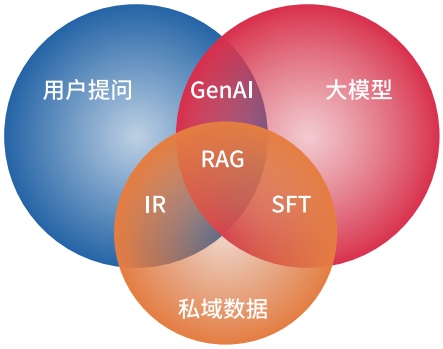
拓尔思 TRS
股票代码 300229

TRS 海贝向量数据库系统

全文向量双擎驱动，助力DeepSeek应用落地

随着DeepSeek的爆火，基于大语言模型（LLM）的人工智能技术正在飞速发展。然而，由于LLM的参数量巨大，导致其训练成本较高、训练时间较长，进而使得数据更新相对缓慢，难以及时学习到最新的知识。此外，由于训练数据的局限性，以及无法访问用户私域信息等原因，LLM在语义理解方面可能会出现不准确的情况，即所谓的“幻觉”。因此，基于大模型的检索增强生成（RAG）已经成为应用落地的通用范式。



混合检索成为趋势

传统搜索依赖关键词字面匹配（如倒排索引），虽效率高但存在语义理解缺陷：如无法处理同义词（“智能电话”vs“智能手机”）、多义词歧义（“Java”含义混淆）及上下文意图（“冷”指温度或病症），且强依赖用户输入精准度。为此，向量搜索通过嵌入技术（如BERT）将文本映射为高维向量，基于语义相似度（如余弦计算）突破字面限制，实现跨语言、多义词区分及多模态扩展。而混合检索进一步融合两者优势：先用关键词快速筛选明确条件（如“2025款手机”），再以向量模型解析模糊语义（“拍照好”关联“高像素”），最后综合排序，成为解决架构复杂性与语义需求矛盾的主流方案。

产品定位

TRS海贝向量数据库系统采用分布式架构设计，通过全文与向量的融合计算，实现关键词匹配、语义搜索及跨模态内容的联合检索。系统支持大规模数据的存储和检索，具备低延迟、高召回率等特性，能够为生成模型提供动态知识补充，满足企业级RAG场景的实时推理与数据扩展需求。

消除幻觉，提供事实支撑　**实时更新知识，突破时间限制**　**垂直领域精通，提升专业能力**

	关键词搜索	向量搜索
核心原理	倒排索引，BM25	嵌入模型（如BERT），向量索引（IVF/HNSW等）
优点	速度快：倒排索引结构成熟，实时性高 精准匹配：支持标红，可解释性强 技术成熟：工具生态完善	语义理解：捕捉上下文和隐含意图 扩展性强：自动关联同义词、近义词、跨语种 多模态支持：可处理文本、图像、音频的跨模态检索
缺点	语义盲区：无法处理同义词、多义词 依赖关键词：需用户精准输入 灵活性差：长尾查询覆盖率低	计算成本高：需GPU加速，检索延迟较高 数据依赖性：依赖训练数据质量 解释性弱：相似度评分缺乏直观依据
典型场景	文档精确检索（如法律条款、代码搜索） 日志分析（如ERROR日志筛选） 数据统计（趋势分析）	语义推荐（如电商“透气鞋”→网面运动鞋） 问答系统（如“如何缓解压力”→心理疏导方法） 跨模态搜索（如图搜商品、音频找歌词）

技术亮点

TRS海贝向量数据库不仅支持原生向量搜索，还支持文本、数值、地理位置信息等多种数据类型，满足企业数据多样性的需求。

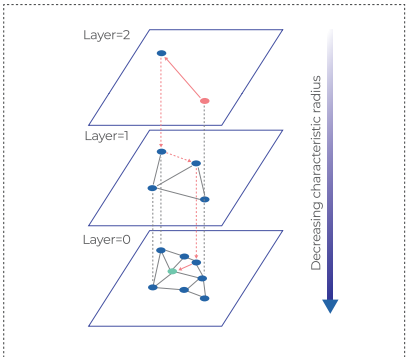
原生向量支持	与文本/数值/地理位置数据统一存储	分布式架构下的自动分片与负载均衡，支持横向扩容

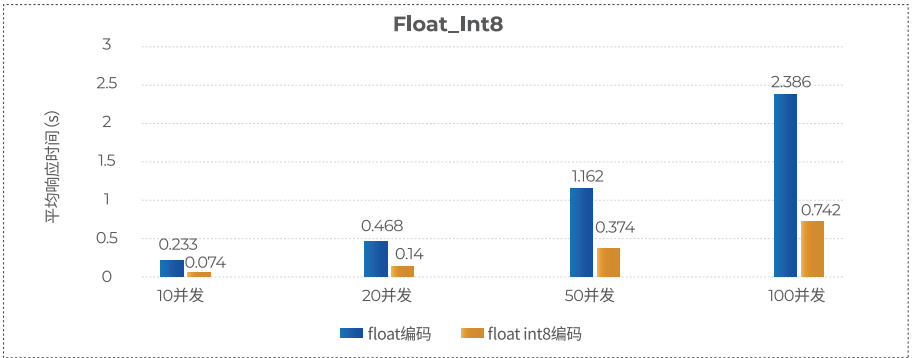
核心能力

高性能检索算法

TRS海贝向量数据库支持高效的HNSW索引，可以支持单机亿级数据毫秒级响应，满足用户海量数据大并发的需求。

近似最近邻搜索（ANN）算法优化	支持HNSW（Hierarchical Navigable Small World）索引	SQ量化，降低资源消耗，提高性能





企业级能力

面对复杂的企业级市场，用户不仅需要考虑数据库的性能，还需要考虑数据的安全性，开发运维的易用性，落地成本以及扩展能力等。TRS海贝向量数据库集成了大量企业级功能，能够满足用户的非功能性需求。

可视化管理台

系统状态监控告警、
系统任务调度、集群管理

安全合规

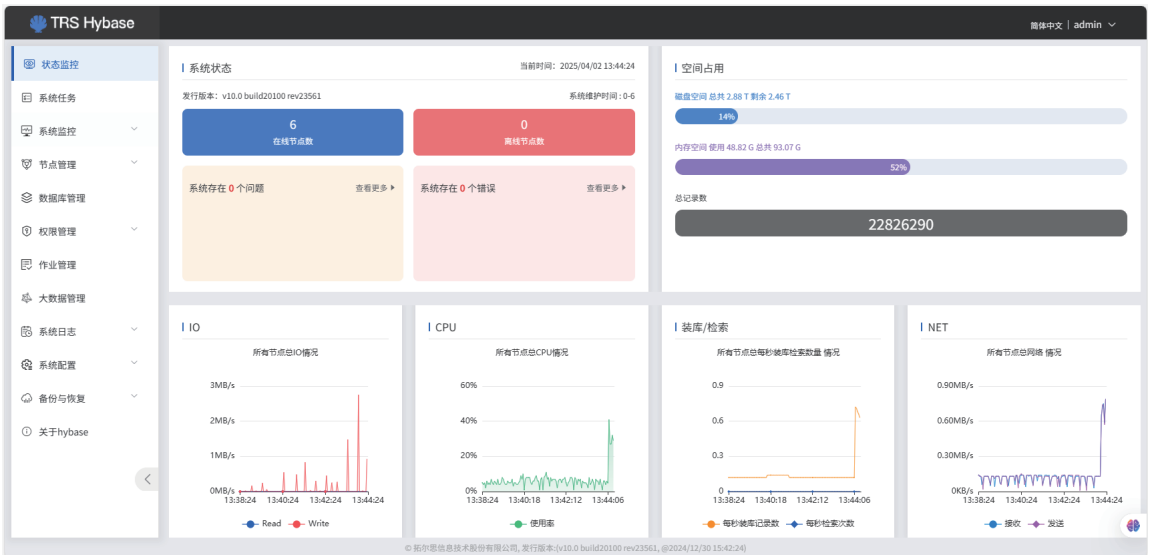
权限控制、数据加密、审计日志

极致扩展

横向扩展至上千节点

数据生命周期管理

读写分离，冷热分区，访问隔离



竞品分析

	海贝Hybase	开源Milvus
向量索引	HNSW，IVF	HNSW，IVF，PQ等
量化	SQ量化	SQ量化、PQ量化
数据类型	全文+标量+向量+GIS	向量+标量
全文检索	全语种分词器，支持多种检索模式（全文、拼音、同义词、简繁转化、词根处理、距离运算等）	2.5版本引入
实时性	强一致性，数据写入即可查询	最终一致性，数据同步存在延时
可视化管理台	支持集群监控，满足日常数据运维需求	简单的管理台，监控依赖第三方组件
集群架构	对等节点机制，架构简单清晰，运维简单	云原生架构，架构复杂，运维复杂度较高
权限管理	完善的权限管理机制	开源版本只有简单的访问控制
生态支持	可视化低代码的ETL工具	依赖第三方组件

基于海贝的RAG落地方案

检索增强生成解决方案 (RAG)

检索增强生成（Retrieval-Augmented Generation，RAG）被认为是当下解决大模型幻觉的最有效手段之一。基于拓尔思海聚数据整合系统（TRS ETL）、拓天大模型以及TRS海贝向量数据库构建的检索增强生成解决方案，具有集成度高，生成效果好等特点。

基于海贝的RAG解决方案

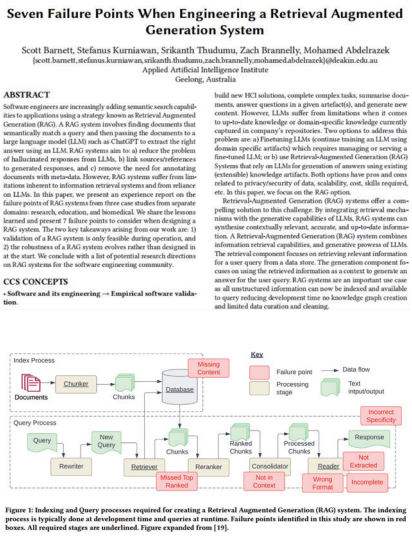
采用TRS海贝向量数据库和RAG解决方案，可以快速搭建智能问答系统，利用索引和搜索算法来快速匹配问题与知识库中的相关内容，从而提供更准确的答案。

TRS RAG解决方案技术架构图

工程化落地方案

RAG工程化落地”一看就会,一用就废”?

数据挑战 数据治理与质量控制	**内容缺失:**当大模型检索不到数据,就开始胡编乱造
	排名缺失:上下文长度限制,排名靠前才会被检索返回
技术挑战 检索架构/效率/精度	**合并缺失:**虽然数据被成功检索,但合并提取时被忽视
	性能瓶颈:难以有效管理和处理大量数据,系统过载
成本挑战 性能与资源的平衡点	**成本控制:**向量搜索成本过高,模型资源消耗不可控
	数据安全:防范恶意输入、确保安全输出,数据防泄漏



Seven Failure Points When Engineering a Retrieval Augmented Generation System

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, Mohamed Abdelrazek
{scott.barnett,stefanus.kurniawan,srikanth.thudumu,zach.brannelly,mohamed.abdelrazek}@deakin.edu.au
Applied Artificial Intelligence Institute
Geelong, Australia

ABSTRACT

Software engineers are increasingly adding semantic search capabilities to applications using a strategy known as Retrieval Augmented Generation (RAG). A RAG system involves finding documents that semantically match a query and then passing the documents to a large language model (LLM) such as ChatGPT to extract the right answer using an LLM. RAG systems aim to: a) reduce the problem of hallucinated responses from LLMs, b) link sources/references to generated responses, and c) remove the need for annotating documents with meta-data. However, RAG systems suffer from limitations inherent to information retrieval systems and from reliance on LLMs. In this paper, we present an experience report on the failure points of RAG systems from three case studies from separate domains: research, education, and biomedical. We share the lessons learned and present 7 failure points to consider when designing a RAG system. The two key takeaways arising from our work are: 1) validation of a RAG system is only feasible during operation, and 2) the robustness of a RAG system evolves rather than designed in at the start. We conclude with a list of potential research directions on RAG systems for the software engineering community.

CCS CONCEPTS

• Software and its engineering → Empirical software validation.

build new HCI solutions, complete complex tasks, summarise documents, answer questions in a given artefact(s), and generate new content. However, LLMs suffer from limitations when it comes to up-to-date knowledge or domain-specific knowledge currently captured in company's repositories. Two options to address this problem are: a) Finetuning LLMs (continue training an LLM using domain specific artifacts) which requires managing or serving a fine-tuned LLM; or b) use Retrieval-Augmented Generation (RAG) Systems that rely on LLMs for generation of answers using existing (extensible) knowledge artifacts. Both options have pros and cons related to privacy/security of data, scalability, cost, skills required, etc. In this paper, we focus on the RAG option.

Retrieval-Augmented Generation (RAG) systems offer a compelling solution to this challenge. By integrating retrieval mechanisms with the generative capabilities of LLMs, RAG systems can synthesise contextually relevant, accurate, and up-to-date information. A Retrieval-Augmented Generation (RAG) system combines information retrieval capabilities, and generative prowess of LLMs. The retrieval component focuses on retrieving relevant information for a user query from a data store. The generation component focuses on using the retrieved information as a context to generate an answer for the user query. RAG systems are an important use case as all unstructured information can now be indexed and available to query reducing development time no knowledge graph creation and limited data curation and cleaning.

Figure 1: Indexing and Query processes required for creating a Retrieval Augmented Generation (RAG) system. The indexing process is typically done at development time and queries at runtime. Failure points identified in this study are shown in red boxes. All required stages are underlined. Figure expanded from [19].

从理论到工程实践,开发和优化一个高效的RAG系统并非易事,RAG系统面临数据、技术、成本等诸多挑战,需要兼顾技术深度与工程实践,从检索架构、检索精度、生成质量到系统稳定性、安全性等,均需精心设计。

工程化:提供RAG工程化落地规划、实施服务

梳理需求 → 盘点数据 → 部署系统 → 测评 → 本地训练 → 使用 → 反馈 → 运营

咨询服务

- 用户算力有多少?
- 用户数据在哪里?
- 应用场景是什么?
- 数据规模有多大?

结合算力、数据、场景提供RAG解决方案

快速部署 24h

- 安装部署
- 数据抽取
- 系统配置
- 运行验证

简化部署流程,提供部署安装服务

RAG工程优化

- 提示词工程
- 嵌入模型优化
- 重排模型优化
- 知识图谱引导

多方式助力提升召回效果,解决RAG常见问题

客户案例

新华社稿件传播分析-基于向量搜索的传播力比对

通过对指定监测范围内的中文网站、国内电子报纸、国内视频网站、移动客户端、英文媒体、新浪微博、微信平台等渠道的图片、视频等融媒体数据进行向量化存储，利用图片视频特征比对技术，获取相关稿件的浏览数、阅读量、评论数、点赞数、观看量等传播互动数据，为传播分析应用和版权监测系统的数据展示、对外数据服务等提供数据支撑。

贵州省智能问答-知识库问答

构建领导人知识库、互动问答知识库、政策文件知识库、办事指南知识库，实现知识库向量化存储。用户问题向量化，与知识库内的向量匹配，理解百姓通俗表达。通过RAG解决生成模型的知识局限性和信息及时性问题，赋能精准化、可信化政务服务智能体。

浦发银行-大模型应用体系建设

通过引入TRS海贝向量数据库，借助其产品能力，形成统一规范和管理体系，以实现多场景下向量数据的存储、管理与检索，推动大模型及知识库应用场景建设，提升金融服务智能化水平。

目标	**构建统一向量数据管理体系，支撑大模型及知识库应用场景**				
优势	**安全可控：**海贝向量数据库作为自主可控的国产化加密数据引擎，支持数据与索引的完全加密，具备金融级安全		**强大的检索能力：**海贝可以支持几乎所有常见数据类型的的快速检索，提供语义搜索、图片搜索、跨模态检索		
产品能力	**数据库基础能力** 多维向量管理 标量数据管理 相似距离算法 数据分区	**数据索引能力** 向量索引 标量索引 增量索引 图索引算法	**多场景检索能力** 精确检索 混合检索 向量检索 融合检索		
	权限管理能力	**集群运维管理能力**	**大模型生态集成能力**	**数据备份恢复能力**	**高可用可扩展能力**
环境	华为HCSO	云上裸金环境	虚拟机环境		

从关键词到语义理解，从文本到多模态

一站式解锁数据价值

拓尔思官方服务号　　拓尔思官方订阅号

服务热线：4006 300229　　E-mail：trs@trs.com.cn　　官网：www.trs.com.cn

总部地址：北京市海淀区建枫路(南延)6号院3号楼