Generative AI

Quantum Insights ⚡

In constantly evolving fields, it is essential to understand what are the main scientific challenges of the moment. This project aims to explore large language models accessible in open-source to *synthesize, summarize, translate and popularize research work*, through the design of a RAG model specialized in the fields of quantum physics research.

Key Features

Research

Extraction of information on scientific research from ArXiv.org

Synthesis

Exploration of topics, creation of explanatory summaries, synthesis, comparison, etc..

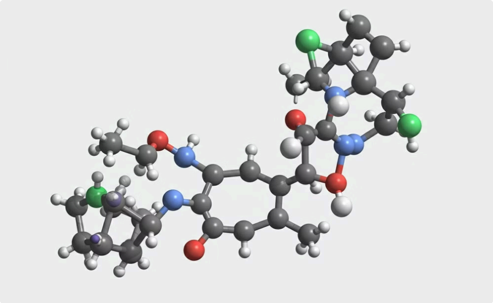
Code Extraction

Identification and explanation of code snippets.

Popularization

Explanations adapted to all levels (from beginner to expert)

Making Science Accessible



Expert Level

Technical details for the most experienced.

Advanced Level

In-depth explanations for professionals in the field.

Intermediate Level

Key concepts for the uninitiated.

Beginner Level

Simple explanation for the general public.

Use cases

Appropriation

Introduction to **basic concepts** and **fundamental principles**.

Explanations of Historical and Recent Discoveries

Presentation of **major scientific breakthroughs**. Overview of the **latest advances** in the field.

Search for Industrial Applications

Exploration of **technological innovations** resulting from this science. Identification of **impacted industrial sectors**.

Learning Mathematical Conventions

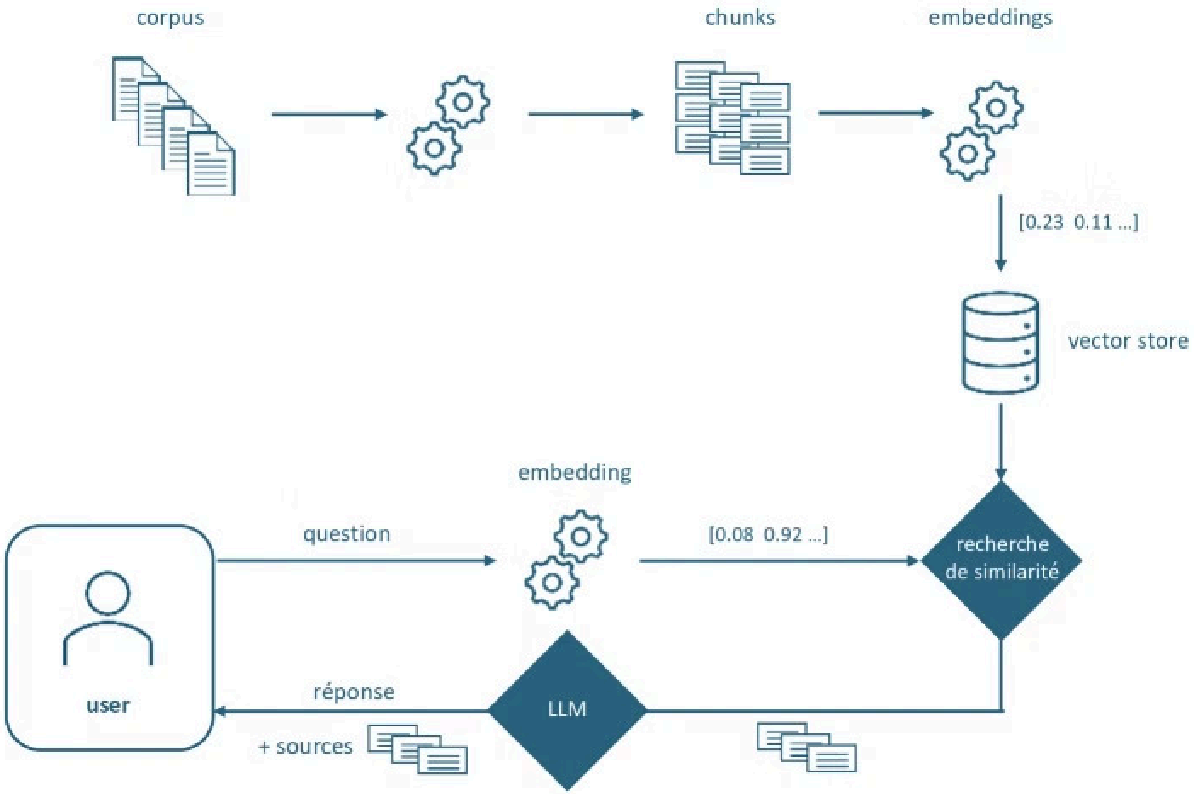
Explanation of **notations** and **formulas inherent** to the field.

Code Snippet Suggestions

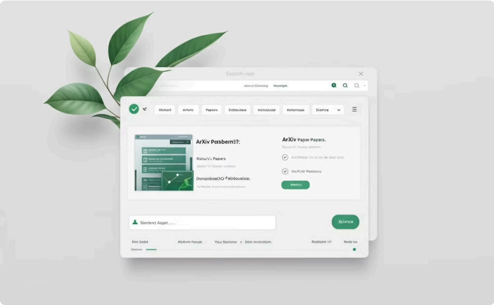
Practical examples to **illustrate the concepts**.

Information on **open-source libraries** and **cloud services** offered.

RAG Methodology



Technology Stack



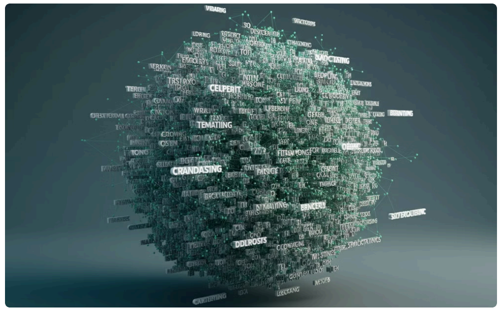
ArXiv Integration

Search and retrieval of the latest scientific articles.

ArXiv API

Document Processing

Structure detection via NLP SpaCy **`en_core_web_sm`** and sentiment analysis via DistilBERT



Embeddings

Generation of document embeddings.

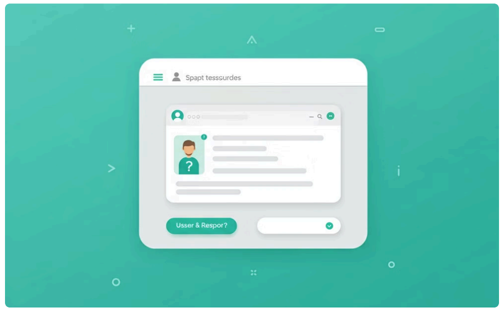
sentence-transformers/

all-MiniLM-L6-v2

Vector Store and RAG System

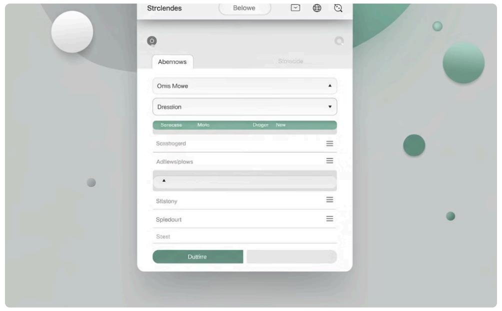
Storing embeddings and performing similarity search.

ChromaDB



Text-to-text

Text generation via LLM. Meta/Llama-2-7b-chat-hf



User Interface

UI/UX for a smooth interaction. Streamlit

CI/CD

Continuous integration and continuous deployment. GitHub Actions Docker

Cloud Deployment

Hosting and deployment of the application. AWS

Remaining work

1 **arXiv Search**

Improve the article search system.

(addressing all research areas)

2 **Conversation**

Improve the UX.

(user session, context window size, re-prompt ...)

3 **Translation**

Evolution of the chat in multiple languages.

(accessibility)

4 **Call for contributions**

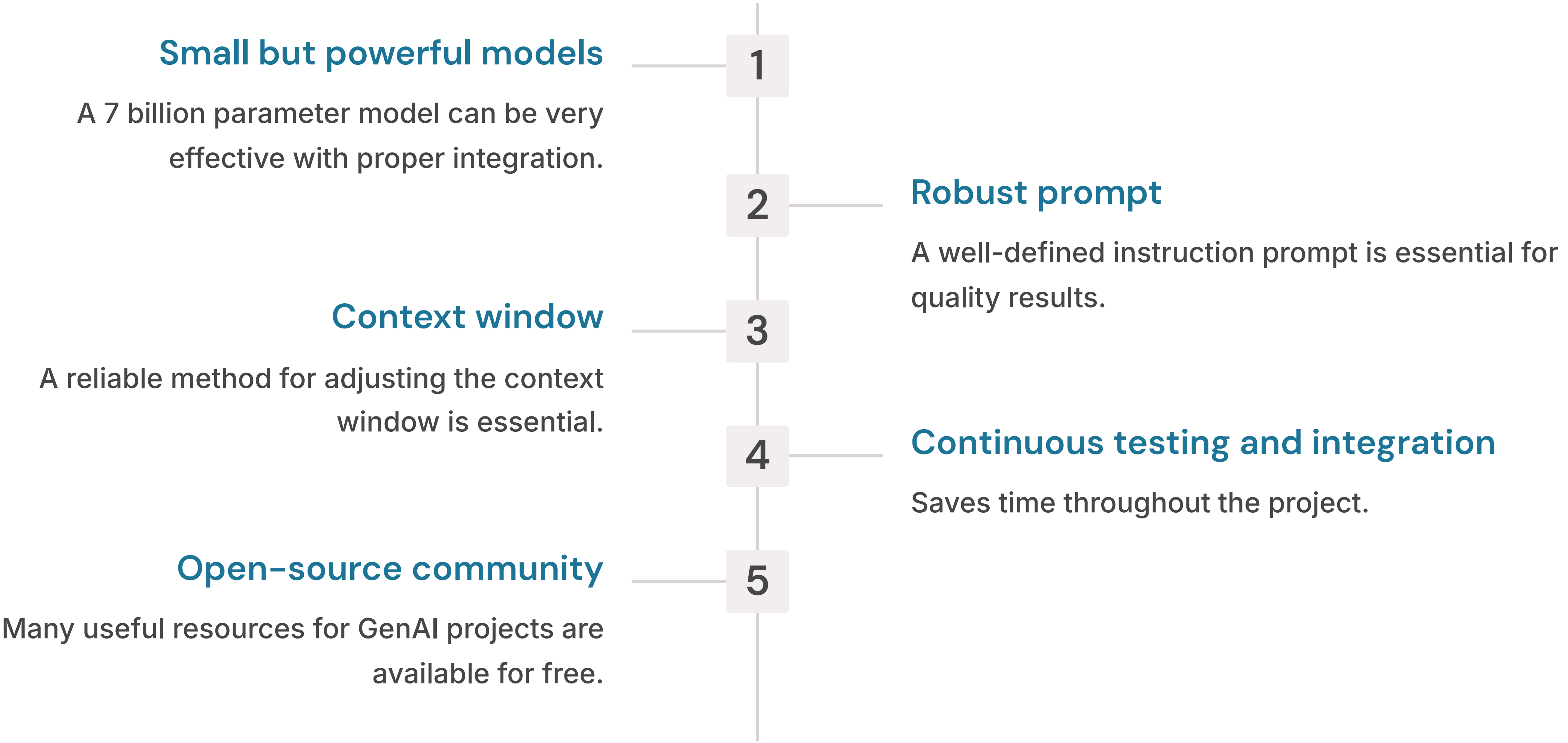
Organization of the repository in open-source.

(allowing contribution)

5 **Hosting**

Creation of a dedicated domain name.

Lessons Learned



1. **Small but powerful models**

   A 7 billion parameter model can be very effective with proper integration.

2. **Robust prompt**

   A well-defined instruction prompt is essential for quality results.

3. **Context window**

   A reliable method for adjusting the context window is essential.

4. **Continuous testing and integration**

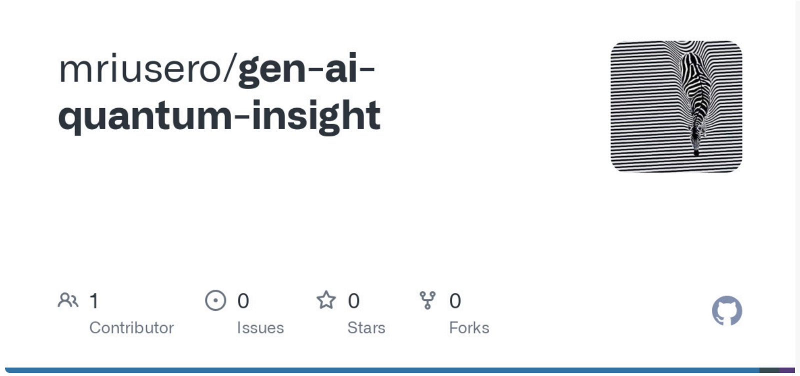
   Saves time throughout the project.

5. **Open-source community**

   Many useful resources for GenAI projects are available for free.

Sources

Repository



mriusero/**gen-ai-quantum-insight**

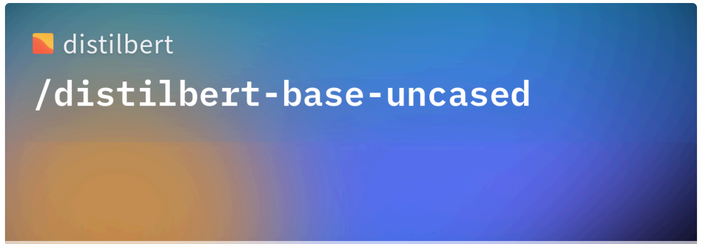
1 Contributor · 0 Issues · 0 Stars · 0 Forks

GitHub

GitHub - mriusero/gen-ai-quantum-insight

Contribute to mriusero/gen-ai-quantum-insight development by creating an account on GitHub.

Open-source models

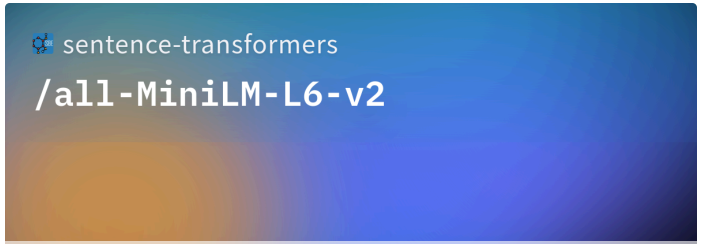


distilbert

/distilbert-base-uncased

distilbert/distilbert-base-uncased · Huggin...

huggingface.co

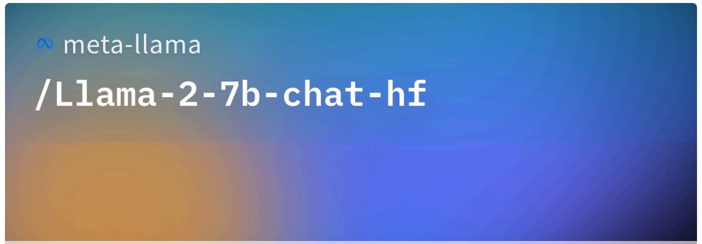


sentence-transformers

/all-MiniLM-L6-v2

sentence-transformers/all-MiniLM-L6-v2 ·...

huggingface.co



meta-llama

/Llama-2-7b-chat-hf

meta-llama/Llama-2-7b-chat-hf · Hugging...

huggingface.co