**Financial Document Analysis with Retrieval-Augmented Generation (RAG) and Finetuning**

Individual Contributions: -

|  |  |  |
| --- | --- | --- |
| **Name** | **BITS ID** | **Contribution** |
| SUBHRANSU MISHRA | 2023AC05489 | 100% |
| DULAL DAS | 2023AC05041 | 100% |
| LAKSHMISRINIVAS PERAKAM | 2023AC05540 | 100% |
| JAWAHARLAL RANJAN | 2023AC05504 | 100% |
| SHAILESH KUMAR SINGH | 2023AC05475 | 100% |

# **Introduction**

This project implements a **Comparative Financial Q&A System** using **Retrieval-Augmented Generation (RAG)** and **Fine-Tuned (FT)** language models to answer questions from company financial statements. The dataset is derived from **Kyndryl Holdings Inc. annual reports for FY 2023 and FY 2024**, converted from PDF to structured text and tables. The pipeline includes **data preprocessing, hybrid retrieval (dense + sparse), cross-encoder re-ranking, and generative response synthesis**. The fine-tuned model leverages the same curated Q&A pairs for domain adaptation. The system is deployed with an interactive **Gradio UI**, enabling seamless comparison of RAG and FT methods on various financial queries.

# **Design & Methodology**

# **Dataset Preparation**

# Extracted data from **Kyndryl Holdings Inc. Annual Reports (FY 2023 & FY 2024)**.

# Converted PDFs to structured tables and cleaned text for Q&A pair generation.

# **Q&A Dataset Creation**

# Generated **curated question-answer pairs** for both Income Statement and Balance Sheet data.

# Created a balanced dataset to ensure coverage of key financial metrics.

# **RAG System**

# **Dense Retrieval:** all-MiniLM-L6-v2 embeddings with **FAISS** vector store.

# **Sparse Retrieval:** Keyword-based search using **BM25**.

# **Hybrid Search:** Combined dense and sparse retrieval for better recall and precision.

# **Re-ranking:** Cross-encoder (ms-marco-MiniLM-L-6-v2) to improve relevance scoring.

# **Response Generation:** Flan-T5-base model for context-grounded answers.

# **Fine-Tuning System**

# Fine-tuned Flan-T5-base using the curated financial Q&A dataset.

# Applied **Mixture-of-Experts adapters** for Balance Sheet and Income Statement specialization.

# **Evaluation & Benchmarking**

# Tested with mandatory and extended financial queries.

# Measured **accuracy, confidence, and latency** for both RAG and FT systems.

# **User Interface**

# Developed a **Gradio-based interactive UI**.

# Supports query input, method selection, and history tracking for comparisons.

# **Test Query Screenshots - UI**

# Below are three test queries executed using different retrieval methods. Each screenshot includes:

A screenshot of a computer

AI-generated content may be incorrect.

# A screenshot of a computer AI-generated content may be incorrect.

# Summary Comparison Table (with Latency / Confidence Score etc.)

**A screenshot of a computer program

Description automatically generated**

*Please check the python notebook for a more detailed view of various steps we captured.*

# 3. Hosted App Link

* **Platform:** Hugging Face
* **Demo Link:** <https://huggingface.co/spaces/Jawahars/flan-t5-base-kyndryl2023-24>
* **Usage Instructions:**
  1. Enter a query in the input box.
  2. Select retrieval method (RAG or Fine-Tuning).
  3. View the generated answer.
  4. View or Delete chat History (The UI is created to replicate an AI-chatbot)

***Please check below the screenshot from live application (some fonts and color might be different compared to local run due to underline platform support) -> Please do contact us if the link is not working or expired, as we are using a limited version of hosting infra.***

A screenshot of a chat

Description automatically generated

# 

# 4. Conclusion

This project successfully demonstrates the design and implementation of a **Comparative Financial Q&A System** using **Retrieval-Augmented Generation (RAG)** and **Fine-Tuned (FT)** approaches. By leveraging **Kyndryl’s FY 2023 and FY 2024 annual reports**, the system accurately handles financial queries through a combination of **hybrid retrieval, cross-encoder re-ranking, and generative modeling**.

The **RAG pipeline** proved effective for flexible, context-grounded responses, while the **fine-tuned model** delivered faster and more domain-adapted answers. The integrated **Gradio interface** provides an intuitive platform to compare both approaches in real-time, enabling interactive evaluation of accuracy, latency, and reliability.

This work highlights how **open-source NLP models** can be adapted for specialized financial domains, providing a scalable foundation for future enhancements such as advanced analytics, additional company datasets, and multi-year financial trend analysis.