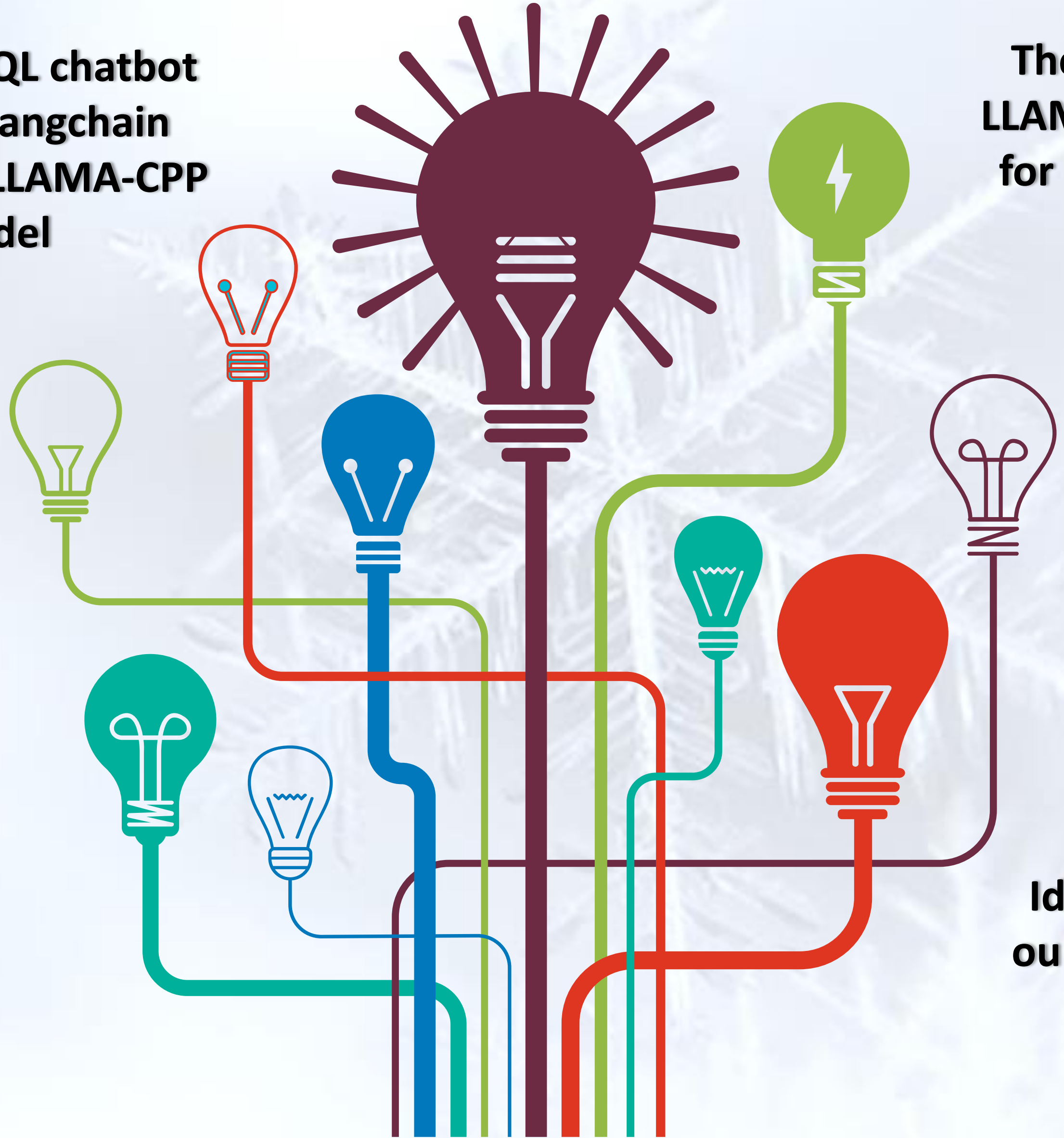
FINQUERY: SQLMetric Chatbot

Overview

- Introduction
- Model Terminology
- Architecture
- Inference Results
- Comparison with AWS Bedrock Bot
- Chatbot functionality
- Limitations

Introducing SQLMetric Chatbot



Developed an SQL chatbot using Python Langchain framework and LLAMA-CPP LLM model

The chatbot utilizes the LLAMA3 8B Instruct model for generating responses

Compared chatbot's performance to Bedrock's LLAMA3 8B Instruct implementation

Our evaluation focused on query accuracy and response relevance

Identified limitations in our chatbot's capabilities

LLM Model Terminology

Parameters

Parameters in a machine learning model include weights and biases that the model learns during training. These values are adjusted to minimize the error in predictions and are critical in defining the model's ability to generalize from training data to unseen data

8B Parameters

A model with 8 billion parameters. Suitable for tasks requiring substantial capacity but with somewhat lower resource requirements compared to larger models

13B Parameters

A model with 13 billion parameters. Offers higher performance and can handle more complex tasks than smaller models

70B Parameters

A very large model with 70 billion parameters. Ideal for highly complex tasks and applications requiring significant computational resources

Quantization

A technique to reduce the size and improve the efficiency of a model by lowering the precision of its weights

Benefits: Reduces model size, speeds up inference, and lowers computational and memory requirements, making models deployable on edge devices or less powerful hardware.

4-bit Quantization: Reduces weight precision to 4 bits, dramatically decreasing model size and computational load while retaining most performance

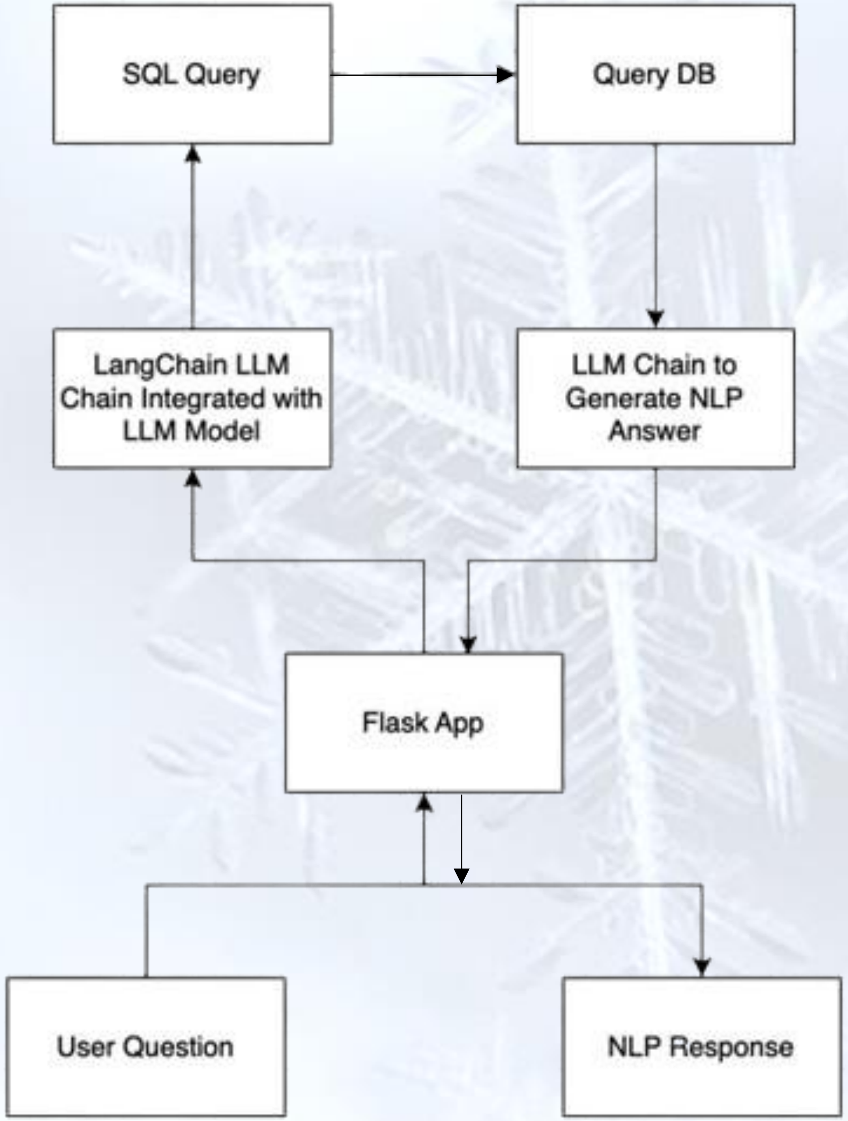
8-bit Quantization: Reduces weight precision to 8 bits, offering a balance between size reduction and maintaining higher accuracy.

Significant reduction in model size and computational needs, but with a higher potential for accuracy loss

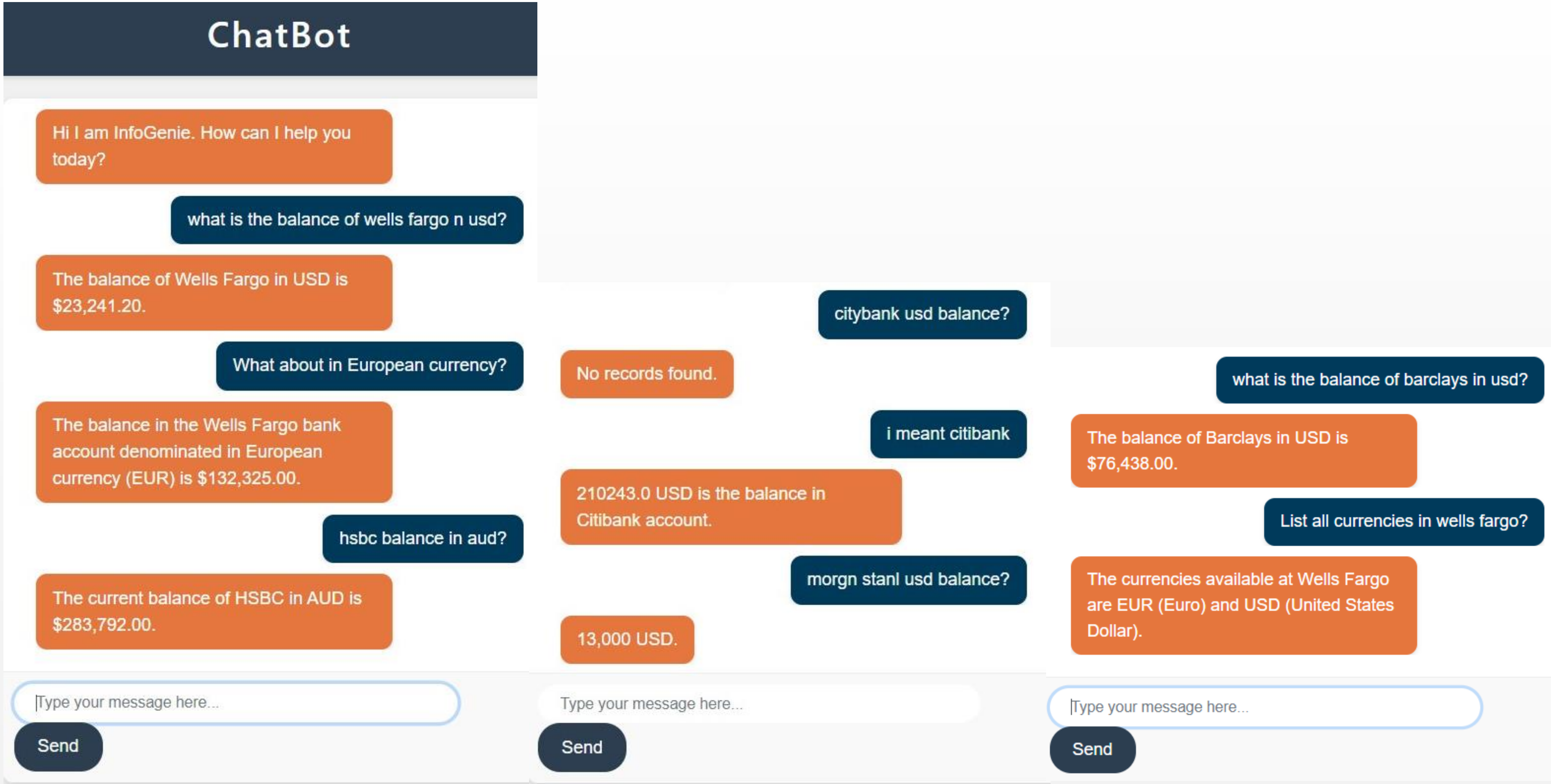
LLAMA3 8B Instruct 4-Bit

8B	8B stands for "8 billion" parameters in the model, indicating its size and complexity. This large parameter count allows the model to capture extensive information and perform well on various tasks
Instruct	Model fine-tuned to follow instructions
GGUF	A file format optimized for efficiently storing and inferring machine learning models
4 KM	Indicates the use of 4-bit K-means quantization, a method that reduces the precision of the model's weights to 4 bits using K-means clustering, making the model more compact and faster to run

Architecture


SQL Query
Query DB
LangChain LLM Chain Integrated with LLM Model
LLM Chain to Generate NLP Answer
Flask App
User Question
NLP Response

Inference Results - LLAMA3 8B Instruct


ChatBot
Type your message here...
Send
Type your message here...
Send
Type your message here...
Send

Hi I am InfoGenie. How can I help you today?

what is the balance of wells fargo n usd?

The balance of Wells Fargo in USD is $23,241.20.

What about in European currency?

The balance in the Wells Fargo bank account denominated in European currency (EUR) is $132,325.00.

hsbc balance in aud?

The current balance of HSBC in AUD is $283,792.00.

citybank usd balance?

No records found.

i meant citibank

210243.0 USD is the balance in Citibank account.

morgn stanl usd balance?

13,000 USD.

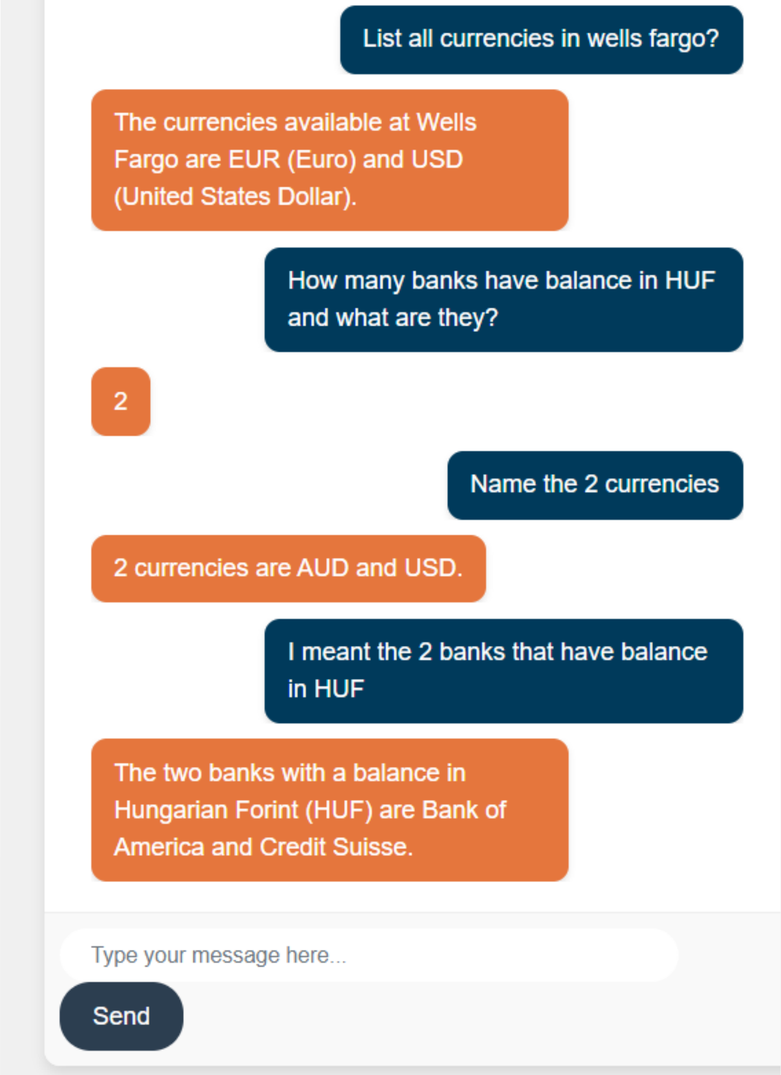
what is the balance of barclays in usd?

The balance of Barclays in USD is $76,438.00.

List all currencies in wells fargo?

The currencies available at Wells Fargo are EUR (Euro) and USD (United States Dollar).

Inference Results - LLAMA3 8B Instruct



List all currencies in wells fargo?

The currencies available at Wells Fargo are EUR (Euro) and USD (United States Dollar).

How many banks have balance in HUF and what are they?

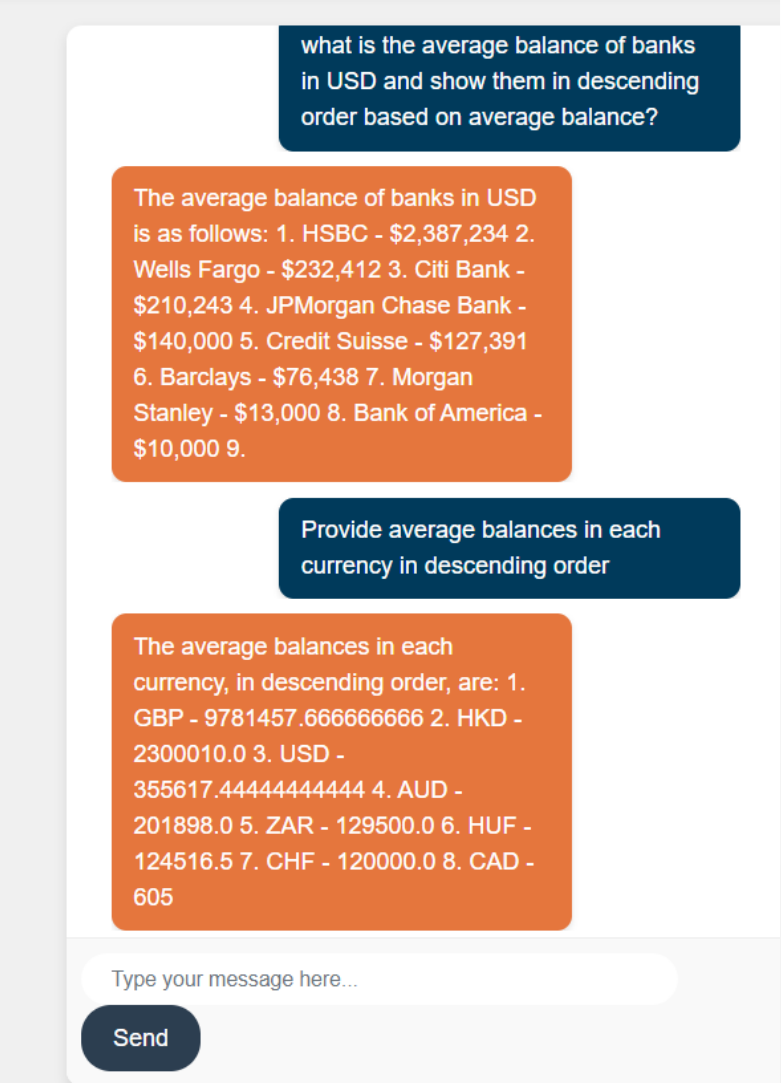
2

Name the 2 currencies

2 currencies are AUD and USD.

I meant the 2 banks that have balance in HUF

The two banks with a balance in Hungarian Forint (HUF) are Bank of America and Credit Suisse.



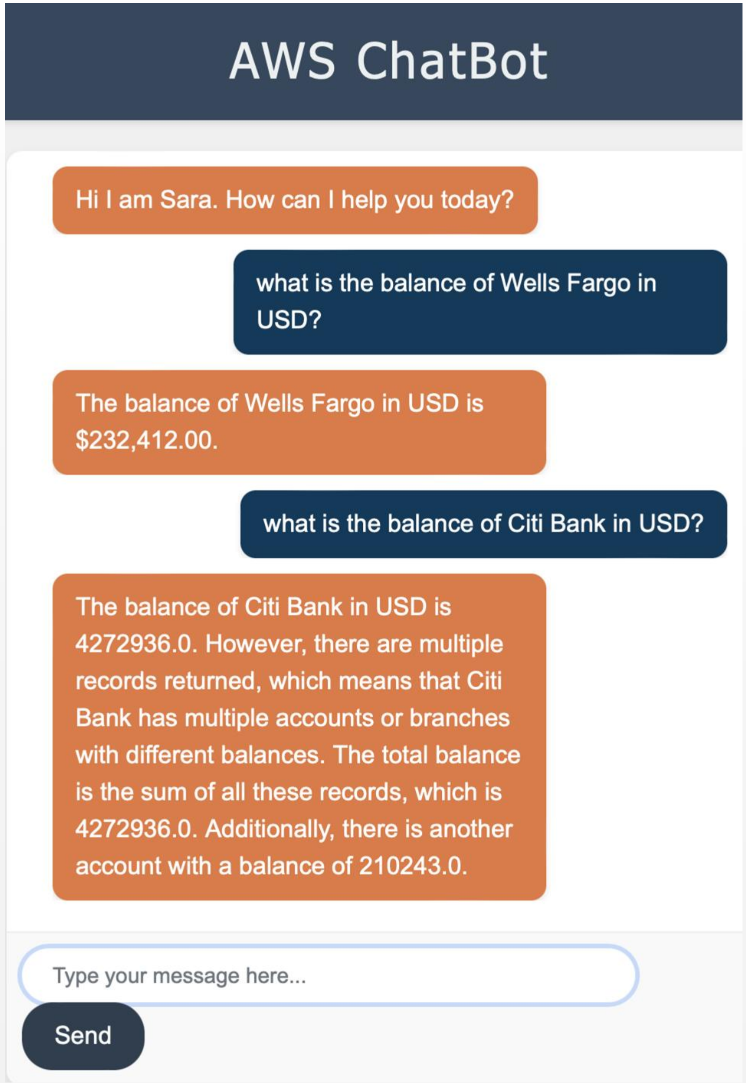
what is the average balance of banks in USD and show them in descending order based on average balance?

The average balance of banks in USD is as follows: 1. HSBC - $2,387,234 2. Wells Fargo - $232,412 3. Citi Bank - $210,243 4. JPMorgan Chase Bank - $140,000 5. Credit Suisse - $127,391 6. Barclays - $76,438 7. Morgan Stanley - $13,000 8. Bank of America - $10,000 9.

Provide average balances in each currency in descending order

The average balances in each currency, in descending order, are: 1. GBP - 9781457.666666666 2. HKD - 2300010.0 3. USD - 355617.44444444444 4. AUD - 201898.0 5. ZAR - 129500.0 6. HUF - 124516.5 7. CHF - 120000.0 8. CAD - 605

Comparison with AWS Bedrock LLAMA3 8B Instruct



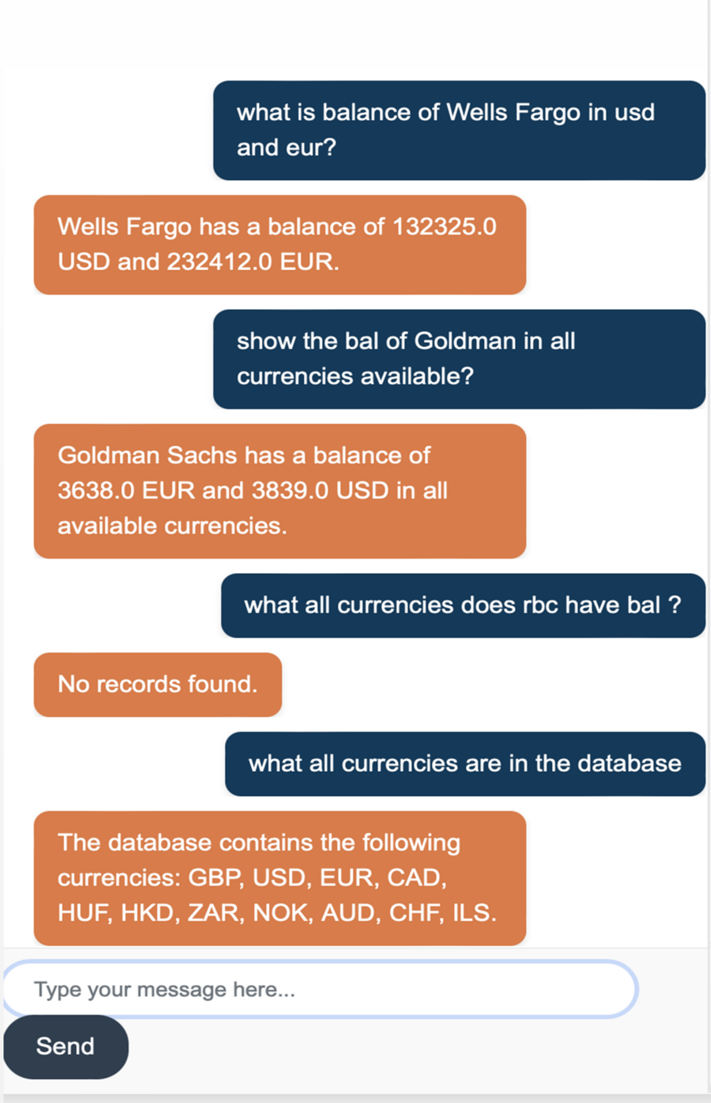
Hi I am Sara. How can I help you today?

what is the balance of Wells Fargo in USD?

The balance of Wells Fargo in USD is $232,412.00.

what is the balance of Citi Bank in USD?

The balance of Citi Bank in USD is 4272936.0. However, there are multiple records returned, which means that Citi Bank has multiple accounts or branches with different balances. The total balance is the sum of all these records, which is 4272936.0. Additionally, there is another account with a balance of 210243.0.



what is balance of Wells Fargo in usd and eur?

Wells Fargo has a balance of 132325.0 USD and 232412.0 EUR.

show the bal of Goldman in all currencies available?

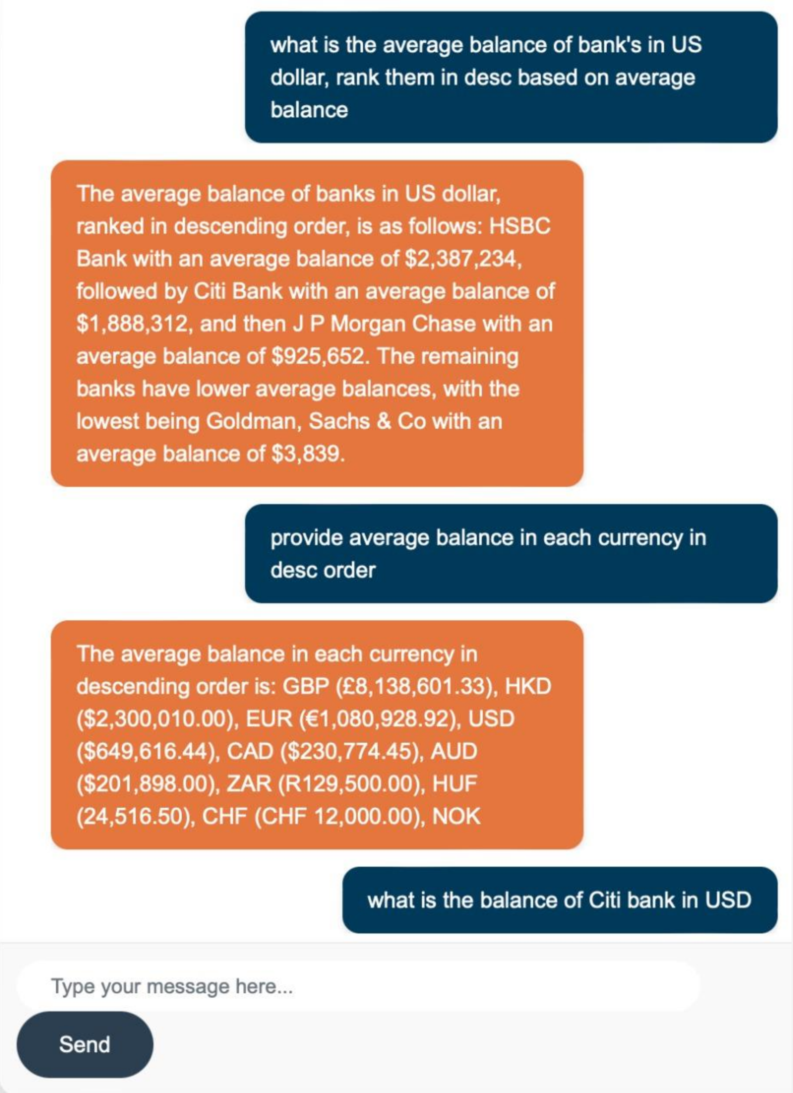
Goldman Sachs has a balance of 3638.0 EUR and 3839.0 USD in all available currencies.

what all currencies does rbc have bal ?

No records found.

what all currencies are in the database

The database contains the following currencies: GBP, USD, EUR, CAD, HUF, HKD, ZAR, NOK, AUD, CHF, ILS.



what is the average balance of bank's in US dollar, rank them in desc based on average balance

The average balance of banks in US dollar, ranked in descending order, is as follows: HSBC Bank with an average balance of $2,387,234, followed by Citi Bank with an average balance of $1,888,312, and then J P Morgan Chase with an average balance of $925,652. The remaining banks have lower average balances, with the lowest being Goldman, Sachs & Co with an average balance of $3,839.

provide average balance in each currency in desc order

The average balance in each currency in descending order is: GBP (£8,138,601.33), HKD ($2,300,010.00), EUR (€1,080,928.92), USD ($649,616.44), CAD ($230,774.45), AUD ($201,898.00), ZAR (R129,500.00), HUF (24,516.50), CHF (CHF 12,000.00), NOK

what is the balance of Citi bank in USD

Chatbot Functionality

Offline Chatbot works on CPU

Query retrieval at a reasonable speed

LLM able to understand reasonable typos from user's query

Able to follow users conversation to a limit

Limitations

Speed	Using GPUs or cloud services will maximize the utilization of LLM models and provide answers within seconds, unlike CPUs which tend to maximize CPU utilization
Accuracy	Limiting ourselves to CPU GGUF quantized models compromises accuracy and restricts prompt engineering improvements needed to enhance accuracies
Hallucination	CPU GGUF models tend to hallucinate(inaccurate answers) more frequently than usual
Latency	The model experiences significant latency and crashes due to maximum CPU utilization giving nonsensical answers
Unreliability	Sometimes, the LLM provides incorrect answers to the same question, but it may give the correct response if prompted again

Thank you