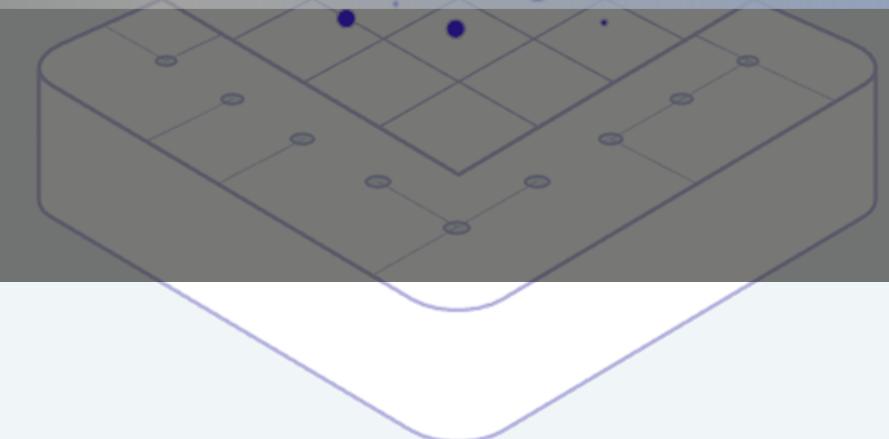
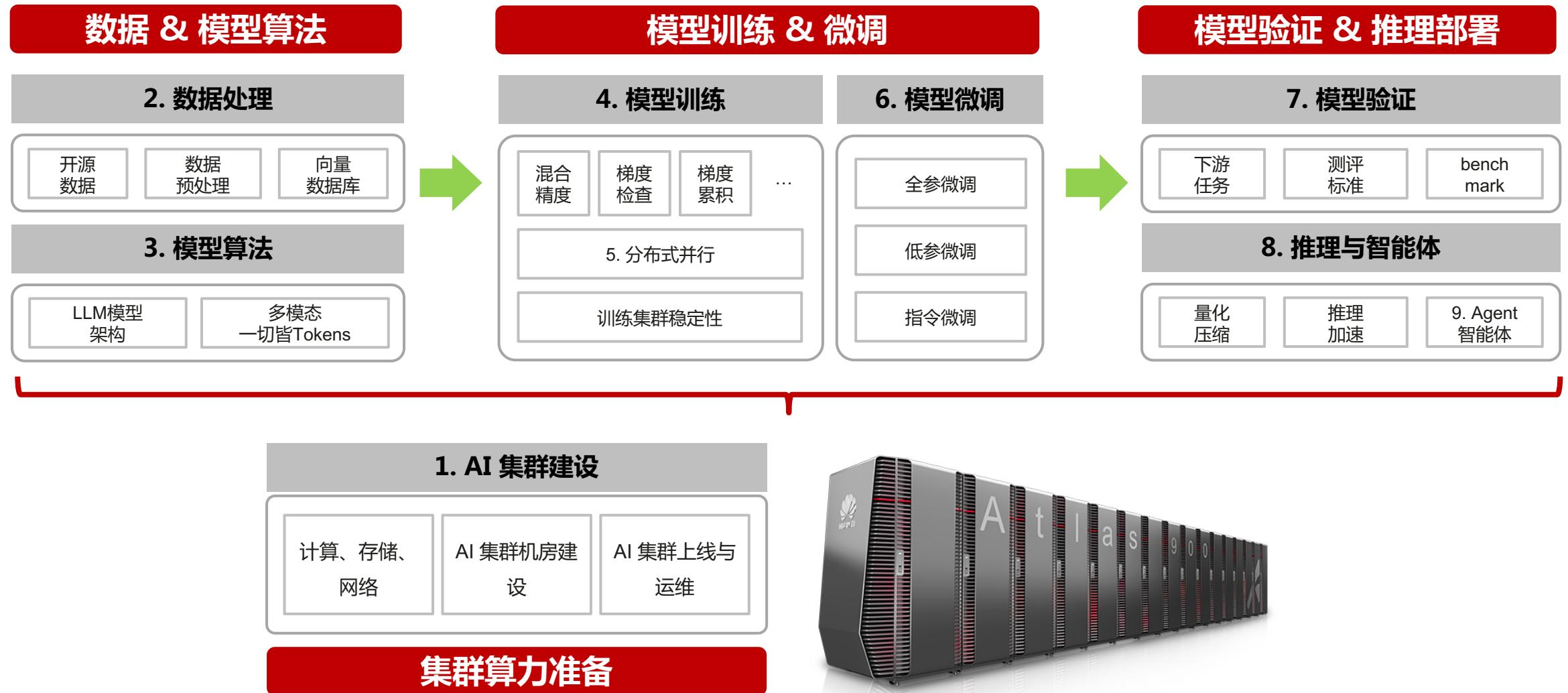


大模型系列 – 数据处理

向量数据库遇大模型



大模型业务全流程

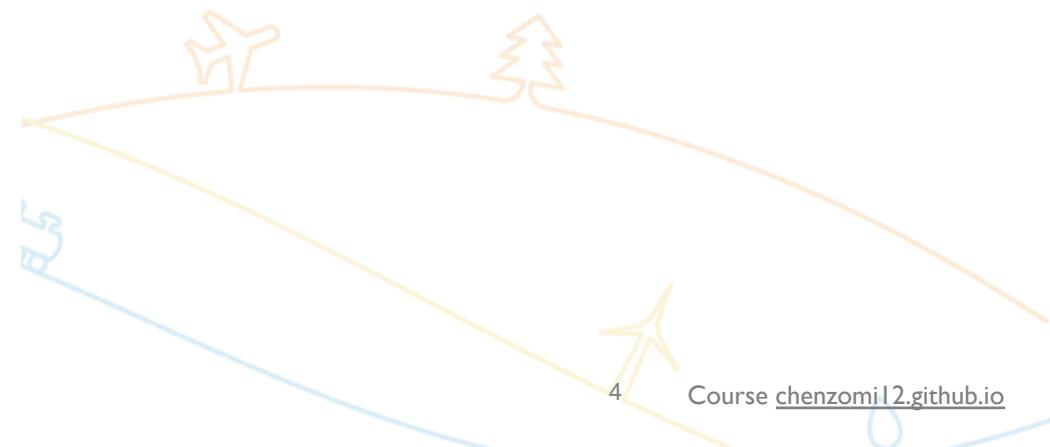
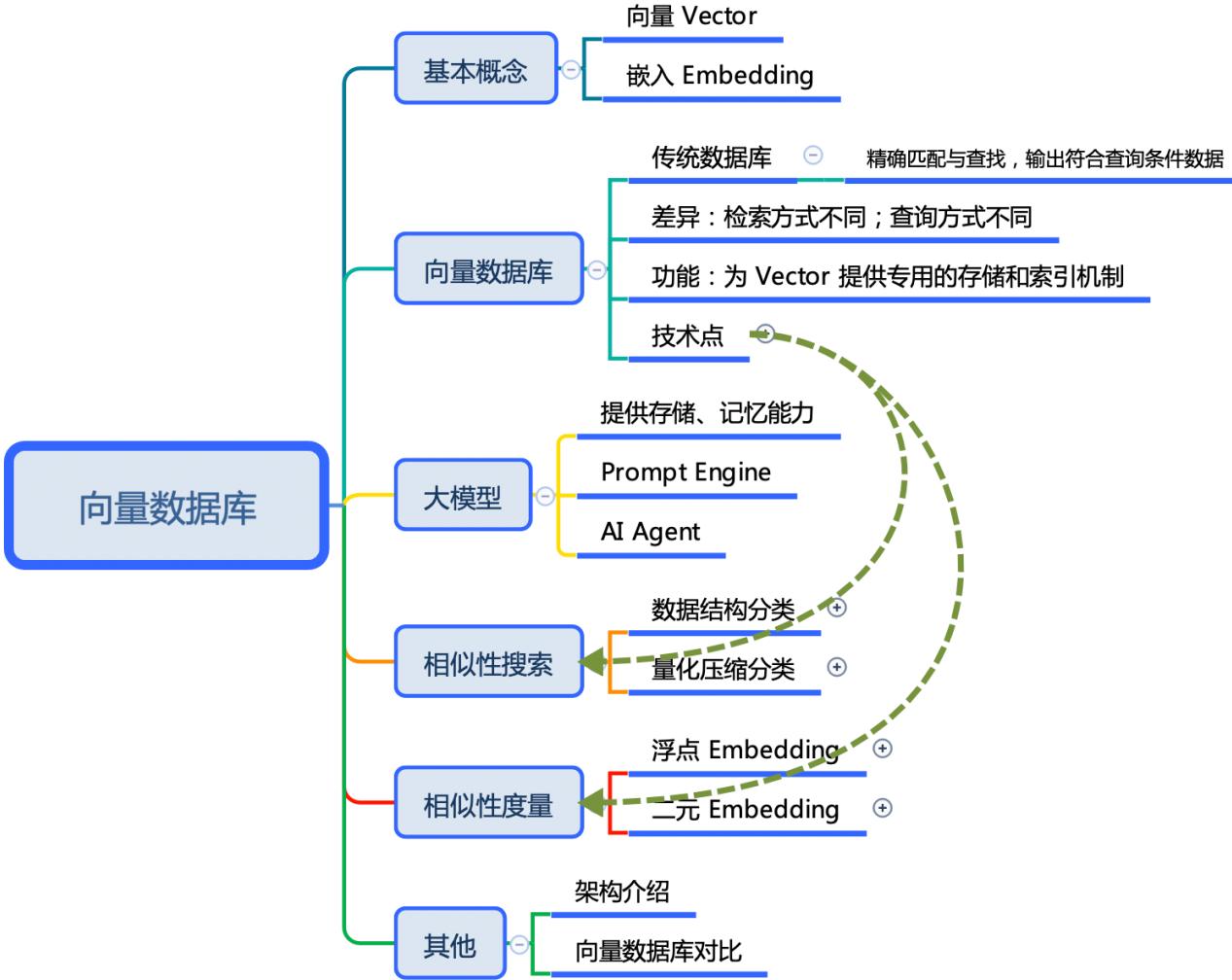


大模型系列 – 数据处理之向量数据库

• 具体内容

- **向量与检索** : 向量 Vector 的表示 -- Embedding 原理
- **向量数据库** : 向量数据库原理、功能、特点 -- Vector-DB 应用场景
- **大模型关系** : 向量数据库遇到大模型 – 大模型与 Vector-DB 应用场景
- **相似性搜索** : K-Means 聚类 -- Faiss 算法 -- PQ 算法 -- IVF 算法 -- HNSW 算法
- **相似性度量** : 欧氏距离 (L2) -- 内积 (IP) -- 其他度量方式
- **通用性架构** : 通用 Vector-DB 架构 -- KDB 架构示例
- **对比与小结** : 业界向量数据库横向对比 -- Vector-DB 小结

大模型系列 – 数据处理之向量数据库



Question ?

1. 不是大模型让向量数据库火起来的嘛，为什么没有大模型相关的场景？
2. 向量数据库的应用场景规模感觉比较有限，增量在哪里？



1. 大模型 & 向量数据库

大模型时代的机遇

- 向量数据库 Embedding 后使用向量化计算高速地处理数据：
 - 大规模、
 - 高维度、
 - 多模态
- 支持复杂查询、分布式，更大规模数据计算。

VS

- LLM 大模型从数据中提取出有用特征和模式：
 - 强大学习和表示能力
 - 处理庞大和复杂的数据
 - 数据中自动生成新的内容
- 通过微调快速提升模型性能和适配下游任务能力

大模型时代的机遇

- LLM 的出现改变了人机交互方式，带来AI技术革命，但仍然存在很多问题：
 - 知识储备能力不足
 - 隐私数据安全性
 - 推理算力成本高
- 如何增强LLM知识的同时保护私域数据安全？
- 通过外置数据和工具来增强LLM能力

Baichuan2-13B-Chat

Baichuan2-13B-Chat为Baichuan2-13B系列模型中对齐后的版本，预训练模型可见Baichuan2-13B-Base

任正非今年几岁了?

任正非出生于1944年，所以他现在是78岁。

Baichuan2-13B-Chat

Baichuan2-13B-Chat为Baichuan2-13B系列模型中对齐后的版本，预训练模型可见Baichuan2-13B-Base

任正非今年几岁了?

任正非出生于1944年，所以他现在是78岁。

任正非明年几岁了?

任正非是1944年出生的，所以明年他将年满80岁。

大模型时代的机遇

- LLM 的出现改变了人机交互方式，带来AI技术革命，但仍然存在很多问题：

- 知识储备能力不足
- 隐私数据安全性
- 推理算力成本高

- 如何增强LLM知识的同时保护私域数据安全？
- 通过外置数据和工具来增强LLM能力

Baichuan2-13B-Chat
Baichuan2-13B-Chat为Baichuan2-13B系列模型中对齐后的版本，预训练模型可见Baichuan2-13B-Base

Baichuan2-13B-Chat

任正非今年几岁了？

任正非出生于1944年，所以他现在是78岁。

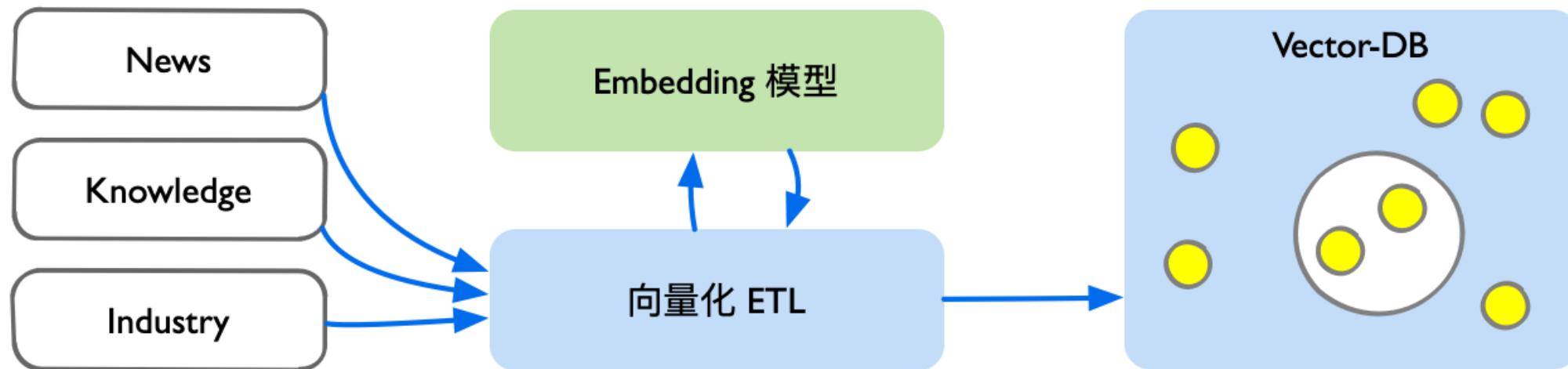
Baichuan2-13B-Chat
Baichuan2-13B-Chat为Baichuan2-13B系列模型中对齐后的版本，预训练模型可见Baichuan2-13B-Base

任正非1944年出生，假设今年是2024年，那么任正非今年几岁了？

如果今年是2024年，那么任正非应该是80岁。

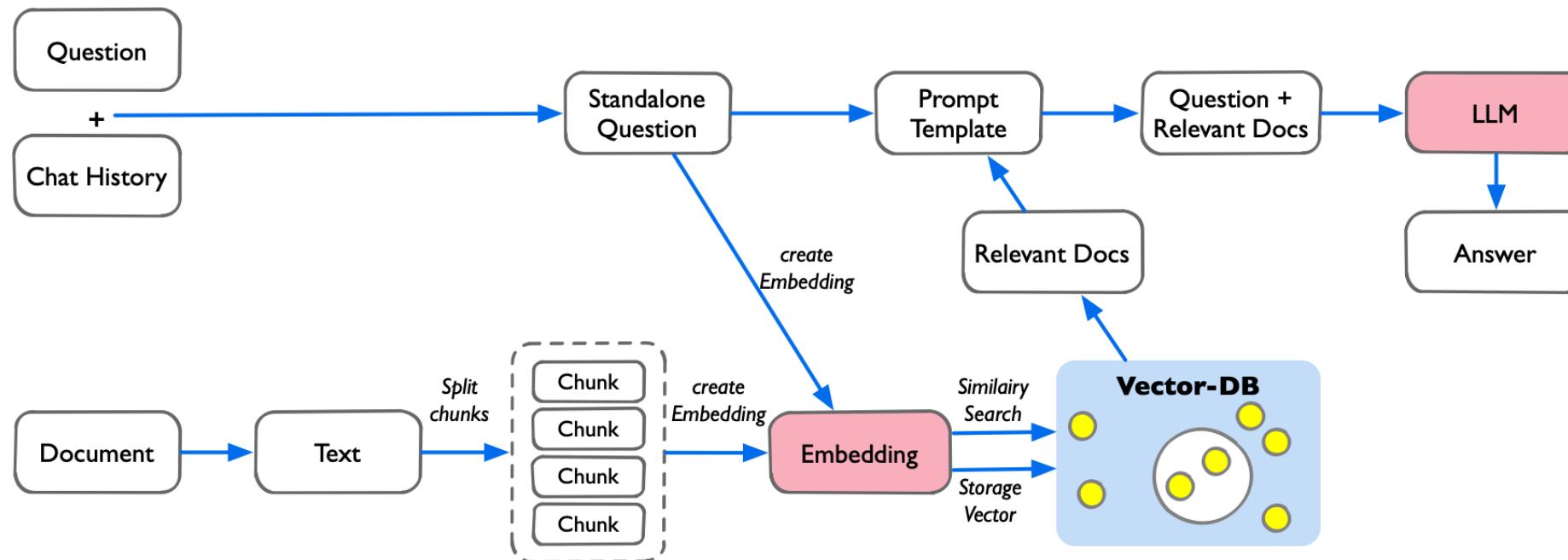
大模型遇到向量检索

- LLM 应用离不开 Prompt Engine , 提示工程离不开向量检索
- 本质语义搜索 , 从海量数据中找到匹配的内容 , 拼接提示词
- 向量数据库对输入数据 Embedding 后 , 使用向量化计算为大模型提供高效的数据存储和查询支撑 ;



大模型遇到向量检索

- LLM 应用离不开 Prompt Engine，提示工程离不开向量检索
- 本质语义搜索，从海量数据中找到匹配的内容，拼接提示词
- 向量数据库对输入数据 Embedding 后，使用向量化计算为大模型提供高效的数据存储和查询支撑；



大模型遇到向量数据库

Vector-DB 提供存储、记忆能力，大模型提供问题处理和分析能力

- 向量数据库 Embedding 后，使用向量化计算为大模型提供高效的数据存储和查询支撑；
- 大模型与 Vector-DB 深度融合应用为通用人工智能（AGI）的实现提供了可靠路径；
- 大模型新一代 AI 处理器，提供数据处理能力；Vector-DB 提供 Memory 存储能力。

大模型遇到向量数据库

Vector-DB 提供存储、记忆能力，大模型提供问题处理和分析能力

- 大模型输入知识被压缩为静态模型参数（13B/70B），模型中参数不会随交互对话记住个人喜好和内容，也不会调用额外知识信息来辅助判断（模型参数在推理阶段不会更新）。
- Vector-DB 意义给大模型提供外部的长期记忆。e.g.，大模型回答不了最新数据，A 只能回答 2021 年前数据。通过外置 Vector-DB 数据库作为大模型 Prompt，回答更加实时内容。

大模型场景 - 向量检索的挑战

场景

- 大模型知识增强和提示词工程 Prompt Engine

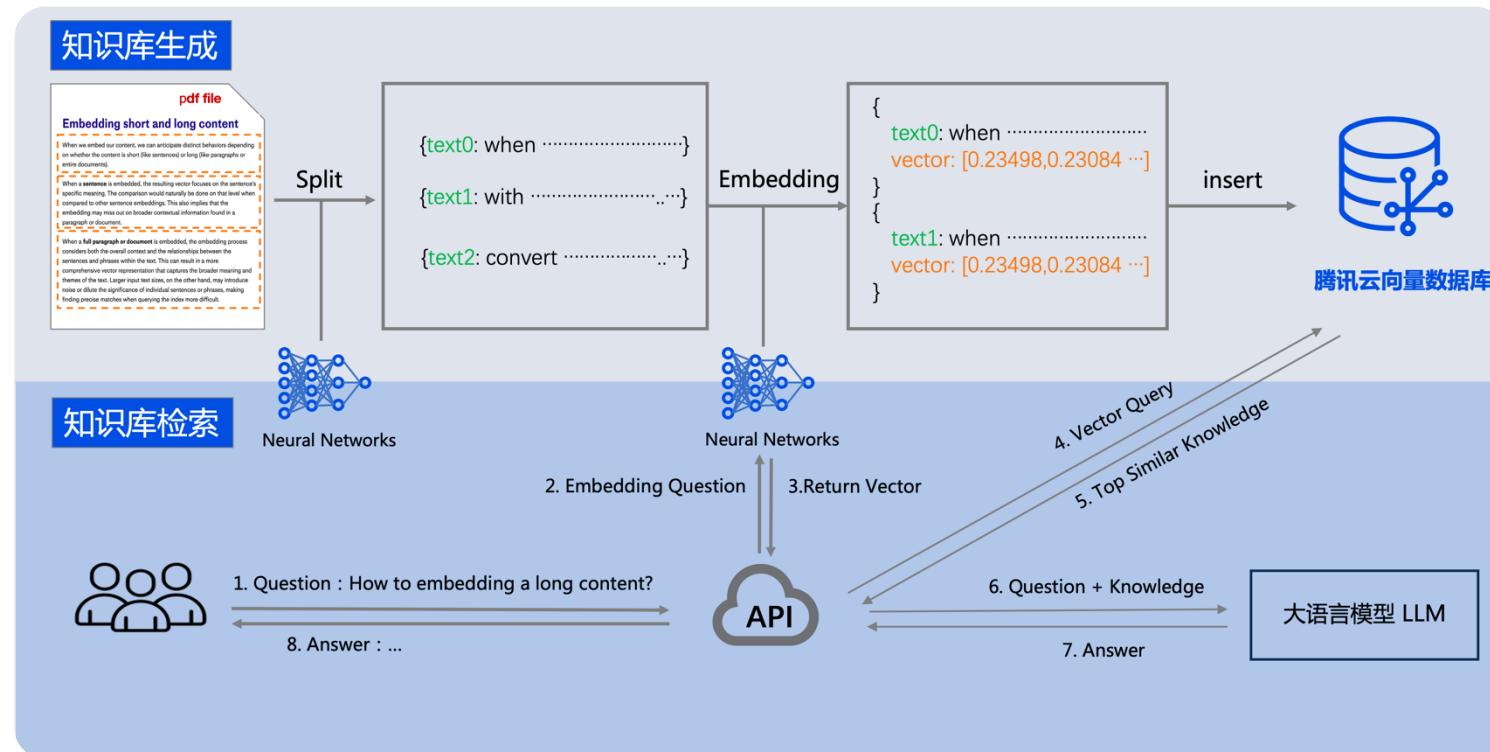
技术挑战

- 百亿数据规模，向量维度较多
- 上千 QPS，需要毫秒级返回
- 系统 99.99% 高可用保障性
- 实时更新的混合场景
- 带 Filter 检索场景

除技术挑战，还要思考如何低成本高弹性的使用资源，如何高效维护 LLM + Vector DB 系统

腾讯云 Vector-DB

- 腾讯云 Vector-DB + 大语言模型 LLM 配合使用。企业的私域数据经过文本分割、向量化后，可以存储在腾讯云 Vector-DB 中，构建起企业专属的外部知识库，从而在后续的检索任务中，为大模型提供提示信息，辅助大模型生成更加准确的答案。



广告时间？

- I. 不存在，只是网上找的资料方便用，直接拿过来介绍而已。

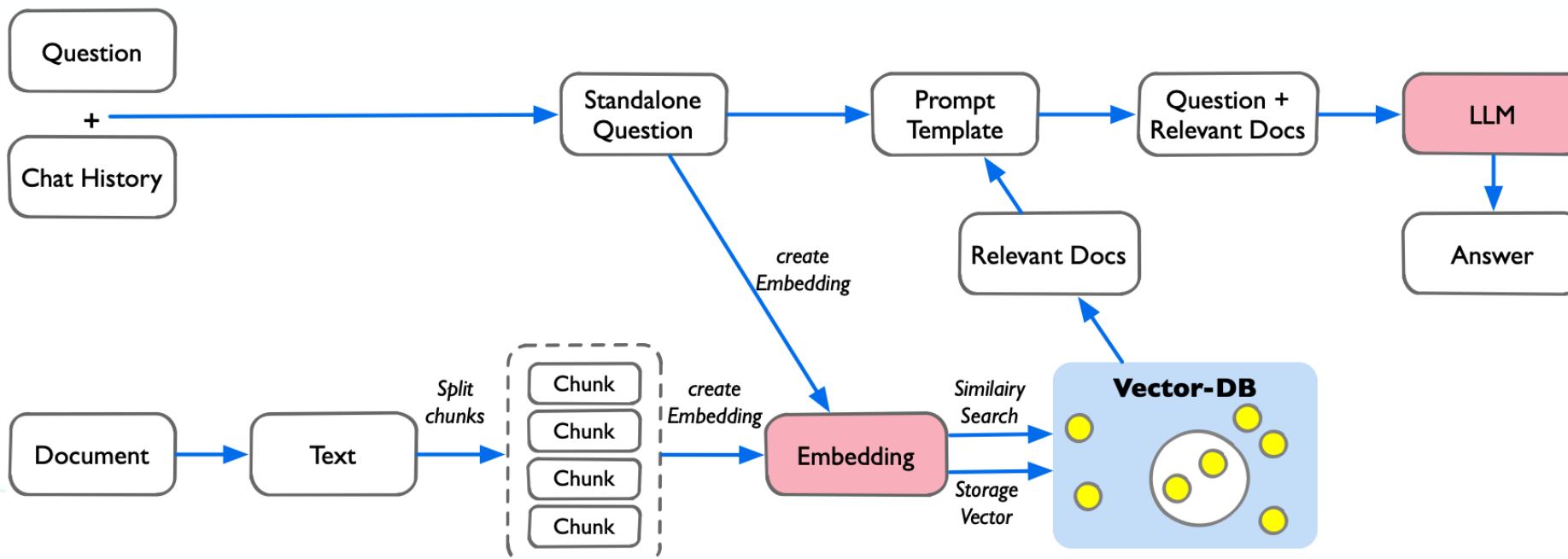


2. Vector DB

大模型应用场景

大模型场景基本应用

1. 超出上下文长度文本划分成 chunks，不同 chunks 转换为向量（embedding）。
2. 输入 prompt 到 LLM 之前，将 prompt 转换为向量（embedding）。
3. 将 prompt 向量进行搜索，寻找到最相似的 chunk 向量。
4. 将最相似的 chunk 向量与 prompt 向量拼接，作为 LLM 的输入。



Vector-DB 对大模型的价值

大模型应用挑战1：幻觉现象

- **作用**：为大模型提供优质 Prompt
- **优点**：回答信息真实可溯源

大模型应用挑战2：推理成本高

- **作用**：相似问答缓存在 Vector-DB 中
- **优点**：高频回答无需重复推理

大模型应用挑战3：记忆能力

- **作用**：为 AI Agent 提供长期记忆 Memory
- **优点**：实现外部记忆功能

大模型应用挑战1：幻觉现象

Baichuan2-13B-Chat

Baichuan2-13B-Chat为Baichuan2-13B系列模型中对齐后的版本，预训练模型可见Baichuan2-13B-Base

Baichuan2-13B-Chat

今天天气怎么样？

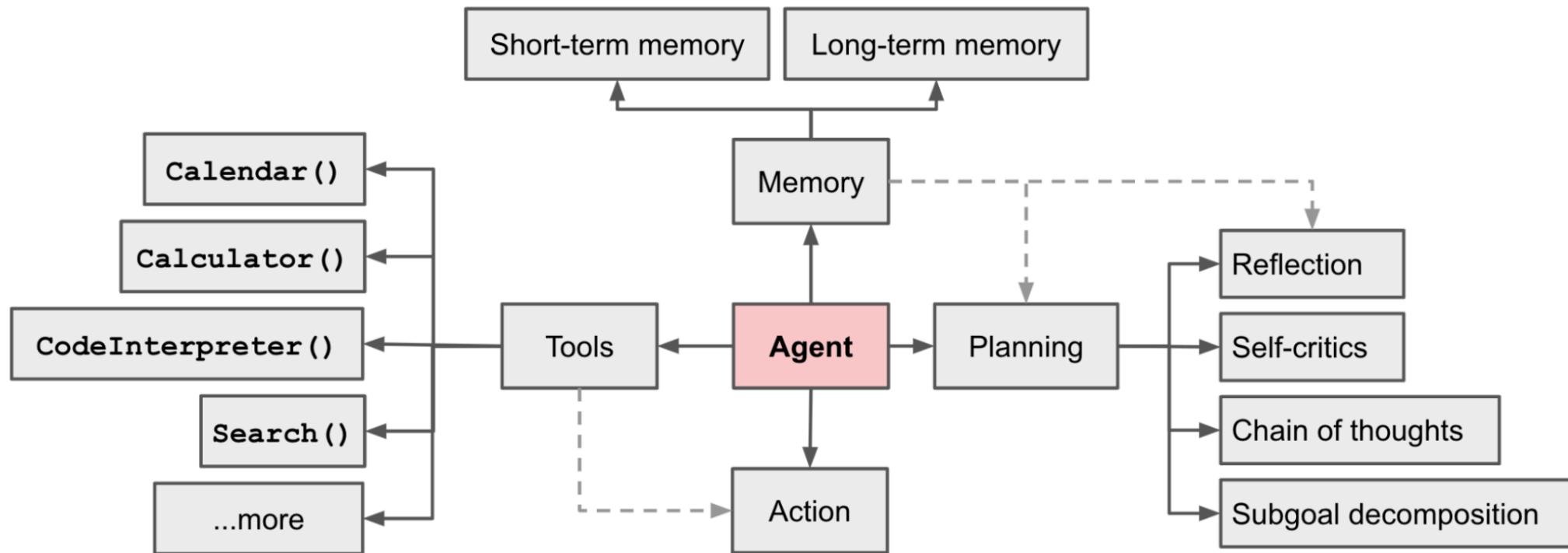
作为一个大语言模型我无法实时获取天气信息。请使用天气应用或搜索引擎查询您所在地区的天气情况。

当前杭州天气一周天气长期天气杭州限行尾号:5和0 多云 4 空气质量 优东北风 3级 湿度:94%紫外线:0级气压1023百帕 体感:冷 2°C
冷 舒适 热 能见度:欠佳 2.8km 好 差 云量:阴 95%。那么杭州当前天气如何？

根据您提供的信息，今天的杭州天气是多云，气温为4°C，空气优良，东北风3级，湿度为94%，紫外线指数为0级，气压为1023百帕。这种天气条件下，感觉较冷。请注意保暖并关注天气变化。

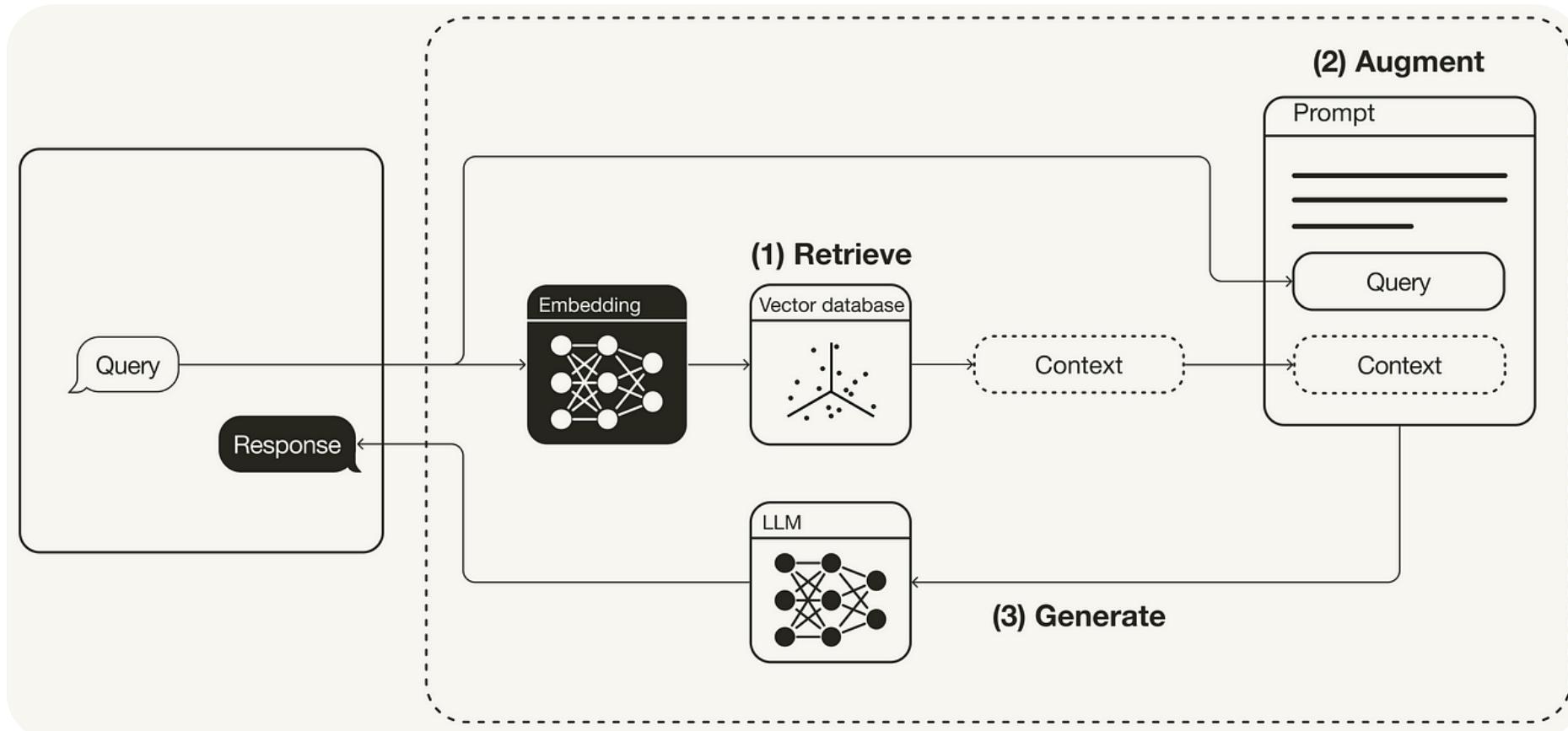
大模型应用挑战3：记忆能力

- AI Agent 关键组成：规划 Planning + 记忆 Memory + 工具 Tools



基于Vector-DB 的 RAG 成为业界标准

- Retrieval Augmented Generation (RAG) 查询增强生成式 AI 应用框架



基于Vector-DB 的 RAG 成为业界标准

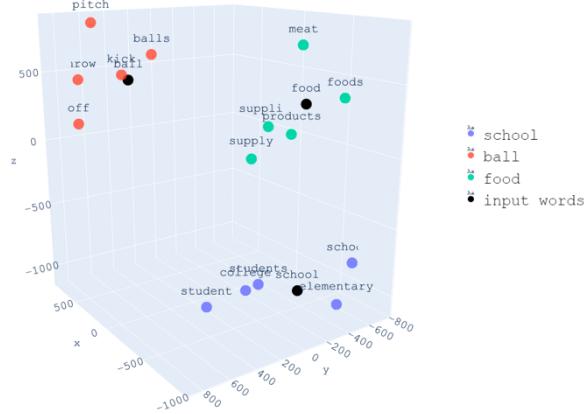
- Retrieval Augmented Generation (RAG) 查询增强生成式 AI 应用框架

	RAG	Fine-Tuning
定义	利用模型以外知识，如用户私域数据、产品信息等进行信息增强	使用较小数据集进行SFT，分为全量微调 FFT 和低参微调 PEFT
优劣	<ol style="list-style-type: none">1. 知识外置，信息实时化2. 私域数据，数据隐私性高3. 数据置信度高，长尾知识好4. 同等质量模型，RAG 模型参数更少	<ol style="list-style-type: none">1. 特定任务效果更好2. 知识更新周期需要等待收集足够数据
挑战	<ol style="list-style-type: none">1. 性能与 Vector-DB 索引强相关2. 召回率影响结果的相关性3. 计算与 IO 密集，性能要求高	<ol style="list-style-type: none">1. 知识参数化，无法获取实时知识2. 难以删除参数化隐私数据3. 保证数据可信度
案例	OpenAI RAG model for QA	Fine-Tuning GPT Models
场景	企业应用：金融、审计等对数据隐私、可信、领域知识、可重复性、时效性要求高的领域。	特定微调任务、推理任务，如 SD 文生图、代码生成等；

基于Vector-DB 的 RAG 两大类型

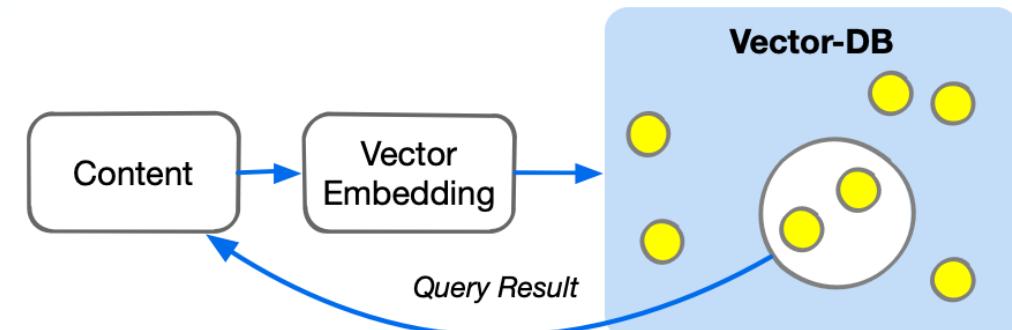
One-Shot REALM

- 查询相似向量，提前组成 Prompt；
- 查询位于输入层，一次推理只执行一次；
- **当前 Vector-DB 主要应用方式**



In-Context REALM

- 提高检索效果，支持超长序列推理；
- 通过向量化查询，降低端到端时延；
- **类似于将推理进行记忆化、状态化；**

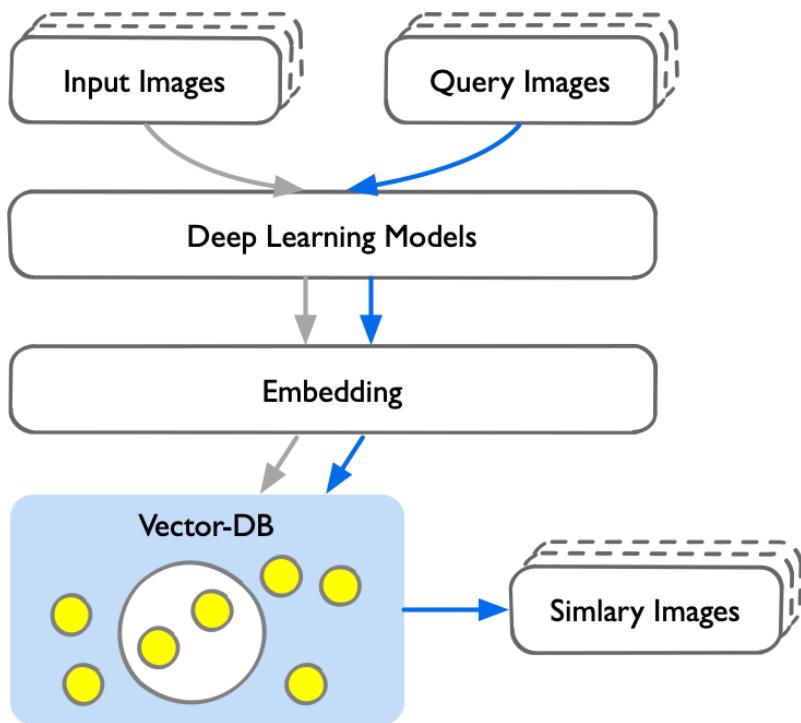


3. Vector DB

其他应用场景

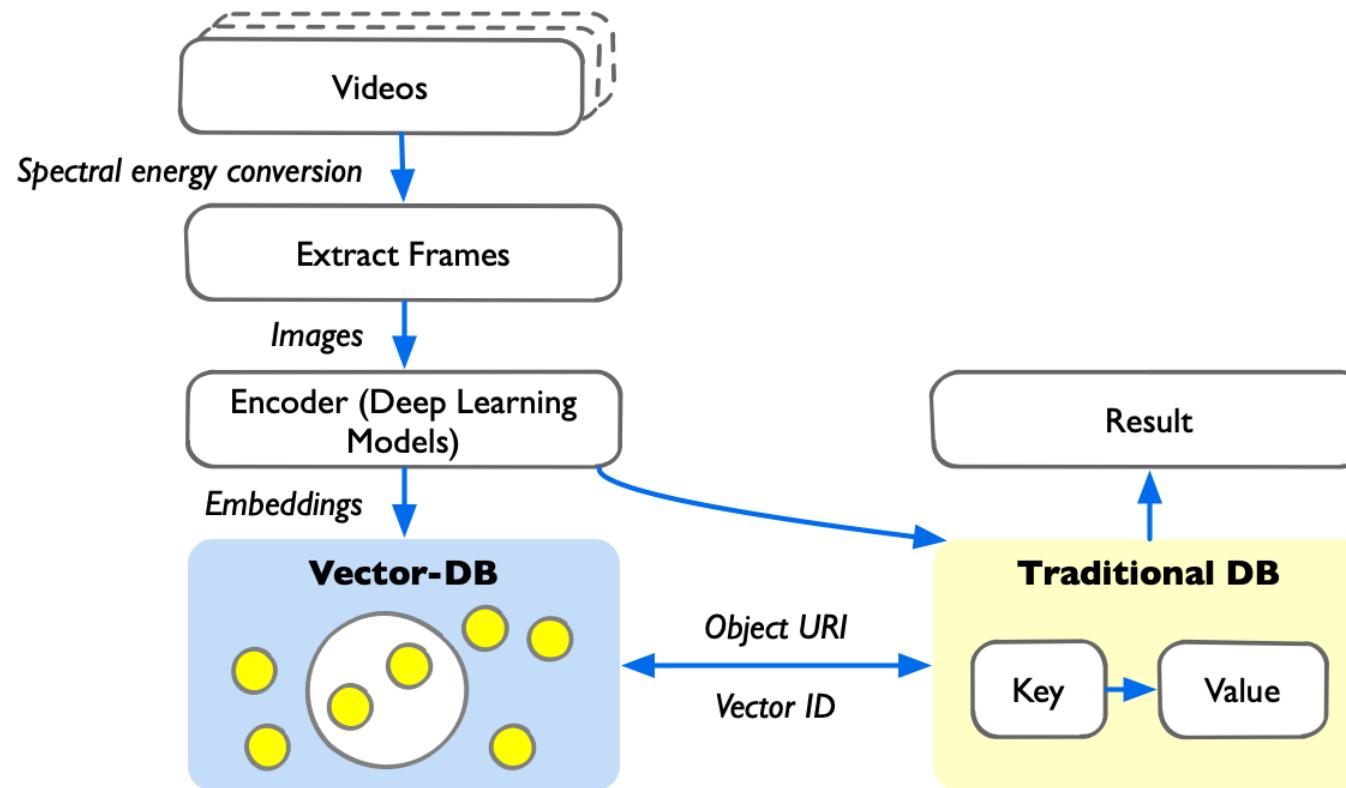
图像相似性搜索

- Google、Baidu 搜索引擎提供按图搜索的选项；
- 电子平台拼多多、淘宝也提供类似功能为在线购物带来的便利，并整合到 APP 应用中；



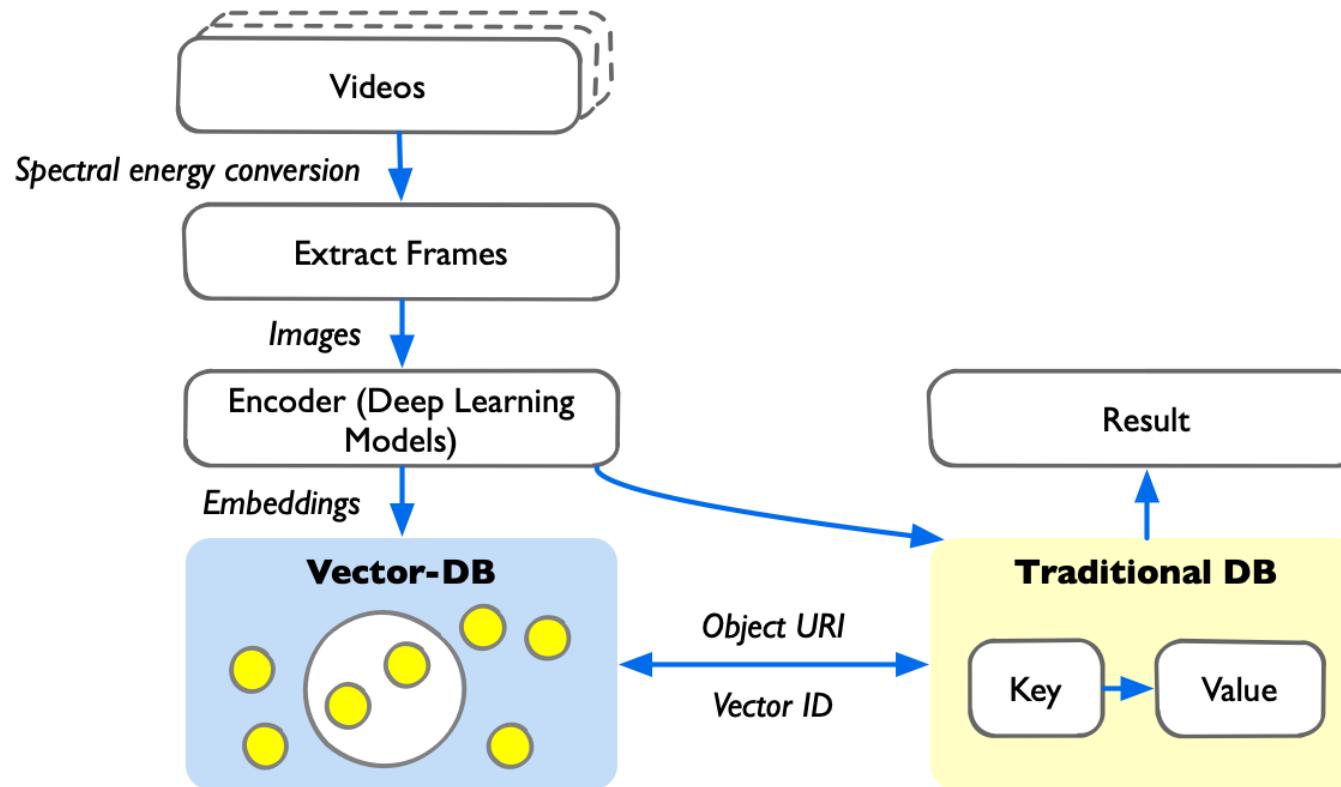
视频相似性搜索

- 将喜欢电影/视频截图发表社交平台。利用视频相似性搜索，获知电影/视频名称并打 TAG。



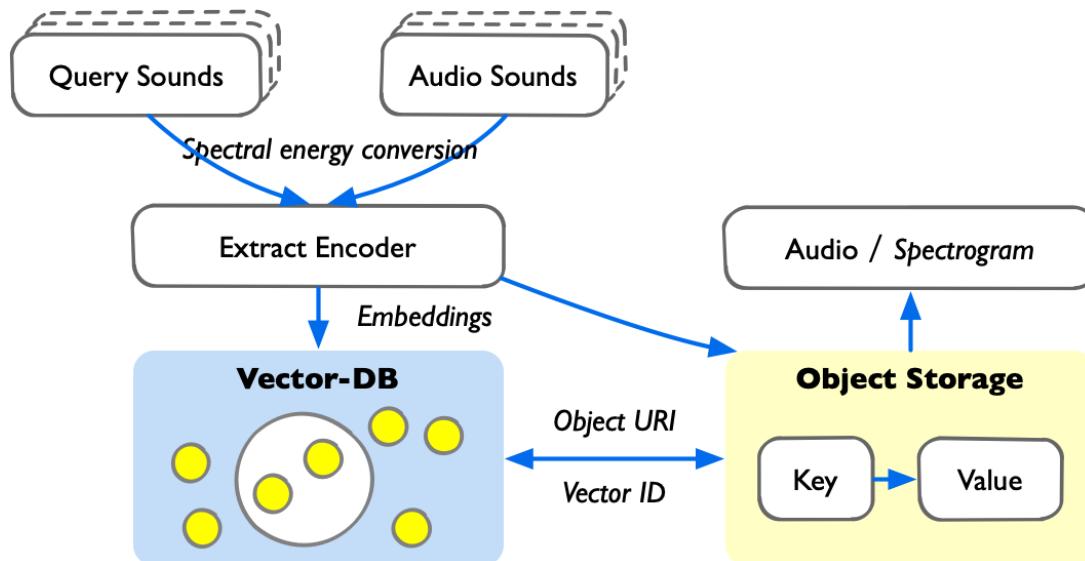
视频相似性搜索

- 技术架构设想：1) 视频中提取关键帧，使用 AI 模型获取关键帧特征向量。2) 特征向量存储在 Vector-DB 并进行搜索，Vector-DB 返回相似向量 ID。3) 将 ID 映射到存储在 MySQL 中的视频。



音频相似度搜索

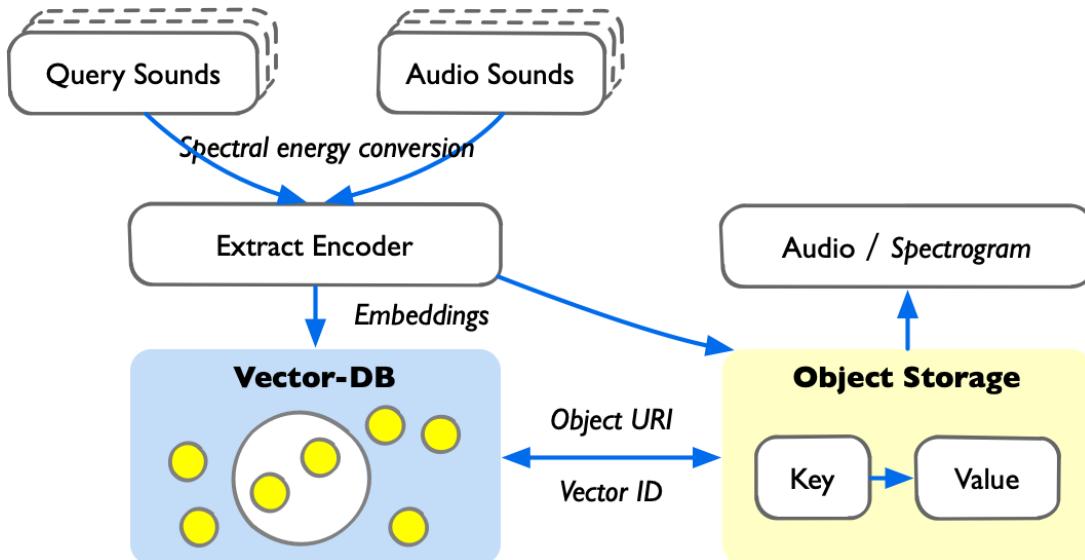
- 技术架构设想：1) 音频片段 PANNs 转换为 Vector；2) 向量存储 Vector-DB 生成唯一 ID；3) Vector-DB 中相似度搜索，查询与 Vector-DB 返回唯一 ID 对应音频数据路径。



视频	Shorts	直播	帖子	播放列表	播客	宣传视频	公开范围	限制	日期 ↓	观看次数	评论数	
Google TPU4光交 16:45							【AI芯片】TPU系列：AI集群超级互联Google TPUv4 光交换机 #AI芯片 #TPU系列 #AI集群 TPU项目开始于201...		发布日期			
Google 05 TPU4芯片 13:23							AI集群超级互联Google TPUv4 #AI芯片 #TPU系列 #... 【AI芯片】TPU系列：AI集群超级互联Google TPUv4 #AI芯片 #TPU系列 #AI集群 TPU项目开始于2014年,简单来...	公开	无	2023年10月22日	219	1
AI集群分布式架构 11:49							大模型在AI集群分布式架构：参数服务器或集合通... 【大模型系列】之AI集群02：大模型在AI集群分布式架构：参数服务器或集合通信架构！在训练大模型需要...	公开	版权	2023年10月12日	302	6
大模型AI集群如何组成？ 14:58							大模型AI集群如何组成？AI集群需要哪些硬件？ #... 【大模型系列】之AI集群01：大模型AI集群如何组成？AI集群需要哪些硬件？在训练大模型需要使用AI集群，通...	公开	版权	2023年10月12日	351	1
大模型AI集群训练时间 & 训练效率如何计算？ 15:44							大模型AI集群训练时间 & 训练效率如何计算？ #大... 【大模型系列】之AI集群01：分享大模型为什么需要使用AI集群，有了AI集群之后，探讨一下什么是AI集群。...	公开	版权	2023年10月11日	396	1
超级计算节点TPU POD介绍 15:19							超级计算节点TPU POD介绍 #AI芯片 #TPU系列 #超... 【AI芯片】TPU系列：超级计算节点TPU POD介绍 TPU项目开始于2014年,简单来说两个原因:1.计算任务不同...	公开	无	2023年10月10日	228	1
南京麒麟上5014度深寒耐候小钢窗 1 15:19							南京麒麟上5014度深寒耐候小钢窗 1 15:19		发布日期	2023年10月10日	338	1
破冰者号 15:19							破冰者号 15:19		发布日期	2023年10月11日	339	1

音频相似度搜索

- 快速查询海量音频数据，并找到相似的声音。e.g.，YouTube 识别视频相似音效减少知识产权侵权；网易云音乐、腾讯音乐的听曲识歌功能；音频数据分类和统计分析。

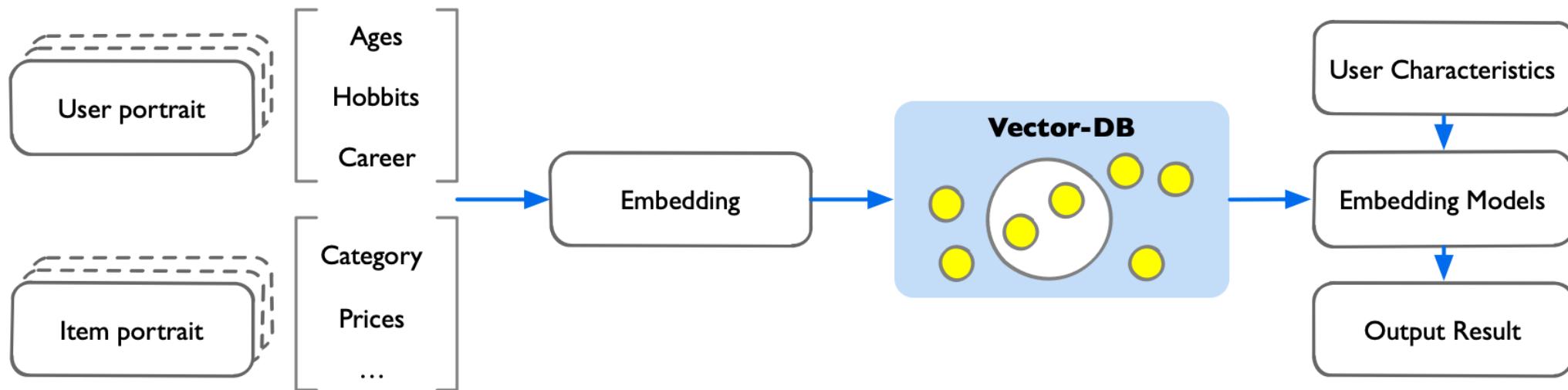


A screenshot of a video search interface, likely YouTube, showing a list of results for a query related to AI chips and TPU. The results are filtered by '视频' (Videos) and show various video thumbnails, titles, and metadata. One specific result is highlighted with a red box, showing details such as '公开' (Public), '版权' (Copyright), and a timestamp of '2023年10月12日' (October 12, 2023). The interface includes standard video search filters like '过滤条件' (Filter Conditions) and sorting options like '日期' (Date) and '观看次数' (View Count).

视频	Shorts	直播	帖子	播放列表	播客	宣传视频
过滤条件						
<input type="checkbox"/> 视频					公开范围	限制
<input type="checkbox"/> Google TPU4对比TPU4.5	16:45				公开	无
<input type="checkbox"/> AI集群超级互联Google TPUv4 #AI芯片 #TPU系列 #AI集群 TPU项目开始于201...	13:23				2023年10月22日	219
<input type="checkbox"/> 大模型在AI集群分布式架构：参数服务器 or 集合通信【大模型系列】之AI集群01：大模型在AI集群分布式架构：参数服务器 or 集合通信...#AI...	11:48				2023年10月12日	302
<input type="checkbox"/> 大模型AI集群如何组成？AI集群需要哪些硬件？#AI...	14:58				2023年10月12日	351
<input type="checkbox"/> 大模型AI集群训练时间 & 训练效率如何计算？#大...	15:44				2023年10月11日	396
<input type="checkbox"/> 超级计算机节点TPU POD介绍 #AI芯片 #TPU系列 #超...	15:19				2023年10月10日	228
发布日期						
观看次数						
评论数						

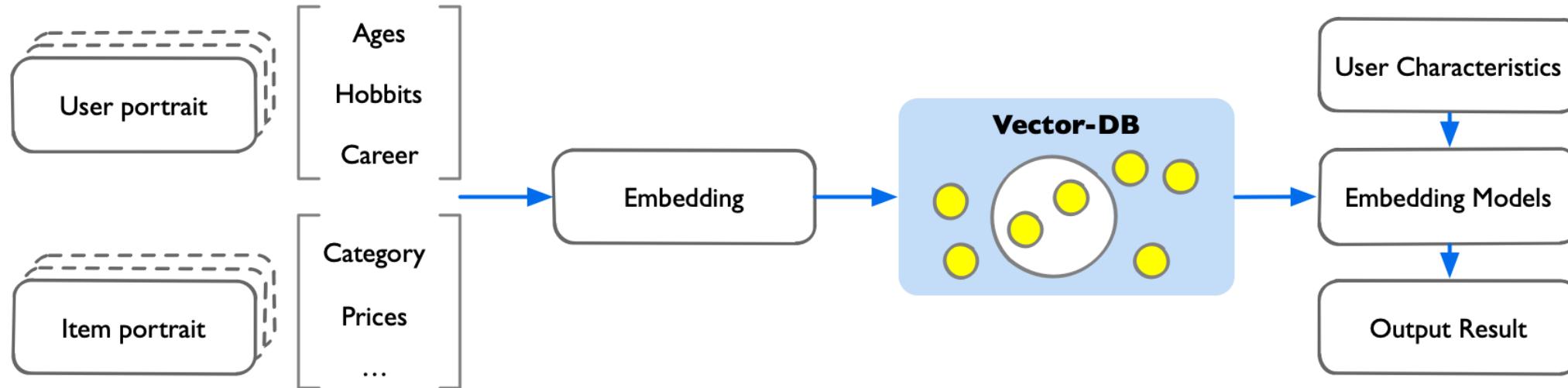
推荐系统 Recommendation System

- 音乐、视频、信息流等各种场景。不需要用户准确描述需求，分析用户行为发现需求和兴趣。



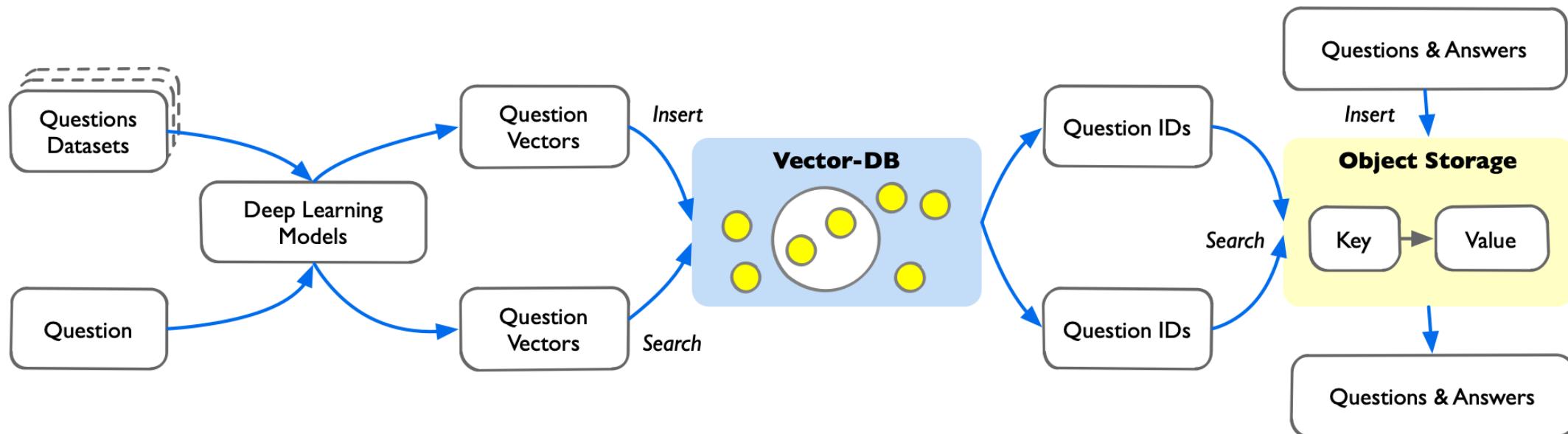
推荐系统 Recommendation System

- **技术架构设想：**1) 准备数据集，包括对物品评分、历史点击记录和物品描述等信息。2) 用户特征 + 物品特征组转换为向量。3) 物品向量存储在 Vector-DB ，使用用户向量进行相似性搜索。4) 用户向量与物品向量相似，Vector-DB 返回商品向量和其 ID 作为推荐结；5) 使用非关系型数据库中物品向量 ID 查询信息。



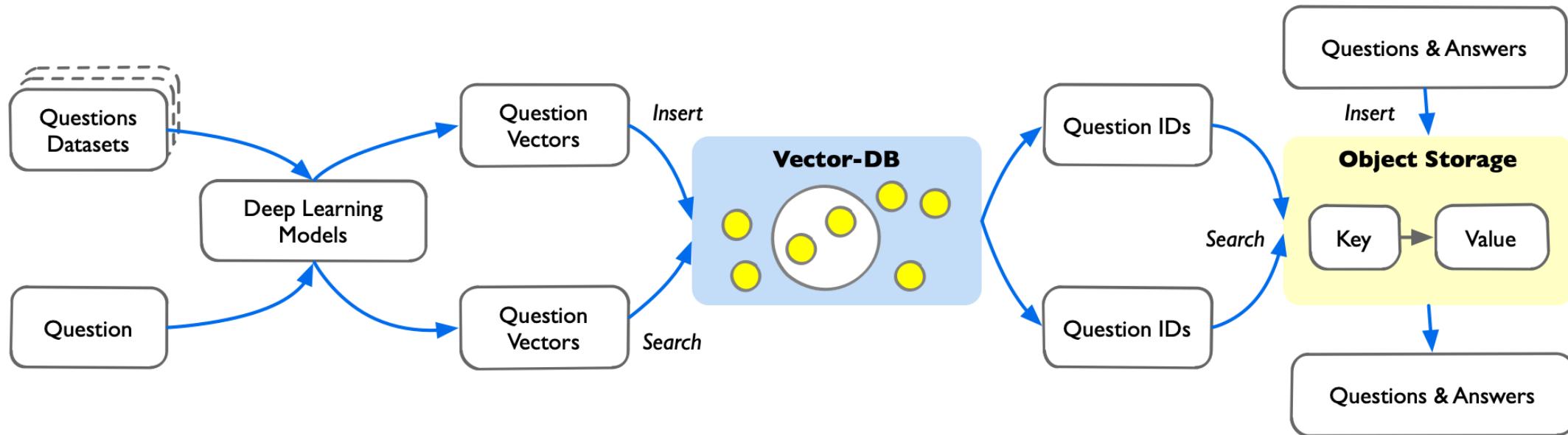
问答系统 QA Chatbot

- 包括在线客服系统、QA 聊天机器人等。QA 系统可归类为：生成式/检索式，单轮/多轮，开放域/特定域。



问答系统 QA Chatbot

- **技术架构设想：**1) 准备问题和相应答案数据集，将问题和答案存储在关系型数据库；2) AI 模型将问题转换为向量，问题向量存储并索引在 Vector-DB；3) 用户输入新问题时将其转换为 Vector，Vector-DB 会搜索与新向量最相似的问题向量与相应答案。



Summary



思考

1. 中小型企业 & AI Agent 中应用发力

- L0 基础大模型开放，缺少规模AI集群中心的中小型企业会基于 L0 进行推理应用；AI Agent与LLM大模型结合外部知识需要结合向量数据库提供长期记忆。

2. 大模型训练降本增效，参与训练和推理过程

- 神经网络/大模型的输入输出为向量，训练过程的数据分类、去重、清洗可以利用向量数据库；推理减少大模型幻觉，充当大模型长期记忆并补充实时性、定制型知识从L0->L2。**是大模型生态落地的重要组成。**

3. 保护数据隐私和安全

- 面向政企等对数据有高度隐私敏感场景，大模型的数据训练不需要结合此数据，采用独立部署的向量数据库具备安全隐私性。

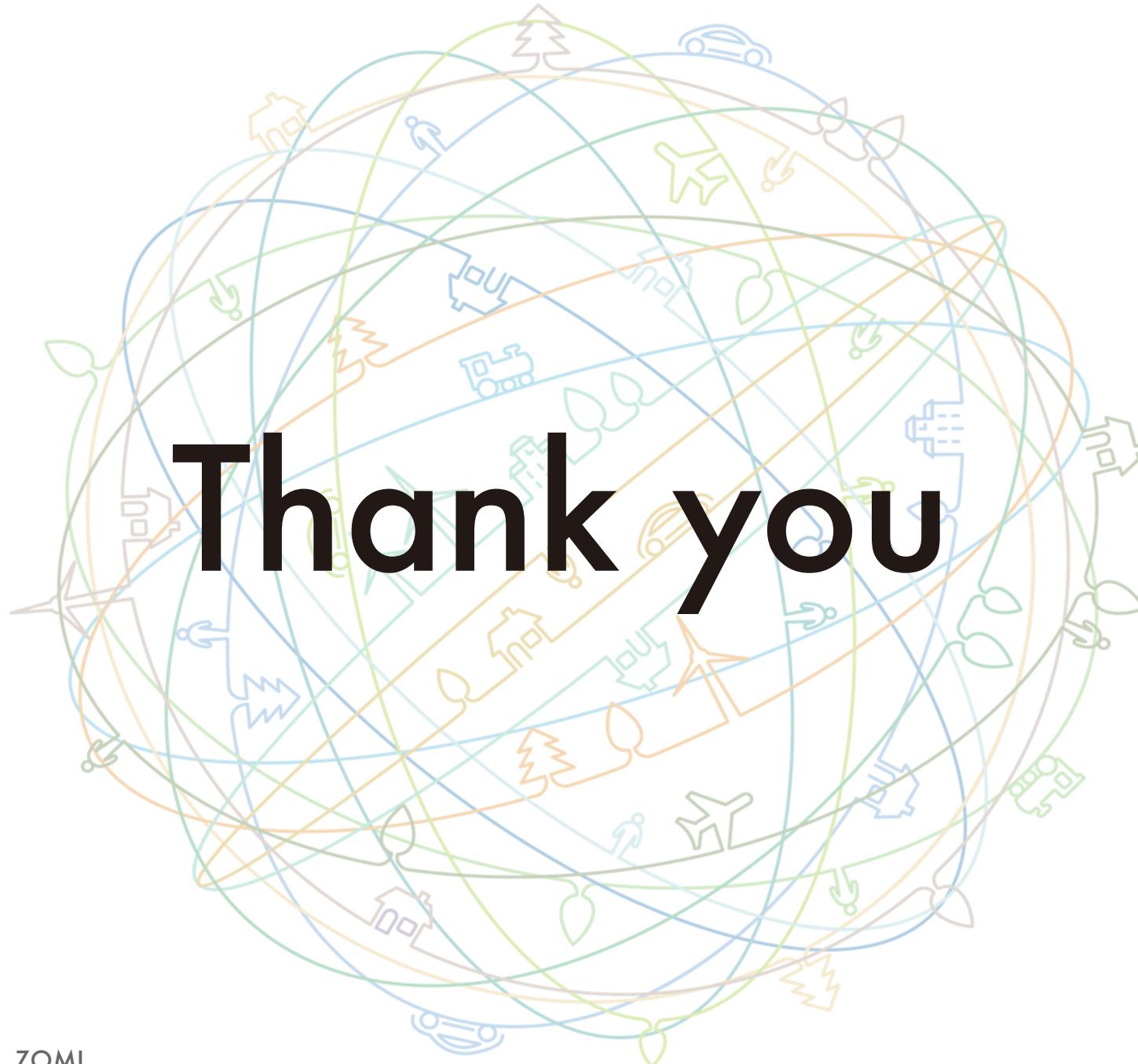
思考

- 当前向量数据库主要作为单独的服务提供，场景价值主要停留在知识库增强和缓存加速上，未来是否有可能内置化到大模型进行训推一体化加速呢？



Reference 引用&参考

1. Maximizing the Potential of LLMs: Using Vector Databases (ruxu.dev)
2. Maglott D , Ostell J , Pruitt KD , Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 2005 Jan 1;33 (Database issue):D54-8.
3. Wang Y , Xiao J , Suzek TO , Zhang J , Wang J , Zhou Z , Han L , Karapetyan K , Dracheva S , Shoemaker BA , Bolton E , Gindulyte A , Bryant SH. PubChem's BioAssay Database. *Nucleic Acids Res.* 2012 Jan;40 (Database issue):D400-12.
4. Tong W , Altman RB. 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinformatics.* 2017 Mar 16;18 (1):302.
5. Zheng S , Shao W , Chen L. UniVec: a database of gene expression vectors for PCA based gene similarity search. *BMC Genomics.* 2017 Dec 6;18 (Suppl 10):918.
6. Manning CD , Raghavan P , Schütze H. *Introduction to Information Retrieval.* Cambridge: Cambridge University Press , 2008.
7. Mikolov , Tomas , et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
8. Andoni , Alexandr , and Piotr Indyk. "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions." *Communications of the ACM* 51.1 (2008): 117-122.
9. Jégou , Hervé , et al. "Product quantization for nearest neighbor search." *IEEE transactions on pattern analysis and machine intelligence* 33.1 (2010): 117-128.
10. Ge , Tiezheng , et al. "Optimized product quantization." *IEEE transactions on pattern analysis and machine intelligence* 36.4 (2013): 744-755.
11. Babenko , Artem , and Victor Lempitsky. "The inverted multi-index." *IEEE transactions on pattern analysis and machine intelligence* 37.6 (2014): 1247-1260.
12. Datar M , Immorlica N , Indyk P , Mirrokni VS. Locality-sensitive hashing scheme based on p-stable distributions. *Proceedings of the twentieth annual symposium on Computational geometry.* 2004 Jun 8:253-62.
13. Muja M , Lowe DG. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1).* 2009 Feb 4:331-40.
14. Jégou , Hervé , et al. "Product quantization for nearest neighbor search." *IEEE transactions on pattern analysis and machine intelligence* 33.1 (2011): 117-128.
15. Chen , Zhenjie , and Jingqi Yan. "Fast KNN search for big data with set compression tree and best bin first." *2016 2nd International Conference on Cloud Computing and Internet of Things (CCIOT).* IEEE , 2016.
16. Dehmamy , Nima , Albert-László Barabási , and Rose Yu. "Understanding the representation power of graph neural networks in learning graph topology." *Advances in Neural Information Processing Systems* 32 (2019).
17. Babenko A , Lempitsky V. The inverted multi-index. *IEEE transactions on pattern analysis and machine intelligence.* 2014 Jun 7;37 (6):1247-60.



把AI系统带入每个开发者、每个家庭、
每个组织，构建万物互联的智能世界

Bring AI System to every person, home and
organization for a fully connected,
intelligent world.

Copyright © 2023 XXX Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.



Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem