# RAG System for AI ACT

## Retrieval-Augmented Generation

CONTENTS

## Introduction

The development of a RAG system provides a robust and scalable solution for the consultation of the European AI ACT (English or Spanish). This architecture not only guarantees fast and accurate access to information, but also offers the ability to generate contextualised and coherent responses, adapting to the specific needs of the user and the domain. The deployment of the RAG infrastructure has been carried out on AWS, which provides high availability, security and scalability of the system.

This document details the specifications and development of a Retrieval-Augmented Generation (RAG) system designed to query the European AI ACT. The system has been developed using LangChain technology, and several advanced technologies have been employed such as Hugging Face embeddings with the all-mpnet-base-v2 model, OpenAI's GPT-4 model, and the Pinecone vector database. The following is a breakdown of the system components, their rationale, and the proposed architecture for implementation on AWS.

## What is a RAG and why is it necessary?

A Retrieval-Augmented Generation (RAG) system combines information retrieval techniques with text generation models to provide accurate and contextually relevant answers to user queries. This type of system is especially useful for querying knowledge databases, such as the European AI ACT, for several reasons:

1. **Fast and Accurate Access to Information**: A RAG can quickly search through large volumes of data and extract the most relevant information, saving time and effort compared to manual search.
2. **Generation of Contextualised Responses**: Using advanced language models, a RAG not only retrieves information, but can also generate responses that integrate various pieces of data in a coherent manner.
3. **Adaptability and Scalability**: These systems can easily adapt to different domains and scale to handle large numbers of simultaneous queries, which is crucial in applications at the European level.

## Justification for the use of technologies

**Hugging Face model: "all-mpnet-base-v2".**

Hugging Face's "all-mpnet-base-v2" model has been selected for its high accuracy in creating text embeddings. This model is known for:

- **High Quality Embeddings**: Provides dense and semantically meaningful vector representations of text, which is essential for efficient information retrieval.
- **Computational Efficiency**: It is optimised for use in high-demand applications, providing a good balance between accuracy and performance.
- **Flexibility**: It can be easily integrated into natural language processing pipelines, such as those provided by LangChain.

**OpenAI GPT-4:**

GPT-4 has been chosen as the text generation model due to its advanced text generation capabilities:

- **Natural Language Understanding**: GPT-4 can understand and generate text in a coherent and contextually relevant way, which is crucial for providing accurate and useful responses.
- **Versatility**: It is capable of handling a wide variety of language processing tasks, from response generation to information synthesis.
- **Ongoing Support**: OpenAI provides ongoing updates and support, ensuring that the model remains at the forefront of AI technology.

**Pinecone**

Pinecone has been selected as the vector database to store the embeddings because of:

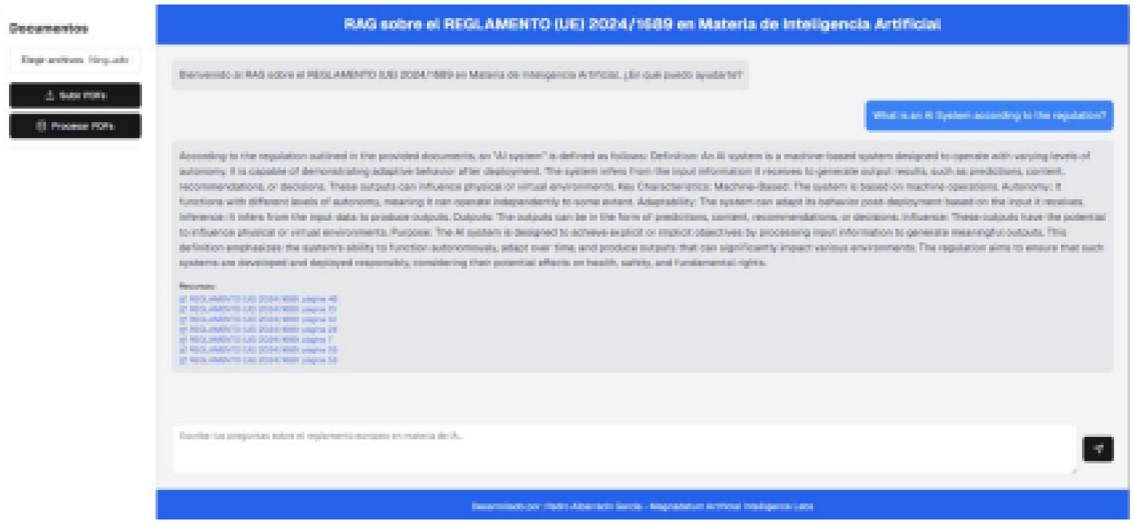
- **Optimisation for Vector Data**: Pinecone is specifically designed to handle large volumes of vector data efficiently.
- **Scalability**: It can scale horizontally to handle increasing volumes of data and queries.
- **Easy Integration**: It offers easy integration with AI and machine learning frameworks, allowing for seamless implementation into the LangChain pipeline.

## System Architecture

The architecture of the RAG system for querying the European AI ACT is designed to be implemented on top of AWS, taking advantage of its scalable and secure services. The proposed architecture is described below:

1. **Data Ingest Layer**:
   - **AWS S3**: For the storage of raw data and reference documents.
   - **AWS Lambda**: To process and transform the data before indexing.
2. **Embeddings Processing Layer**:
   - **EC2 Instances**: Virtual machines configured to run Hugging Face's "all-mpnet-base-v2" model, generating embeddings from the data.
   - **Pinecone**: Storage of the generated embeddings, facilitating quick retrieval.
3. **Generation and Response Layer**:
   - **OpenAI GPT-4**: Access to the GPT-4 model for the generation of responses based on the information retrieved.
   - **AWS API Gateway and Lambda**: To handle user requests and coordinate calls to the Hugging Face and GPT-4 models.
4. **Frontend and User Layer**:
   - **AWS Amplify**: For the development and deployment of the user interface, providing an interactive and responsive experience.
5. **Security and Monitoring Layer**:
   - **AWS IAM**: Identity and access management to secure resources.
   - **AWS CloudWatch**: Monitoring and logging to maintain uptime and detect potential problems.

## Functionality of the RAG System



RAG sobre el REGLAMENTO (UE) 2024/1689 en Materia de Inteligencia Artificial

Bienvenido al RAG sobre el REGLAMENTO (UE) 2024/1689 en Materia de Inteligencia Artificial. ¿En qué puedo ayudarte?

What is an AI System according to the regulation?

According to the regulation outlined in the provided documents, an "AI system" is defined as follows: Definition: An AI system is a machine-based system designed to operate with varying levels of autonomy. It is capable of demonstrating adaptive behavior after deployment. The system infers from the input information it receives to generate output results, such as predictions, content, recommendations, or decisions. These outputs can influence physical or virtual environments. Key Characteristics: Machine-Based: The system is based on machine operations. Autonomy: It functions with different levels of autonomy, meaning it can operate independently to some extent. Adaptability: The system can adapt its behavior post-deployment based on the input it receives. Inference: It infers from the input data to produce outputs. Outputs: The outputs can be in the form of predictions, content, recommendations, or decisions. Influence: These outputs have the potential to influence physical or virtual environments. Purpose: The AI system is designed to achieve explicit or implicit objectives by processing input information to generate meaningful outputs. This definition emphasizes the system's ability to function autonomously, adapt over time, and produce outputs that can significantly impact various environments. The regulation aims to ensure that such systems are developed and deployed responsibly, considering their potential effects on health, safety, and fundamental rights.

Recursos:

- REGLAMENTO (UE) 2024/1689 página [illegible]
- REGLAMENTO (UE) 2024/1689 página [illegible]
- REGLAMENTO (UE) 2024/1689 página [illegible]
- REGLAMENTO (UE) 2024/1689 página [illegible]
- REGLAMENTO (UE) 2024/1689 página 7
- REGLAMENTO (UE) 2024/1689 página [illegible]
- REGLAMENTO (UE) 2024/1689 página [illegible]

Escribe tus preguntas sobre el reglamento europeo en materia de IA...

Desarrollado por Pedro Albarracín García - Magnadatum Artificial Intelligence Labs

This RAG (Retrieval-Augmented Generation) system is designed to facilitate the consultation and understanding of the **European Artificial Intelligence Regulation (AI Act)**, providing an interactive and efficient tool to access relevant information from the regulation and other related documents.

**Main functionality:**

1. **Intelligent Consultation of the IA Act**:

   The system allows users to ask specific questions about the content of the regulation, providing detailed and accurate answers. These answers are generated through a combination of advanced linguistic processing and extraction of key content from the document.

2. **References to Original Documentation**:

   Each response includes direct links to the relevant sections of the regulation or other documents. This allows users to verify the information provided or explore the full context of the response.

3. **Uploading and Indexing of Documents**:

   In addition to the AI Act, the system allows users to upload their own documents in PDF format. These documents are automatically indexed for integration into the system, expanding the range of possible queries and customising the accessible content.

4. **PDF file processing**:

Users can process multiple uploaded documents to be analysed together, allowing for complex questions covering information from multiple sources.

5. **User-Friendly and User-Oriented Interface**:

The interface is designed for ease of use, with clear options such as "**Upload PDFs**" and "**Process PDFs**", as well as a field for entering questions. The system generates comprehensive answers that not only resolve specific questions, but also help to understand key concepts.

**Benefits:**

**- Time saving**: Users do not need to read the entire regulation; they can obtain specific information instantly.

- **Reliability**: By linking directly to the original documentation, it ensures that information is always verifiable.

**- Adaptability**: The ability to integrate other documents makes it possible to customise the use of the system to suit specific needs.

- **Accessibility**: The design of the system is oriented to facilitate its use even for users without technical or legal experience.

Furthermore, this RAG system does not just answer the questions entered by the user, but uses advanced techniques to enrich the context and provide more accurate and relevant answers. This process is explained in detail below:

### Internal Generation of Contextualisation Questions

When the user enters a question into the system, the model performs additional processing which includes the following steps:

1. **Analysis of the User Question**:

The system analyses the initial question to identify the keywords, the central theme and the possible scope of the query. This includes detecting whether the question is generic, specific or ambiguous.

2. **Generation of Internal Questions**:

   Based on the user's question, the system automatically formulates **three additional questions** with the aim to:

   - **Broaden the context:** Explore related aspects that may be relevant to the response.
   - **Specify terms:** Disambiguate concepts in case the question is vague or imprecise.
   - **Fill possible gaps:** Ensure that important information related to the topic is not omitted.

For example, if the user asks *"How does the AI Act affect autonomous systems?"*, the model could internally ask questions such as:

- *What defines the AI Act as an autonomous system?*
- *What specific obligations does the IA Act impose on these systems?*
- *What exceptions or flexibilities does the regulation offer for stand-alone systems?*

3. **Collection of Internal Responses**:

   Each of these internal questions is queried within the indexed content (uploaded documentation and regulations). This generates multiple pieces of information that are integrated into a single answer block.

4. **Generation of the Final Response**:

   The final response presented to the user combines:

   - Information derived directly from the initial question.
   - Additional context obtained from internal questions.
   - Specific references to the sections of the regulation or indexed documents that support the answer.

**Benefits of this approach:**

**- Increased Accuracy**: By exploring multiple dimensions of the topic, it ensures that the answer is as complete and relevant as possible.

- **Automatic Disambiguation**: If the original question is vague, internal questions help the model better interpret the user's intent.

- **Context Enrichment**: The user not only gets a direct answer, but also complementary information that can broaden their understanding.

**Practical Example of the Process:**

Suppose the user asks: *"What responsibilities do suppliers have under the AI Act?"*, the model will internally generate questions such as:

- *What are the key definitions of a supplier under the regulation?*
- *What specific responsibilities apply to suppliers during the lifecycle of the IA system?*
- *What sanctions apply if these responsibilities are not fulfilled?*

In this way, the system not only answers the user's question, but ensures that the answer includes critical details and is supported by the relevant sections of the regulation.

This innovative approach ensures that the user obtains high quality information, optimising both the accuracy and usefulness of the answers.

In short, this RAG acts as an expert virtual assistant on the European AI Act, offering an interactive, efficient and verifiable way to access regulations and other supporting documentation. It is an ideal tool for legal professionals, technicians and managers interested in understanding and applying the European AI regulation.