



腾讯云 **ES8** 新一代高性能高精度 向量检索引擎

7月27日, 2024

黄国航, 腾讯 ES 专家工程师



议程



微信读书的 **AI** 问书系统以及背后的技术架构



腾讯云 **ES** 集成混元大模型的一站式 **RAG** 技术平台



腾讯云 **ES** 在向量检索引擎的自研内核优化

通用大语言模型LLM 已火爆席卷全球



.....

通用大语言模型

幻觉?

数据实效性?

安全?

腾讯云
ES
RAG

公网最新数据
企业内部数据
版权数据
.....

微信读书“AI问书”简介

快乐和心理变化，唯有此，产品才能有

复制

划线

写想法

书摘

查询

AI

AI问书

当前，我国正处于消费结构升级的发展时期，新的消费热点不断出现，同时，互联网的快速发展也给消费领域带来了

如果我们在阅读过程中遇到不懂的名词或者片段，就可以选中后点击“AI问书”，“AI问书”会在相关资料中即可在快速获得相关内容的概念解释，同时还可以找到这本书中所有关于你这个名词的相关片段，来帮助你加深理解。

消费结构升级

消费结构升级是指各类消费支出在消费总支出中的结构升级和层次提高，它直接反映了消费水平和发展趋势。消费结构升级是一个复杂的过程，受多方面条件的制约，包括经济水平、社会产品供给状况、居民收入状况、国家消费政策、居民消费习惯以及国家消费环境等①。

一、消费结构升级的特征

1. 消费结构升级的结果就是消费者生活和工作方式的重大变化甚至是变革②。

2. 消费结构升级不仅将催生新供给、新产业、新业态，还将带来较高收入的就业机会③。

二、消费结构升级的影响因素

1. 消费结构升级是一个十分复杂的过程，受多方面条件的制约，包括经济水平、社会产品供给状况、居民收入状况、国家消费政策、居民消费习惯以及国家消费环境等①。

2. 人口对经济结构的影响，除了体现为人口年龄结构引发储蓄率、投资率变化，同时体现在对消费结构升级的

以下为书籍引源

① 引自白玉苓《消费心理学》

消费心理学是一门建立在经济学、管理学、心理学等基础上的应用型学科，它以市场活动中消费者心理现象...

② 引自杨柳等《现代广告学原理与实务》

消费心理学是研究人们在生活消费过程中，在日常购买行为中的心理活动规律及个性心理特征的学问

③ 引自刘善武等《商业活动中的影响力法则》

消费心理学作为消费经济学的一个重要分支，研究消费者的心理，对消费者来说，可以提高消费的效益，而...

④ 引自书胜编著《消费行为心理学》

消费心理学旨在理解消费者的需求、动机和行为，以及如何影响消费者购买和消费决策

⑤ 引自傅师申《纺织服装营销学》

作为研究人类心理活动规律的一门科学，消费心理学对人类各种心理活动的研究内容，主要包括对人类心理...

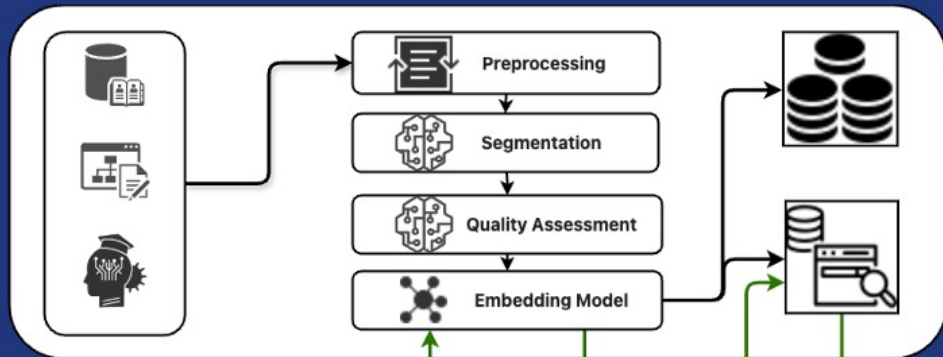
⑥ 引自陈志霞《社会心理学》

消费心理学研究消费者在消费过程中的各种社会心理现象和规律，主要涉及消费者偏好、消费决策的过程及...

C

微信读书业务的 AI 问书背后的 RAG 架构以及初期面临的挑战

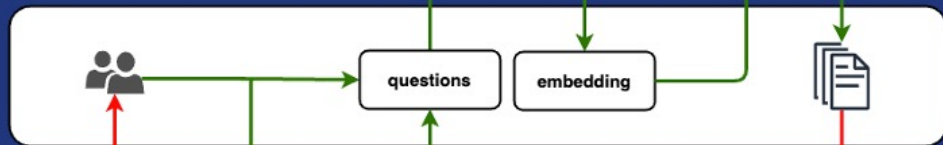
1. 书籍处理



超 10 亿级的向量规模

开发、运维成本高

2. 用户提问



ToC 业务，稳定性要求高，5个9

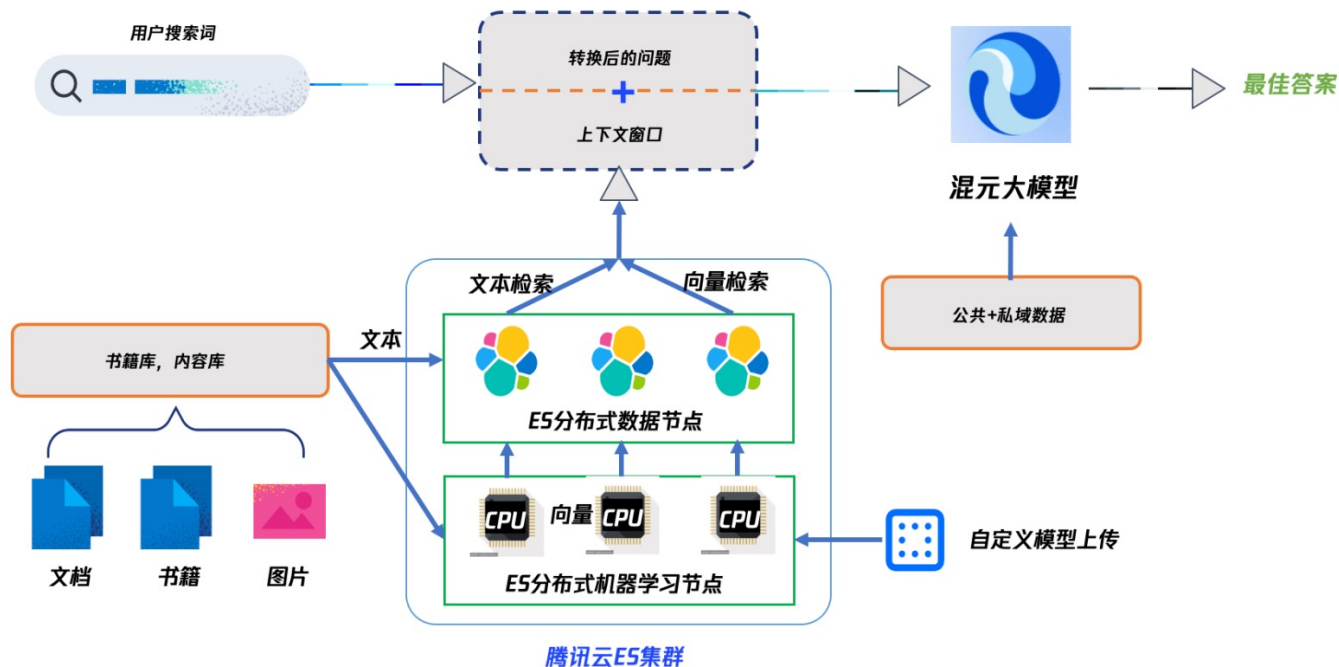
性能要求高，延时要求 $\leq 50\text{ms}$

3. 大模型回答



通用大模型幻觉，安全性等问题

腾讯云 ES 为微信读书提供的解决方案： 集成混元大模型的一站式 RAG 技术平台



主要能力：
内嵌向量化模型
文本向量混合搜索
多路召回
安全高可用
Kibana 可视化
集成混元一站式
--- 模型仓库
--- 模型精调
--- 模型部署

腾讯云ES为微信提供的向量多路召回混合检索



微信读书为何会选择腾讯云ES？区别于传统RAG的优势在哪？



开发实现效率高

一站式技术方案，上手容易，
研发实现高效

运维投入低

基于ES独立架构即可完成向量生成->
向量存储->向量检索的全流程操作。



高准确率

文本+向量的混合搜索，有效
应对各场景下的检索需求。

查询效率高

高并发场景下，十亿级向量检索查询延
迟低至毫秒级。



存储成本低

支持LZ4、Deflate，ZSTD等
压缩算法、有效降低存储成本。

稳定可靠

自研基于内存的熔断限流策略，有效应
对高并发与大查询，为业务保驾护航。



腾讯云 ES 通过信通院检索增强生成 RAG 技术标准

中国信通院云计算与大数据研究所大数据与智能化部依托中国通信标准化协会大数据技术标准推进委员会（CCSA TC601），联合40余家企业的70余位专家共同编制完成了《检索增强生成（RAG）技术要求》标准。

RAG 标准包含知识库构建能力、知识检索能力、内容生成能力、质量评估能力、平台能力等五大能力域，17个能力子域，50个能力项。

知识库构建能力	知识检索能力	内容生成能力	质量评估能力	平台能力
数据读取	查询优化	Prompt优化	评估指标	用户管理
数据预处理	检索能力	大模型生成	评估方法	数据权限管理
内容增强	检索结果优化	生成内容优化	评估数据集	模型权限管理
索引构建				资源及任务管理
知识库管理				日志管理

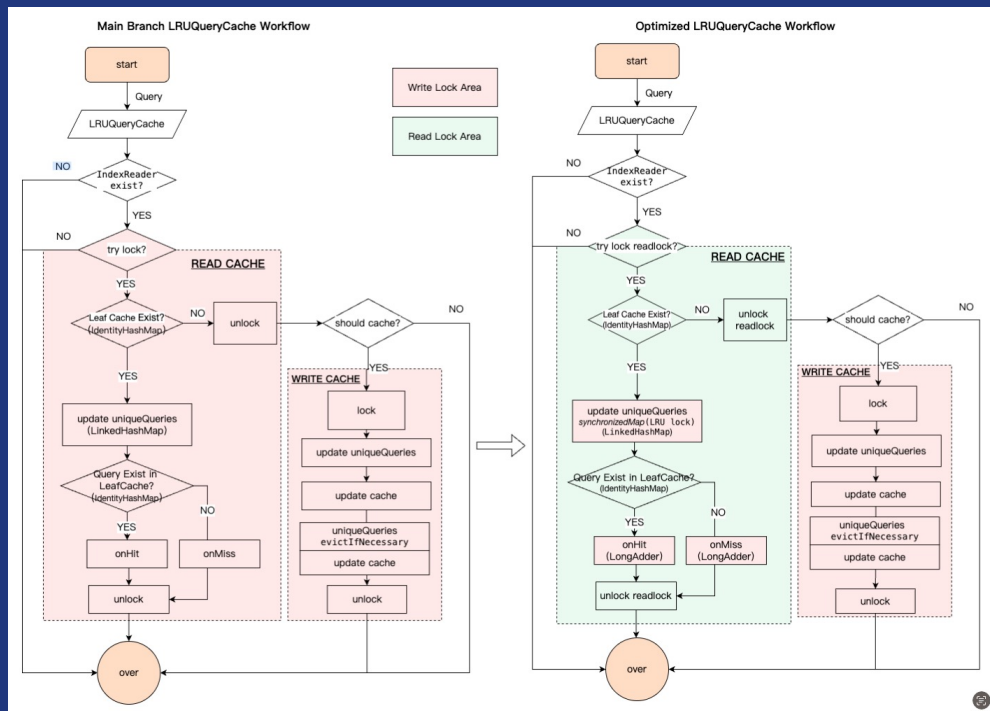
腾讯云 ES 是首个通过 RAG 权威标准认证的企业

腾讯云 ES 是国内公有云首个从自然语言处理，到向量生成/存储/检索，并与大模型集成的端到端一站式技术平台

腾讯云ES 在向量检索领域的四大技术亮点



腾讯云ES 在向量检索引擎的自研内核优化



doc count	field cardinality	term 数量	优化前 QPS	优化后 QPS	提升
3000 0000	1000 000	1000 0	160	473	191.9 %

PR已合并:

<https://github.com/apache/lucene/pull/13306>

腾讯云ES 在向量检索引擎的性能优势

数据集	测试 集群 规格	维度	数据量 分片和副本	索引类 型	召回率	Milvus	腾讯云ES 8.13.3 自研增强版	
						QPS	QPS	P90耗时
sift-128-euclidean	4CP U 8GB 3节 点	128维	100W; shardNum =1; replicaNu m=2.	HNSW; m=16; efCons truntion =200.	99%	8653	18479	4ms
chinese100w-768- angular		768维				2166	4586	5ms
gist-960-euclidean		960维				480	4136	9ms

腾讯云ES 跟紫霄GPU 相结合

紫霄 V1 具有**高效能、高吞吐、高带宽**等特点，设计算力规格 NVIDIA A10 相当，显存带宽比 A10 高30%。

在常见中小模型上，紫霄相比 NVIDIA T4通常有 100% 以上的性能提升，相比 NVIDIA A10 有 **20%+** 的性能优势。



**THANK
YOU**