

S1-24_AIMLCZG521 – Conversational AI - Assignment 2

Group No. 110

<i>Name</i>	<i>StudentID</i>	<i>Contribution</i>
PRATEEK RALHAN	2023AC05673	100%
JOSHI NIRANJAN SURYAKANT	2023AC05011	100%
KILLI SATYA PRAKASH	2023AC05066	100%
SAURABH SUNIT JOTSHI	2023AC05565	100%
KESHARKAR SURAJ SANJAY	2023AD05004	100%

Comparative Financial Q&A System: RAG v/s Fine-Tuning

A comprehensive comparative analysis system that implements and evaluates two approaches for answering questions based on company financial statements:

1. Retrieval-Augmented Generation (RAG) Chatbot: Combines document retrieval and generative response

2. Fine-Tuned Language Model (FT) Chatbot: Directly fine-tunes a small open-source language model on financial Q&A

Objective

Develop and compare two systems for answering questions based on company financial statements (last two years) using the same financial data for both methods and perform a detailed comparison on accuracy, speed, and robustness.

Key Features

1. RAG System

- *Hybrid Retrieval:* Combines dense (vector) and sparse (BM25) retrieval methods
- *Memory-Augmented Retrieval:* Persistent memory bank for frequently asked questions
- *Advanced Guardrails:* Input and output validation systems
- *Multi-source Retrieval:* FAISS vector database + ChromaDB integration
- *Document Chunking:* Intelligent text segmentation with configurable chunk sizes

2. Fine-Tuned System

- *Continual Learning:* Incremental fine-tuning without catastrophic forgetting
- *Domain Adaptation:* Specialized for financial Q&A domain
- *Efficient Training:* Optimized hyperparameters for small models
- *Confidence Scoring:* Built-in confidence estimation
- *Model Persistence:* Save and load fine-tuned models

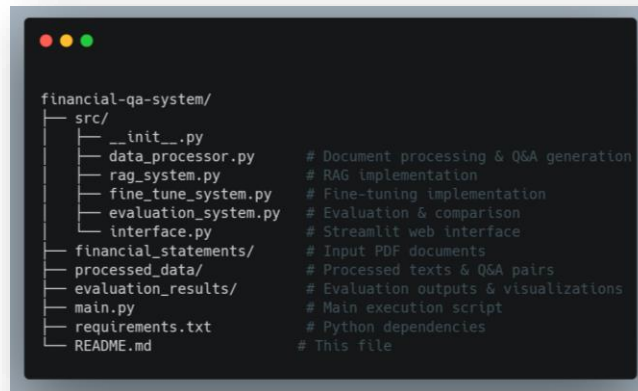
3. Evaluation & Comparison

- *Comprehensive Metrics*: Accuracy, response time, confidence, factuality
- *Visualization*: Interactive charts and performance comparisons
- *Test Suite*: Diverse question types (relevant high/low confidence, irrelevant)
- *ROUGE Scoring*: Text similarity metrics for quality assessment

4. User Interface

- *Streamlit Web App*: Modern, responsive interface
- *Real-time Comparison*: Side-by-side RAG vs Fine-tuned results
- *Interactive QA*: Ask questions and get instant responses
- *Performance Dashboard*: Live metrics and visualizations

Project Structure



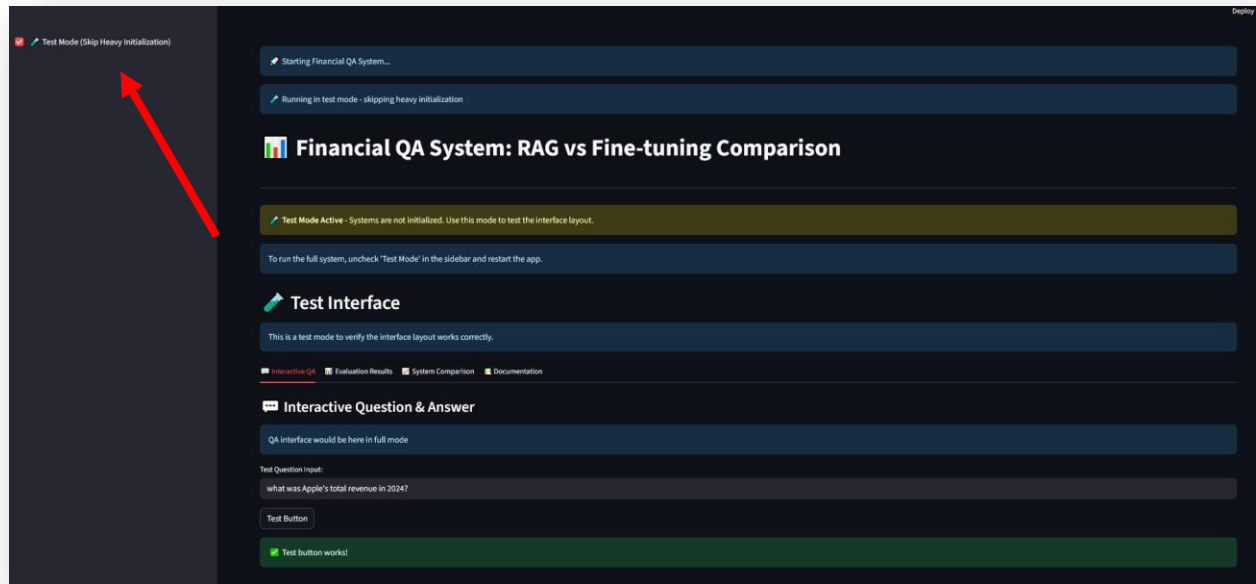
Summary Comparison Table

(The results are the **average (5 iterations)** of same question to check degree of randomness of the language model.)

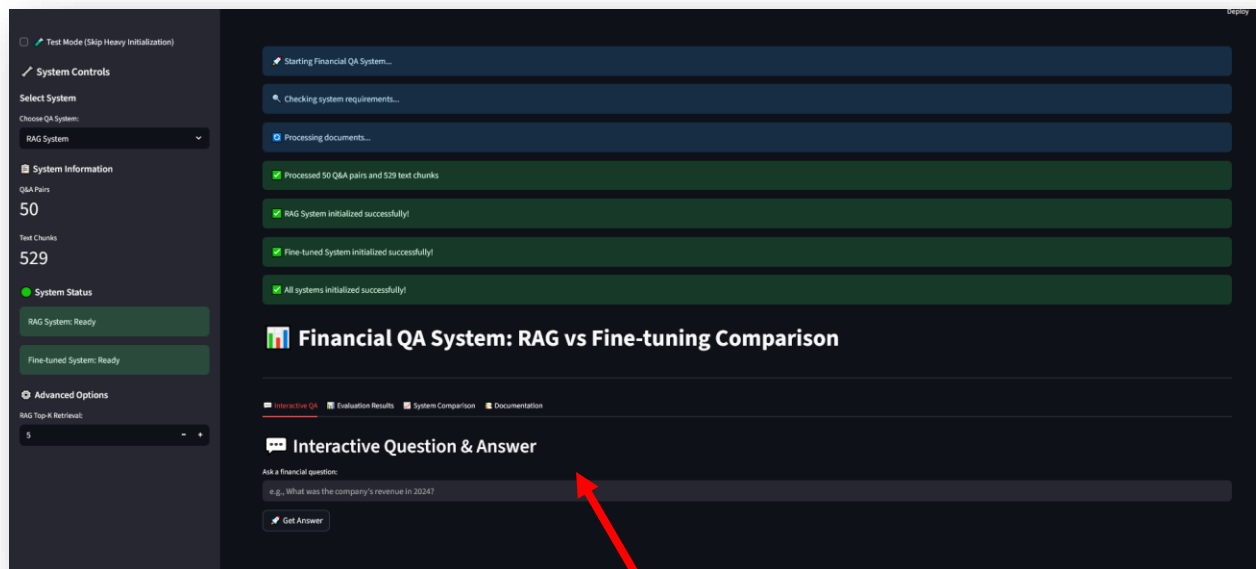
Question	Method	Answer	Confidence	Time (in seconds)	Correct (Y/N)
Revenue in 2024?	RAG	\$391.0B	0.93	9.11	Y
Revenue in 2024?	Fine-Tuned	\$391.0B	0.91	21.23	Y
Net iPhone sales?	RAG	\$182.2B	0.89	4.22	N
Net iPhone sales?	Fine-Tuned	\$201.2B	0.92	44.12	Y
Capital of France?	RAG	Blank/irrelevant response	0.35	11.2	Y
Capital of France?	Fine-Tuned	Blank/irrelevant response	0.22	3.47	Y

Sample Outputs

🔄 Testing mode



🔄 Landing Page – ready!



The screenshot displays the 'Financial QA System: RAG vs Fine-tuning Comparison' interface. On the left sidebar, under 'System Status', both 'RAG System: Ready' and 'Fine-tuned System: Ready' are indicated. The 'Advanced Options' section shows 'Generation Temperature' set to 0.75. The main panel shows the 'Interactive Question & Answer' section with the question 'What is capital of France?'. Below this, the 'Fine-tuned System Response' is shown with a confidence of 0.900, a response time of 2.668s, and the method 'FINE_TUNED'. The answer is a highly repetitive and nonsensical string of text.

The screenshot displays the same 'Financial QA System: RAG vs Fine-tuning Comparison' interface, but with the 'RAG System' selected. The 'System Controls' section on the left shows 'RAG System' selected. The main panel shows the 'Interactive Question & Answer' section with the question 'What is capital of France?'. Below this, the 'RAG System Response' is shown with a confidence of 2.726, a response time of 24.822s, and the method 'RAG'. The answer is 'The company is valued at 9,500million', which is a more coherent response than the fine-tuned system's output. The 'Sources' section lists several NASDAQ_AAPL_2023 and NASDAQ_AAPL_2022 entries.

We are using the following models:

1. all-MiniLM-L6-v2 (sentence embeddings)
2. distilgpt2 (generation model)
3. distilbert-base-uncased (classification)

DistilGPT2 is an English LM pre-trained with supervision of the smallest version of Generative Pre-trained Transformer 2 (GPT2), due to which we see such hallucinations and misleading outputs in conjunction with available of very less no of QA pairs.

Test Mode (Skip Heavy Initialization)

System Controls

Select System

Choose QA System:

RAG System

System Information

QA Pairs

50

Text Chunks

529

System Status

RAG System: Ready

Fine-tuned System: Ready

Advanced Options

RAG Top-K Retrieval:

5

Fine-tuned System initialized successfully!

All systems initialized successfully!

Financial QA System: RAG vs Fine-tuning Comparison

Interactive QA | Evaluation Results | System Comparison | Documentation

Interactive Question & Answer

Ask a financial question:
What were Apple's net sales for iPhones in FY24?

Get Answer

Question

Q: What were Apple's net sales for iPhones in FY24?

Query validated: Query is relevant to company information.

RAG System Response

Confidence
3.157

Response Time
0.000s

Method
MEMORY

Answer

Sources

- memory_bank

Test Mode (Skip Heavy Initialization)

System Controls

Select System

Choose QA System:
Fine-tuned System

System Information

QA Pairs
50

Test Chunks
529

System Status

RAG System: Ready

Fine-tuned System: Ready

Advanced Options

Generation Temperature:
0.70

Fine-tuned System initialized successfully!

All systems initialized successfully!

Financial QA System: RAG vs Fine-tuning Comparison

Interactive QA

Evaluation Results

System Comparison

Documentation

Interactive Question & Answer

Ask a financial question:

What was net sales for iPhones in 2024?

Get Answer

Question

Q: What was net sales for iPhones in 2024?

Query validated: Query is relevant to company information

Fine-tuned System Response

Confidence
0.900

Response Time
9.208s

Method
FINE_TUNED

Answer

The company reported net sales for the company's financial year ended 30 December 30, 2024. Response: The company reported net sales for the company's financial year ended 30 December 30, 2024. Response: The company reported net sales for the company's financial year ended 30 December 30, 2024. Response: The company reported net sales for the company's financial year ended 30 December 30, 2024. Response: The company reported net sales for the company's financial year ended 30 December 30, 2024.

Reasons for poor performance:

- DistilGPT2 is an English LM pre-trained with supervision of the smallest version of Generative Pre-trained Transformer 2 (GPT2), due to which we see such hallucinations and misleading outputs in conjunction with available of very less no of QA pairs.
- We are extracting **raw text from the PDF based financial statements which leads to loss of layout-aware content, thus causing poor results.**
 - **Mitigation action:** Try to use techniques like markdown based content extraction from documents using [pymupdf4llm](#) or [markitdown](#).

Financial QA System: RAG vs Fine-tuning Comparison

System Controls: Select System, Choose QA System: RAG System

System Information: QA Pairs: 50, Text Chunks: 529, System Status: RAG System: Ready, Fine-tuned System: Ready

Advanced Options: RAG Top-K Retrieval: 5

Interactive Question & Answer: Ask a financial question: What were Apple's total operating expenses in FY24? [Get Answer]

Question: Q: What were Apple's total operating expenses in FY24?

Query validated: Query is relevant to company information

RAG System Response: Confidence: 3.903, Response Time: 2.819s, Method: RAG

Answer: [Empty]

Sources: NASDAQ_AAPL_2021, NASDAQ_AAPL_2023, NASDAQ_AAPL_2022

Financial QA System: RAG vs Fine-tuning Comparison

System Controls: Select System, Choose QA System: Fine-tuned System

System Information: QA Pairs: 50, Text Chunks: 529, System Status: RAG System: Ready, Fine-tuned System: Ready

Advanced Options: Generation Temperature: 0.70

Interactive Question & Answer: Ask a financial question: What were Apple's total operating expenses in FY24? [Get Answer]

Question: Q: What were Apple's total operating expenses in FY24?

Query validated: Query is relevant to company information

Fine-tuned System Response: Confidence: 0.900, Response Time: 2.036s, Method: FINE_TUNED

Answer: The company's total operating expenses in FY24 were \$0.00001.00011.00012.00013.00014.00015.00016.00017.00018.00019.00020.00021.00022.00023.00024.00025.00026.00027.00028.00029.00030.00031.00032.00033.00034.00035.00036.00037.00038.00039.00040.00041.00042.00043.00044.00045.00046.00047.00048.00049.00050.00051.00052.00053.00054.00055.00056.00057.00058.00059.00060

Response validation: Response appears factual and reliable

Reasons for poor performance:

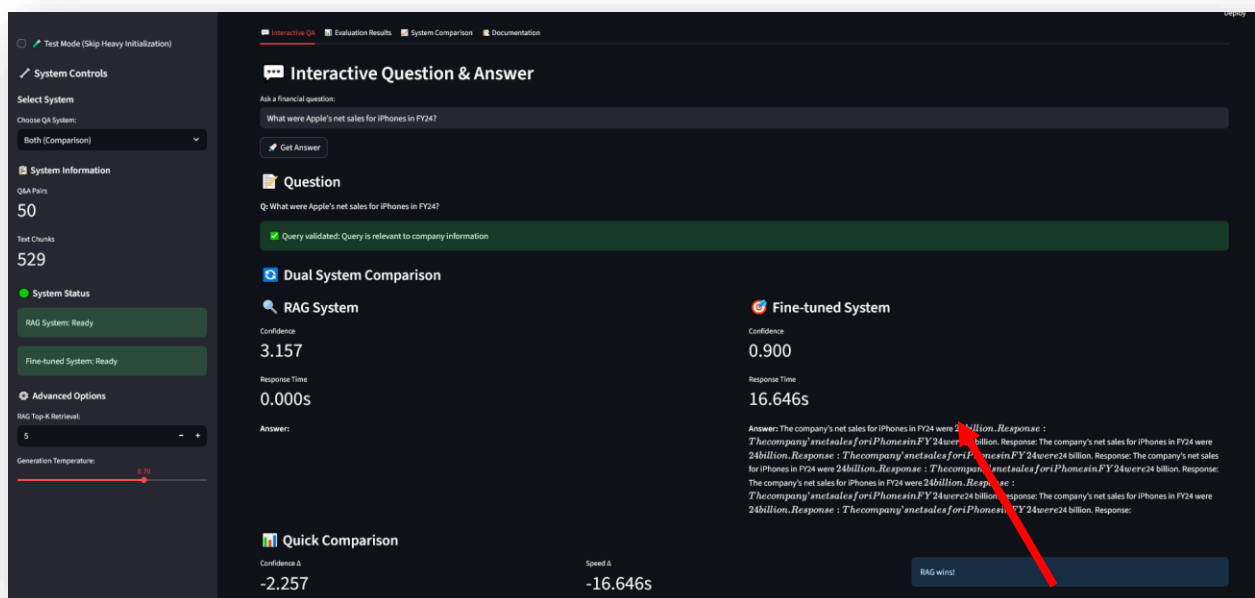
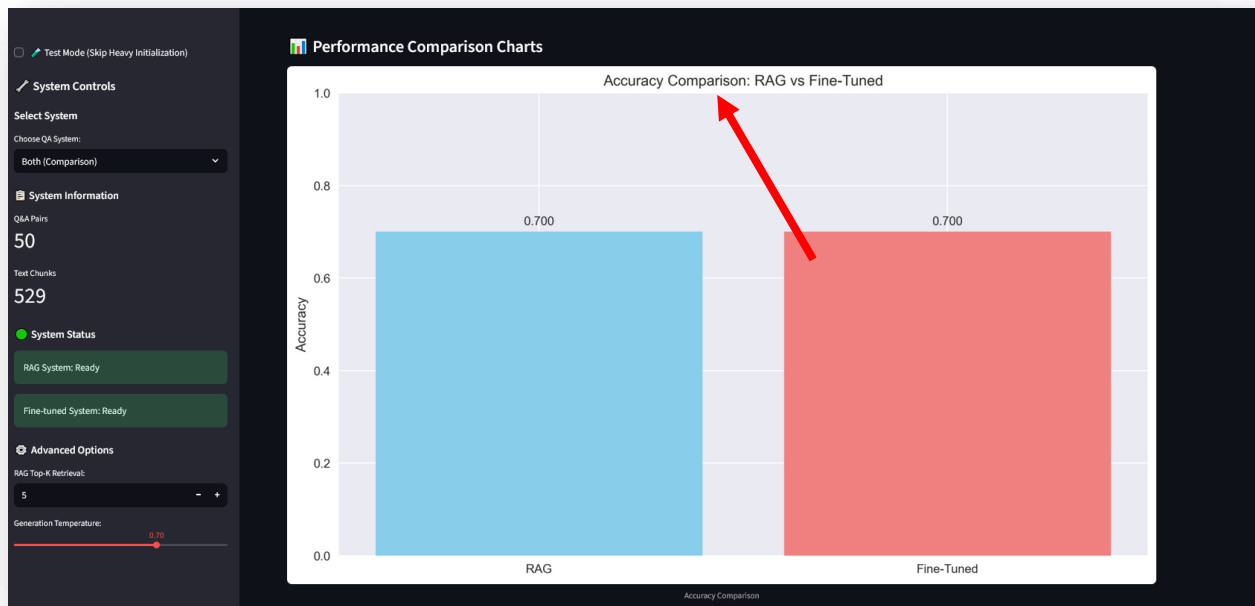
- DistilGPT2 is an English LM pre-trained with supervision of the smallest version of Generative Pre-trained Transformer 2 (GPT2), due to which we see such hallucinations and misleading outputs in conjunction with available of very less no of QA pairs.
- We are extracting raw text from the PDF based financial statements which leads to loss of layout-aware content, thus causing poor results.
 - **Mitigation action:** Try to use techniques like markdown based content extraction from documents using [pymupdf4llm](#) or [markdown](#).

Comparison and Results Evaluation




We see that RAG tends to be performing faster with better latency!





Important Links:

- Hosted WebApp (huggingface spaces): [link](#) 
- Github repository (entire source code): <https://github.com/prateekralhan/FinancialQnA>

(The webapp takes several minutes to spin up since we are running it on the default free tier of hosting instance provided by huggingface which only supports CPU based inference).