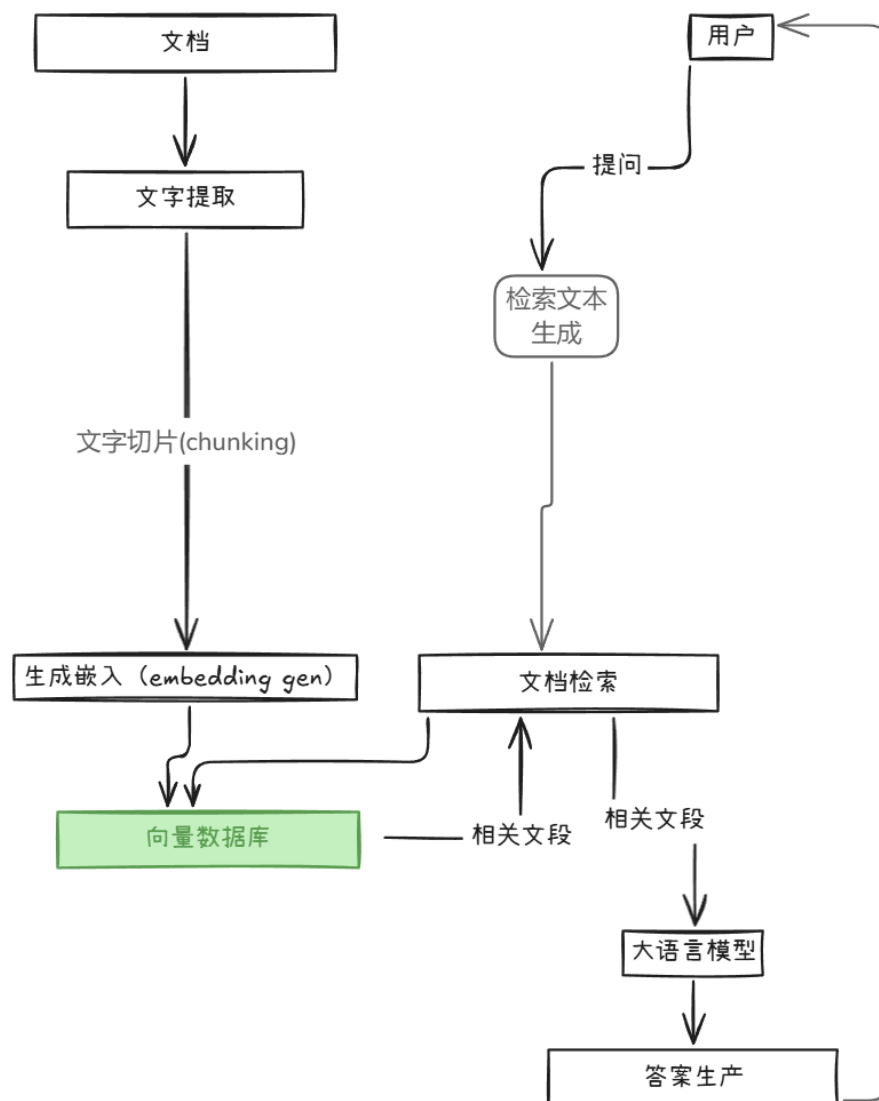


## 12.11 开始

慢慢进入 AI 时代了。基于文档的问答系统是很多高级智能体应用的基础。即使是基础的问答系统，在教育领域也有很大的实用价值。

因此，想做一个基于 rag 的问答系统。

系统架构如下：



因时间关系，选择尝试完成向量数据库部分。数据库可以增删查改向量；这一部分决

定了整个问答系统的效率和准确性。我们使用 c++ 实现。其他功能使用 python 实现。

使用 cpython 连接成一个整体，使得应用能够跨语言运行。

尝试拆解向量数据库。它的基本功能是增删查改。

面对的数据规模？考虑一个常见的情况，向量数据库存储一本教科书的文本。以严蔚敏《数据结构》（C 语言版）为例子，本书字数为？

**责任编辑：**范素珍

**责任印制：**王秀菊

**出版发行：**清华大学出版社

<http://www.tup.com.cn>

**地 址：**北京清华大学学研大厦 A 座

**邮 编：**100084

**社 总 机：**010-62770175

**邮 购：**010-62786544

**投稿与读者服务：**010-62776969, c-service@tup.tsinghua.edu.cn

**质 量 反 馈：**010-62772015, zhiliang@tup.tsinghua.edu.cn

**印 刷 者：**北京密云胶印厂

**装 订 者：**北京市密云县京文制本装订厂

**经 销：**全国新华书店

**开 本：**185×260 **印 张：**21.75 **字 数：**493 千字

**附光盘** 1 张

**印 次：**2011 年 5 月第 34 次印刷

**印 数：**598001~668000

**定 价：**30.00 元

---

**产品编号：**025648-01/TP

查看版权页，本书字数为 493 千字，即  $493 * 10^3 = 4.93 * 10^5$ 。

拟定使用 openai 的 text-embedding-3-large 模型生成 embedding，生成后的向量维度为 3072.

input\_max\_token 8192

dimension 3072

更多的技术参数。

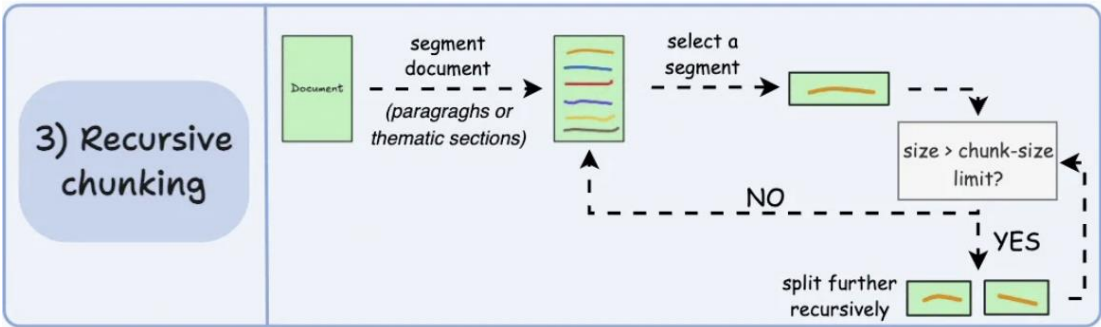
若使用阿里的 text-embedding-v3，有三种输出维度。

模型名称	向量维度	最大行数	单行最大处理Token 数	支持语种
text-embedding-v3	1024 768 512	6	8192	中文、英语、西班牙语、法语、葡萄牙语、印尼语、日语、韩语、德语、俄罗斯语等 50+语种

## Chunking 方法

拟定使用递归切分法，将文本切分成 chunks。

原理和示意图：



首先，基于内在的分隔符（如段落或章节）进行切分。

然后，如果某个切片的大小超过预定义的切片大小限制，就将其进一步分割。如果切片符合大小限制，则不再进行切分。

输出结果可能如下所示：

#### Paragraph 1

Artificial intelligence is transforming industries by automating processes, enhancing decision-making, and providing insights through data analysis. Machine learning, a subset of AI, enables systems to learn and improve from experience without explicit programming. Deep learning, a branch of machine learning, uses neural networks with multiple layers to model complex patterns in data.

#### Paragraph 2

AI is also improving natural language processing, enabling applications like chatbots and virtual assistants.

如上所示：

- 首先，我们定义了两个切片（紫色的两个段落）。
- 接下来，第 1 段被进一步分割成较小的切片。

与固定大小的切片不同，这种方法也保持了语言的自然流畅性，并保留了完整的思想。

不过，在实现和计算复杂性方面有一些额外的开销。

先实现一个 testbench。

我需要输入数据，以及样例输出。

为了验证系统的正确性，首先使用经过验证的暴力算法来生成对应输入的答案。

输入数据从哪儿来？

应该从实际的文档中来。