云原生向量数据库 PieCloudVector

助力多模态大模型 AI 应用

墨天轮

目录

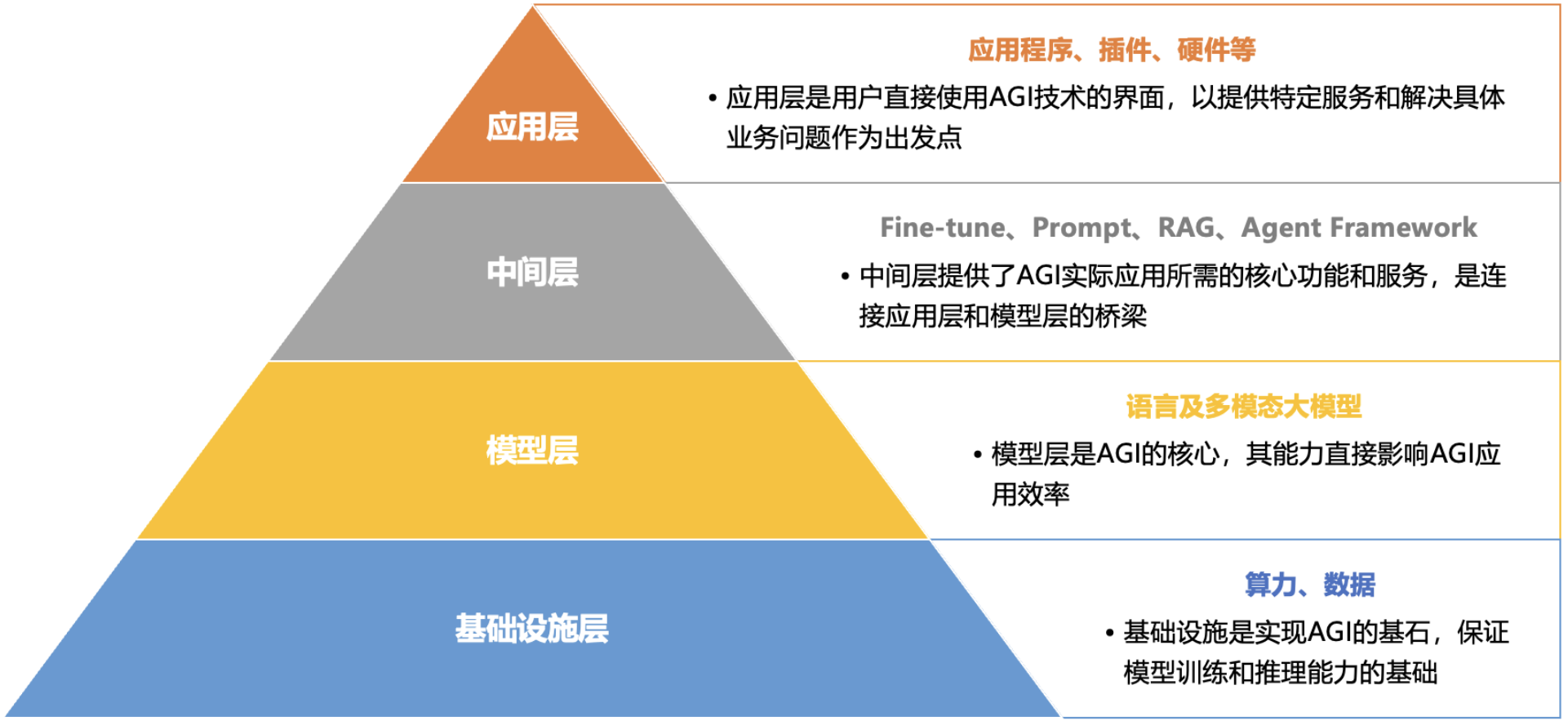
中国AGI市场增长趋势预测

- InfoQ研究中心预计，2030年中国AGI应用市场规模将达到4543.6亿元人民币。
- 2024-2027中国AGI应用市场将经历过速启动期;每年市场增速都将超过100%，2028年起，市场将进入快速成长期，年市场增速保持在50%以上。并于2027年突破千亿人民币市场规模。
- InfoQ研究中心认为，中国AGI应用市场规模发展将由企业市场引领主导，到2030年企业市场规模预计达到3024.6亿元人民币。

说明：数据来自InfoQ研究中心

国内AGI市场分层

- 中国AGI市场自下向上分为基础设施层、模型层、中间层和应用层四层，这四层结构共同构成了中国AGI市场的技术框架。



应用程序、插件、硬件等

- 应用层是用户直接使用AGI技术的界面，以提供特定服务和解决具体业务问题作为出发点

Fine-tune、Prompt、RAG、Agent Framework

- 中间层提供了AGI实际应用所需的核心功能和服务，是连接应用层和模型层的桥梁

语言及多模态大模型

- 模型层是AGI的核心，其能力直接影响AGI应用效率

算力、数据

- 基础设施是实现AGI的基石，保证模型训练和推理能力的基础

典型厂商

出门问问、商汤科技、美图、钉钉、WPS等

阿里巴巴、腾讯、网易、京东、字节跳动、OpenPie等

百度、阿里巴巴、科大讯飞、腾讯、智谱AI、百川智能、零一万物等

中国电信、中国移动、中国联通、华为、浪潮等

AI Agent推动AI迅速发展

- AI Agent正逐渐成为探索的核心路径。随着时间的推移，大模型的一些局限性开始显现，尽管大模型在模仿人类认知方面取得了显著进步，但要达到真正的通用智能，仍需克服重重困难。因此，AI Agent作为新的研究方向，开始受至越来越多的关注。

简化任务执行，拓宽大模型应用范围

AI Agent将简化用户与大模型的互动，允许用户只需指定目标，即可驱动大模型完成任务，人与AI将形成更紧密的合作体系。尽管AI Agent目前主要处理简单任务，但功能性正不断增强，AI Agent在多个领域的应用已经开始，预示着未来将成为AI应用层的基本架构。

个人助理	软件开发	交互式游戏	预测性分析
自动驾驶	智能客服	金融管理	智慧城市
任务管理	生产制造	文档处理	科学研发
市场营销	教育辅导	质量管理	...

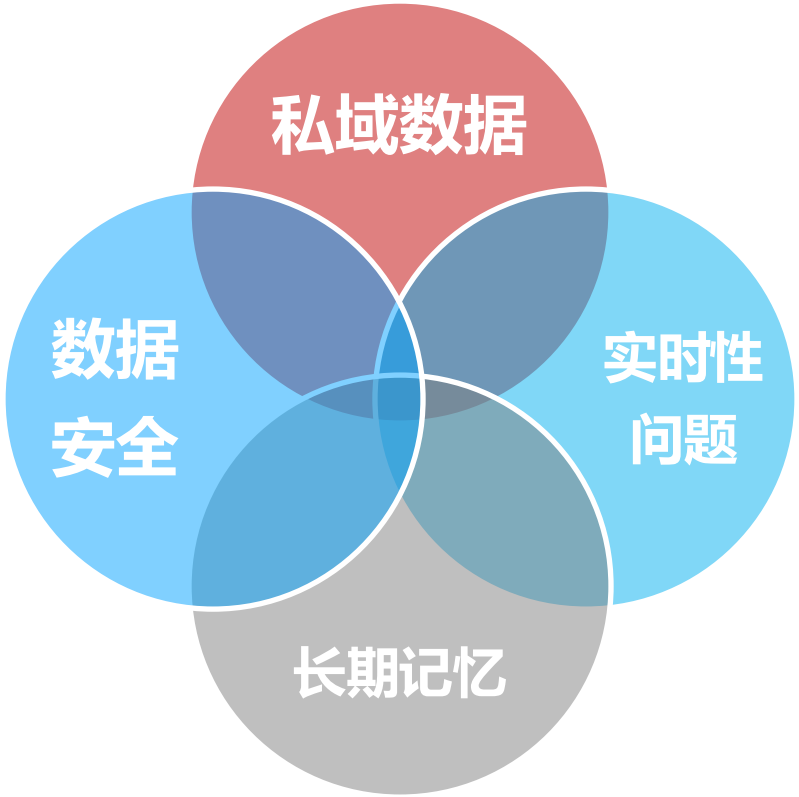
垂直行业是AI Agent最先实现应用的领域

在特定垂直领域中，AI Agent实现应用的优势主要集中于环境高度适应性。AI Agent依赖于对环境的反馈，因此企业环境中的特定场景为AI Agent提供了理想的应用背景，便于建立起对特定垂直领域的深入认知。

垂直行业专业知识	更易理解特定行业的术语、规则、任务、实践路径、甚至行业“黑话”，可提供高度定制化的解决方案。
合规性与规范	更易掌握特定行业的法规和标准，确保其应用的行业合规性相关要求。
业务流程理解	更易掌握并适应特定行业的业务流程，与企业的系统实现无缝集成。

大模型时代向量数据库的必要性

大模型在训练过程中并未接触过企业的私域数据和特定业务场景，因此，它们无法完全满足企业实际需求，也无法优化企业的具体业务流程。可以将其与企业内部的特定知识和数据进行整合。这种融合不仅降低了算力门槛，还大大提高了模型在特定应用场景中的准确性和可用性。

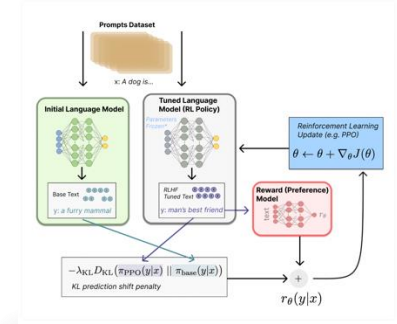


在很多应用场景中，特别是涉及敏感信息的企业应用，数据隐私是一个不可忽视的问题。通过在本地或专有云上部署大模型，并结合向量数据库，企业可以在不暴露任何敏感信息的前提下，充分利用模型的计算能力。

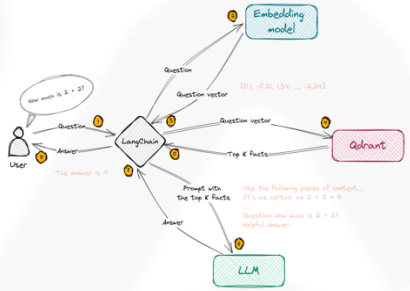
大模型通常需要处理海量的数据，如果不能实时更新和查询，其应用价值就会大打折扣。向量数据库通过其高效的索引和检索能力，可以实时地存储和更新模型的向量信息。这不仅大大提高了模型的响应速度，还使其能够准确地反映最新的数据状态。

传统的数据库由于其设计限制，难以支持模型的动态调整。而向量数据库则通过持久化存储向量信息，为大模型提供了一种形式的“长期记忆”。这使得模型能够根据历史数据和最新信息做出更加精准的预测和决策。

解决方法



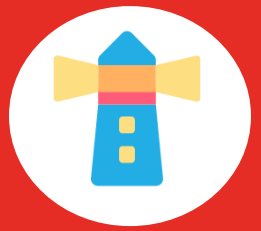
模型微调



基于向量数据库的检索

向量计算引擎的核心能力要求

文字，语音和图像通常会通过内嵌(embedding)操作转换成高维向量，如何快速而准确地对海量向量数据进行检索是一个巨大的技术挑战。向量数据库需要采用更高级的技术和算法来解决这一问题。

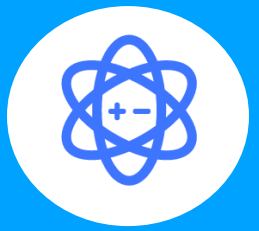


最近邻搜索

最近邻搜索（k-NN）是向量数据库要解决的核心问题，即在给定向量数据集中找到与之距离最小的K个向量。简单的全局搜索与向量维度和总数据量成正比，对于大数据集显然需要更高效的搜索方法。

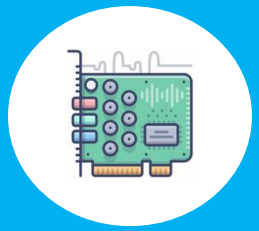
数据结构

在传统的数据库中，B树和哈希表是最常用的数据结构。然而，在向量数据库中，由于需要处理的向量数据通常是高维的，因此需要使用更加复杂的数据结构，如R树、M树等。这些数据结构能够更有效地组织和存储高维数据，从而提高检索效率。



近似检索

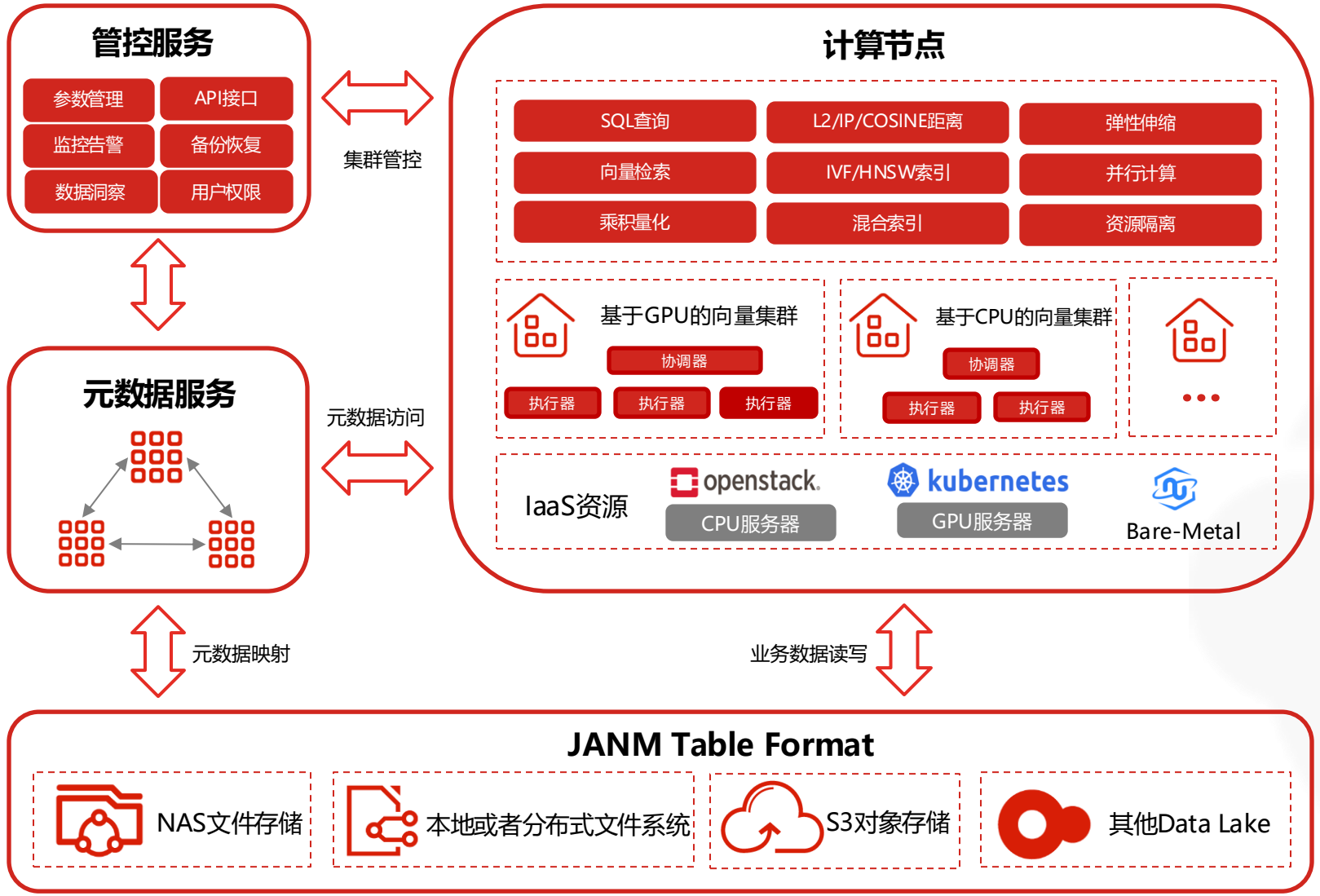
如果可以接受近邻索的精度(recall)有一定程度的损失，那么有一类算法可以大幅提升检索效率，这一类算法我们通常称为近似检索(ANN)算法，常见的如IVF/HNSW等，目前没有一个通用算法能在任意数据集上达到所有指标(recall/qps/内存)均最优，一般都需要做取舍以达到整体平衡。



硬件加速

把计算量非常大的工作分配给专门的硬件来处理以减轻CPU的工作负载，向量数据的计算可借助新硬件进行加速，如GPU、FPGA等，把常见的KNN/ANN算法、PQ算子、Index算法进行优化和集成，由专有硬件进行执行，做到从CPU的Offload。

云原生向量数据库PieCloudVector



基于postgres内核：

- 单机/分布式部署
- 完整ACID
- 向量标量混合查询
- SQL/REST/Python接口
- 兼容Langchain等主流框架

内置模型服务：

- 丰富的模型算法，可根据需求扩展
- 可集成LLM，如ChatGLM、LLaMA等

索引管理：

- 支持主流向量索引
- 索引缓存加速

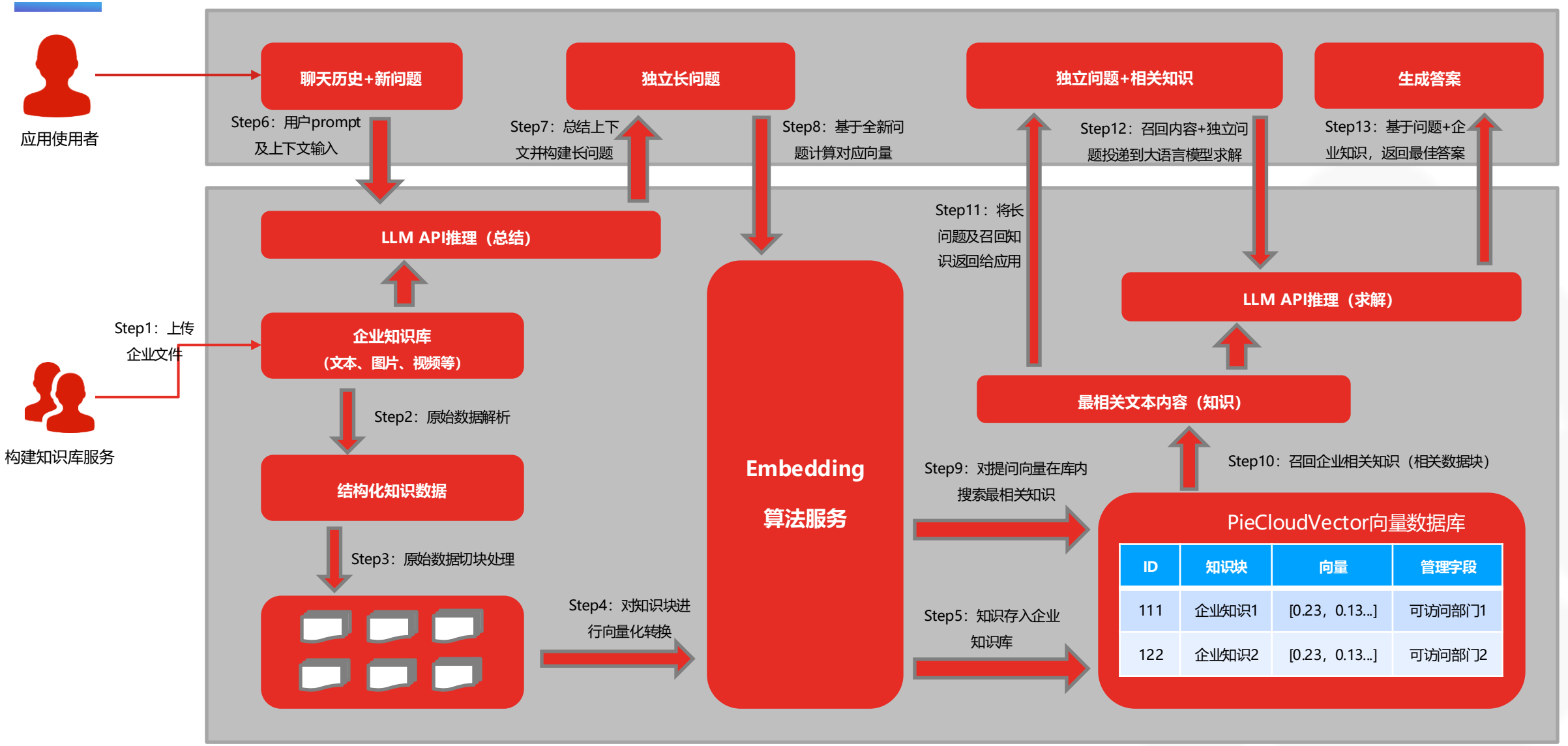
向量检索：

- 支持主流的ANN算法
- 近似向量搜索KNN-ANN，可牺牲部分精度加速搜索
- 支持CPU和GPU加速

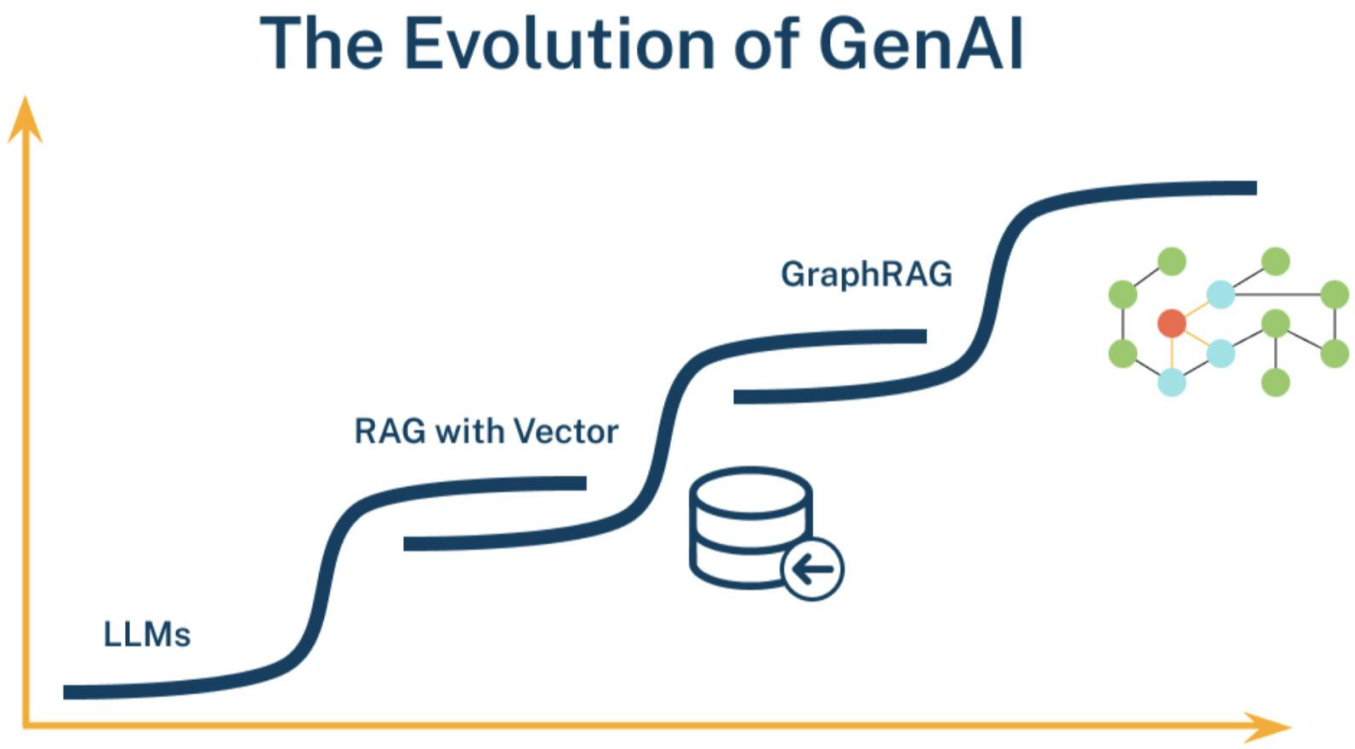
数据存储：

- 原始数据
- 向量数据
- 向量压缩
- 支持Json格式数据类型

RAG工作流程


应用使用者
聊天历史+新问题
Step6：用户prompt及上下文输入
独立长问题
Step7：总结上下文并构建长问题
Step8：基于全新问题计算对应向量
独立问题+相关知识
Step12：召回内容+独立问题投递到大语言模型求解
生成答案
Step13：基于问题+企业知识，返回最佳答案
LLM API推理（总结）
Step11：将长问题及召回知识返回给应用
LLM API推理（求解）
Step1：上传企业文件
企业知识库
（文本、图片、视频等）
构建知识库服务
Step2：原始数据解析
结构化知识数据
Step3：原始数据切块处理
Step4：对知识块进行向量化转换
Embedding
算法服务
Step9：对提问向量在库内搜索最相关知识
Step5：知识存入企业知识库
最相关文本内容（知识）
Step10：召回企业相关知识（相关数据块）
PieCloudVector向量数据库
ID
知识块
向量
管理字段
111
企业知识1
[0.23，0.13...]
可访问部门1
122
企业知识2
[0.23，0.13...]
可访问部门2

基于向量RAG问题



•知识库内容缺失：现有的文档其实回答不了用户的问题，系统有时被误导，给出的回应其实是“胡说八道”，理想情况系统应该回应类似“抱歉，我不知道”。

•TopK截断有用文档：和用户查询相关的文档因为相似度不足被TopK截断，本质上是相似度不能精确度量文档相关性。

•上下文整合丢失：从数据库中检索到包含答案的文档，因为重排序/过滤规则等策略，导致有用的文档没有被整合到上下文中。

•有用信息未识别：受到LLM能力限制，有价值的文档内容没有被正确识别，这通常发生在上下文中存在过多的噪音或矛盾信息时。

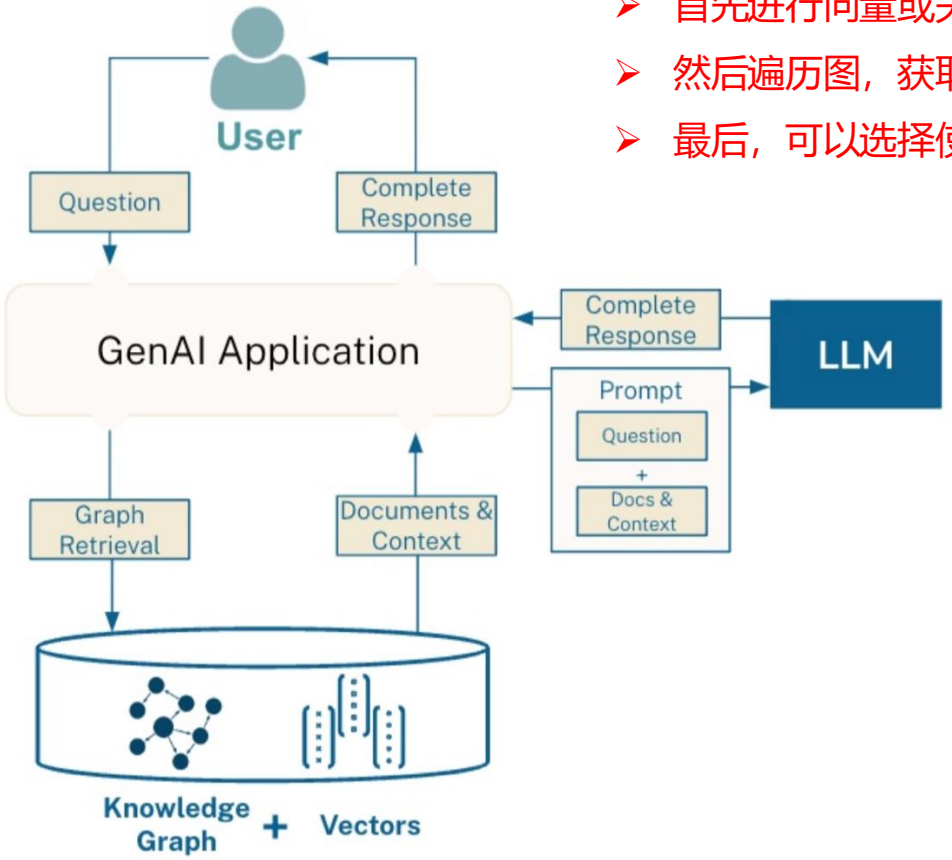
•提示词格式问题：提示词给定的指令格式出现问题，导致大模型/微调模型不能识别用户的真正意图。

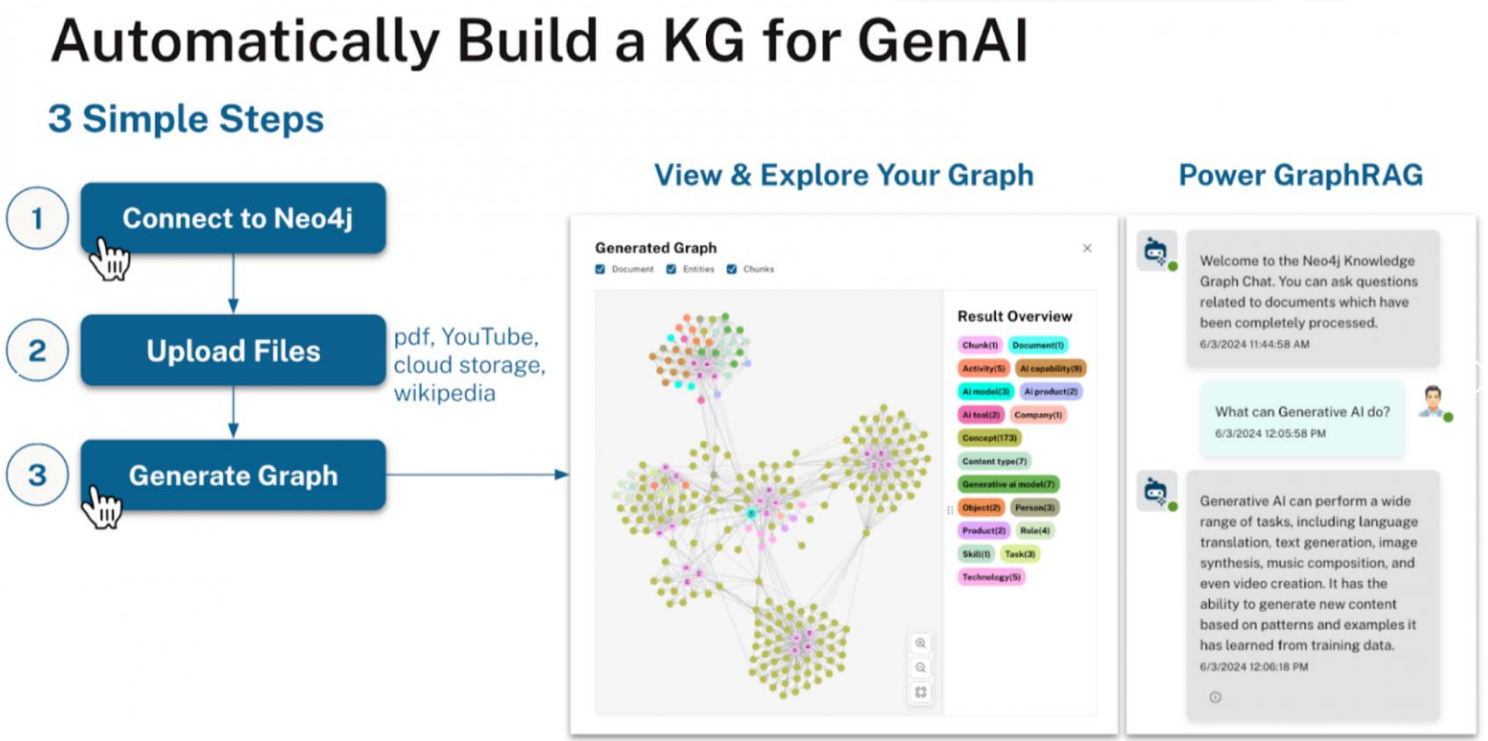
•准确性不足：LLM没能充分利用或者过度利用了上下文的信息，比如给学生找老师首要考虑的是教育资源的信息，而不是具体确定是哪个老师。另外，当用户的提问过于笼统时，也会出现准确性不足的问题。

•答案不完整：仅基于上下文提供的内容生成答案，会导致回答的内容不够完整。比如问“文档 A、B和C的主流观点是什么？”，更好的方法是分别提问并总结。

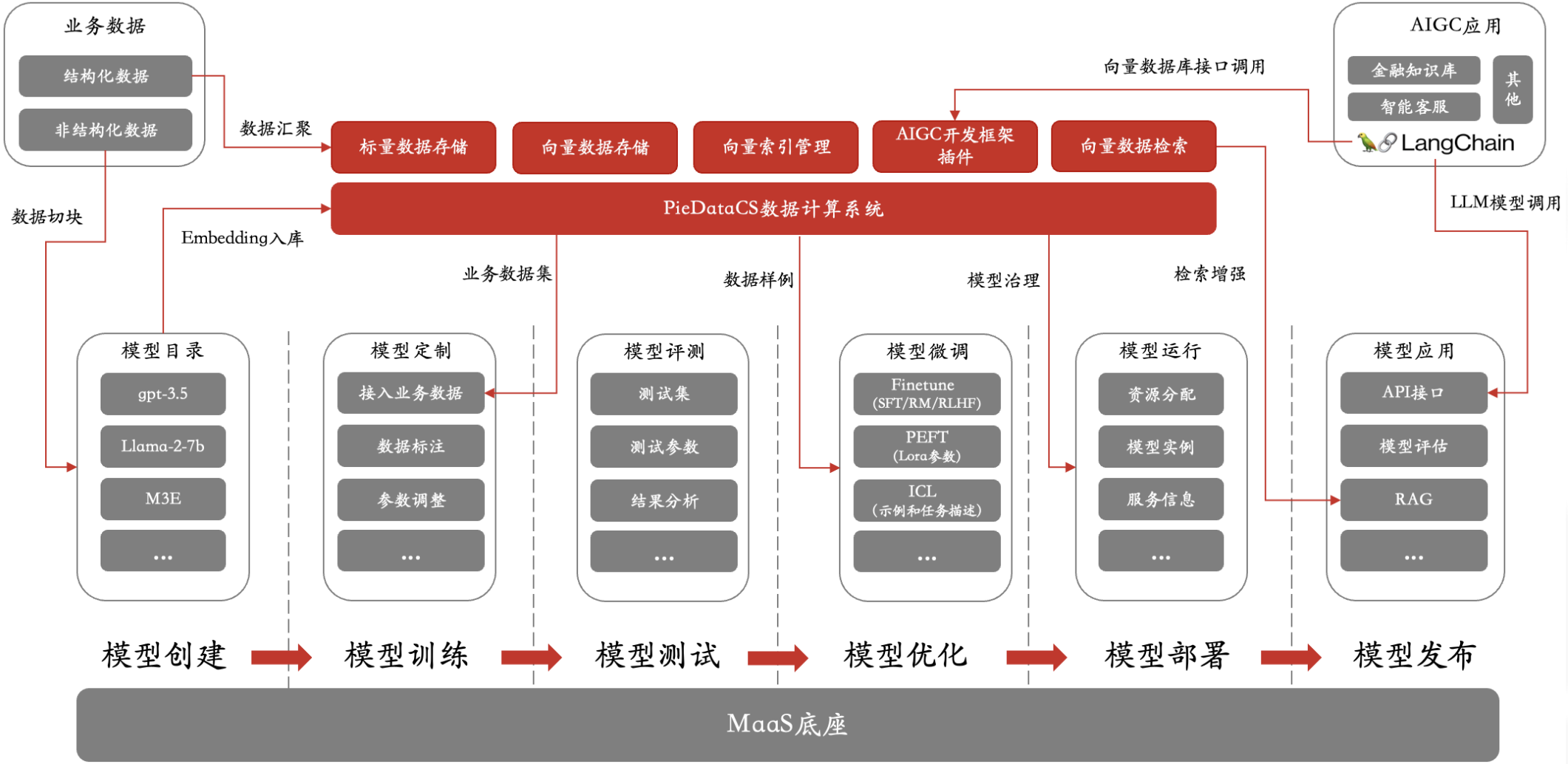
新一代GraphRAG架构

- 首先进行向量或关键词搜索，以找到一组初始节点。
- 然后遍历图，获取这些节点相关的信息。这可以通过图数据库中的查询来实现，比如使用图遍历算法。
- 最后，可以选择使用基于图的排名算法(如 PageRank)对文档进行重新排名。

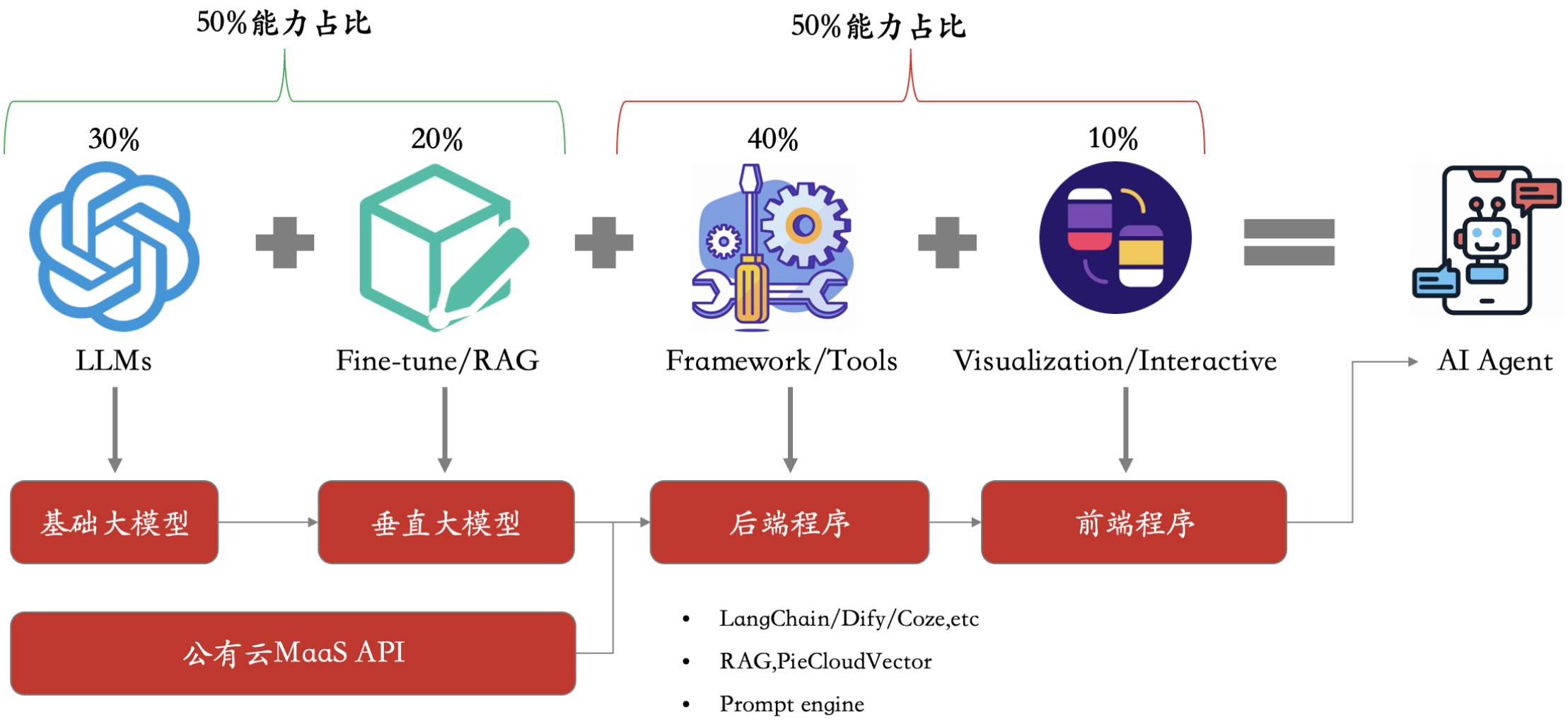




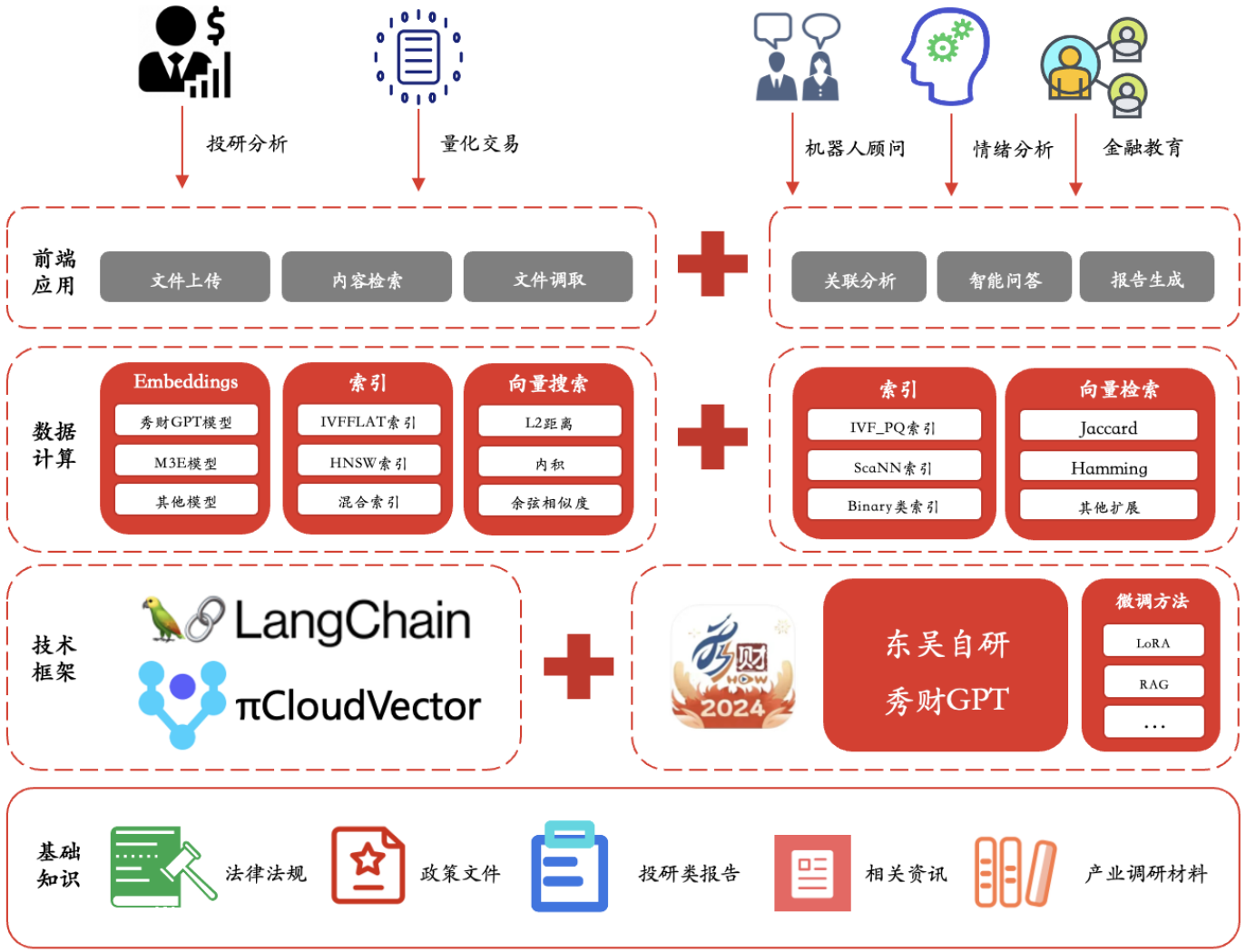
AIGC全生命周期管理



AIGC应用组织


50%能力占比
50%能力占比
30%
20%
40%
10%
LLMs
Fine-tune/RAG
Framework/Tools
Visualization/Interactive
AI Agent
基础大模型
垂直大模型
后端程序
前端程序
公有云MaaS API
LangChain/Dify/Coze,etc
RAG,PieCloudVector
Prompt engine

某金融客户AIGC应用实践



投研分析：

针对日常工作中的法律法规、政策文件以及投研报告进行快速检索，形成对应的分析报告，为客户提供投资相关的数据支撑。

量化交易：

基于 GPT 的大量金融数据训练，可以发现事件的情绪对资产的影响模式，可将这种模式用到量化交易策略中，由 GPT 实时产生交易信号以自动执行交易。这种数据驱动的量化交易策略可以更快地对市场变化进行响应。

机器人顾问：

根据个人需求和偏好实时提供金融建议，提高了获取建议的便利程度，降低了获得服务的成本。GPT可以学习大量历史案例和研究报告，在此基础上，它可以总结出投资策略与建议。用户可以直接使用这些建议，或根据实际情况进行适当调整，这可以极大提高工作效率。

金融市场情绪分析：

对投资相关的言论和情绪表达等数据进行深度分析，获取市场情绪的指标，帮助投资者更好的把握市场走势，制定合理的交易策略，避免情绪犯错。

金融教育：

形成基础金融知识和产品知识库，针对用户和内部员工传授相关的投资策略和产品功能，用户可以与GPT交互，询问投资工具、风险和回报等具体问题，让用户对投资原理有更深的了解。同时也可以让员工更快速的掌握更复杂的金融产品和工作技能。

业务运营报告：

结合内容运营的需求，把行情类的数据进行智能化和个性化创作，定时为客户推送实时行情、财经类的资讯以及精准的市场数据和分析，提升内容创作效率。

联合某高校打造多模态数据分析课程

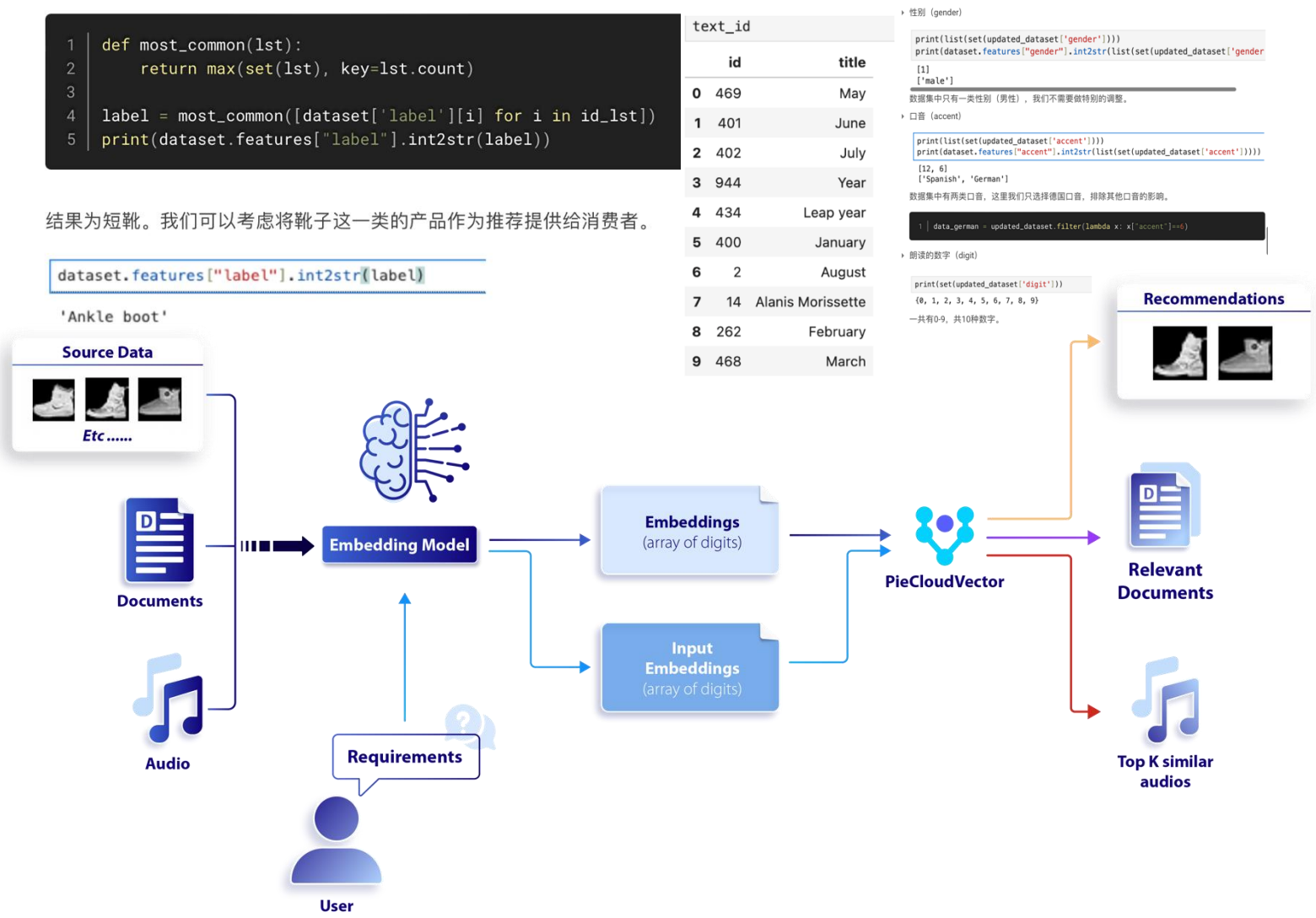
```python
def most_common(lst):
    return max(set(lst), key=lst.count)

label = most_common([dataset['label'][i] for i in id_lst])
print(dataset.features["label"].int2str(label))
```

结果为短靴。我们可以考虑将靴子这一类的产品作为推荐提供给消费者。

```
dataset.features["label"].int2str(label)
'Ankle boot'
```

	id	title
0	469	May
1	401	June
2	402	July
3	944	Year
4	434	Leap year
5	400	January
6	2	August
7	14	Alanis Morissette
8	262	February
9	468	March



文本数据分析

1. 加载Wikipedia数据集，该数据集包括id、url、title、text等字段内容，数据Embedding后写入PieCloudVector；
2. 选取有关四月的维基百科英文文本，通过sentence_transformers 工具，采用 paraphrase-MiniLM-L6-v2模型算法进行Embedding，得到一个384维的向量；
3. 向PieCloudVector发送 query 来查询，使用 L2 Distance 寻找最相似的10条文档。

图片数据分析

1. 加载图片数据，该数据集包含了服装图片、类型等数据，数据Embedding后写入PieCloudVector；
2. 选取一张鞋子图片，通过Embedding后得到一个768维的向量；
3. 向PieCloudVector发送 query 来查询最相似（与目标数据向量距离最近）的10个单品。这里我们计算距离使用的算法为 L2 Distance。

音频数据分析

1. 加载音频数据，该数据包含了不同口音、来自不同地区、性别各异的个体使用英语朗读数字的音频数据；
2. 选择一段音频数据，采样率为4000，音频向量的长度在3000左右；
3. 向PieCloudVector发送 query 来查询最相似的音频，采用IP算法返回的结果更为准确，判断标准为
   a) 性别
   b) 口音
   c) 朗读的数字

备注：一份音频数据中，包含音频文件路径、音频波形矩阵，以及波形所对应的采样率。数据集中波形的采样率为48000，较高的采样率虽然更精准，但也会导致矩阵较大（一个矩阵中有超过3万个数字），为之后的计算带来负担。

关注产品公众号

随时获得产品动态

加入技术交流群

获得更多技术干货

墨天轮

乐知乐享，同心共济。

知行合一，不负所托！