

Leveraging Custom Large Language Models for Economic Insights

Nik Ahmad Akram & Eilyn Chong

BANK NEGARA MALAYSIA

The views expressed are solely the responsibility of the author and should not be interpreted as reflecting the views of or the endorsement of any specific software by the Central Bank of Malaysia, or of anyone else associated with the Central Bank of Malaysia.

Motivation

Data-Rich World: Why We Need New Methods for Extracting Meaningful Insights

- In today's **data-rich environment**, economists have access to more data than ever before, often stored in accessible formats. However, they typically deal with complex **tabular data** that comes **in various schemas**, making it **challenging to extract** meaningful **insights** efficiently.
- The challenge is **to find an easier way to get insights** from these large and complex datasets. **Current methods** often require a lot of effort and **specialized skills in data analysis**. This study aims to **make information retrieval more straightforward** through natural language queries, potentially making the process easier for economists and decision-makers.

Global Datasphere

The global datasphere expected to double by 2026 relative to the end of 2022

2022: 101 Zettabytes

2023: 123 Zettabytes

2026: 221 Zettabytes

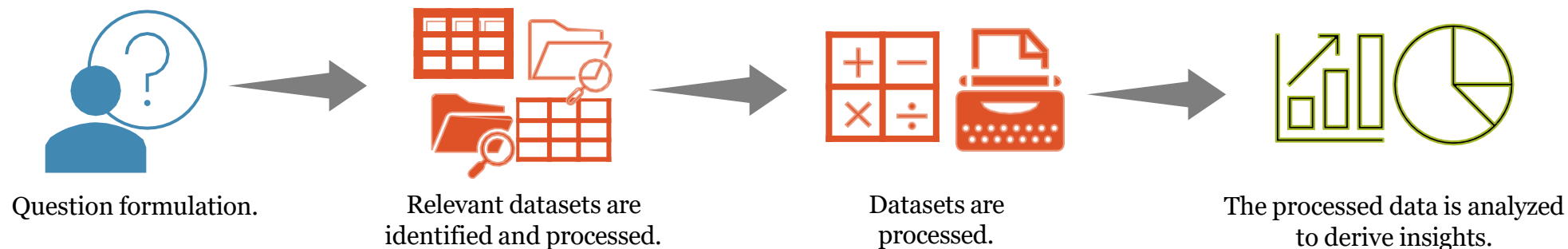
Sources: AVANT by Avison Young, IDC Global DataSphere Forecast (2022-2026)

Addressing Dataset Navigation Challenges

- In our organization, **we have curated dataset specifically for economists** and presented it through a **data dashboard**. However, users still **face challenges navigating** this dashboard due to the **vast datasets** and the **varied questions different users have**, each requiring **specific data manipulation** to provide accurate insights.
- With the rise of AI, there's growing interest in implementing **Large Language Models (LLMs) to address these challenges**. However, **general LLMs** are too broad and not suited for the nuanced, specific questions economists ask. This situation calls for **a specialized LLM framework** designed to meet the unique needs of our users

Navigating Complex Data to Drive Informed Decisions

Streamlining Economic Analysis: From Questions to Insights in a Data-Rich World



Challenges in Handling Economic Data

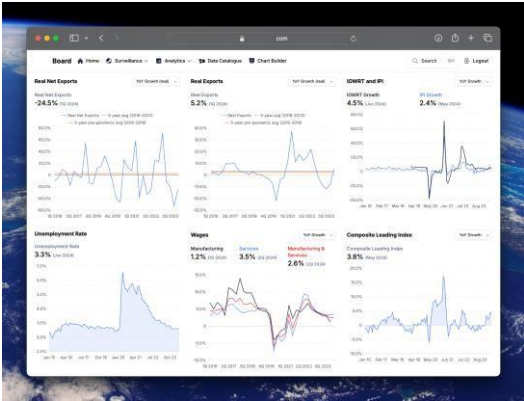
Dataset Formats: Economic datasets are often presented in tabular forms, requiring specific expertise and tools to manage effectively.

Structured Workflow: Economists adhere to a systematic process that necessitates precise data manipulation to generate insights.

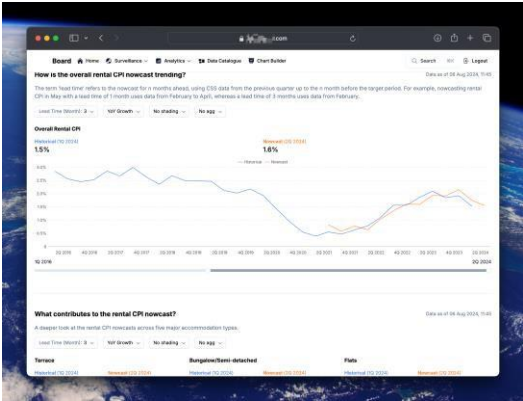
Context Dilution: Directly inputting large datasets into a general LLM can dilute context, resulting in less accurate outputs.

Dashboard: Curated Dataset for Economists' Analysis

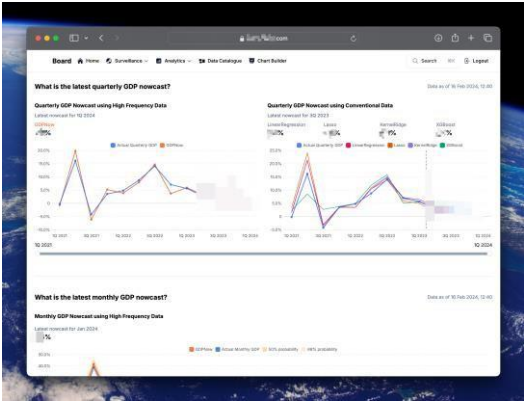
Economics Indicator Dashboard in BNM



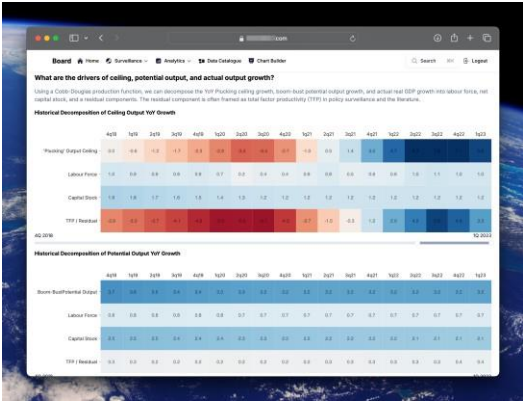
Comprehensive Overview: Multiple Economic Indicators on a Single Dashboard Page



Nowcasting Rental CPI: Analysing Current Trends in the Rental Market



GDP Forecast: Leveraging High-Frequency Data for Enhanced Accuracy



Heatmap Visualizations

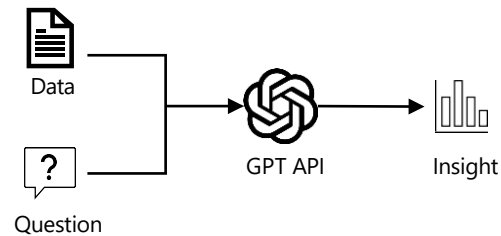
Highlights

- **100+ Charts:** Visualizations to track key economic indicators.
- **50+ Datasets:** Comprehensive data coverage for detailed analysis.
- **Daily Updates:** Ensures the latest information is always available.
- **Web-Based Access:** Optimized for both laptop and tablet use.
- **API Access:** Allows users to directly extract data for custom analysis.
- **Interactive:** Users can easily filter data by date and other parameters for customized insights.

Conceptual Framework & Proposed Solutions

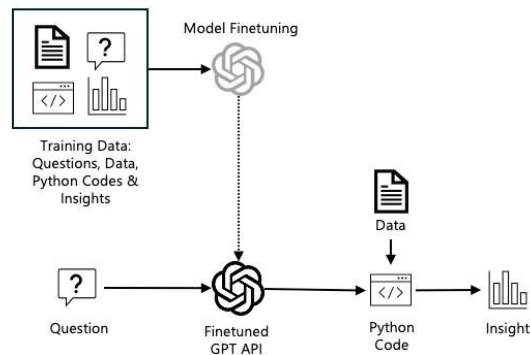
Initial Approach Using General-Purpose LLM

This method involves directly providing both the question and the tabular data in the prompt to the model.



Enhanced Approach with Fine-Tuned LLM

The model was trained with sample data to understand the dataset structure, allowing it to generate Python code for data processing during inference, thus improving efficiency and accuracy.



Proposed Solutions:

- **Model Fine-Tuning:** The LLM is fine-tuned using training data that includes typical questions, data structures, Python code, and the format in which economists usually present answers. This fine-tuning is done initially and updated periodically as new types of questions emerge.
- **Python Code:** When a user asks a question, the model generates the appropriate Python code tailored to the query.

Data itself is not fed into the LLM, which helps prevent context dilution that could occur if the model had to handle large datasets directly from the query

This process ensures that the model provides correct responses to specific, nuanced questions posed by economists.

Sample Results #1 : What is the GDP level for the agriculture sector in Q1 2022?

- For the purpose of this performance evaluation, we took a dataset of Malaysia's GDP time series data from an open data portal www.data.gov.my.
- The dataset consists of GDP level for every quarter since 2015 Q1 until 2024 Q1. The dataset also includes the components of GDP both from the expenditure and production approaches.

General Model Output:

Answer: RM 24,833 million



Enhanced Approach Output (Python Code)

Answer: 24833000000



```
# Filter the data for Q1 2022
q1_2022_data = data[(data["date"] == "01/01/2022") &
                    (data["chart_type"] == "real_sa")]
# Extract the GDP level for the agriculture sector in Q1 2022
gdp_agri_q1_2022 = q1_2022_data["supply_agri"]

gdp_agri_q1_2022 : 24833000000
```

Both approaches are able to produce the correct answer, as it only requires data extraction.

Sample Results #2 : What is the highest YoY growth in the manufacturing sector in 2021?

General Model Output:

Answer: 12.35%



Enhanced Approach Output (Python Code)

Answer: 27.2345795387164



```
# Calculate year-on-year (YoY) growth for the manufacturing sector
data["supply_manufacturing_yoy"] =
data["supply_manufacturing"].pct_change( periods=4 ) * 100
# Filter the data for the year 2021
data_2021 = data[ data["date"].dt.year == 2021 ]
# Find the highest YoY growth in the manufacturing sector in 2021
max_yoy_growth_2021 = data_2021["supply_manufacturing_yoy"].max()

# max_yoy_growth_2021 = 27.2345795387164
```

The general model is unable to provide the correct answer because calculating YoY growth involves specific computational steps that it cannot execute directly.

Sample Results #3 :How has the manufacturing sector's GDP evolved from 2015 to 2023?

General Model Output:

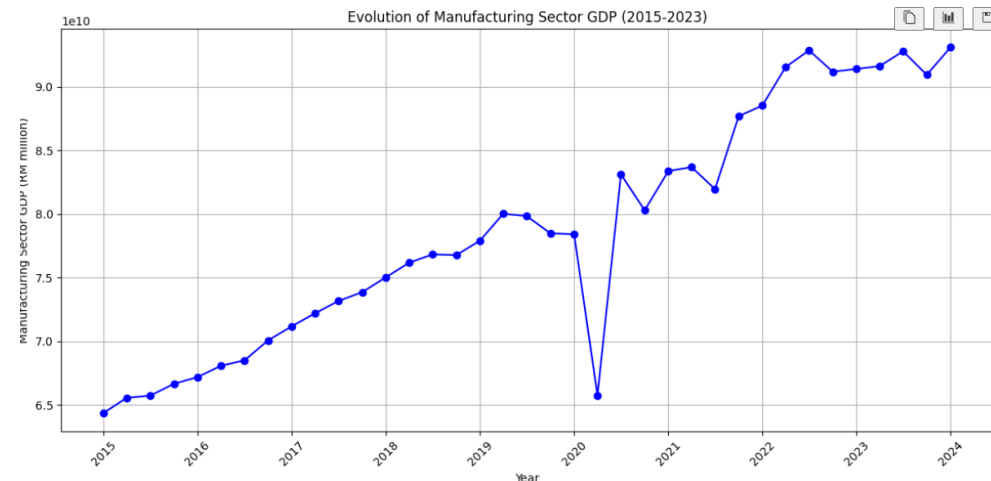
Answer: Increased from 158,444,000,000 in 2015 to 236,823,000,000 in 2023.



Enhanced Approach Output (Python Code)



```
# Extract the data for the manufacturing sector GDP from 2015 to 2023
manufacturing_gdp = data[["date", "supply_manufacturing"]]
# Convert the date to datetime format for better visualization
manufacturing_gdp["date"] = pd.to_datetime(manufacturing_gdp["date"],
format="%d/%m/%Y")
# Plot the evolution of the manufacturing sector GDP from 2015 to 2023
plt.figure(figsize=(12, 6))
plt.plot(manufacturing_gdp["date"],
manufacturing_gdp["supply_manufacturing"], marker="o", linestyle="-",
color="b")
```



The general-purpose model provides a simple numerical answer, while the enhanced approach generates code for a line chart, offering a more comprehensive and insightful representation of the data. This enhanced approach provides a more relevant response, illustrating the trend over time.

Discussion & Conclusion

- **Domain Specific:** A custom-trained LLM excels at understanding and interpreting domain-specific language, making data analysis more intuitive and efficient for economists.
- **Context Dilution:** The model generates Python code instead of directly handling large datasets, reducing context dilution and ensuring accurate data processing.
- **Data Governance:** For future work, we plan to incorporate open-source LLM models like LLaMA to create a secure, on-premises framework that better protects data privacy.
- **Enhanced Efficiency:** We aim to add a table retrieval mechanism to automatically select the most relevant table from large datasets, further improving efficiency and accuracy.

Challenges:

- **Reduced Generalization:** Specialization may limit versatility, leading to errors with unfamiliar queries. Expanding the training dataset is essential to cover a broader range of tasks.
- **Code Errors:** Generated code might fail due to errors or missing libraries. Robust validation and error-handling are essential.

thank you ~!

email: ahmad.akram@bnm.gov.my