证券研究报告|行业专题报告
计算机行业
行业评级 强于大市（维持评级）
2023年5月8日

华福证券
HUAFU SECURITIES CO.,LTD.

向量数据库-大模型引发爆发式增长

证券分析师：
钱劲宇 执业证书编号：S0210522050004

投资要点

核心观点：

- **向量数据库为大模型提供记忆，是大模型应用的刚需工具。**

在大模型应用中，不断涌现的B端对专用数据的需求、C端对个性化与自动化的需求，带来给大模型增加记忆功能的刚性需求。向量数据库因为可以为大模型提供记忆而需求倍增，AutoGPT更是把对向量数据库需求量推到了更高的水平。

- **向量数据库的竞争格局：大模型厂商不构成竞争，以专业厂商为主**

大模型的训练和推理本身只涉及embedding模型，不需要向量数据库，因此大模型厂商不形成直接竞争。专业向量数据库厂商当前以Zilliz、Pinecone等为主，4月以来海外多家知名向量数据库创业企业陆续传出融资喜讯。

- **市场空间：预计2025年向量数据库占非结构化数据处理需求约三成，数据向量化后存储将带来较大膨胀**

中国数据库市场规模2022年约300亿人民币，预计到2025年将达到约500亿人民币。随着非结构化数据的增加，非关系型数据库的营收占比预计将逐年提升。向量数据库同样用于非结构化数据的处理分析需求，我们推测到2025年其将占非结构化数据处理需求约三成，同时数据向量化后相比传统非结构化数据存储有较大膨胀，因此其价格将会数倍于传统的非关系型数据库产品。

投资建议：建议关注星环科技。星环科技主营数据库和大数据产品，公司研发实力雄厚有相关技术储备。

风险提示：技术研发不及预期、市场开拓不及预期。

目录

从向量搜索到向量数据库，向量伴随着ML/AI模型的发展而发展

概念	定义	备注
向量	为AI理解世界的通用数据形式，是多模态数据的压缩	虽然大模型端到端呈现的都是文字文本，但模型实际接触和学习数据是向量化文本，因为文本本身直接作为数据维度太高，机器学习效率太低
Embedding	将文字文本转化为保留语义关系的向量文本	即embedding模型对自然语言的压缩和总结，将高维数据映射到低维空间
向量搜索	在海量存储的向量中找到最符合要求的Top N个目标。向量搜索是模糊匹配，返回的是相对最符合要求的N个数据，并没有精确标准答案	传统数据库索引是精确匹配，也就是说，传统数据库中的数据要么符合查询要求/返回数据，要么不符合查询要求/无数据返回
向量数据库	用以高效存储和搜索向量。保证100%信息完整的情况下，通过向量嵌入函数(embedding)精准描写非结构化数据的特征，从而提供查询、删除、修改、元数据过滤等操作。	传统数据库无法满足此类操作和需求，只能实现部分向量数据的存储，且无法高效搜索向量

资料来源：36Kr，华福证券研究所

从向量搜索到向量数据库：虽然向量数据库当前最主要的使用场景和火热原因是因为它为大模型提供记忆。但其实向量的embedding结构和向量搜索算法之前就已经出现，只是在大模型出现之前，向量搜索需求只存在于大厂，大厂自研向量搜索算法即可，且没有很强的向量存储需求。

伴随着 ML/AI 模型发展到今天，一直到大模型场景下，向量搜索和存储的需求才真正开始爆发式增长。向量数据库因为可以为大模型提供记忆而需求倍增，AutoGPT更是把对向量数据库需求量推到了更高的水平，AutoGPT从一开始就是采用了OpenAI API+ Pinecone的模式。

通过向量数据库实现有记忆交互是大模型的应用趋势

在大模型的应用中，不断涌现出**B端对专用数据的需求**、**C端对个性化与自动化的需求**，带来给大模型增加记忆功能的刚性需求，相关产品需求量快速增长。

	交互形式	相关产品
无记忆交互	Zero-shot Prompt：直接对话 Few-shot Prompt：带事例对话	ChatGPT等大模型
有记忆交互	根据外部存储增强模型记忆：Retrieval Augmentation	OpenAI Retrieval Plugin、Langchain、LlamaIndex等辅助工具 Pinecone、Zilliz/Milvus等向量数据库产品

资料来源：36Kr，华福证券研究所

无记忆交互：在初始的LLM中，世界知识和语义理解被压缩为静态参数，模型不会随着交互记住用户的聊天记录和喜好，也无法调用额外知识信息来辅助判断，因此模型只能根据历史训练数据回答问题，并且经常产生幻觉，给出与事实相悖的答案。

一个解决方法是在 Prompt 中将知识告诉模型，但是这往往受限于 token 数量，在 GPT-4 之前一般是 4000 个字的限制，且不经济，而且过多不相干信息还可能导致幻觉。

有记忆交互：当模型需要记忆大量的聊天记录或行业知识库时，可将其储存在向量数据库中，后续在提问时将问题向量化，送入向量数据库中匹配相似的语料作为prompt，向量数据库通过提供记忆能力使prompt更精简和精准，从而使返回结果更精准。

向量数据库为大模型提供记忆——有记忆交互原理

有记忆交互原理/大模型记忆增强原理：

- *Step 1——语料库准备：*

将与行业相关的大量知识或语料上传至向量数据库，储存为向量化文本；

- *Step 2 ——问题输入：*

输入的问题被Embedding引擎变成带有向量的提问；

- *Step 3 ——向量搜索：*

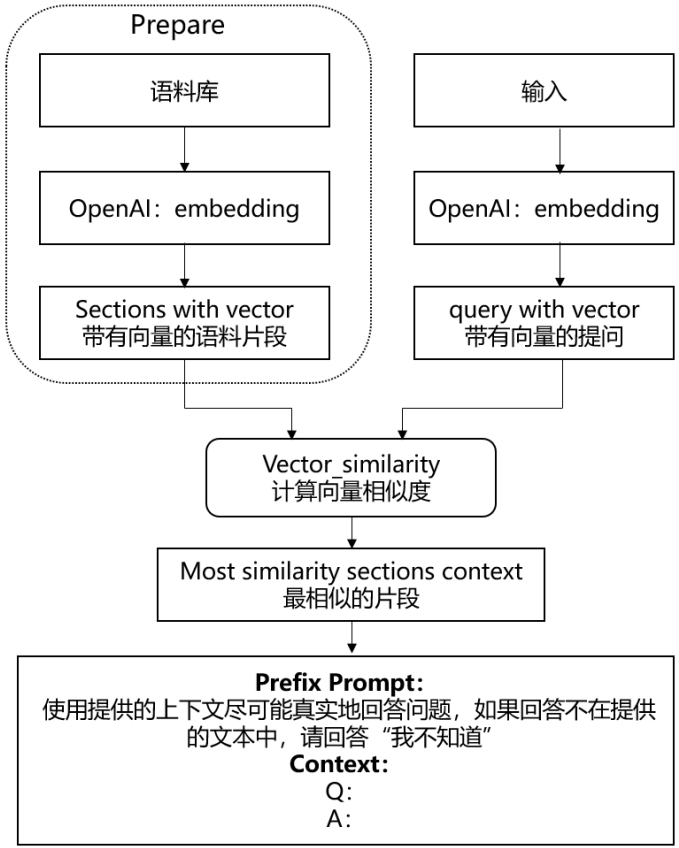
向量化问题进入提前准备好的向量数据库中，通过向量搜索引擎计算向量相似度，匹配出Top N条语义最相关的Facts（向量数据库是模糊匹配，输出的是概率上最近似的答案）

- *Step 4 ——* ***Prompt优化：***

输出的Top N条Facts，和用户的问题一起作为prompt输入给模型。

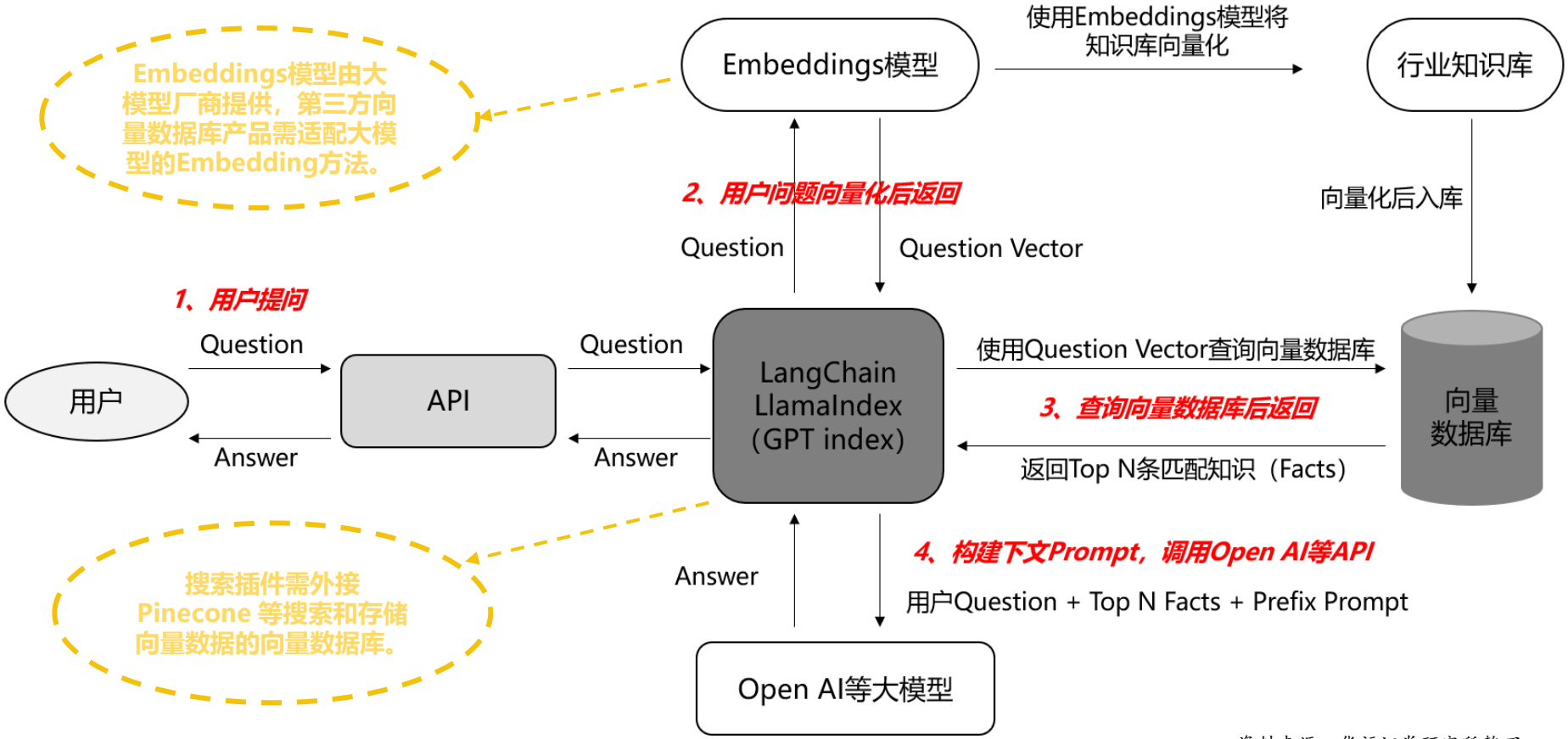
- *Step 5、结果返回：*

有记忆交互下得到的生成内容更精准且缓解了幻觉问题。



资料来源：华福证券研究所整理

向量数据库为大模型提供记忆——有记忆交互流程



资料来源：华福证券研究所整理

向量数据库厂商与大模型厂商的竞争格局

- 大模型厂商：大模型的训练和推理本身只涉及embedding模型，不需要向量数据库，不形成直接竞争。以OpenAI为例，在 Plugin 的项目代码中，OpenAI推荐使用 Pinecone 等作为插件搜索和存储向量数据用的数据库。
- 初创团队：海外的以Zilliz、Pinecone等为主，随着大模型带来的应用需求市场，4月以来海外知名向量数据库创业企业陆续传出融资喜讯。

公司	成立	总部	产品特点	融资情况
Zilliz (Milvus)	2017年	美国 (中国团队)	开源。云原生向量数据库，最成熟、可扩展、行业公认的“最快的向量数据库”，可提供“万亿向量数据集上的毫秒搜索”	近年获得高瓴、淡马锡等知名机构多轮投资
Pinecone	2019年	美国	可实现完全托管与可扩展，开箱即用	2023年4月28日获1亿美元B轮融资
Weaviate	2019年	荷兰	开源。用户 self-host，公司仅提供支持性的服务。自主控权权大	2023年4月22日获5000万美元B轮融资
Qdrant	2021年	德国	基于Qdrant向量搜索的人脸识别技术领先	2023年4月19日获 750 万美元种子轮融资
Chroma	2021年	美国	开源。产品更轻量级，提供轻量级的便捷封装	2023年4月6日获1800万美元种子轮融资

资料来源：公司官网、企查查、墨天轮、华福证券研究所

向量数据库的商业模式与市场空间

- **定价**：使用价格由用户的数据存储量和使用时长一起决定。以Pinecone为例，按存储量0.025美元/GB/月，按使用时长0.1-1美元/小时（依算力等级而定）。

- **成本**：最便宜的GPT 3.5-turbo api调用是0.002美金/每1000 tokens，数据向量化仅为0.0004美金/每1000 tokens，便宜一个数量级，考虑到其对搜索精准度的提升，经济价值非常高。

- **市场空间**：

1、中国数据库市场规模2022年约300亿人民币，预计到2025年将达到约500亿人民币。

2、用于处理多模态/非结构化数据的非关系型数据库在营收占比约三成，预计未来占比随着非结构化数据的增加逐年提升（据IDC预测到2025年中国的数据量将增长到48.6ZB，其中80%是非结构化数据）。

3、向量数据库同样用于非结构化数据的处理分析需求，但其主要应用在于生成应用厂商接入大模型推理的场景，我们推测到2025年其将占非结构化数据处理需求约三成，同时向量数据库由于其独特的向量化数据的存储形式，数据向量化后相比传统的非结构化数据存储会带来较大膨胀，因此其价格将会数倍于传统的非关系型数据库产品。

投资建议

➢ 建议关注星环科技：

- 我们认为向量数据库会在 Infra 领域占有重要位置，甚至 LLM + Vector DB 的组合会对传统的非结构化数据库产生冲击，因为这一组合对非结构型数据的理解和利用效率会显著高于传统的数据索引和存储模式。

- 星环科技主营各类数据库和大数据产品，公司研发实力雄厚有相关技术储备。

风险提示

- 技术研发不及预期：

向量数据库目前仍然以海外产品为主，国内厂商存在技术推进不及预期的风险。

- 市场开拓不及预期：

向量数据库的应用直接受大模型应用影响，尽管目前国内大模型市场发展迅速，但发展时间仍较短，向量数据库在大模型中的商用落地如果推进缓慢则会影响相关公司市场开拓。

分析师声明及一般声明

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

一般声明

华福证券有限责任公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，该等公开资料的准确性及完整性由其发布者负责，本公司及其研究人员对该等信息不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，之后可能会随情况的变化而调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

在任何情况下，本报告所载的信息或所做出的任何建议、意见及推测并不构成所述证券买卖的出价或询价，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。投资者应自行决策，自担投资风险。

特别声明及投资声明评级

特别声明

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	评级	评级说明
公司评级	买入	未来6个月内，个股相对市场基准指数涨幅在20%以上
	持有	未来6个月内，个股相对市场基准指数涨幅介于10%与20%之间
	中性	未来6个月内，个股相对市场基准指数涨幅介于-10%与10%之间
	回避	未来6个月内，个股相对市场基准指数涨幅介于-20%与-10%之间
	卖出	未来6个月内，个股相对市场基准指数涨幅在-20%以下
行业评级	强于大市	未来6个月内，行业整体回报高于市场基准指数5%以上
	跟随大市	未来6个月内，行业整体回报介于市场基准指数-5%与 5%之间
	弱于大市	未来6个月内，行业整体回报低于市场基准指数-5%以下

备注：评级标准为报告发布日后的6~12个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中，A股市场以沪深300指数为基准；香港市场以恒生指数为基准；美股市场以标普500指数或纳斯达克综合指数为基准（另有说明的除外）。

诚信专业 发现价值

联系方式

华福证券研究所 上海
公司地址：上海市浦东新区浦明路1436号陆家嘴滨江中心MT座20楼
邮编：200120
邮箱：hfyjs@hfzq.com.cn