

# GROUP 1

## SEMANTIC FORCE

FOR EBA 5004 PRACTICAL LANGUAGE PROCESSING (PLP) PROJECT MODULE

**YATHARTH MAHESH SANT** | A0286001R

**KRISTOFER ROOS** | A0285949A

**GOH MIN HUA** | A0285810A

**TAN LI MING** | A0027883W

**CHUA KIAN YONG KENNY** | A0056377W

# PROJECT CONTEXT AND SPONSOR



## Sponsor Introduction

- Collaboration with a top strategic consulting firm
- Focus on tackling large-scale data challenges



## SemanticForce Purpose

- Enables efficient data summarization
- Empowers quick insight extraction from complex reports



## Firm's Involvement

- Crucial input on tool's design and usability
- Guided feature refinement for practical consulting needs



## Deployment

- To be deployed for immediate use in real scenarios
- Integration session scheduled post-exams for operational setup

# BUSINESS PROBLEM



## Multiple Data Sources

- Managing diverse data sources leads to complexity.
- Difficulty in extracting actionable insights from overwhelming information.



## Time Constraints

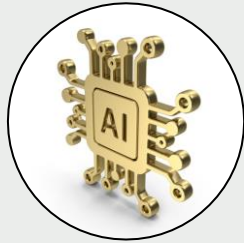
- Delays in decision-making due to time-consuming analysis.
- Inability to respond promptly to rapidly changing market conditions.



## Manual Processes

- Manual data tasks are prone to errors and inconsistencies.
- Lack of scalability and productivity in manual processes.

# SYSTEM FEATURES



## SoTN models for different queries

- Llama2 finetuned to recognize user intent
- T5 models finetuned for financial reports and news summarization
- Hybrid fusion RAG answer accurately to query with little to no hallucination



## Intuitive User Interface

- Clean and intuitive interface for effortless usage and navigation
- Interactive conversational AI facilitating natural exchanges to retrieve key information



## Prompt Insights Extraction

- Effortlessly retrieve essential financial data without the need to manually sift through reports.
- Gain dynamic insights from the latest news with our sentiment analysis.

# SYSTEM ARCHITECTURE OVERVIEW

## User-friendly interface

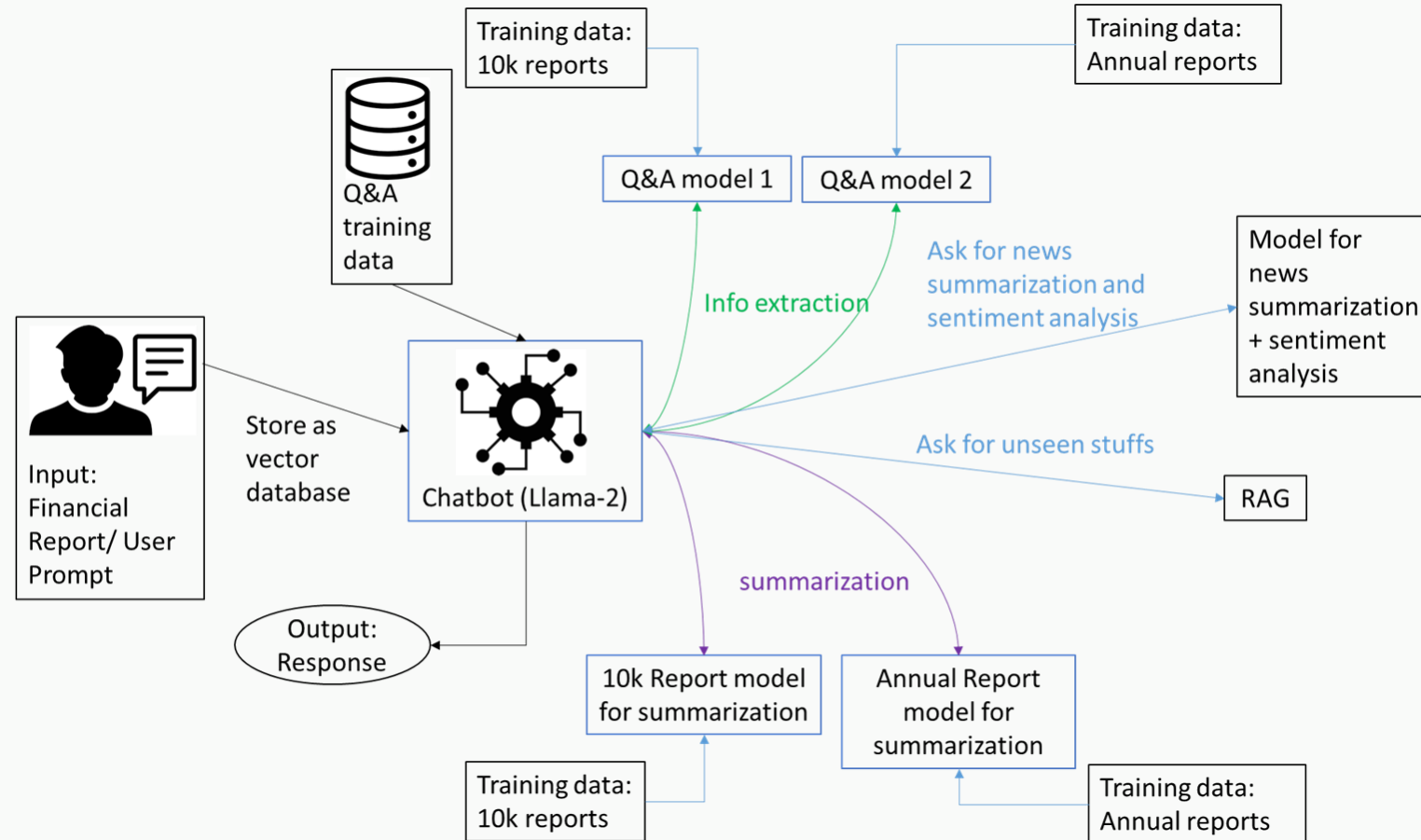
- Webpage is clear such that instructions are given, functions of webpage like uploading files and saving chat history and interacting with chatbot are very visible

## Robust Chatbot

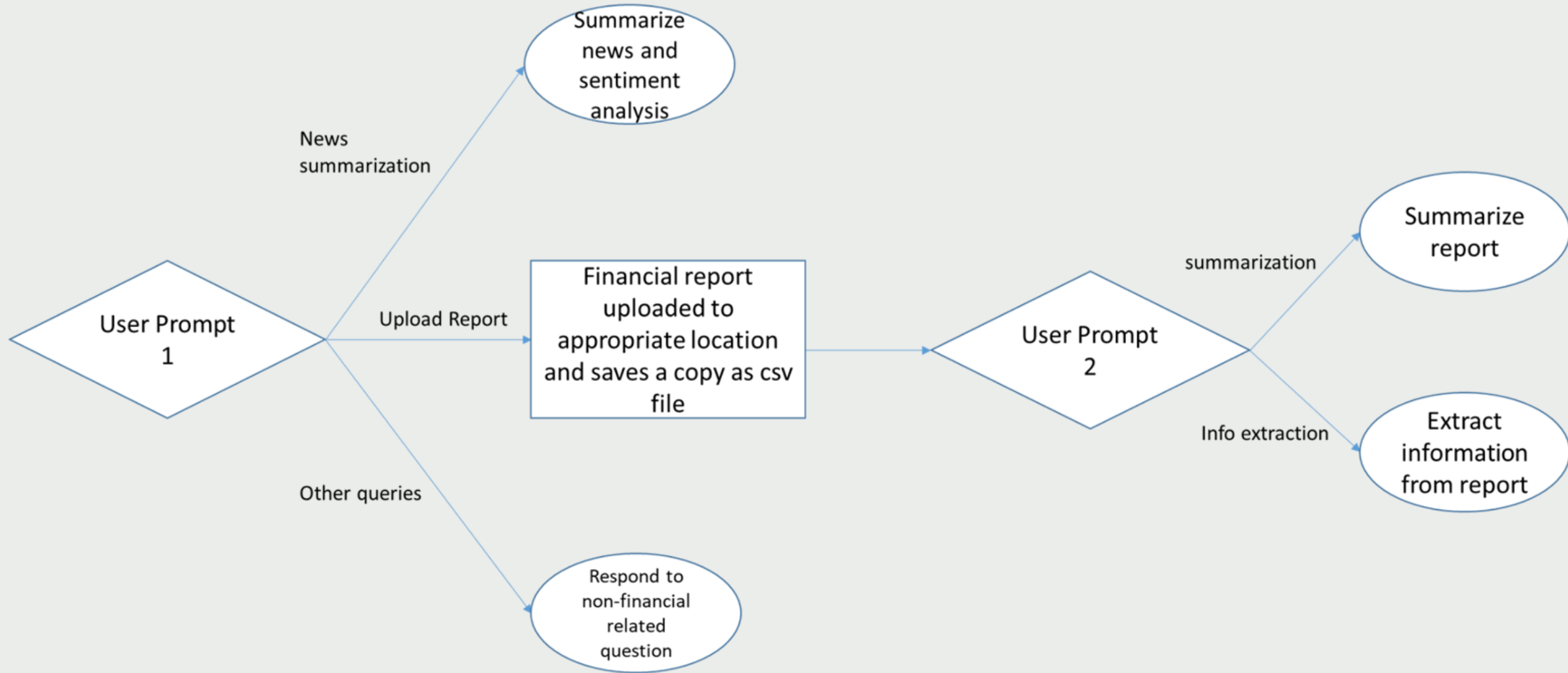
- Can identify user intent and map to different finetuned LLMs with high accuracy
- Will always send to RAG for any questions that fall 'out of the circle'

## Smart RAG

- Documentation consists of Wikipedia pages, which have a plethora of data



# USER WORKFLOW



# TECHNICAL STACK



## Frontend Dev

- Web Application: Streamlit
- Interaction: Frontend communicates with Backend via functions imported from Python files



## Backend Dev

- Web Application: Python, Huggingface, Langchain
- Llama2-7b: Processes user inputs and manages chatbot interactions
- LLMs and RAG: Process query with accurate response. Scrapes for news and generates sentiment analysis. Summarize financial report and news.



## Databases

- Store model's weights as binary file
- Store financial reports as vector embeddings
- Store news as csv files



## Security

- API Security: Tokens from Gemini

# DATASETS



## 10K Reports

- 132 reports from 44 companies between 2021-2023
- PDF to Text Conversion via pdfplumber
- Segmented into distinct items for effective text length management
- Gemini used to establish target summary standard



## News

- News articles related to 47 companies were extracted over four time periods from Jan-Apr 2024.
- Extracted news articles form training data for subsequent model training for text summarization.



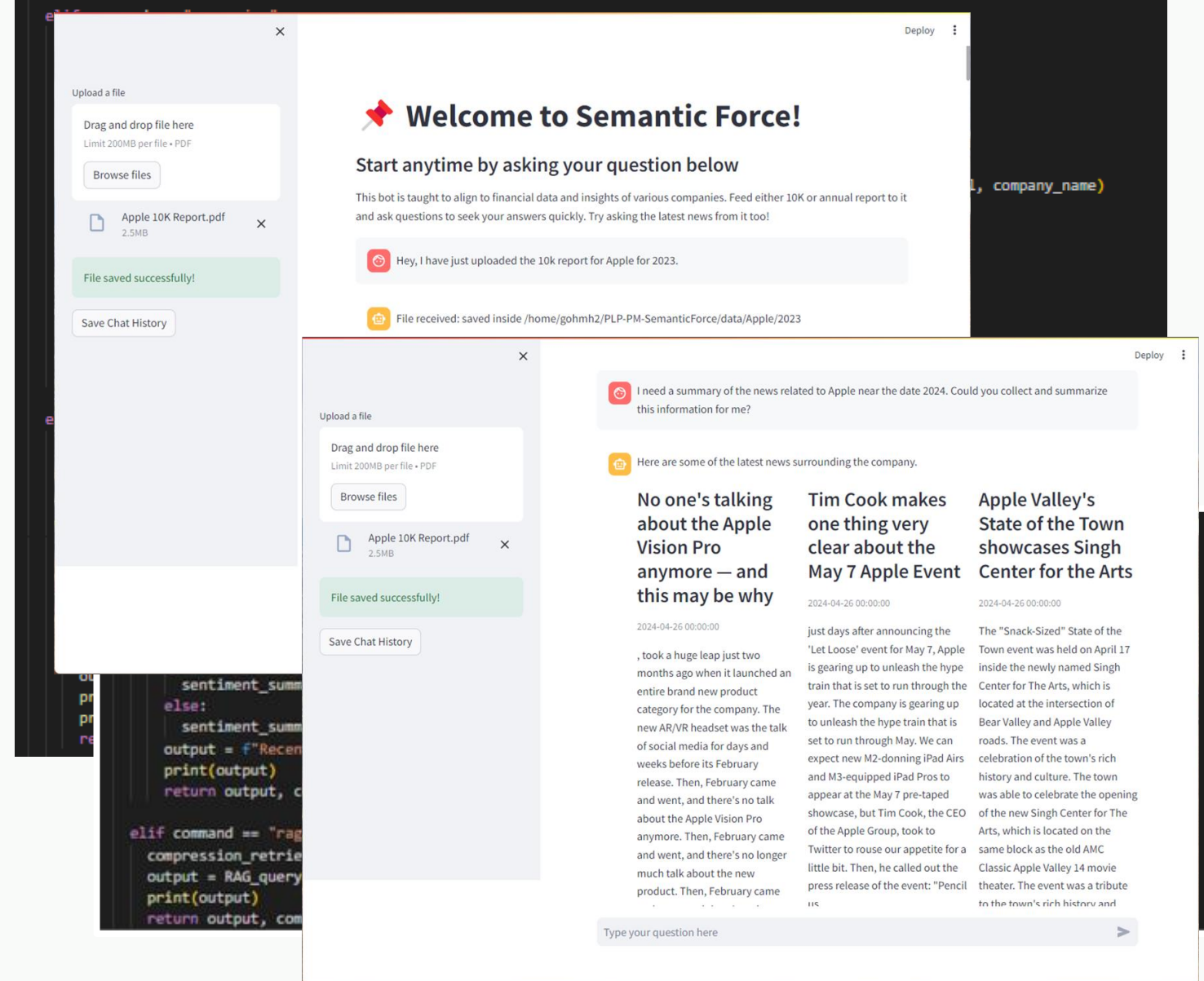
## Annual Report

- Sourced from over 100 annual reports, totaling 160,000+ words
- Manual extraction of 1000+ article-observation pairs
- Automated data extraction enhanced by RPA and custom macros



# LLM CHATBOT

- Llama2-7b finetuned to detect user intent, type of report, the year date and type of financial metric from user query
- Training dataset generated with the help of ChatGPT
- Maps user intent to different appropriate models
  - T5 Base Models for Financial Report Summarization, news summarization and sentiment analysis
  - RAG to deal with user queries that can't be mapped



# 10K REPORTS (T5)

## T5 Model

1. Training Time Benefits: Longer training improves all performance metrics, showing the value of extended learning sessions.
2. Summary Length Influence: Shorter summaries enhance relevance and coherence, as evidenced by higher Rouge scores.
3. Model Size Effectiveness: The larger T5 model consistently outperforms the base model, especially in capturing the essence of longer texts.
4. Stable Learning Rate: A consistent learning rate ensures fair comparison and reliable interpretation of model improvements over different setups.

| Model    | Layers                  | Hidden | Heads | Parameters |
|----------|-------------------------|--------|-------|------------|
| T5-base  | 12 encoder + 12 decoder | 768    | 12    | 220M       |
| T5-large | 24 encoder + 24 decoder | 1024   | 16    | 770M       |

| Model Name | Epochs | MAXLE N | SUMMA RYLEN | Bleu   | Rouge1 | Rouge2 | RougeL |
|------------|--------|---------|-------------|--------|--------|--------|--------|
| T5-base    | 5      | 512     | 150         | 0.2843 | 0.5167 | 0.3223 | 0.4174 |
| T5-base    | 10     | 512     | 150         | 0.3018 | 0.5703 | 0.3744 | 0.4657 |
| T5-base    | 5      | 256     | 256         | 0.2312 | 0.5413 | 0.3503 | 0.4392 |
| T5-base    | 10     | 256     | 256         | 0.2382 | 0.5925 | 0.3956 | 0.4888 |
| T5-large   | 4      | 512     | 256         | 0.2843 | 0.5167 | 0.3223 | 0.4174 |
| T5-large   | 10     | 512     | 256         | 0.2616 | 0.6209 | 0.4179 | 0.52   |

# 10K REPORTS (OTHER MODELS)

## Performance Comparison with Other Models:

- GPT2-Summarizer, GPT2-small, GPT2-medium, and BART models explored for summarization.
- GPT2 models showed inferior performance compared to T5, exhibiting coherence issues.
- BART model demonstrated significant improvement in summary quality over GPT2 models. But still not as good as T5

| Model       | Layers                    | Hidden | Heads | Parameters |
|-------------|---------------------------|--------|-------|------------|
| Gpt2-small  | 12                        | 768    | 12    | 117M       |
| Gpt2-medium | 24                        | 1024   | 16    | 345M       |
| BART-Base   | 12 encoder+<br>12 decoder | 768    | 12    | 139M       |

| Model Name      | Epochs | Learning Rate | Rouge 1 | Rouge 2  | Rouge L | Rouge L Sum |
|-----------------|--------|---------------|---------|----------|---------|-------------|
| GPT2_Small      | 1      | 1.00E-04      | 0.06496 | 0.004792 | 0.05722 | 0.06613     |
| GPT2_Small      | 4      | 1.00E-04      | 0.09658 | 0.009822 | 0.08103 | 0.09385     |
| GPT2_Small      | 10     | 1.00E-04      | 0.1071  | 0.01403  | 0.0921  | 0.5042      |
| GPT2_Medium     | 4      | 1.00E-04      | 0.05618 | 0.003363 | 0.05012 | 0.4102      |
| GPT2_Medium     | 10     | 1.00E-04      | 0.1812  | 0.03258  | 0.1397  | 0.5284      |
| GPT2_Summarizer | 0      | NA            | 0.3157  | 0.1315   | 0.2062  | 0.2306      |
| BART            | 10     | 1.00E-04      | 0.2852  | 0.2195   | 0.2641  | 0.2667      |

## Summarization:

- T5-base and BART models explored for summarization.
- Comparing Rouge-L scores, T5-base (8 epochs) seem to demonstrate the best performance among this limited test models.

## Info Extraction:

- Hugging Face all-MiniLM-L6-v2 model for contextual embeddings
- FAISS library for fast and accurate information extraction
- Question-answering framework using the Langchain library

| Model     | Layers                  | Hidden | Heads | Parameters |
|-----------|-------------------------|--------|-------|------------|
| T5-base   | 12                      | 768    | 12    | 220M       |
| BART-Base | 12 encoder + 12 decoder | 768    | 12    | 139M       |

| Model Name | Epochs | Learning Rate | Max Length | Summary Length | Bleu   | Rouge 1 | Rouge 2 | Rouge L | Rouge L Sum |
|------------|--------|---------------|------------|----------------|--------|---------|---------|---------|-------------|
| T5-base    | 8      | 1.00E-04      | 512        | 150            | 0.1204 | 0.4213  | 0.1686  | 0.2625  | 0.2620      |
| T5-base    | 5      | 1.00E-04      | 512        | 150            | 0.1179 | 0.4237  | 0.1664  | 0.2597  | 0.2591      |
| T5-base    | 2      | 1.00E-04      | 512        | 150            | 0.1119 | 0.4173  | 0.1659  | 0.2640  | 0.2639      |
| BART       | 10     | 1.00E-04      | 512        | 256            | 0.0900 | 0.4098  | 0.1659  | 0.2532  | 0.2525      |
| BART       | 5      | 1.00E-04      | 512        | 256            | 0.13   | 0.4261  | 0.1627  | 0.2585  | 0.2579      |

**T5 Model Utilization:** Selected for its flexibility in handling complex texts and pre-training on extensive corpora, essential for understanding nuanced financial narratives

**Optimal Configuration:** Discovered the best performing parameter configuration at a learning rate of 1e-4, with 256 max and summary lengths, and 10 epochs to ensure detailed and efficient summarization.

**Deployment and Accessibility:** Deployed on Hugging Face, supplemented with a custom PDF parser and summary combiner

| Model   | Layers | Hidden | Heads | Parameters |
|---------|--------|--------|-------|------------|
| T5-base | 12     | 768    | 12    | 220M       |

| Final Parameter | Value |
|-----------------|-------|
| Learning Rate   | 1e-4  |
| Max Length      | 256   |
| Summary Length  | 256   |
| Epochs          | 10    |
| Seed            | Fixed |

| Metric      | Value | Description                                       |
|-------------|-------|---|
| BLEU Score  | 0.23  | Indicates good linguistic accuracy.               |
| ROUGE-1     | 0.55  | Reflects high overlap of unigrams with reference. |
| ROUGE-2     | 0.30  | Shows overlap of bigrams with reference.          |
| ROUGE-L     | 0.43  | Captures longest common subsequence score.        |
| ROUGE-L Sum | 0.43  | Summarization specific ROUGE-L.                   |

# Retrieval Augmented Generation (RAG)

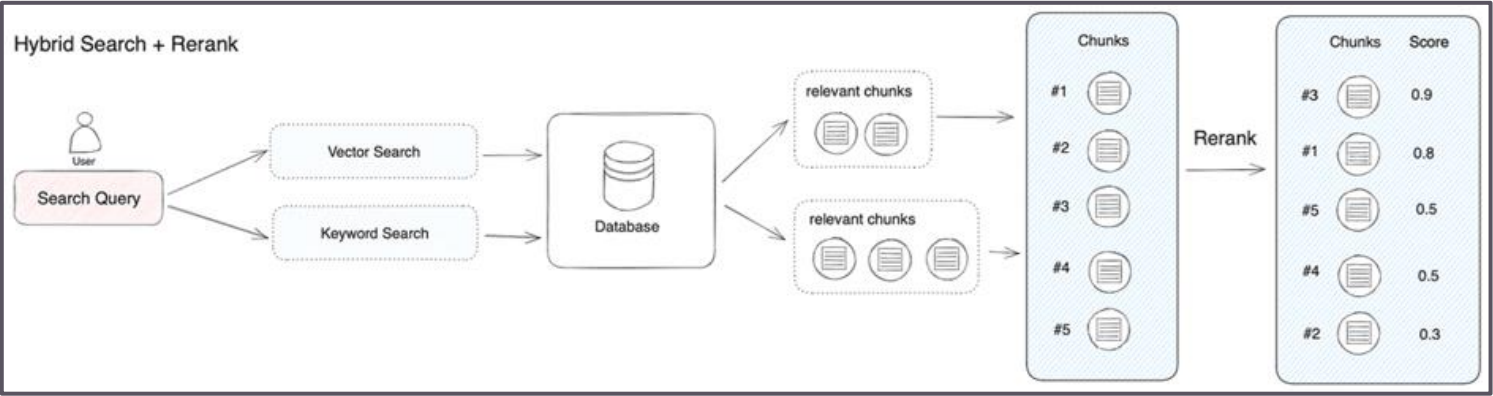
**Robust against Hallucinations yet**

**Flexible:** Documentation comprises of wikipedia pages to entertain various questions and up-to-date information minimizes chances of wrong answer.

**Hybrid Search:** Semantic and keyword search combined query for more accurate candidates.

**Reranking:** Uses MS Marco BERT cross-encoder perform attention across query and document to give score for relevancy.

**Evaluation of RAG using RAGAs:** Gold standard ground truth generated by Gemini. Overall accuracy attained is **0.89**.

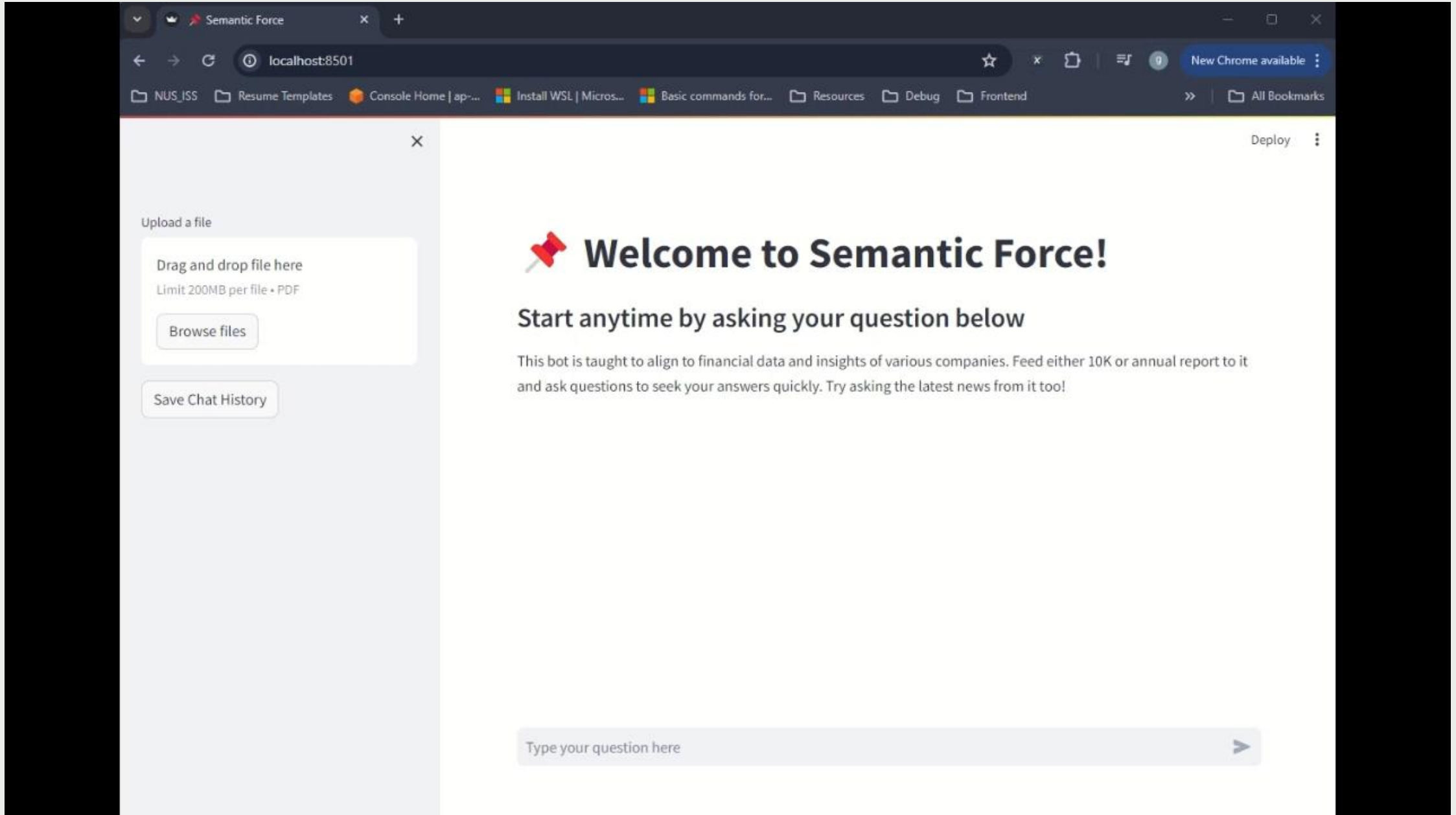


## Example of 1 output from RAGAs

| Question  | Context  | Answer  | Ground Truth   | Context Precision | Context Recall | Faithfulness | Answer Relevancy |
|---|--|---|--|-------------------|----------------|--------------|------------------|
| Why has Apple been so successful in the tech industry ? | "Apple's focus on design and user experience, its tightly integrated ecosystem of hardware and software, and its powerful brand image have all contributed to its success. The company has also made strategic acquisitions, such as Beats Electronics, that have expanded its offerings." | Apple's success stems from its innovative products, strong brand, strategic acquisitions, and emphasis on design and user experience. | Apple has been successful due to a combination of factors, including innovative products, strong marketing and brand recognition and strategic acquisitions and its focus on design and user experience. | 0.87              | 1.00           | 0.78         | 0.93             |



# PRODUCT DEMO



The screenshot shows a web browser window with the address bar at `localhost:8501`. The browser's bookmark bar contains several links: `NUS_ISS`, `Resume Templates`, `Console Home | ap-...`, `Install WSL | Micros...`, `Basic commands for...`, `Resources`, `Debug`, and `Frontend`. A notification for "New Chrome available" is visible in the top right corner of the browser.

The application interface is divided into two main sections. On the left is a sidebar with a close button (`X`) and a "Deploy" button with a dropdown menu. The sidebar contains an "Upload a file" section with a text prompt "Drag and drop file here" and a subtext "Limit 200MB per file • PDF". Below this is a "Browse files" button. At the bottom of the sidebar is a "Save Chat History" button. The main content area features a large heading "Welcome to Semantic Force!" accompanied by a red pushpin icon. Below the heading is a subheading "Start anytime by asking your question below" and a paragraph of text: "This bot is taught to align to financial data and insights of various companies. Feed either 10K or annual report to it and ask questions to seek your answers quickly. Try asking the latest news from it too!". At the bottom of the main area is a text input field with the placeholder "Type your question here" and a right-pointing arrow button.

# CHATGPT OR OTHER LLM USAGE



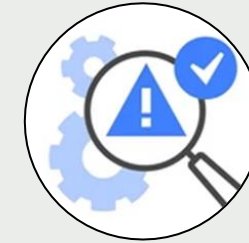
## Project Outcomes using LLMs

- Summarization with T5 and BART
- User Intent Recognition with Llama2
- Gold Standard Summaries with Gemini



## For Training and Deployment

- ChatGPT augmented Llama2's training dataset
- Gemini established vector database for RAG

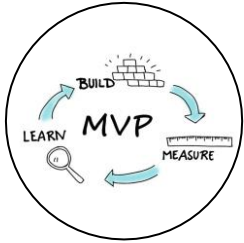


## For Troubleshooting

- Technical/Coding troubleshooting with ChatGPT. Challenge though in addressing issues with new models like Gemini (sought forums support instead)



# CONCLUSION



## MVP Success

- Efficiently extract key business metrics
- Summarise complex financial reports and news
- Analysing public perception



## Alignment with Sponsor Goals

- Enhances analytical processes
- Enable actionable insights efficiently



## Future Enhancements

- Potential to integrate OCR
- Personalised dashboards
- Expanded data sources



# Thank you!

GROUP 1

SEMANTIC FORCE

**FOR EBA 5004 PRACTICAL LANGUAGE PROCESSING (PLP)**

**YATHARTH MAHESH SANT** | A0286001R

**KRISTOFER ROOS** | A0285949A

**GOH MIN HUA** | A0285810A

**TAN LI MING** | A0027883W

**CHUA KIAN YONG KENNY** | A0056377W