

BCG-X Challenge 2024



*“Using Generative
AI to provide a
sustainable future”*

Authors



Marcelo
Data Scientist



Vitor
Data Engineer



Matheus
Software Engineer



Melissa
Data Scientist



Motivation

The Problem

Brazil ranks fourth among the **top 10** countries with the **highest cumulative emissions of CO₂ from fossil fuel combustion, deforestation, and land-use change.**



<https://adit.com.br/mudancas-climaticas-e-desenvolvimento-urbano-impactos-e-desafios/>

Go to the Market



"In society, there is a perception that climate change is something that exists. People have heard about it and consider it a fact. In politics, it is different; it enters a complex universe involving various levels and issues, with climate change being one of them."

Ph.D. Jean P. Ometto - Senior Researcher at the Brazilian National Institute for Space Research and Dean of the Earth System Science Centre.

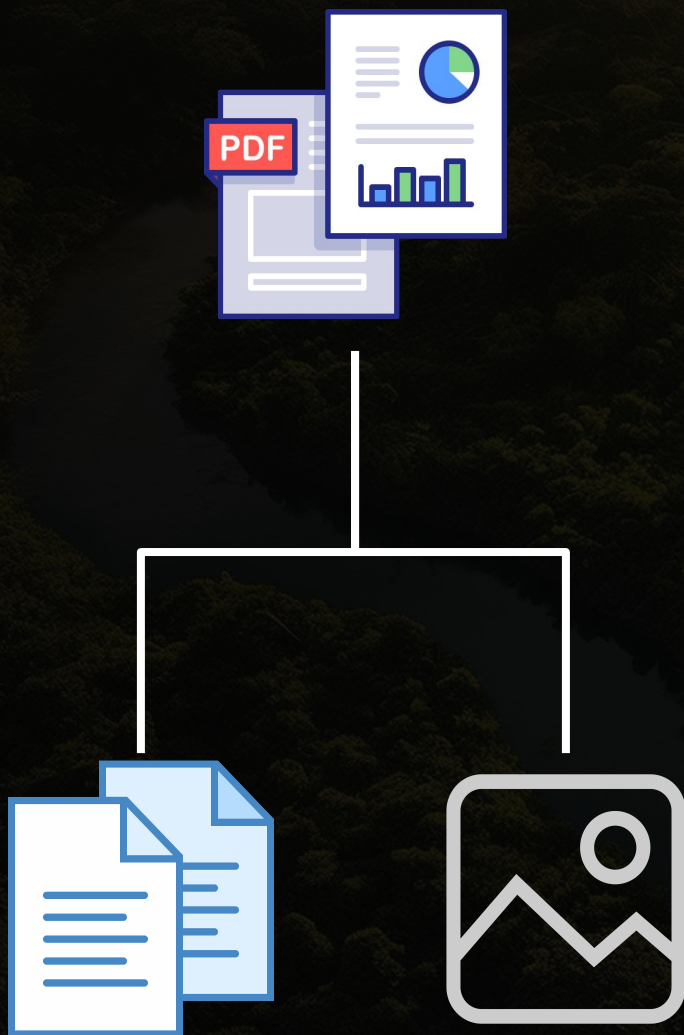
"Promoting effective communication between public managers, technical reports, and the community can facilitate the incorporation of different perspectives into the decision-making process."

Ph.D. Evaldo Luiz Gaeta Espindola - Senior researcher in ecotoxicology and aquatic ecosystem ecology, conducting research, mentoring, and water resource management, as well as contributing to environmental policies.





Data Preparation



Extract text from the PDF files

PDF files are loaded and processed using PyMuPDF which is the most efficient library available nowadays, the PDFs texts are extracted and sent to a list.

Extract images from the PDF files

PyMuPDF is also used to extract images from the PDF files, this is done because it was identified that several images contain important informations that can be used to offer more quality information to the LLM.

The texts inside the images are then extracted using pytesseract and loaded to a list with their metadata.

Clean the text extracted

Once all the information is extracted from the files and stored in lists, it's time to clean all the texts using regex and applying treatments like:

- Remove useless pages, such as summary;
- Remove line breaks, tabs and replace multiple spaces with a single space;
- Remove meaningless strings of letters;
- Remove unwanted characters.

This process guarantee that the LLM will only receive quality data and nothing that could trigger hallucinations.

Generate embeddings and indexes

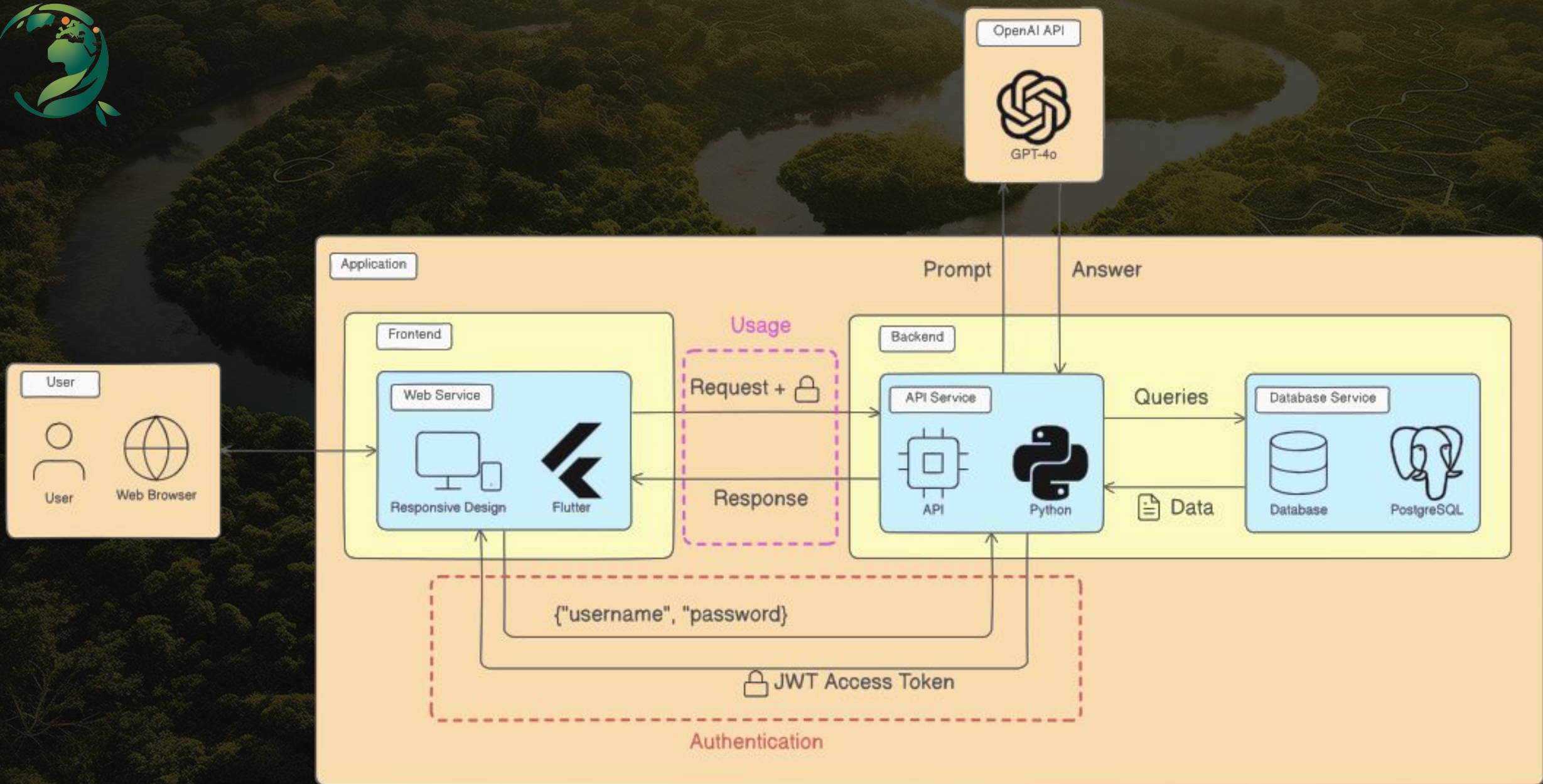
After the texts are fully cleaned their splitted into chunks with suitable size and the embeddings are generated using the openai API key.

The intention of separating the text from the images text is so that it's possible to personalize the chunks size to both sources of information, and that allows a better similarity search.

Then, using FAISS (Facebook AI Similarity Search), from Meta, the embeddings are stored in a vector database and create the indexes that increase the performance of the similarity searches.



Architecture



Key Concepts

JWT Authentication

Stateless, reduces the overhead added by sessions and removes the need to provide credentials on every request

Flutter

Responsive design, manages the state of its components.

FastAPI

Handles dependency injection of the database session and user authentication.

PostgreSQL

Stores user and chat data and implements ownership mechanisms through relationships.



Chatbot

Functionalities

Code Structure

LangChain Functionalities

- Concise Code Structure
- Useful Methods
- Easy Implementation and Orchestration

Question Classification

Filter as a Question Classification using LLM

- More Human like Interaction
- Simple Questions are faster processed
- Exclude the necessity of a RAG mechanisms for simple interactions

Memory Mechanism

Memory for handling Chat History

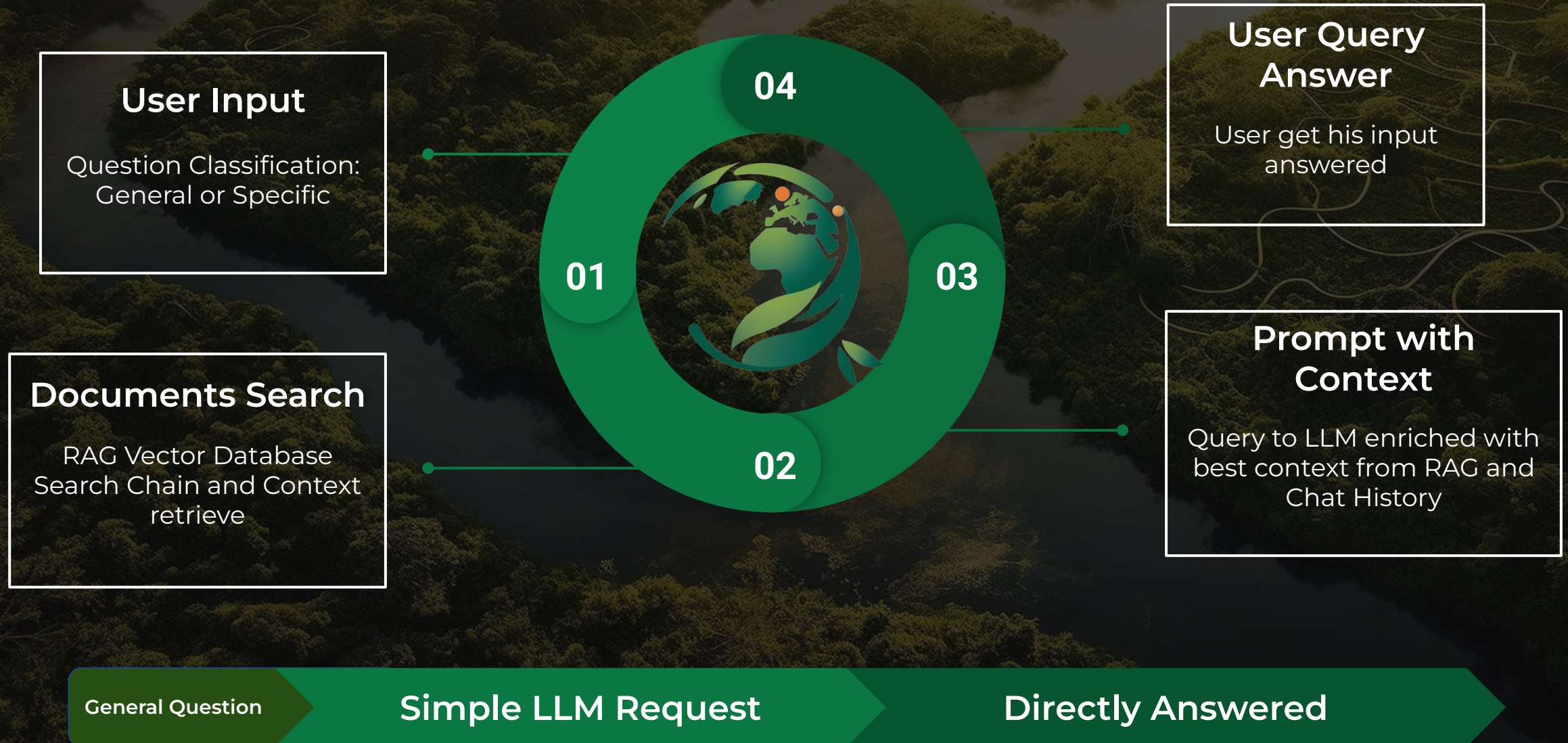
- Chat History used to improve answer quality
- Combination of Strategies for better results
- Maximum of tokens not reached

Guard Rails

Steps that prevent Hallucinations

- Default Message for minimum similarity score
- Temperature, Frequency and Presence Penalty parameters

ChatBot Chain Workflow





Results

Ragas and Test Dataset

Dataset for Testing

Default Dataset with 100 question generated by IA and answers validated with ambiental specialists from USP - São Carlos

Ragas Pipeline

Use of RAGAS - A library created to help validate query results based on a dataset with the questions and the answer considered the Ground Truth.

RAGAS also has a gamma of metrics for us to use as comparison for parameters variation tests

Faithfulness

RAGAS Metrics Range: [0,1]

Measures how faithful the response is to the retrieved documents' content. Higher scores indicate that the response is more grounded in the correct information.

0,74

Answer Relevancy

Assesses how well the response aligns with the user's question. High scores mean the answer is directly relevant to the query.

0,87

Answer Similarity

Compares the chatbot's response to an ideal reference answer, assessing the similarity in both content and structure.

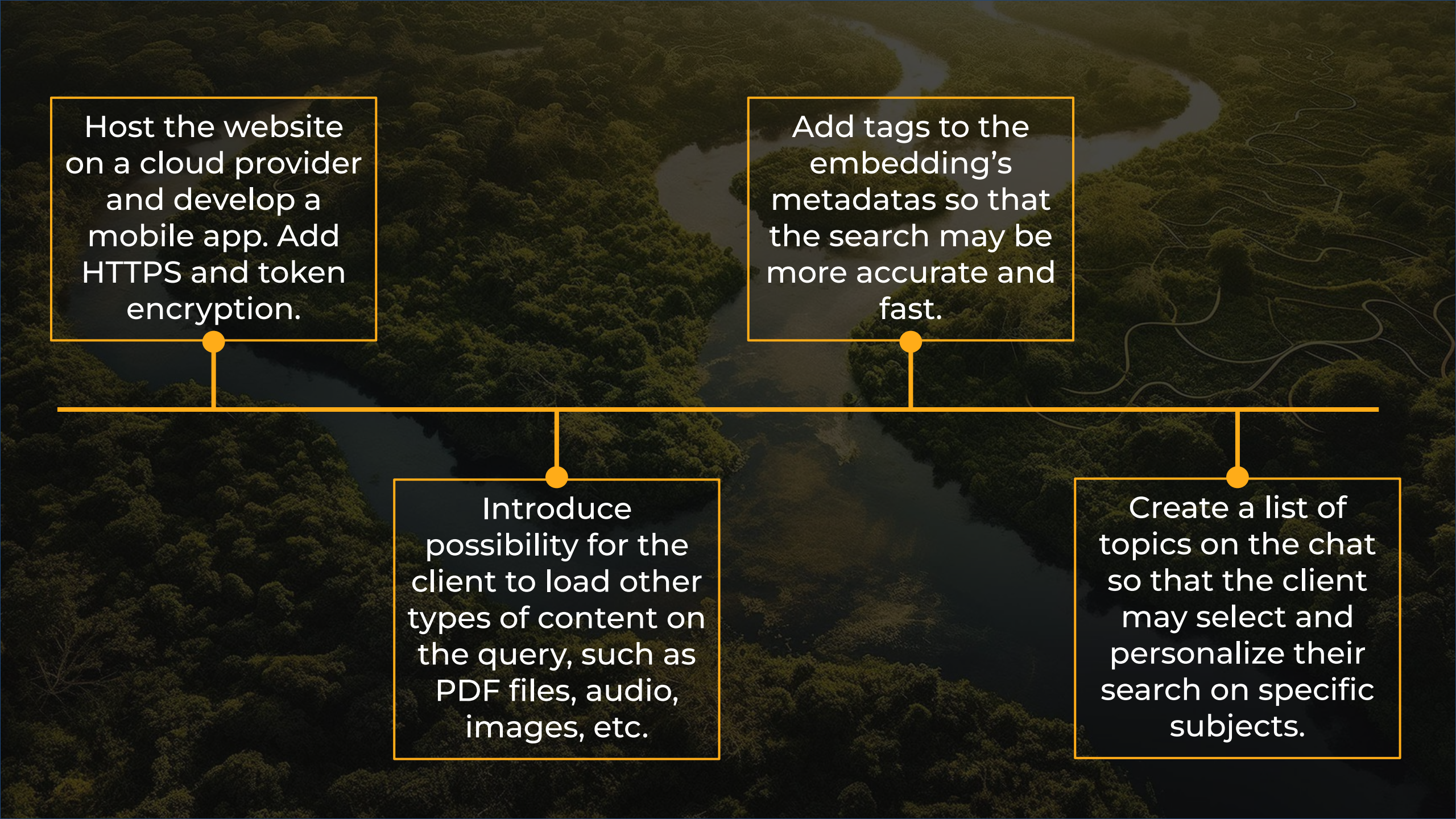
0,91



Demonstration



Next Steps



```
graph TD; A[Host the website on a cloud provider and develop a mobile app. Add HTTPS and token encryption.] --- B[Add tags to the embedding's metadata so that the search may be more accurate and fast.]; A --- C[Introduce possibility for the client to load other types of content on the query, such as PDF files, audio, images, etc.]; A --- D[Create a list of topics on the chat so that the client may select and personalize their search on specific subjects.];
```

Host the website on a cloud provider and develop a mobile app. Add HTTPS and token encryption.

Add tags to the embedding's metadata so that the search may be more accurate and fast.

Introduce possibility for the client to load other types of content on the query, such as PDF files, audio, images, etc.

Create a list of topics on the chat so that the client may select and personalize their search on specific subjects.



Thank You

