

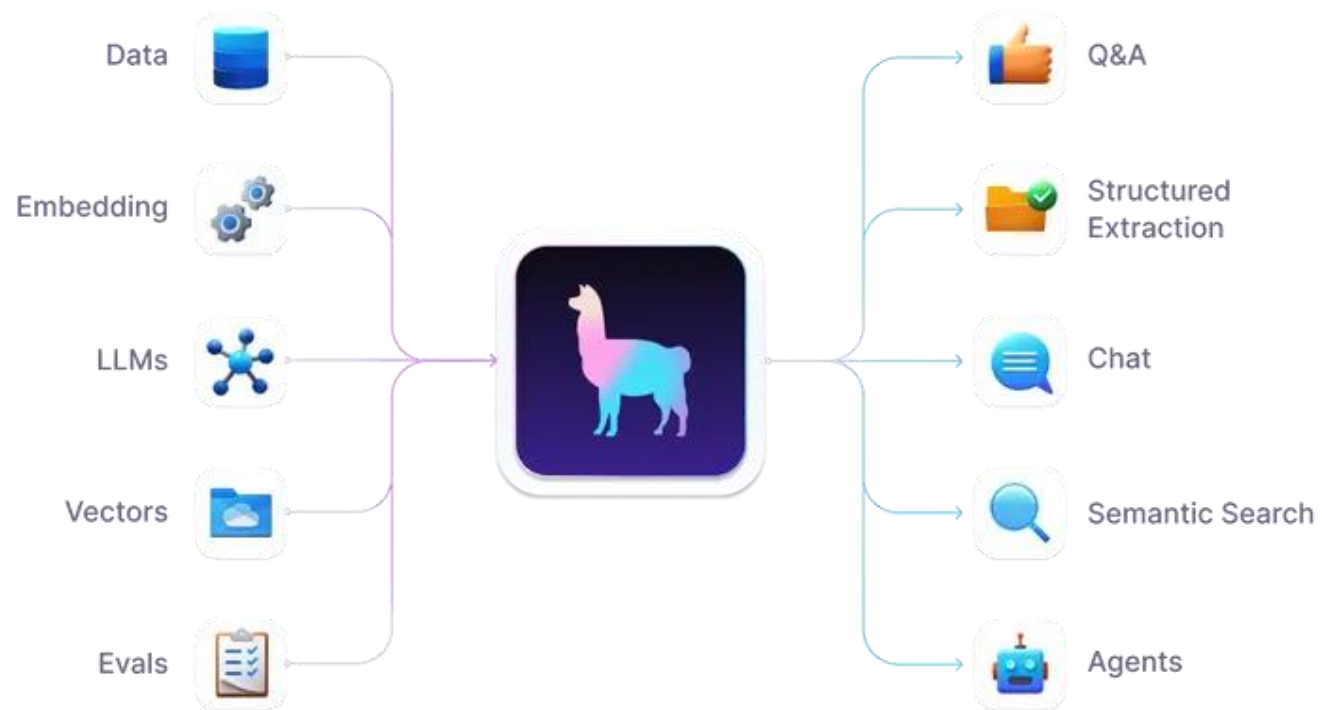
基于 LlamaIndex 构建 RAG 问答系统

智能架构 Kevin






LlamaIndex




Loading 1

Load in 160+ data sources and data formats, from unstructured, semi-structured, to structured data (API's, PDF's, documents, SQL, etc.)

 [EXPLORE MORE >](#)


Indexing 2

Store and index your data for different use cases. Integrate with 40+ vector store, document store, graph store, and SQL db providers.

 [EXPLORE MORE >](#)


Querying 3

Orchestrate production LLM workflows over your data, from prompt chains to advanced RAG to agents.

 [EXPLORE MORE >](#)

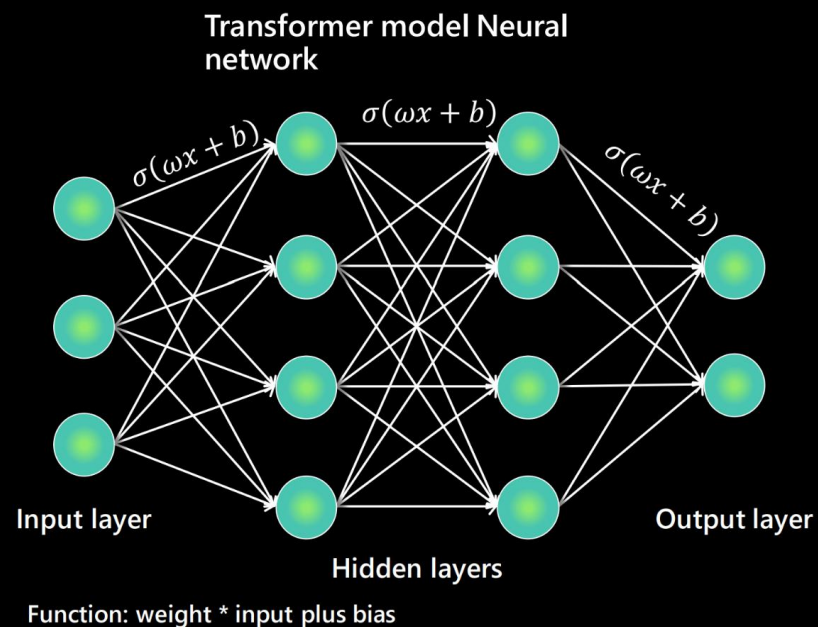
Evaluating 4

Evaluate the performance of your LLM application with a comprehensive suite of modules. Measure retrieval and LLM response quality. Effortlessly integrate with observability partners.

 [EXPLORE MORE >](#)

The inherent limitations of LLM

How large are they?



BERT Large - 2018

345M

GPT2 - 2019

1.5B

GPT3 - 2020

175B

Turing Megatron NLG
2021

530B

GPT4 - 2023

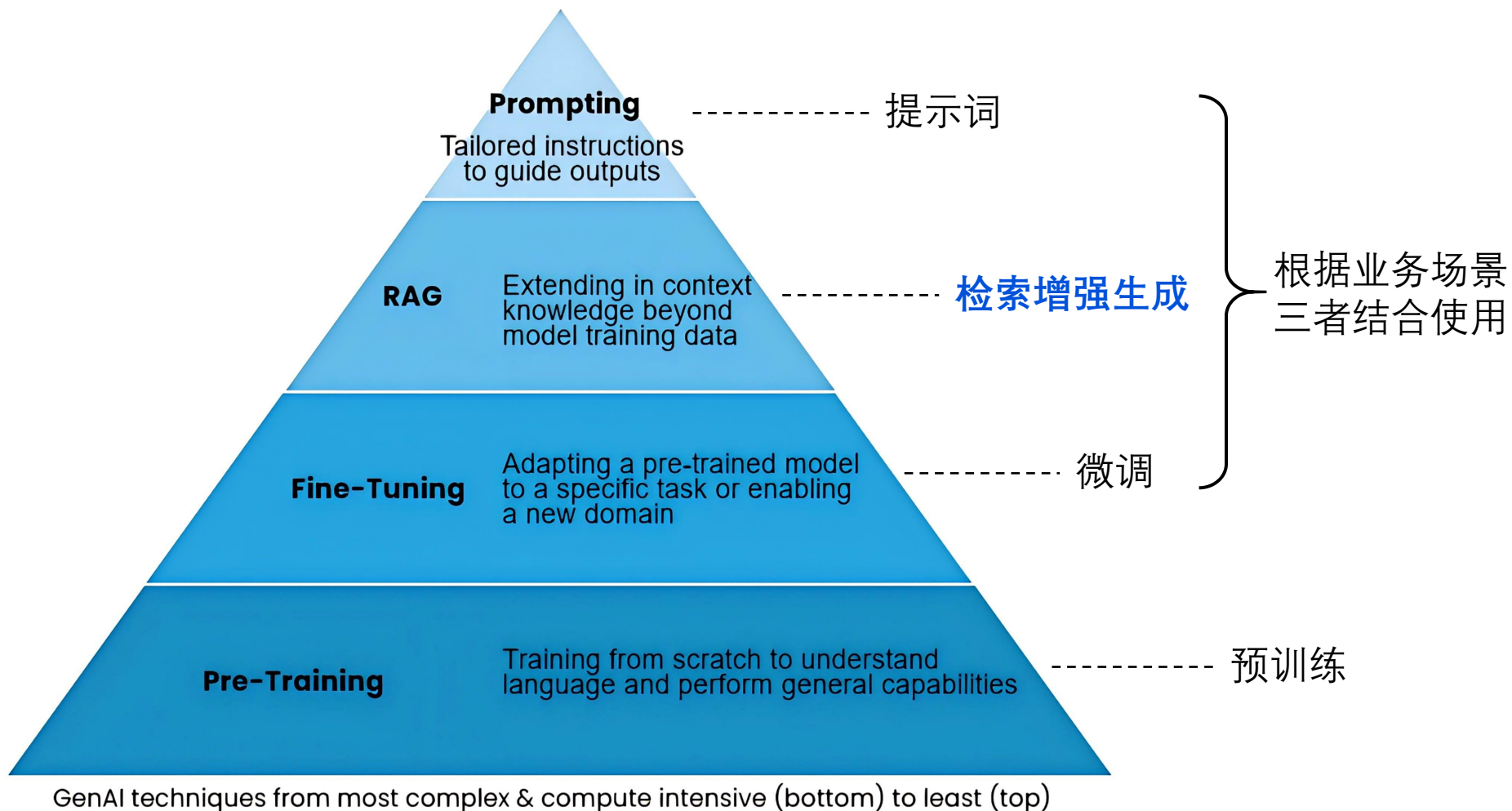
1.4T (estimated)

A large language model (LLM) is a type of AI that can process and produce natural language text. It learns from a massive amount of text data such as books, articles, and web pages to discover patterns and rules of language from them.

LLM Illusion

1. The knowledge of LLM is not up-to-date.
2. LLM may not know your private domain knowledge or business knowledge.

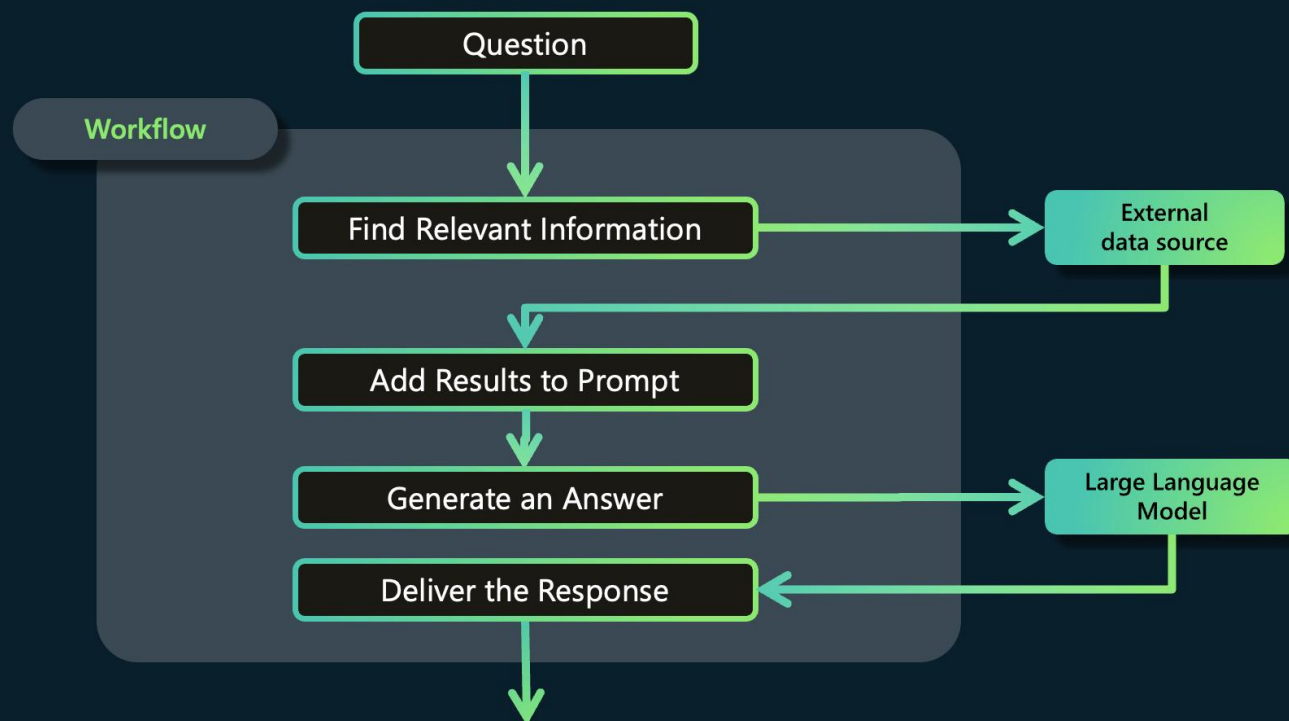
LLMs 应用落地核心技术



Add knowledge, In-Context-Learning



Retrieval Augmented Generation

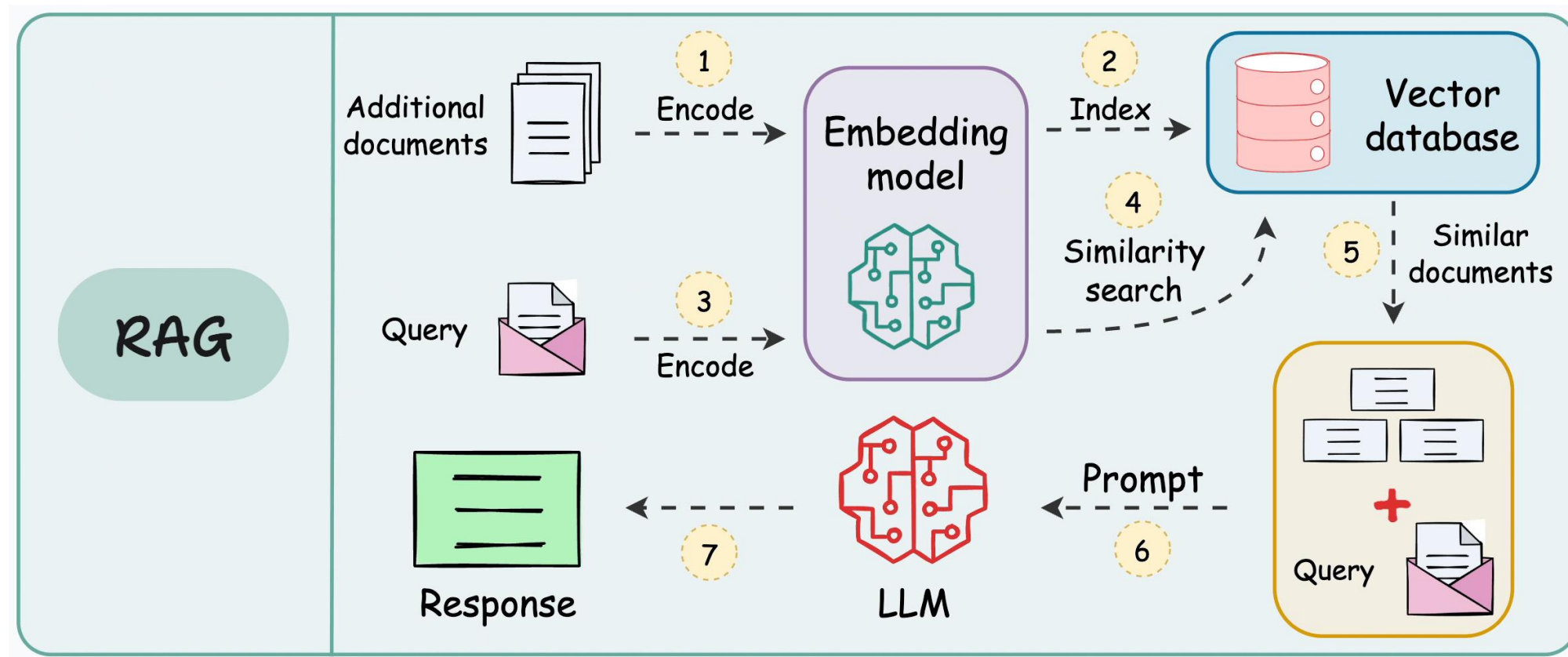


RAG combines a language model with a search system to provide more accurate and detailed information.

Here are the steps needed:

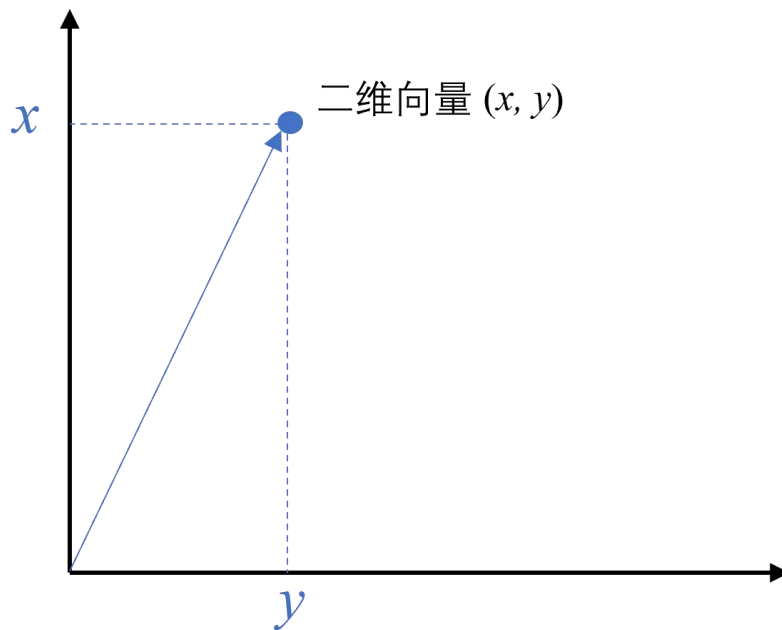
1. Ask a Question
2. Find Relevant Information
3. Choose the Best Bits
4. Generate an Answer
5. Deliver the Response

RAG Workflow

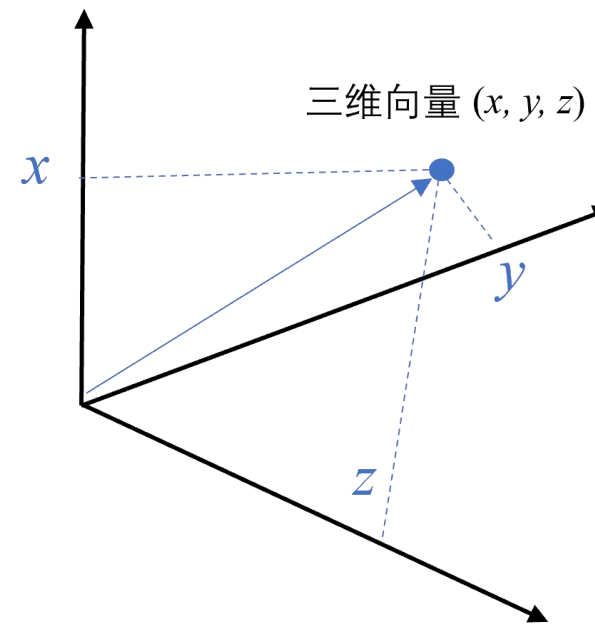


Vector

向量是一种有大小和方向的数学对象，它可以表示为从一个点到另一个点的有向线段。



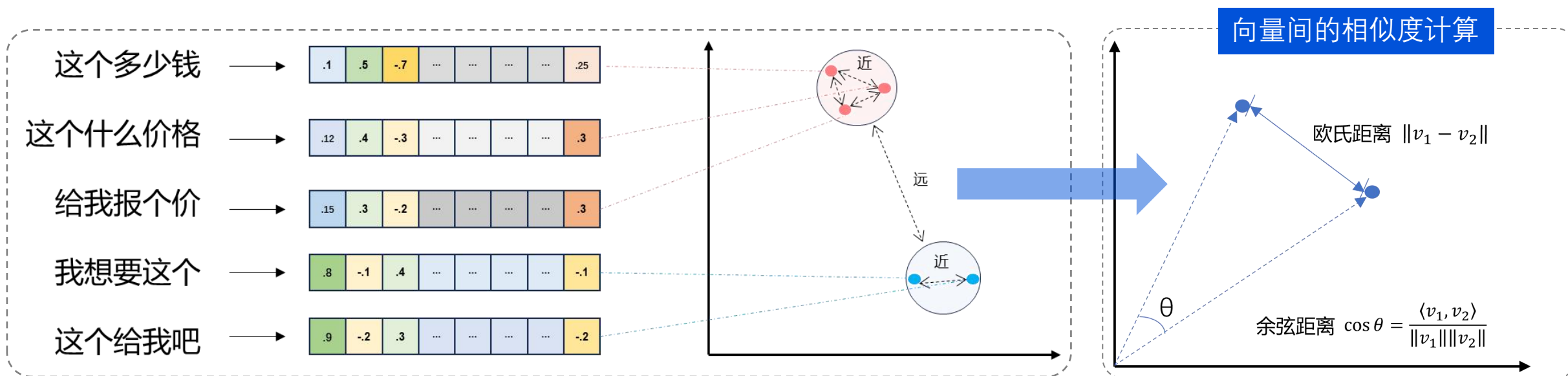
二维空间中的向量可以表示为 (x,y) ，表示从原点 $(0,0)$ 到点 (x,y) 的有向线段。



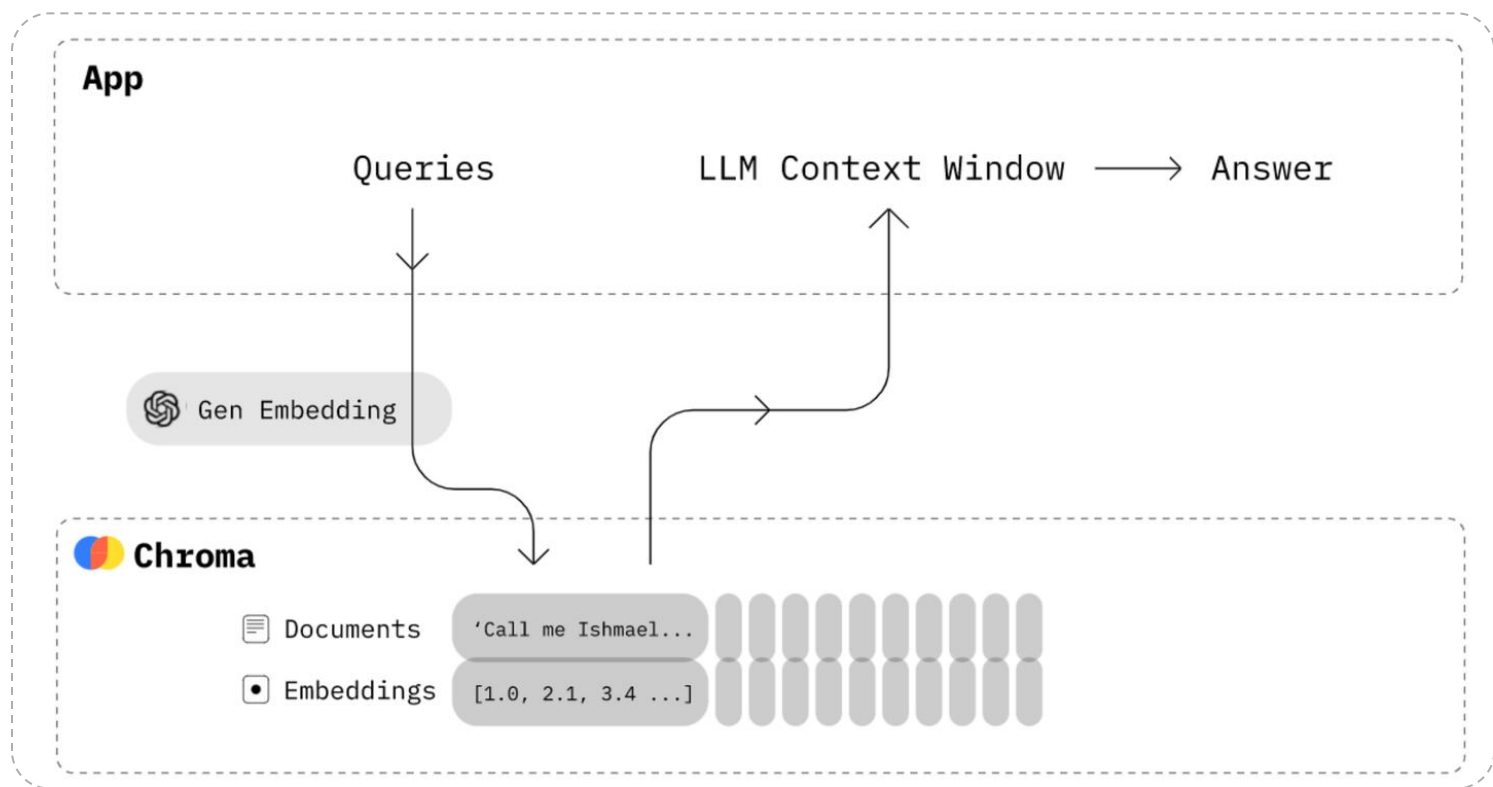
以此类推，可以用一组坐标 $(x_0, x_1, \dots, x_{N-1})$ 表示一个 N 维空间中的向量， N 叫向量的维度。

Embeddings

1. 将文本转成一组 N 维浮点数，即**文本向量**又叫 Embeddings
2. 向量之间可以计算距离，距离远近对应**语义相似度**大小



Vector Database



FAISS: Meta 开源的向量检索引擎

Chroma: 开源向量数据库, 同时有云服务

Pinecone: 商用向量数据库, 只有云服务

Milvus: 开源向量数据库, 同时有云服务

Weaviate: 开源向量数据库, 同时有云服务

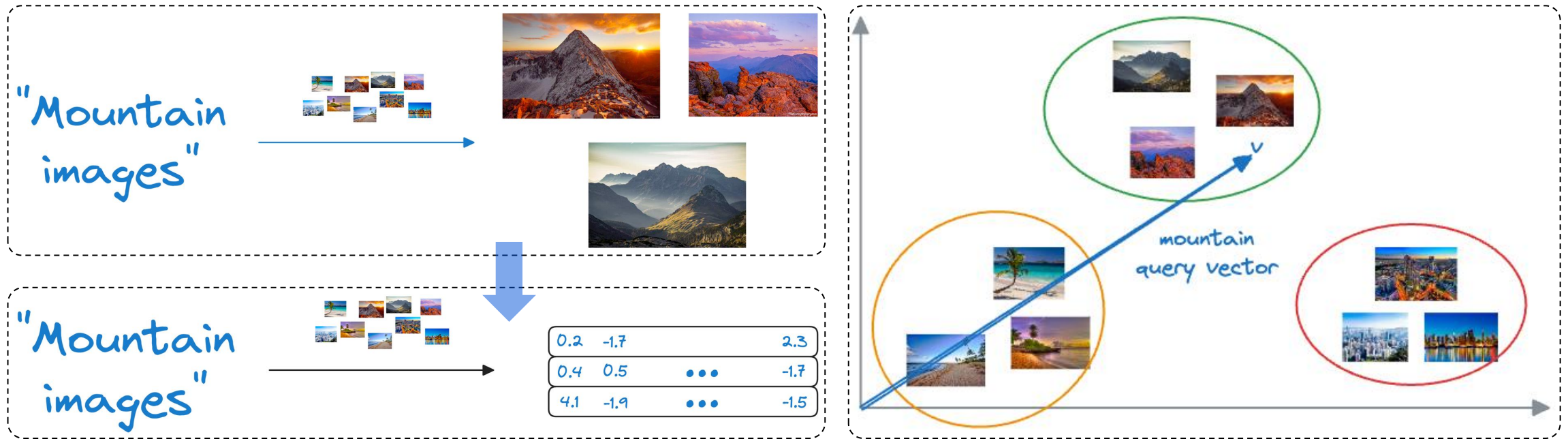
Qdrant: 开源向量数据库, 同时有云服务

PGVector: Postgres 的开源向量检索引擎

RediSearch: Redis 的开源向量检索引擎

ElasticSearch: 也支持向量检索

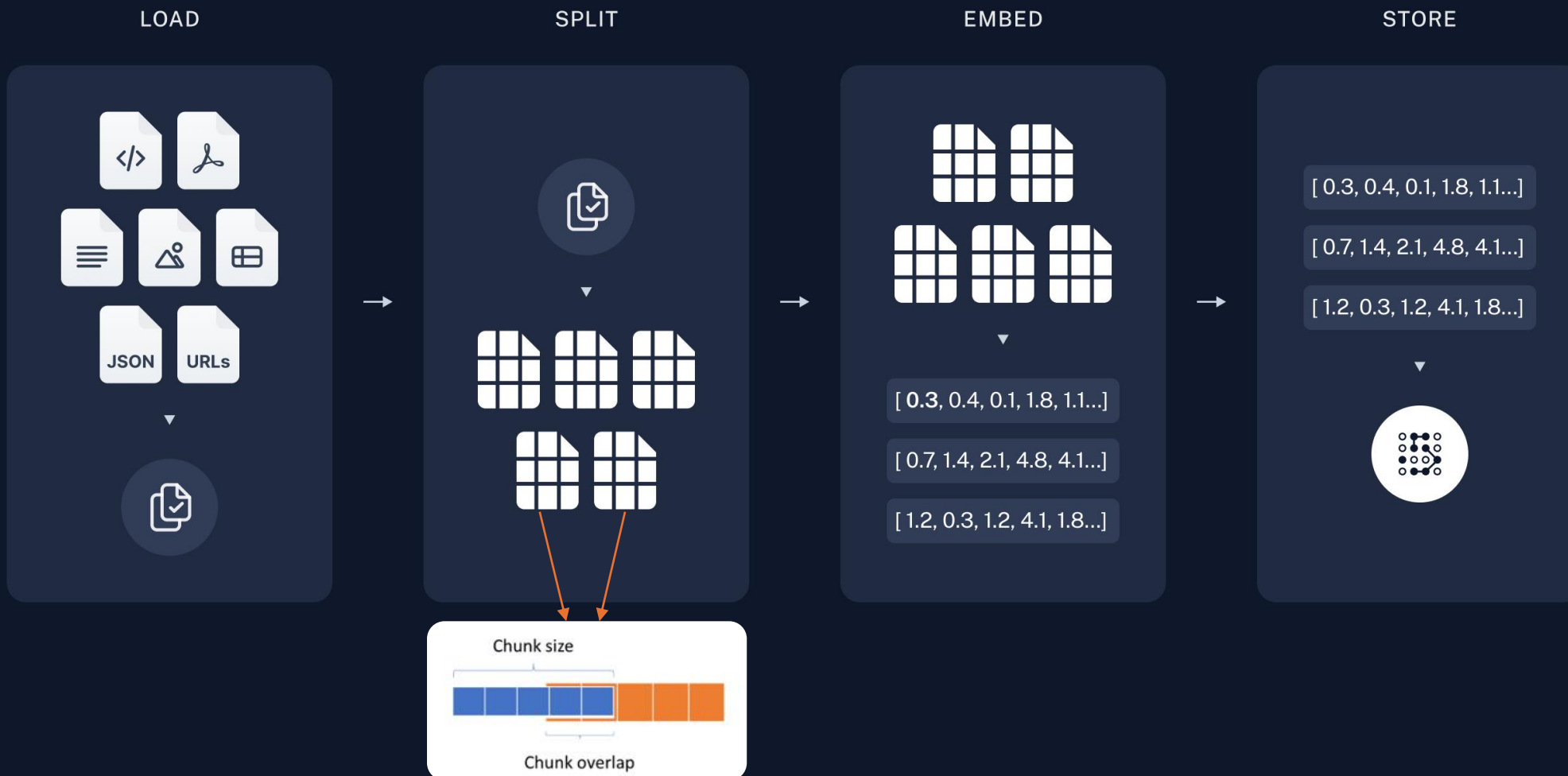
Querying a vector database



Approximate nearest neighbors (ANN)

Step 1 - Indexing

Offline processing



Split

- 简单文本处理

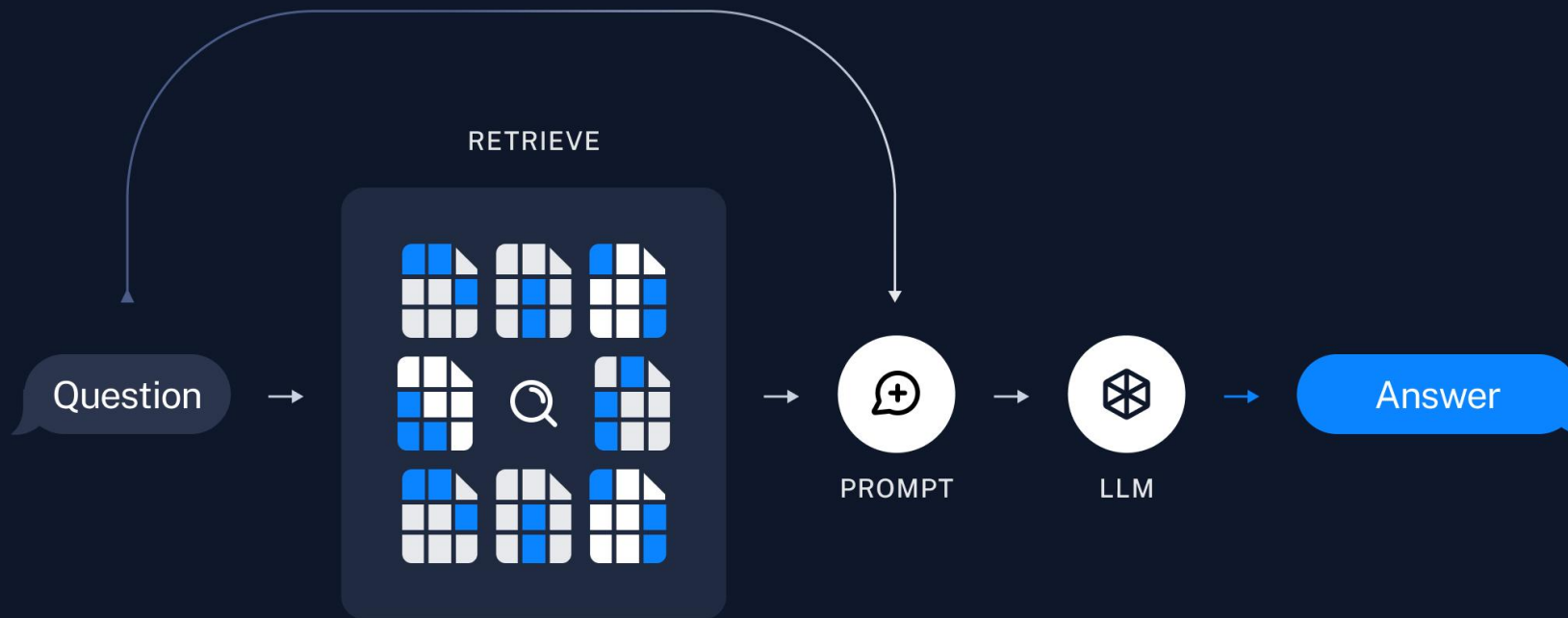
- LangChain `RecursiveCharacterTextSplitter`
 - `["\n\n", "\n", " ", ""]`

- 复杂文本处理

- 基于NLP篇章分析 (discourse parsing) 工具
 - 提取段落之间的主要关系, 把所有包含主从关系的段落合并成一段
- 基于BERT中NSP (next sentence prediction) 训练任务
 - 设置相似度阈值 t , 从前往后依次判断相邻两个段落的相似度分数是否大于 t , 如果大于则合并, 否则断开。

Step 2 - Retrieval & Generation

Online processing



Prompt template in RAG

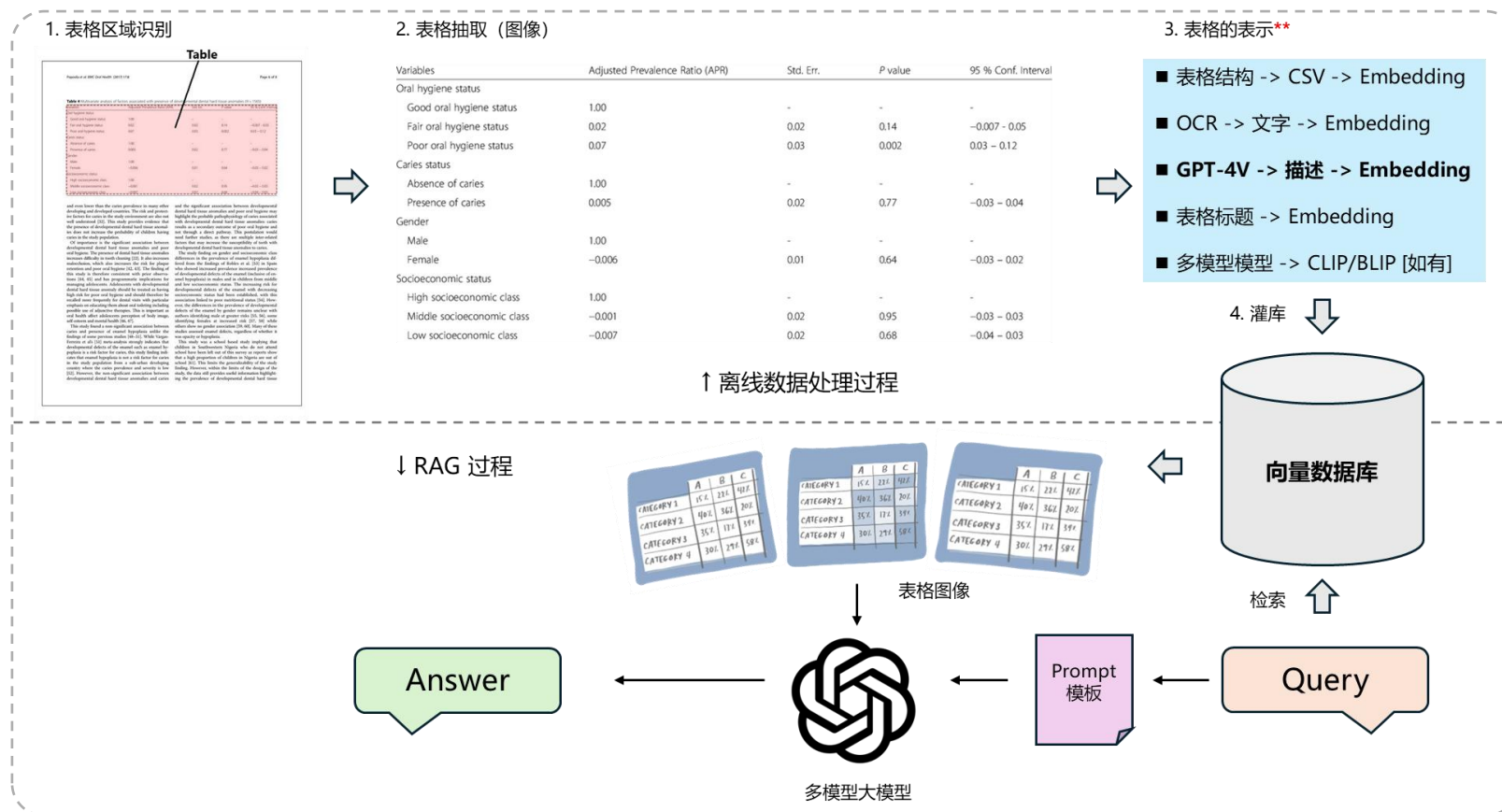
```
prompt_template = """  
你是一个问答机器人。  
你的任务是根据下述给定的已知信息回答用户问题。
```

```
  
已知信息：  
{context} # {context} 就是检索出来的文档
```

```
  
用户问：  
{question} # {question} 就是用户的问题
```

```
  
如果已知信息不包含用户问题的答案，或者已知信息不足以回答用户的问题，请直接回复"我无法回答您的问题"。  
请不要输出已知信息中不包含的信息或答案。  
请用中文回答用户问题。  
"""
```

PDF 中的表格怎么处理



处理步骤:

1. 将每页PDF转成图片
2. 识别文档（图片）中的表格
3. 基于 GPT-4 Vision API 做表格问答
4. 用 GPT-4 Vision 生成表格（图像）描述，并向量化用于检索

如何参与 AI 革命



使用 AI 产品

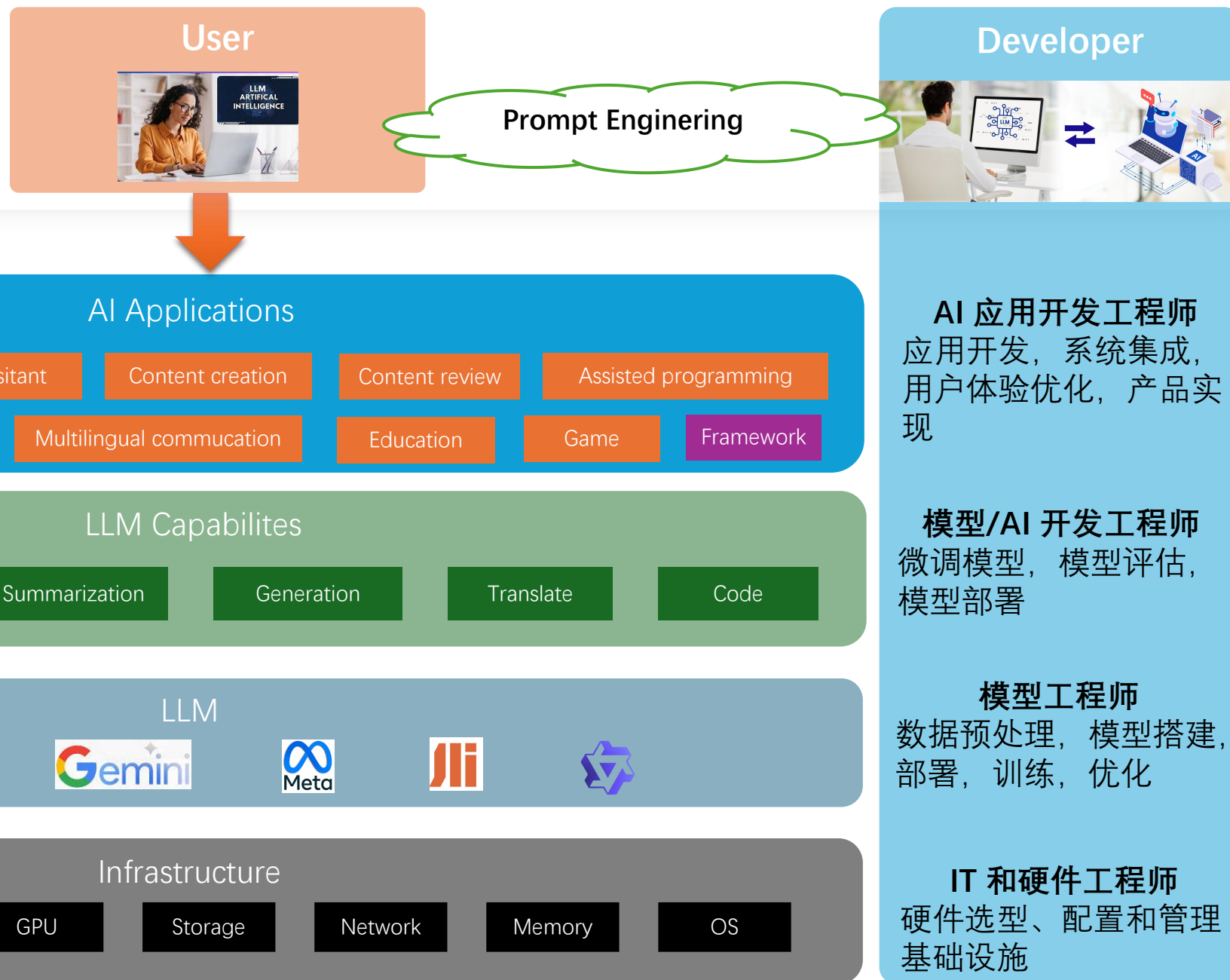
提供生产效率
节省人力
改变创造力



构建 AI 产品

构建 AI 能力
构建 AI 服务
开发 AI 应用

Use or Build





End



Thanks