

RAG

?????? - Retrieval Augmented Generation

RAG ?????????????LLM????????????????/???hallucination???????RAG
?????????retrieval?????????generation???????????????????????????????????? LLM
????????????????

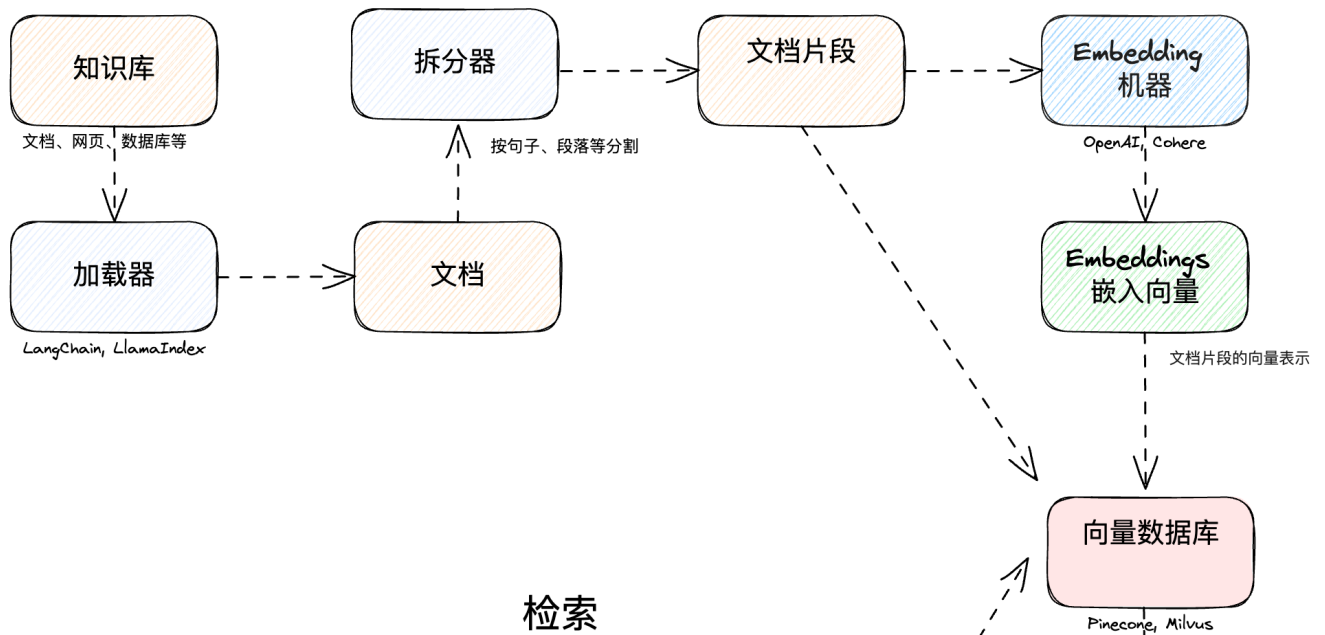
RAG ???

- ?? AI ??
- ?????????
- ???????
- ???????

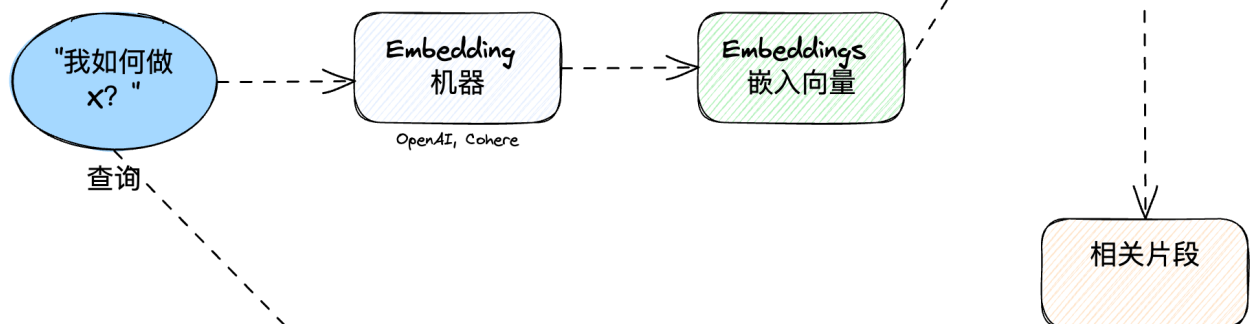
??????

Retrieval-Augmented Generation 检索增强生成

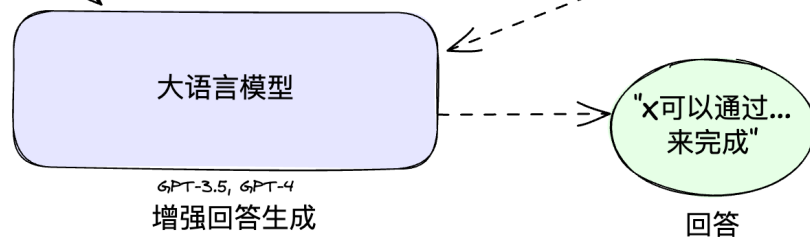
索引



检索



生成



Tutorials

- [ollama + Langchain + Gradio RAG ?????](#)

- [A flexible Q&A-chat-app for your selection of documents with langchain, Streamlit and chatGPT | by syrom | Medium](#)
- [????4?????AI?????ChatGPT?????|???? BusinessNext \(bnext.com.tw\)](#)
- [????PDF Chatbot with Llama3 & RAG?? #chatbot #chatgpt #llama3 #rag #chatpdf - YouTube](#)
- [?????https://github.com/Shubhamsaboo/awesome-llm-apps](#)
- [Easy AI/Chat For Your Docs with Langchain and OpenAI in Python](#)
- [RAG????16?????RAG](#)
- [YT:RAG????16?????RAG - YouTube](#)
- [?? LLM ????-Day26-? Langchain ?? PDF ???? - iT ??:???????? IT ???? \(ithome.com.tw\)](#)
- [RAG????LangChain + Llama2 |?????LLM | by ChiChieh Huang | Medium](#)
- [Python RAG Tutorial \(with Local LLMs\): AI For Your PDFs - YouTube](#)
- [? PDF ??????????????](#)

Embedding/Rerank Models

- [???????](#)
- ??
 - [BCEmbedding](#)
 - HuggingFace: https://huggingface.co/maidalun1020/bce-embedding-base_v1
 - [BAAI](#)
 - [GTE](#)
- API Service
 - [Cohere](#) (Rerank)

Vector Databases

- [Qdrant](#) - ??????????????????????GUI?????
 - [???Qdrant????????????? – AI StartUps Product Information, Reviews, Latest Updates](#)
- [Chroma](#)
 - Doc: [? Getting Started | Chroma Docs](#)
 - [Chroma??????. ?????????????????????? | by Lemooljiang | Medium](#)
 - [Chroma with Docker](#)
- [VectorAdmin](#) - ?????????? (???????? OpenAI)

- GitHub: <https://github.com/Mintplex-Labs/vector-admin>
- YT: [VectorAdmin | The universal GUI for vector databases - YouTube](#)
- [VectorAdmin in Docker](#)
- [Pinecone](#) (Cloud)
 - [Introducing Pinecone Inference to streamline your AI workflow | Pinecone](#)
 - [multilingual-e5-large - Pinecone Docs](#)
- [Supabase](#) (Cloud)
- [Astra DB](#) (Cloud)
 - Doc: [Quickstart](#) | [Astra DB Serverless](#) | [DataStax Docs](#)

Advanced RAG

- [RAG ??? | 7 ?????? | ??? LLM. ?? LLM + RAG ?????????????? RAG ??... | by ChiChieh Huang | Medium](#)
- ReRank
 - [RAG ?????ReRank????????????? | DataAgent](#)
- [Advanced RAG: MultiQuery and ParentDocument | RAGStack | DataStax Docs](#)
- [Advanced Retrieval With LangChain \(ipynb\)](#)
- [Advanced RAG Implementation using Hybrid Search, Reranking with Zephyr Alpha LLM | by Nadika Poudel | Medium](#)
- [Five Levels of Chunking Strategies in RAG | Notes from Greg's Video | by Anurag Mishra | Medium](#)
- Chunking/Splitting
 - [Mastering RAG: Advanced Chunking Techniques for LLM Applications - Galileo \(rungalileo.io\)](#)
 - [5 Levels Of Text Splitting \(ipynb\)](#)
 - [??] [Semantic Chunking](#)
 - [???????? RAG ? Chunking ?????](#)
 - [Chunking Evaluation](#)
 - Online Tools
 - [Online Text Splitter](#)
 - [ChunkViz](#)
- [Advanced RAG: Query Expansion](#)

RAG Projects

- [Dot](#)

- [ragapp](#)
- [RAGFlow](#)
 - YT: [RAGFlow???????? - YouTube](#)
- [R2R](#)
- [Easy-RAG](#)
- [Langchain-Chatchat](#)
- [kotaemon](#)

Danswer

[Danswer](#) is the AI Assistant connected to your company's docs, apps, and people. Danswer provides a Chat interface and plugs into any LLM of your choice. Danswer can be deployed anywhere and for any scale - on a laptop, on-premise, or to cloud.

- GitHub: <https://github.com/danswer-ai/danswer>
- Doc: <https://docs.danswer.dev/introduction>

Embedchain

Embedchain streamlines the creation of personalized LLM applications, offering a seamless process for managing various types of unstructured data.

- Doc: [? Quickstart - Embedchain](#)
- GitHub: <https://github.com/embedchain/embedchain>

GraphRAG

??GraphRAG
 ???

- [Get Started](#)
- GitHub: <https://github.com/microsoft/graphrag>
- YT: [Microsoft GraphRAG | ????????RAG???????????????? - YouTube](#)
- GitHub: [GraphRAG Local with Ollama and Gradio UI](#)
- YT: [????RAG?GraphRAG?????????Gemma 2+Nomic
 Embed?????????GraphRAG+Chainlit+Ollama??? #graphrag #ollama #ai - YouTube](#)
- GitHub: [GraphRAG + AutoGen + Ollama + Chainlit UI = Local Multi-Agent RAG
 Superbot](#)

neo4j

- Doc: [GenAI Ecosystem - Neo4j Labs](#)
- ???: [??? AI ??????GraphRAG ???????????? LLM ???? | T?? \(techbang.com\)](#)
- [NeoConverse - Graph Database Search with Natural Language - Neo4j Labs](#)
- LangChain: [Enhancing RAG-based application accuracy by constructing and leveraging knowledge graphs \(langchain.dev\)](#)
- [Build a Question Answering application over a Graph Database | ??? LangChain](#)
- LangChain: <https://neo4j.com/labs/genai-ecosystem/langchain/>
- <https://github.com/neo4j-labs/llm-graph-builder>
- ipynb:
https://github.com/tomasonjo/blogs/blob/master/llm/enhancing_rag_with_graph.ipynb

Verba

Verba is a fully-customizable personal assistant for querying and interacting with your data, either locally or deployed via cloud. Resolve questions around your documents, cross-reference multiple data points or gain insights from existing knowledge bases. Verba combines state-of-the-art RAG techniques with Weaviate's context-aware database. Choose between different RAG frameworks, data types, chunking & retrieving techniques, and LLM providers based on your individual use-case.

- Github: [Retrieval Augmented Generation \(RAG\) chatbot powered by Weaviate](#)
- [Weaviate](#) is an open source, AI-native vector database
 - Doc: [Quickstart Tutorial | Weaviate - Vector Database](#)
- Video: [Open Source RAG with Ollama - YouTube](#)

PrivateGPT

- [Introduction – PrivateGPT | Docs](#)
- GitHub: <https://github.com/zylon-ai/private-gpt>
- Video: [PrivateGPT 2.0 - FULLY LOCAL Chat With Docs \(PDF, TXT, HTML, PPTX, DOCX, and more\) - YouTube](#)
- Video: [Installing Private GPT to interact with your own documents!! - YouTube](#)

LLMWare

The Ultimate Toolkit for Enterprise RAG Pipelines with Small, Specialized Models.

- [Home Ilmware | Ilmware \(ilmware-ai.github.io\)](#)
- GitHub: <https://github.com/ilmware-ai/ilmware>

talkd/dialog

Talkd.ai—Optimizing LLMs with easy RAG deployment and management.

- [talkd/dialog | dialog](#)
- GitHub: <https://github.com/talkdai/dialog>

RAG ??

?????Generation???

- ????Faithfulness?
?????? RAG
??
??
- ??????Answer Relevancy?
??
????????????????????????
- ??????Answer Correctness?
?????????????????????“????”??
??

?????Retrieval???

- ??????Context Recall?
??
????????????????????
- ??????Context Precision?
?????????RAG?????????????????????????????RAG????????????????????????
?????????RAG??

URLs

- Ragas - [? Get Started | Ragas](#)
- [LLM Hallucination Index RAG Special - Galileo - Galileo \(rungalileo.io\)](#)