG**

Phase 1.2: **Question Answering**

Phase 2.1: **Follow-up Questions**

Phase 2.2: **Data Ingestion & Production**

- Further improve RAG performance by distorting the data

- Try more QA reading comprehension model and compare with the current models e.g. LLama 70B

- Provide insights and recommendations based on user query and settings
- Incorporate intelligent filtering of KPIs from Yahoo Finance

- Improve efficiency of data ingestion
- Look for initial user feedback for further iterations

Thank You

Appendix

Next Steps

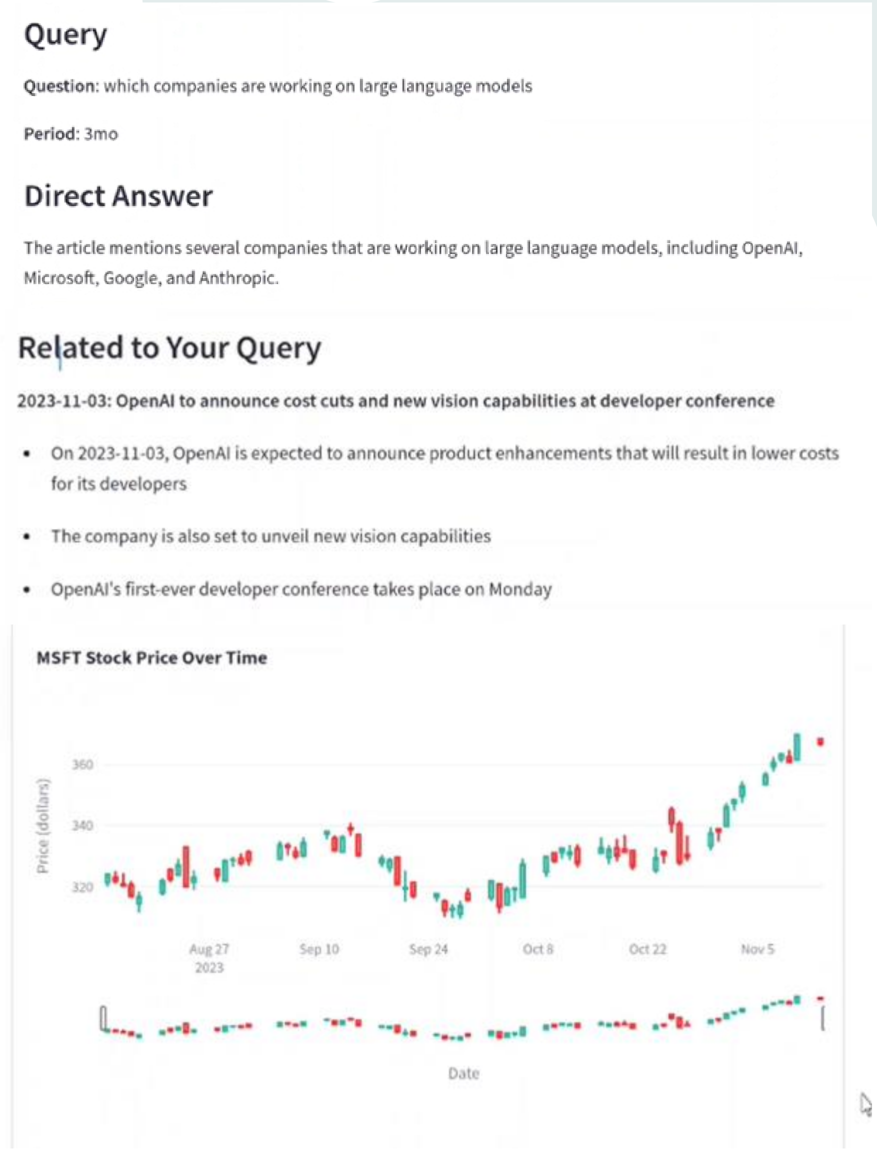
Performance Optimization – Distorting the RAG data,

1. We can continously monitor common questions by the users and add more data for those queries in the RAG dataset.
2. We can try more bigger QnA models

More features:

1. In addition to answering questions, we can add a feature of recommending investments ideas after having observed user's queries and using KPI from yahoo finance.
2. We can look into improving the data ingestion by using BIG data platforms for distributed computing

An AI driven personal investment assistant, FinBot, can empower retail investors



Query

Question: which companies are working on large language models

Period: 3mo

1

Direct Answer

The article mentions several companies that are working on large language models, including OpenAI, Microsoft, Google, and Anthropic.

2

Related to Your Query

2023-11-03: OpenAI to announce cost cuts and new vision capabilities at developer conference

- On 2023-11-03, OpenAI is expected to announce product enhancements that will result in lower costs for its developers
- The company is also set to unveil new vision capabilities
- OpenAI's first-ever developer conference takes place on Monday

3

MSFT Stock Price Over Time

1. Direct QA over News

- **What? :** Generate direct and concise answers to user queries on business and financial news
- **Why? :** Up-to-date information and saves time

2. Key-points Summaries

- **What? :** Extract relevant key-points over multiple topics
- **Why? :** Broader awareness outside user's expertise

3. Intelligent Stock Screener

- **What? :** Intelligently provide key metrics commonly used to identify opportunities
- **Why? :** Identify potential investment opportunities

Tuning Performance of Embedding Model

Data Details

- Data Sizes:
  - Train: 28K pairs of question and answers
  - Eval: 2.5K pairs of question and answers
  - Test: 3.4K pairs of question and answers
- Data Augmentation:
  - GPT-Turbo-3.5
  - "You are … that will re-phrase given text into a formal tone found in news articles or financial reports … never add any new information to the given text ..."
  - Removed unassociated questions and answers

Tuning Details

- Sentence transformer approach
  - Classify one positive pair per batch
  - Other pairs in batch negative
  - Soft-max with input based on vector distance
- Batch size of 32 with learning rate of 0.0001

Performance

Metric@3	**Base Model**	**Tuned Model**
Accuracy	0.81	0.83
Precision	0.58	0.62
Recall	0.44	0.46
NDCG	0.66	0.70
MRR	0.74	0.77
MAP	0.62	0.66

Tuning of the LLMs

Summarization Data Details

- Data Sizes:
  - Train: 77K context and key-points pair after augmentation
  - Test: 2K context and key-points pairs
- Other Augmentations
  - Filtered out articles with irrelevant key-points
    - Current stock trend or other information easily acquired from screener
  - Added noise-only context with randomly picked questions
  - Retained original as well as augmented data

QA Data Details

- Data Sizes:
  - Train: 20K context with Q/A after augmentation
  - Test: 2K context and key-points pairs
- Data Mixes
  - ~5K entries in TAT-QA after filtering
  - Added 1K entries from FINQA
  - Mixed in ~5K data points from WebGLM QA
- Augmentation
  - Total of ~10K entries of same Q/A but shuffled/randomized contexts

Infrastructure

- LORA based fine-tuning
  - Reduced rank of blocks to 16
  - Lead to < 1% parameters that needs to be tweaked
- Tuning with learning rate of 0.0001, and effective batch size of 8
  - Half-precision training and final weights
  - Tensor and pipeline parallelism with sharded weights via DeepSpeed
- Model served with PagedAttention and tensor parallelism via VLLM

An AI driven personal investment assistant, FinBot, can empower retail investors

Direct QA over News

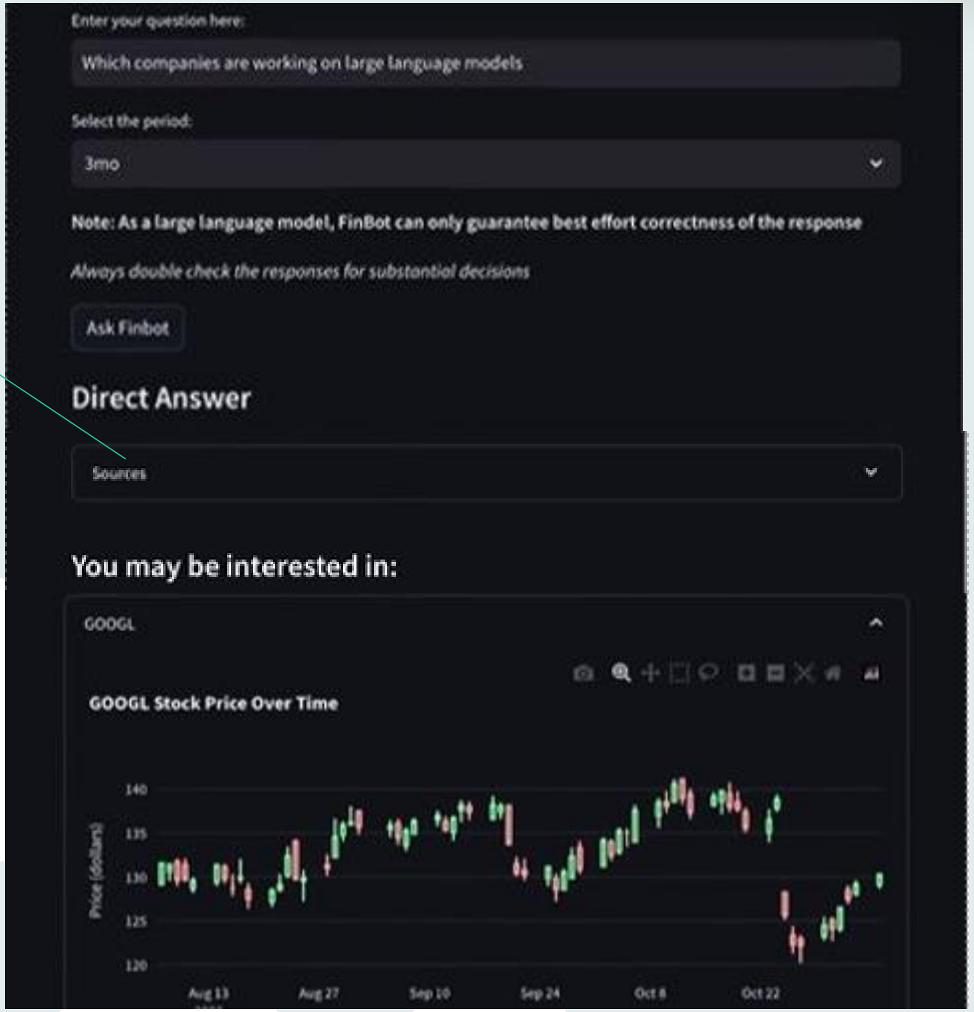
- Generate direct and concise answers to user queries on business and financial news
- Users can get up-to-date information without going through many articles
- Saves time so user can conduct more research on investment

Key-points Summaries

- Extract relevant key-points over multiple topics
- Users can become aware of big pictures outside of their expertise
- Reduces information asymmetry that may causes the user to miss opportunities

Intelligent Stock Screener

- Intelligently provide key metrics commonly used to identify opportunities based on query and output
- Users can easily identify potential investment opportunities without going out of FinBot
- Improves QoL when using the product



Enter your question here:

Which companies are working on large language models

Select the period:

3mo

Note: As a large language model, FinBot can only guarantee best effort correctness of the response

Always double check the responses for substantial decisions

Ask Finbot

Direct Answer

Sources

You may be interested in:

GOOGL

GOOGL Stock Price Over Time