Using RAG and LangChain to Drive Chatbots for Interactive Economic Data Analysis and Insight*

Seah Boon Keong, Eilyn Chong, Max Chew

Central Bank of Malaysia
Monetary Policy Department

(*) The views and conclusions presented in this paper are exclusively those of the author(s) and do not necessarily reflect the position of the Central Bank of Malaysia or its Board members.

- Motivation

- Methodology and Approach

- Usage

- Future Work

- Conclusion

Advancing Economic Data Access via Innovation in Chatbot Technology

- **Problem Statement:** Users often struggle with fragmented and complex economic datasets, requiring specialised knowledge to interpret.

- **Solution:** Our chatbot provides an intuitive interface that simplifies data access and analysis, making it easier for users to gain insights.

Democratize accessibility to vital economic data, extending the benefits to a broader user base including policymakers, researchers, financial analysts.

Ingest a wide array of metadata sourced from numerous authoritative institutions.

Offer an interactive and user-friendly interface to explore and analyze intricate economic datasets, thereby making complex data more comprehensible

Utilizes advanced technologies such Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) and Dynamic Charting capabilities

Interactive Economic Data Analysis and Insight: RAG-LLM

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

3
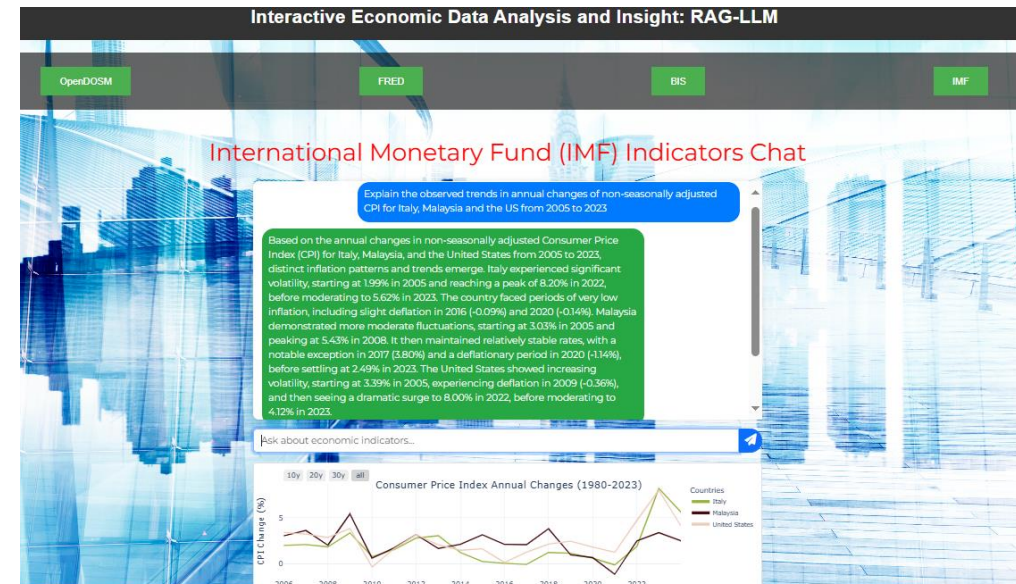
Data Sources

- We source publicly available economic datasets from authoritative institutions such as BIS, FRED, and IMF, as well as OpenDOSM for Malaysia's data

Types of Data

- The system ingests from various economic datasets, including macroeconomic indicators and financial statistics.

- It supports: CSV, Parquet, JSON

- The system will retrieve the datasets based on the URLs specified in the metadata, either from the returned API call or from links of CSV or Parquet files.

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

1. Using Retrieval Augmented Generation (RAG) Architecture

Utilizing **relevant information** effectively enhances the output from large language models. Details in the following slide.
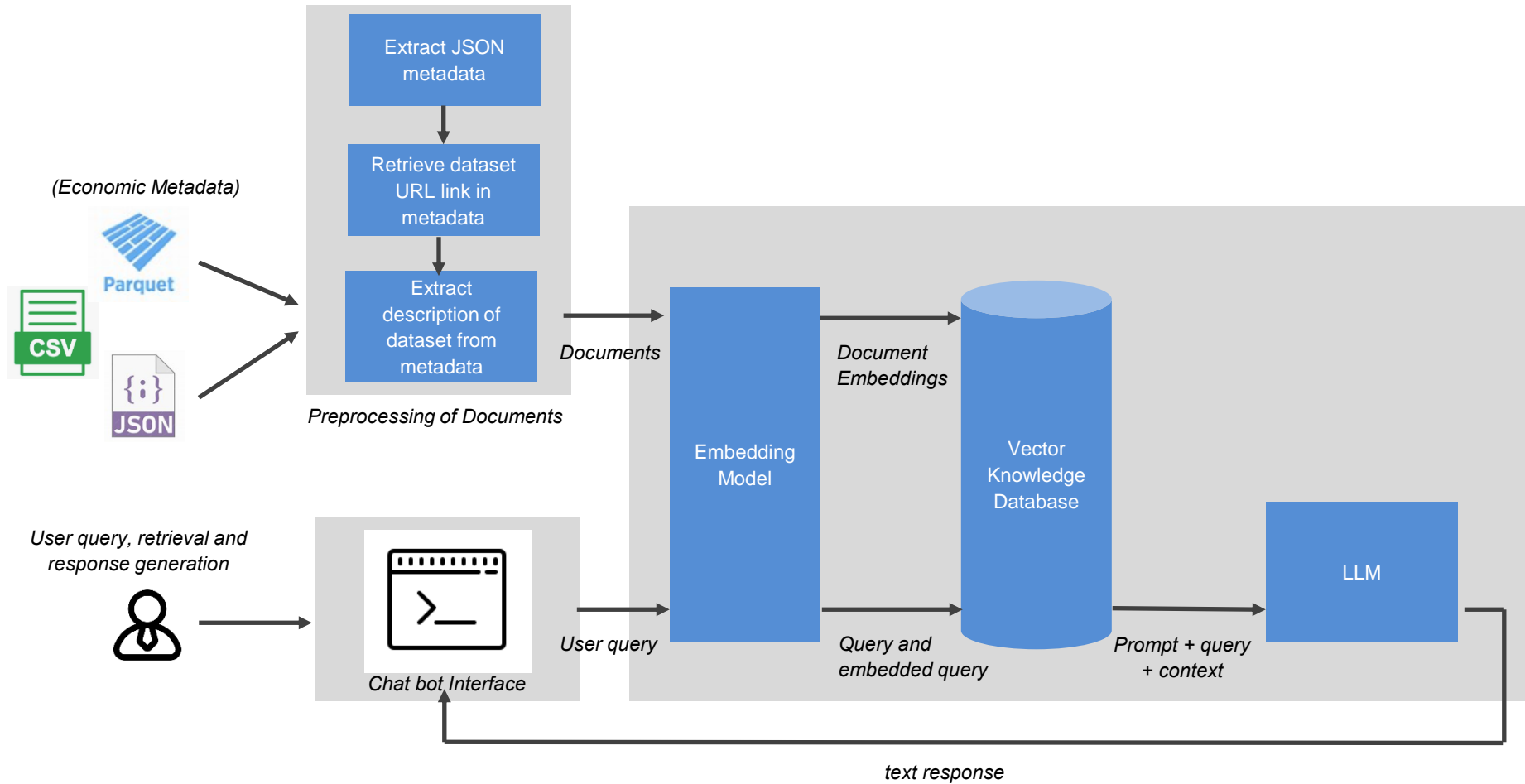
2. Data Source Integration

As a proof of concept, this prototype application ingests **datasets** from FRED, BIS, IMF and OpenDOSM for enriched analysis.

3. Dynamic Charting Techniques

Employing **interactive visuals** allows for better data interpretation and user engagement.

Methodology and Approach

Extract JSON metadata → Retrieve dataset URL link in metadata → Extract description of dataset from metadata

Preprocessing of Documents

(Economic Metadata)

Parquet / CSV / JSON

Documents → Embedding Model → *Document Embeddings* → Vector Knowledge Database

User query, retrieval and response generation

Chat bot Interface

User query → Embedding Model → *Query and embedded query* → Vector Knowledge Database → *Prompt + query + context* → LLM

text response

Blend of several AI components in which the key elements are as follow:

Large Language Model (LLM)
- Core of RAG
- Understands query context and generates responses.
- Trained based on the datasets provided
- Uses LangChain LLM.

External Knowledge Database (Vector DB)
- Vector database – a collection of Datasets retrieved from OpenDOSM, BIS, IMF, FRED.
- Uses FAISS Vector DB

Embedding Model
- Converts queries and data into vector representations.
- Uses GPT-4o mini with text-embedding-ada-002 for high-dimensional data.

Retrieval Mechanism
- Uses embeddings to find relevant information.
- Ensures data matches query context and intent.

Integration and Response Generation
- Integrates retrieved data into responses.
- Provides contextually relevant and corresponding chart

Comparison of FAISS, Chroma, and PineCone based on performance, scalability, ease of use, in-memory storage, API simplicity, and cost efficiency.

Reason for FAISS selection: **Performance and Cost Efficiency**

Feature	FAISS	Chroma	PineCone
Performance	High	Moderate	High
Scalability	High	High	High
Ease of Use	Moderate	High	High
In-Memory Storage	No	Yes	Yes
API Simplicity	Moderate	High	High
Cost Efficiency	High	Moderate	Moderate

Source:
Pinecone vs FAISS, Pinecone vs Chroma

An innovative chatbot system that leverages Retrieval-Augmented Generation (RAG) and LangChain to provide an intuitive, interactive interface for exploring and analysing economic datasets. - **Use case sample 1**
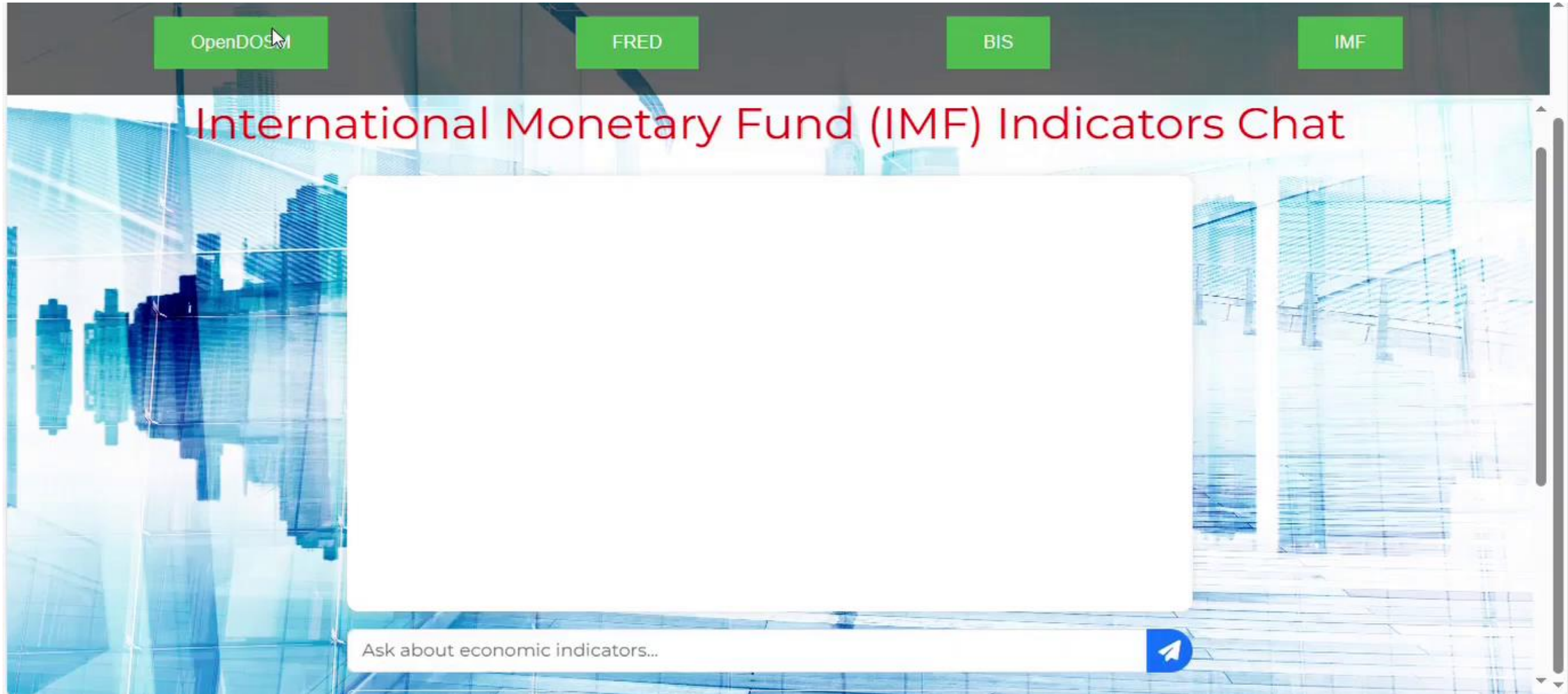
Usage

An innovative chatbot system that leverages Retrieval-Augmented Generation (RAG) and LangChain to provide an intuitive, interactive interface for exploring and analysing economic datasets. - **Use case sample 2**

An innovative chatbot system that leverages Retrieval-Augmented Generation (RAG) and LangChain to provide an intuitive, interactive interface for exploring and analysing economic datasets. - **Use case sample 3**
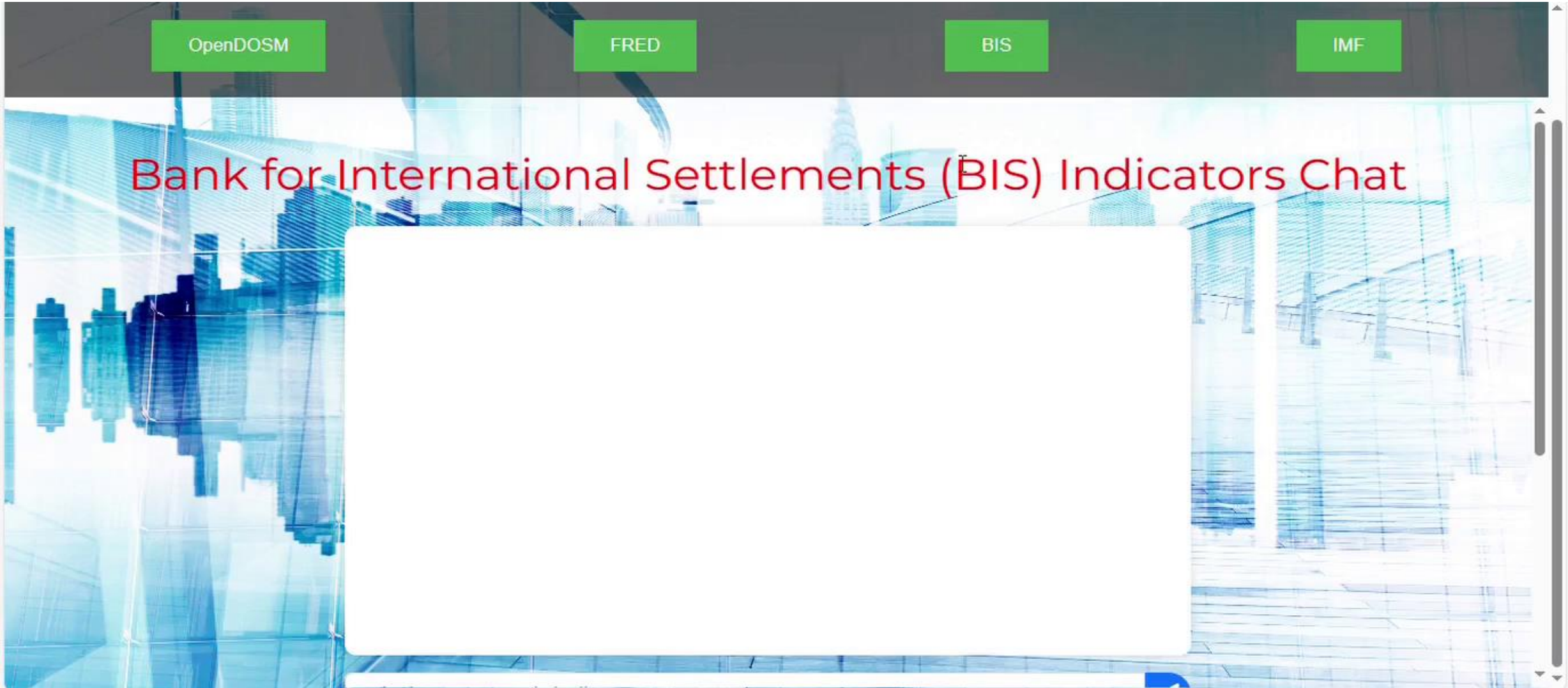
Potential Improvements (Prototype - Proof of Concept)

- This prototype (PoC) serves as a foundation for future enhancements. Key areas for improvement include:

  - **Expanding Data Sources**
    - We aim to integrate additional datasets from sources such as BIS, IMF, and FRED. This expansion will significantly enhance the chatbot's ability to deliver comprehensive and insightful analyses.

  - **Enhancing Dynamic Charting**
    - To provide users with a richer and more versatile visualization experience. We plan to expand it by including more chart types, such as bar charts, heatmaps, and scatter plots.

- In summary, our chatbot system democratizes access to economic data, making analysis more accessible and efficient.

- Facilitate faster exploration of economic data to gather data-driven insights and improving efficiency of economic analysis.

Thank you for your attention!

Seah Boon Keong (Ph.D) seah.keong@bnm.gov.my

Eilyn Chong eilyn@bnm.gov.my

Max Chew kai.xi@bnm.gov.my

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

4th BIS-IFC Workshop on Data Science in Central Banking, 18-20 Feb 2025