



利用 MCP 玩转 Milvus 向量数据库



01

Milvus 向量数据库简介

02

Milvus MCP Server 使用介绍

03

Milvus 落地案例分享

01

Milvus 向量数据库简介



Zilliz——向量数据库的开拓者和全球领先者

2017

创始于

\$113 M

融资总额

100+

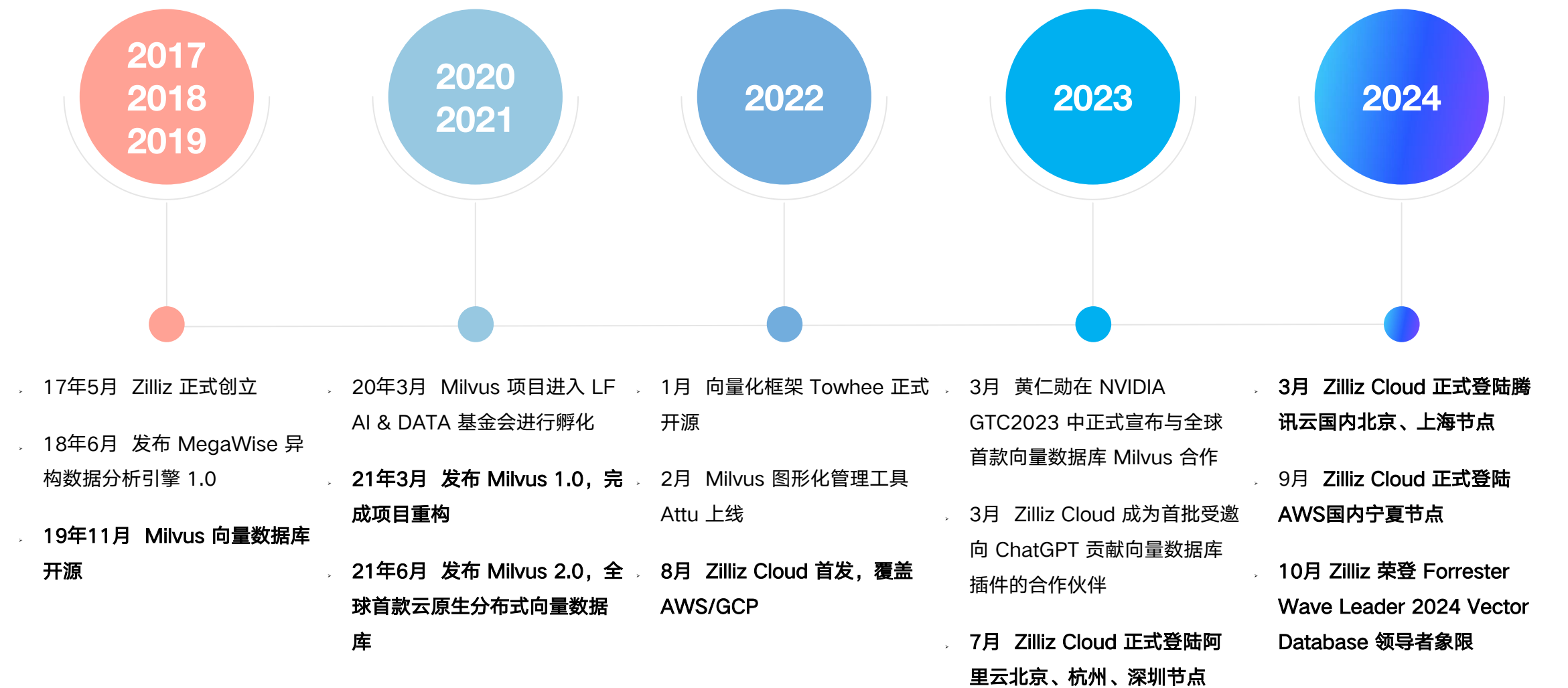
员工



作为全球最受欢迎的开源向量数据库 Milvus 的创造者，Zilliz 提供面向 AI 应用的新一代数据库技术，帮助企业便捷开发 AI 应用。以 AI 民主化为使命，Zilliz 致力于简化 AI 数据管理基础设施，通过向量数据库赋能更多企业。



发展历程



Milvus——全球最受欢迎的开源向量数据库

与超过 100000 家企业用户打磨 6 年，是全球范围内最成熟的分布式向量数据库



GitHub

10000+



Enterprise users

35K



stars

70M



docker pulls

3.1K



forks

DB-Engine最新排名



Rank			DBMS	Database Model	Score		
Aug 2024	Jul 2024	Aug 2023			Aug 2024	Jul 2024	Aug 2023
1.	1.	1.	Oracle +	Relational, Multi-model i	1258.48	+18.12	+16.39
2.	2.	2.	MySQL +	Relational, Multi-model i	1026.86	-12.60	-103.59
3.	3.	3.	Microsoft SQL Server +	Relational, Multi-model i	815.18	+7.52	-105.64
4.	4.	4.	PostgreSQL +	Relational, Multi-model i	637.39	-1.52	+17.01
5.	5.	5.	MongoDB +	Document, Multi-model i	420.98	-8.85	-13.51
6.	6.	6.	Redis +	Key-value, Multi-model i	152.71	-4.06	-10.26
7.	7.	↑ 11.	Snowflake +	Relational	135.97	-0.56	+15.34
8.	8.	↓ 7.	Elasticsearch	Search engine, Multi-model i	129.83	-0.99	-10.09
9.	9.	↓ 8.	IBM Db2	Relational, Multi-model i	123.00	-1.40	-16.23
10.	10.	10.	SQLite +	Relational	104.79	-5.16	-25.13
... ..							
79.	↑ 82.	↑ 103.	TiDB +	Relational, Multi-model i	3.59	+0.13	+0.14
80.	↑ 84.	↓ 73.	Microsoft Azure Table Storage	Wide column	3.43	+0.01	-2.33
81.	81.	↓ 59.	Microsoft Azure Data Explorer	Relational, Multi-model i	3.42	-0.15	-3.64
82.	↓ 79.	↓ 72.	SAP SQL Anywhere	Relational	3.41	-0.18	-2.38
83.	↑ 85.	↑ 90.	ArangoDB +	Multi-model i	3.40	-0.02	-0.97
84.	↑ 89.	↑ 187.	Milvus +	Vector	3.38	+0.26	+1.96
85.	↓ 83.	↓ 77.	Apache Accumulo	Wide column	3.32	-0.14	-2.14
86.	86.	↓ 78.	HyperSQL	Relational	3.21	-0.07	-2.03
87.	↑ 88.	↑ 130.	Pinecone	Vector	3.14	+0.01	+0.53
88.	↓ 87.	↑ 107.	InterSystems IRIS +	Multi-model i	3.09	-0.15	-0.17

Forrester WAVE

Zilliz delivers a cutting-edge, high-performance database designed for cloud scalability.

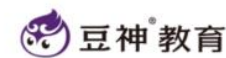
Known as the creator of Milvus, the leading open-source vector database, Zilliz excels in managing **vast amounts** of vector data. It leverages optimized storage, efficient management, and search capabilities. Zilliz is at the forefront of innovation, delivering **exceptional speed and efficiency in vector processing** and search to support **real-time AI applications**. With a customer centric approach, Zilliz integrates feedback from its robust community to continuously enhance its roadmap with new features and functionality. Zilliz's superior roadmap focuses on distributed index, advanced reranking of results, administration, broader data security certifications, multicloud support, and data intelligence.



我们的全球用户



我们的亚洲用户



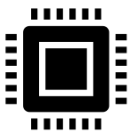
向量数据管理的主要挑战

• 有效存储



向量数据多为浮点或者二进制数据；数据压缩率低，存储成本高

• 高效计算

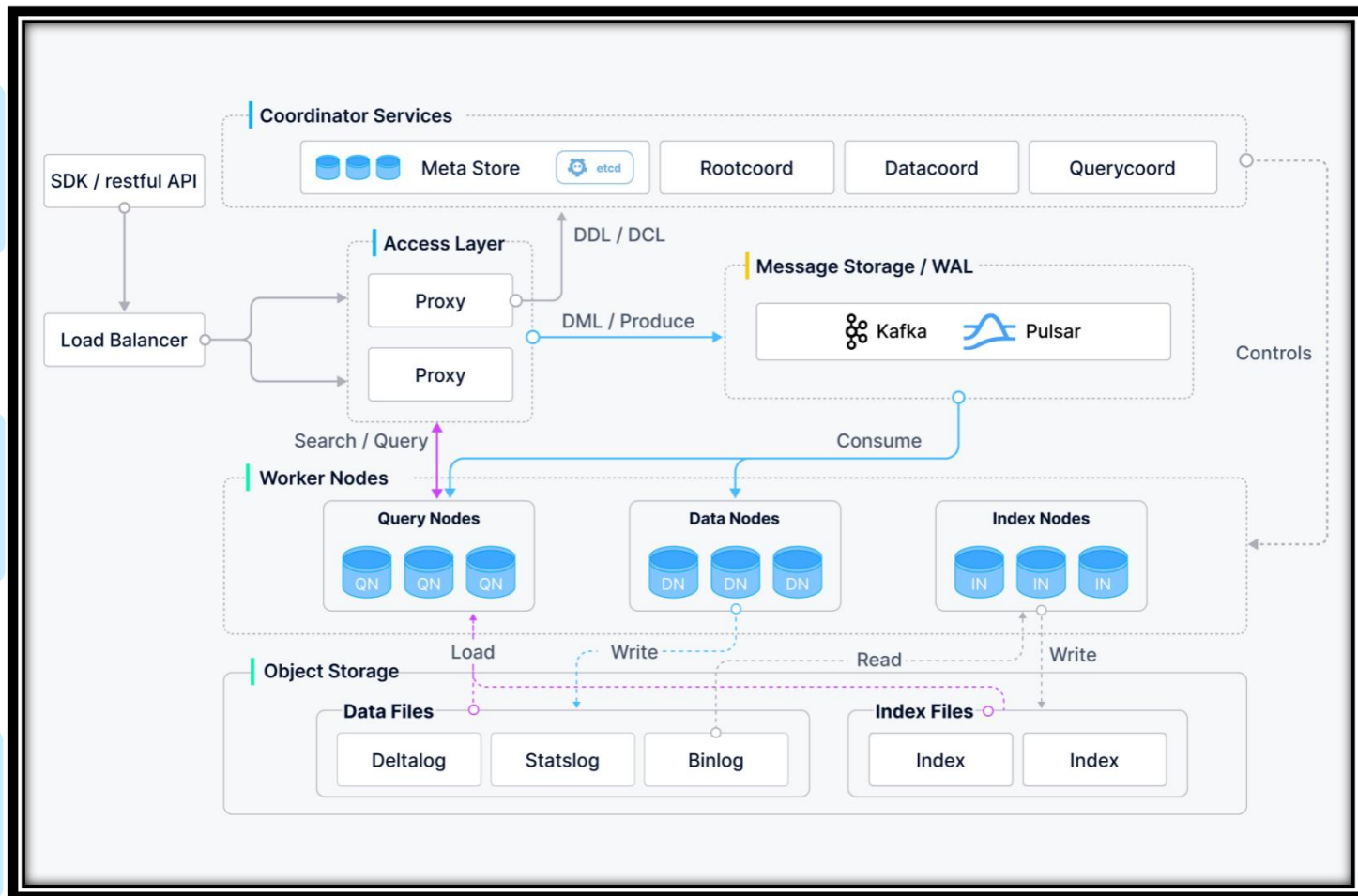


向量数据计算复杂度高，往往需要异构计算的能力

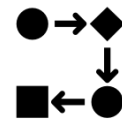
• 索引繁多



有树、图、哈希和倒排等多种向量索引，索引管理和使用成本高



• 扩展性



非结构化数据的快速增长对系统的扩展性要求越来越高

• 低延迟



在线业务对检索功能的毫秒级响应需求

• 混合查询



标量和向量混合存储查询的需求

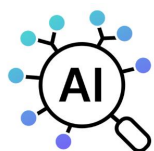
Zilliz - 构建开源+云的非结构化数据处理方案

Open Source

围绕 LLM 的开源非结构化数据解决方案



Deep
Searcher



+



zilliz
Cloud

Zilliz Cloud

开箱即用的全托管
向量检索服务



“开源+云”双擎联动

全球服务覆盖

安全合规，遍布全球的向量数据库基础设施

● AWS ● Google Cloud ● Azure ● Alibaba Cloud ● Tencent Cloud



Zilliz Cloud 功能架构

接口层

向量数据库 API

管控 API

功能层

向量数据库
框架

商业版
引擎

系统与数据
安全

多AZ高可用

监控告警

多实例
类型

访问控制

审计

组织与权限
管理

扩缩容与
弹性能力

备份恢复

鉴权

智能调优

生态工具
对接

基础设
施层



计算实例适配与
软硬联合调优

多层存储适配与
数据布局优化

网络控制

资源管理调度

负载均衡

池化

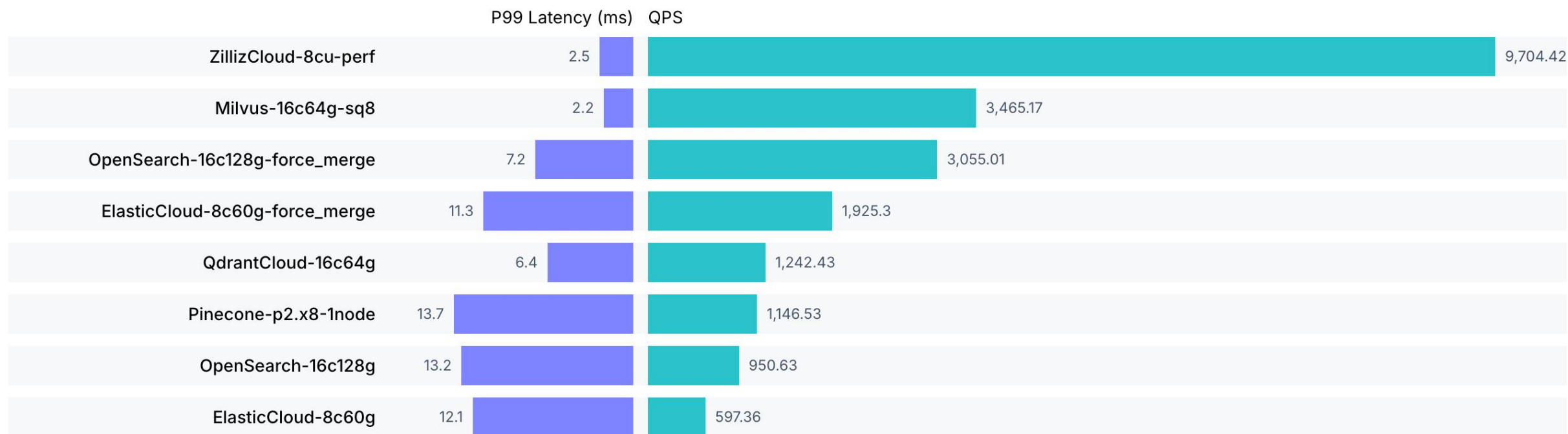
Zilliz VS Milvus 性能差异

Zilliz Cloud 相比 Milvus, 平均 QPS 提升 **2~5** 倍

Vector Search Latency and QPS at \$1,000 Monthly Cost

Dataset

1M Dataset



查看更多: <https://github.com/zilliztech/VectorDBBench>

Zilliz VS Milvus 性能差异

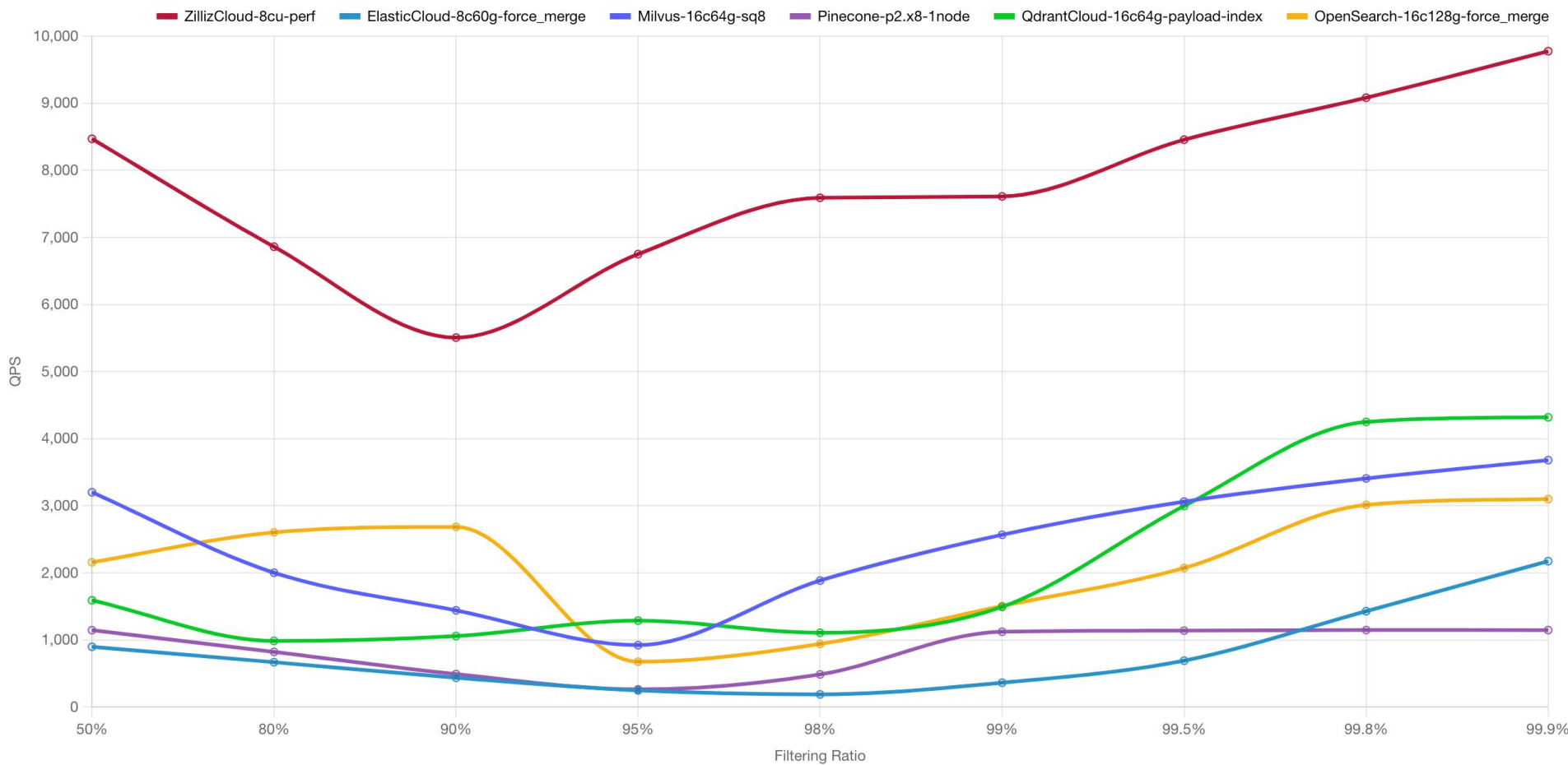
Filtering Performance

Dataset

1M Dataset

Plot values

☒ QPS ☐ P99 Latency ☐ Recall



查看更多: <https://github.com/zilliztech/VectorDBBench>

Zilliz Cloud - 助力全球企业构建云上全托管向量检索服务

维护成本低

low maintenance cost

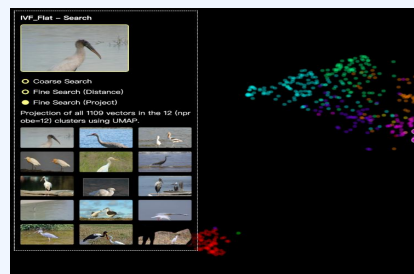
- 一键创建实例资源
- 动态扩缩容
- 无感升级迭代
- 完善的监控报警
- 多云支持



使用门槛低

Low threshold for use

- 免费试用实例
- 可视化界面
- 多语言SDK
- 丰富的生态支持
- 数据迁移



丰富的企业级特性

Enterprise-level features

- 7 * 24服务支持
- 99.95 SLA保障
- 数据备份, 订阅
- 组织架构管理
- Dedicated Cloud



安全放心

Data Security

- RBAC权限管理
- TLS, 白名单
- PrivateLink
- 审计日志
- SOC2合规认证



02

Milvus MCP Server 介绍



3 ↑ MCP Server

mcp-server-milvus

- <https://github.com/zilliztech/mcp-server-milvus>



zilliz-mcp-server

- <https://github.com/zilliztech/zilliz-mcp-server>

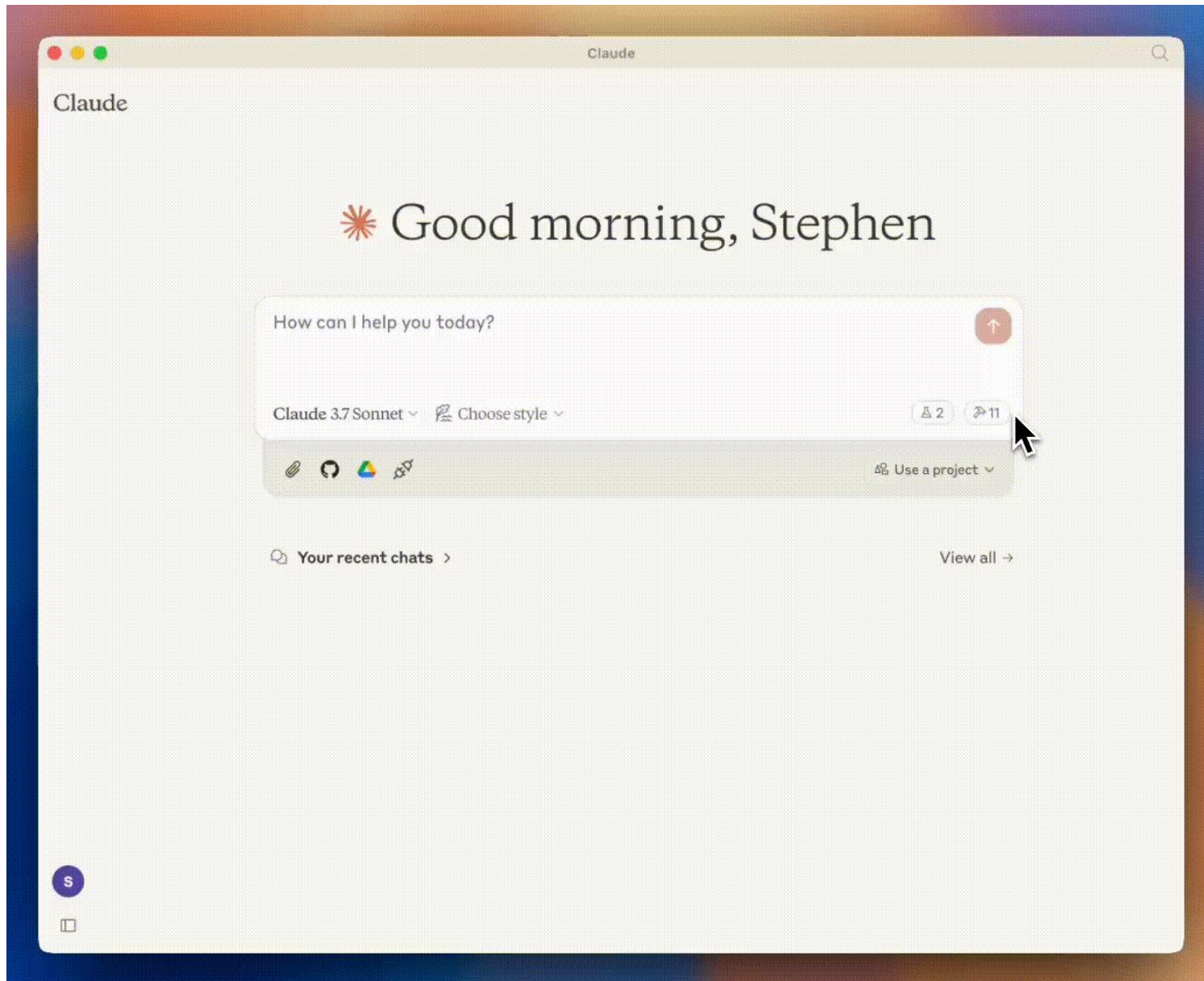


claude-context

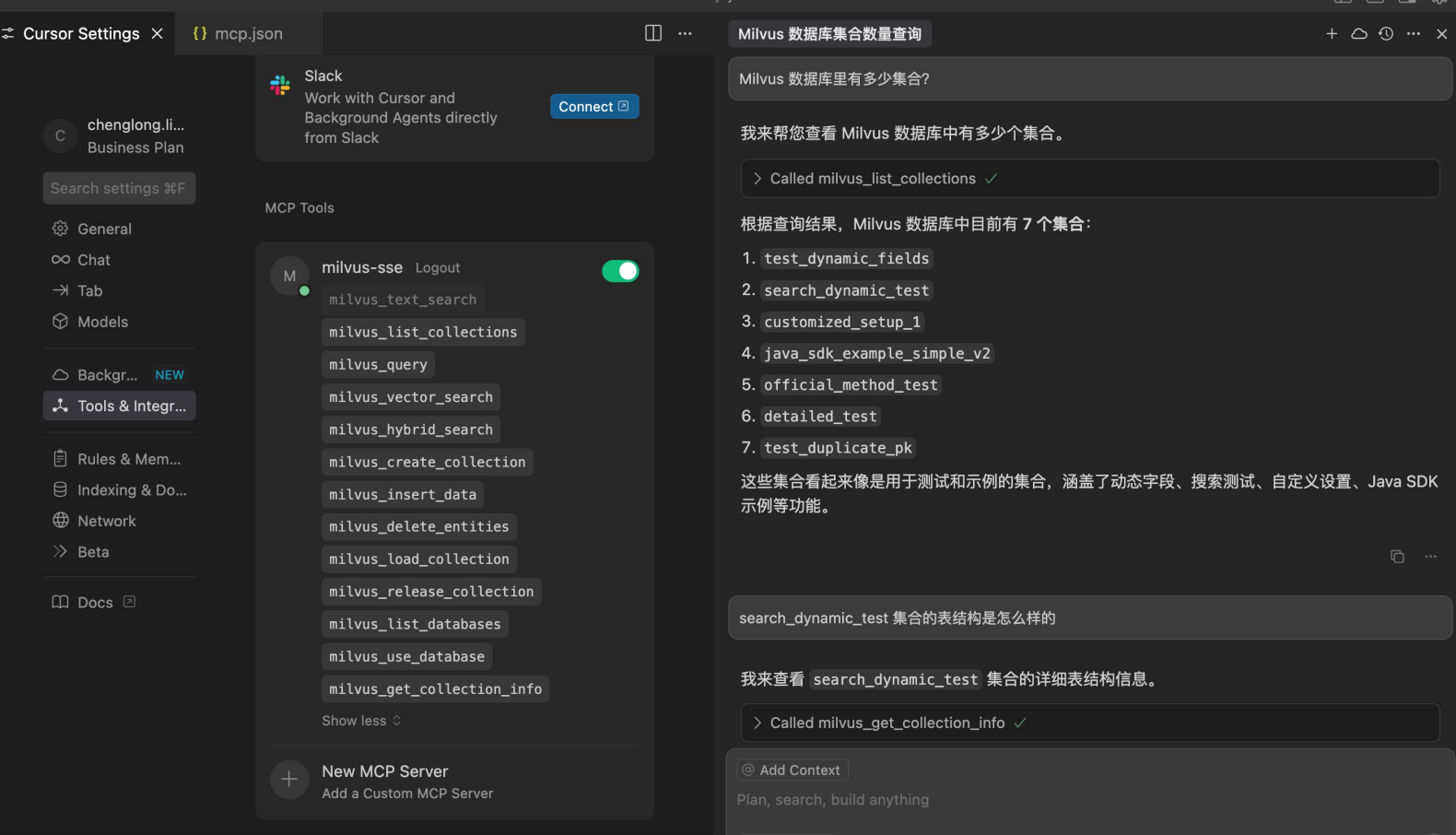
- <https://github.com/zilliztech/code-context>



Demo—Milvus MCP



Demo—Milvus MCP



Milvus 数据库集合数量查询

search_dynamic_test 集合的表结构是怎么样的

我来查看 search_dynamic_test 集合的详细表结构信息。

> Called milvus_get_collection_info ✓

search_dynamic_test 集合的表结构如下：

集合基本信息

- 集合名称: search_dynamic_test
- 自动ID: 启用
- 分片数: 1
- 分区数: 1
- 动态字段: 启用
- 一致性级别: 2

字段结构

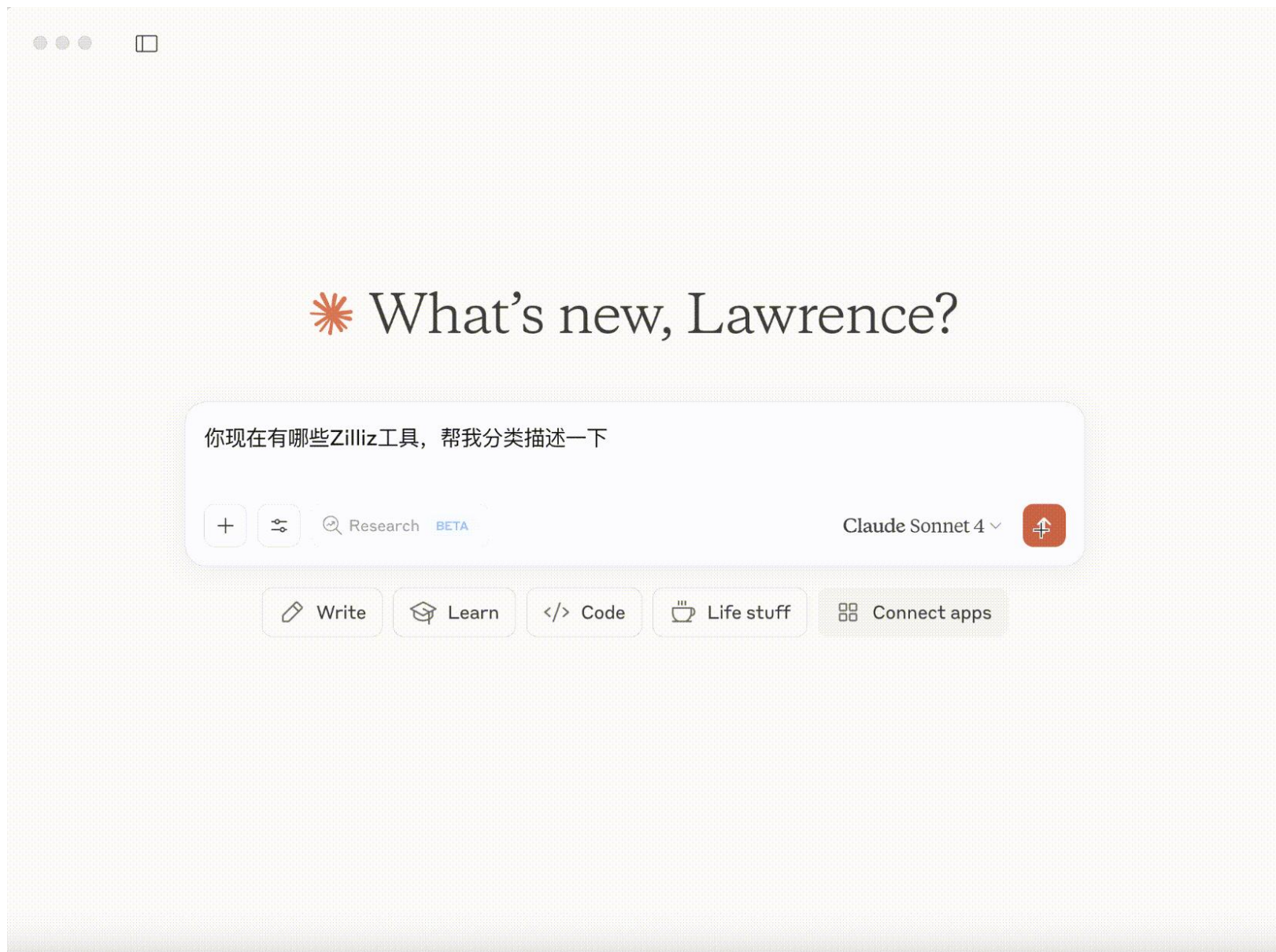
1. id 字段

- 字段ID: 100
- 类型: 5 (INT64)
- 主键: 是
- 自动ID: 是

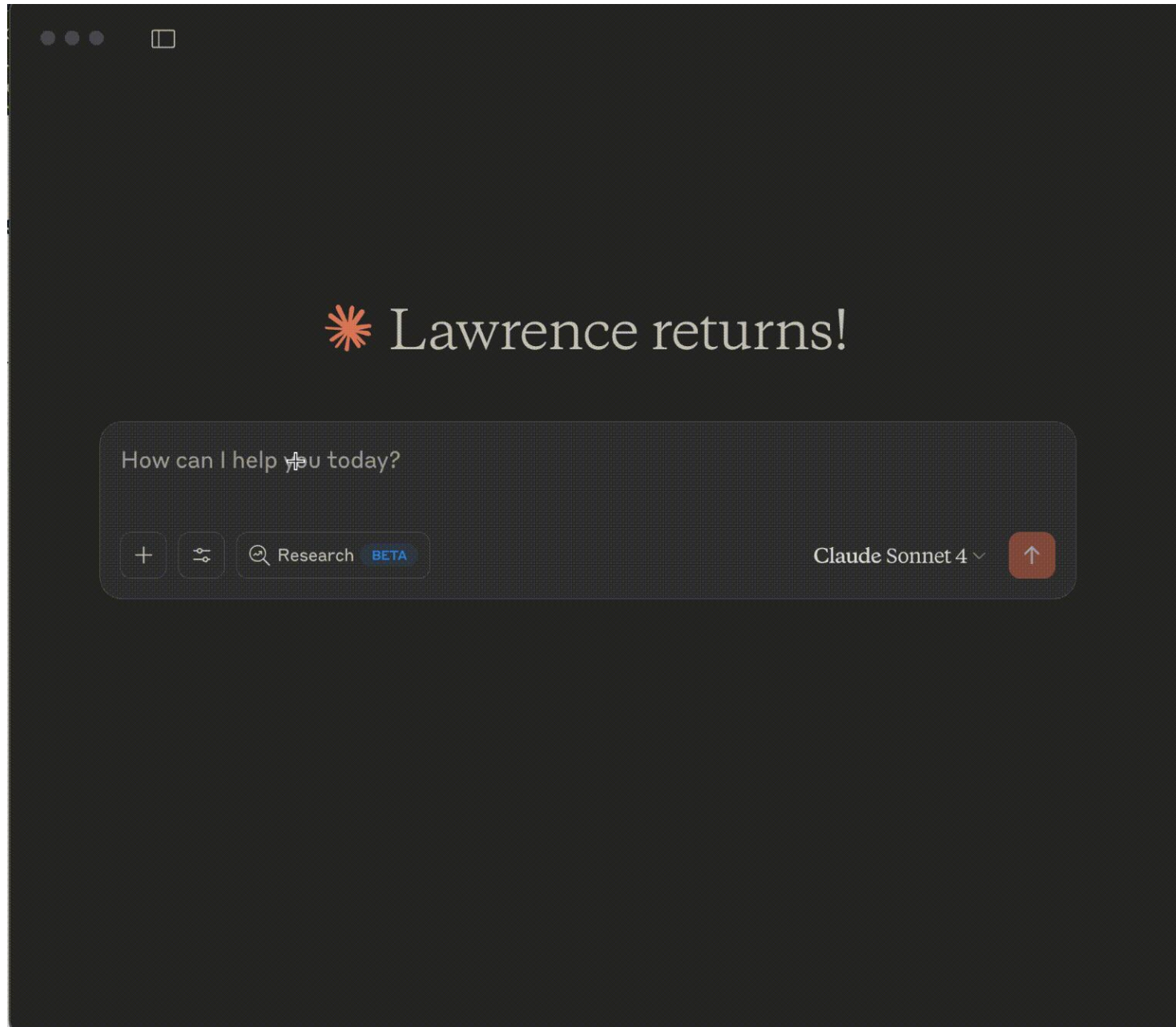
@ Add Context

Plan, search, build anything

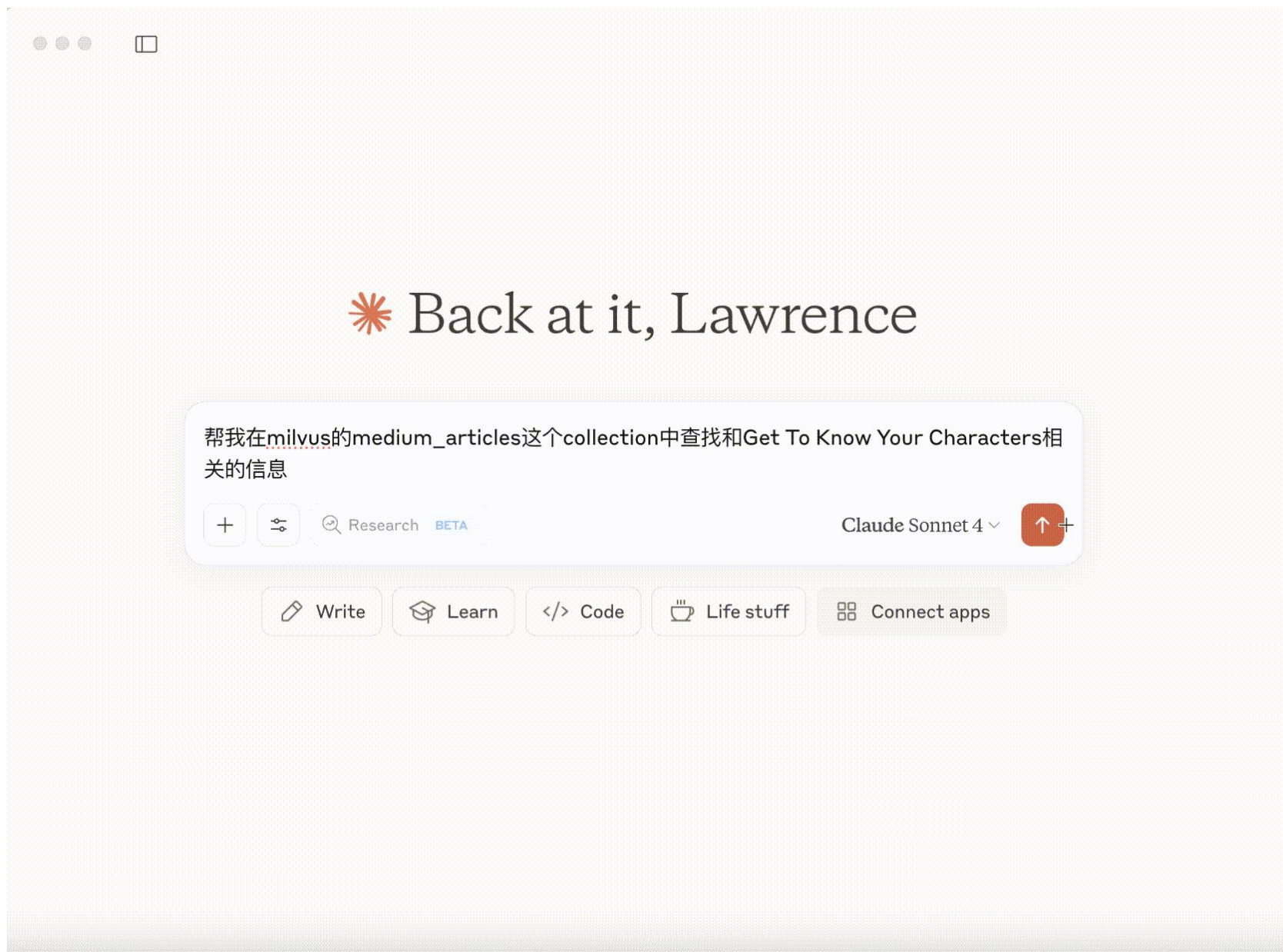
Demo—Zilliz MCP



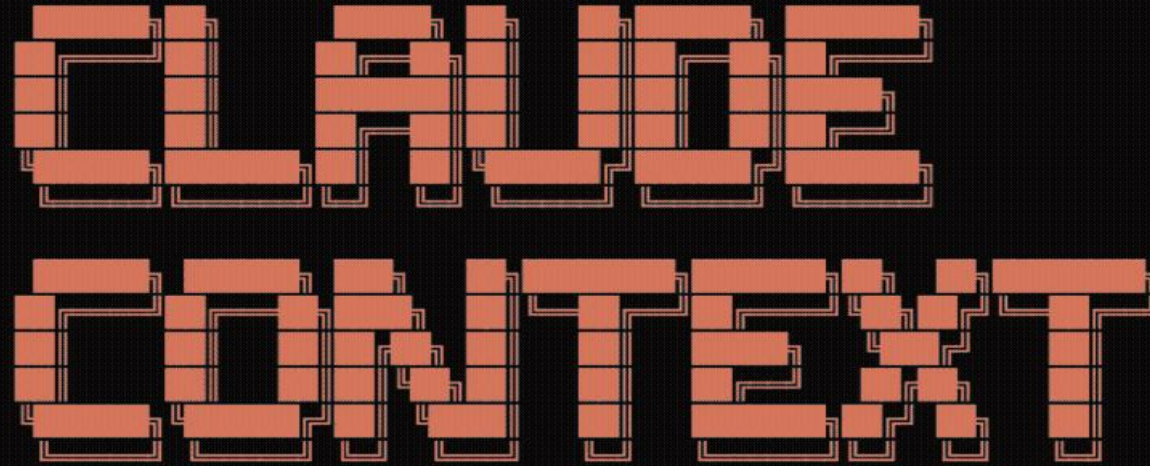
Demo—Zilliz MCP



Demo—Zilliz MCP



Demo—Claude Context



CLAUDE
CONTEXT

The image features the words "CLAUDE" and "CONTEXT" in a stylized, orange, blocky font. The letters are composed of thick orange strokes with a slightly irregular, hand-drawn appearance. The text is set against a solid black background. Behind the letters, there are faint, white, circuit-like patterns that resemble a grid of interconnected lines, giving the impression of a digital or neural network structure. The overall aesthetic is modern and tech-oriented.

Usage

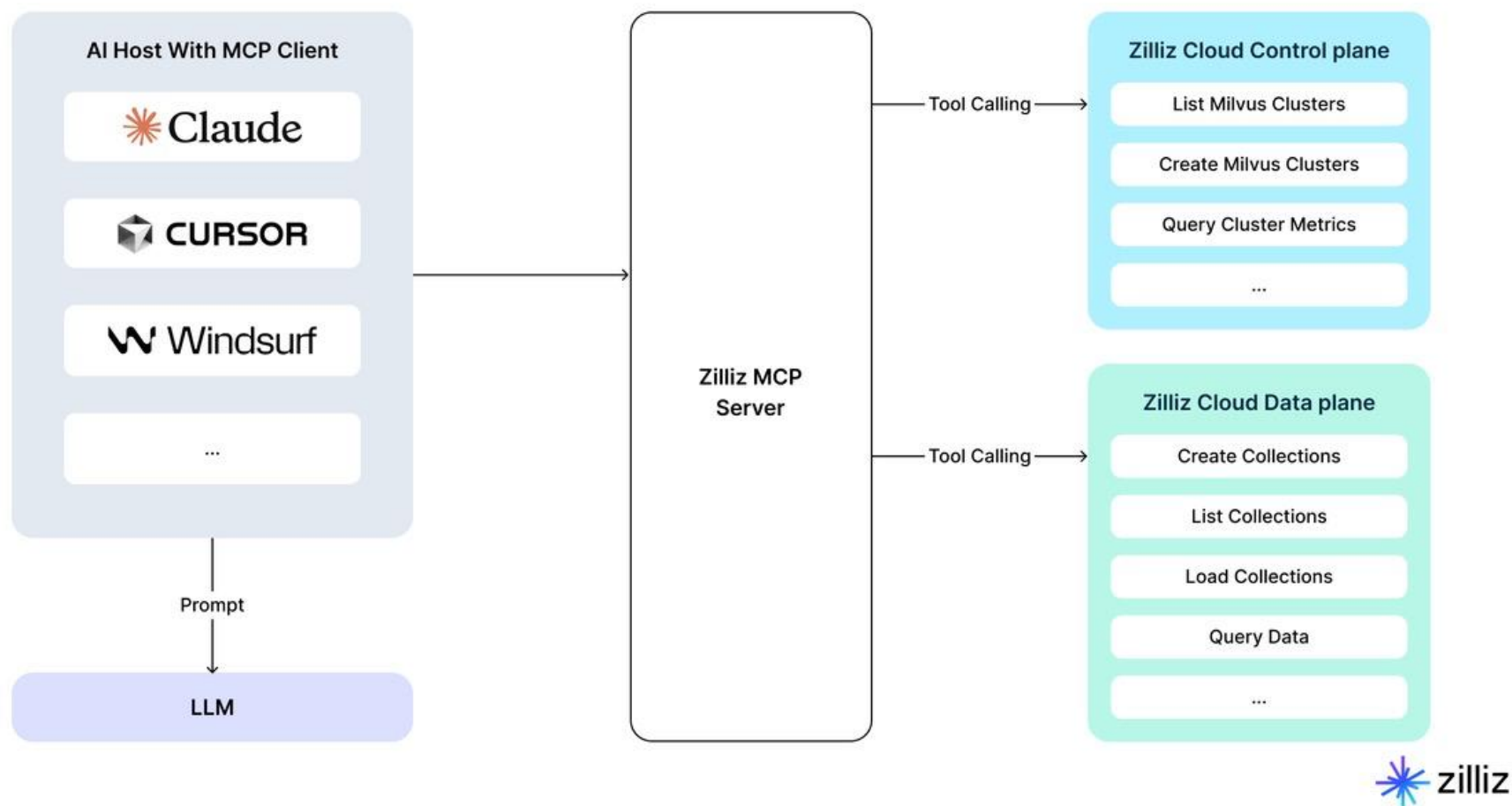
- `uv run src/mcp_server_milvus/server.py --sse --milvus-uri http://localhost:19530 --port 8000`

```
{
  "mcpServers": {
    "milvus-sse": {
      "url": "http://your_sse_host:port/sse",
      "disabled": false,
      "autoApprove": []
    }
  }
}
```

```
{
  "mcpServers": {
    "milvus": {
      "command": "/PATH/TO/uv",
      "args": [
        "--directory",
        "/path/to/mcp-server-milvus/src/mcp_server_milvus",
        "run",
        "server.py",
        "--milvus-uri",
        "http://127.0.0.1:19530"
      ]
    }
  }
}
```

工作原理——Zilliz MCP

How Zilliz MCP Server Works



Claude Context or Grep

- token消耗高：

每次查询都要传输大量无关代码到 LLM。Token 消耗巨大，成本直线上升。

- 时间成本大：

AI 需要反复试探和搜索，用户等待时间过长。开发效率大大降低。

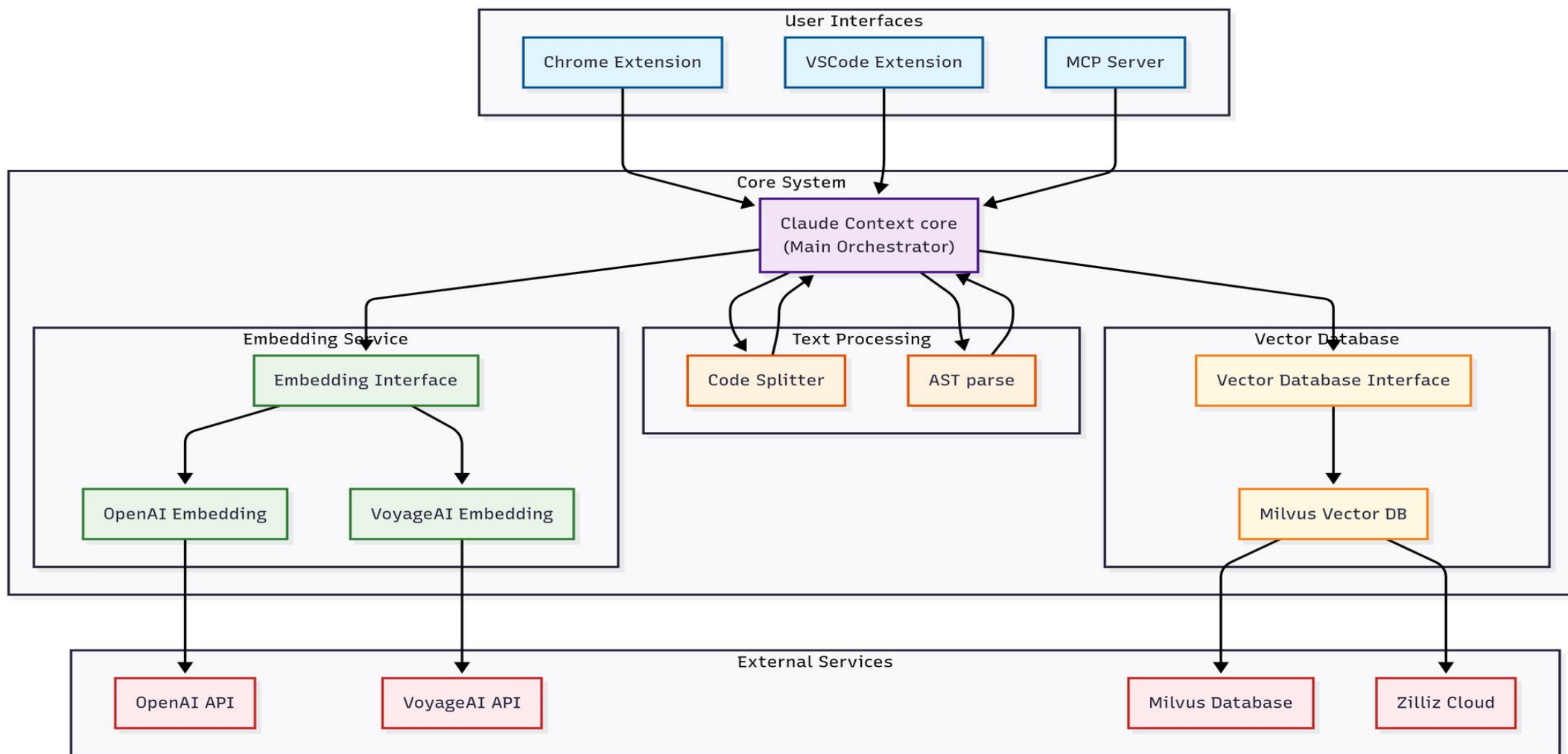
- 关键词搜索无法做上下文语义感知：

传统 grep 只能匹配字面意思，无法理解代码的语义关系和上下文含义。就像在大海里捞针，全靠运气。

工作原理——Claude Context

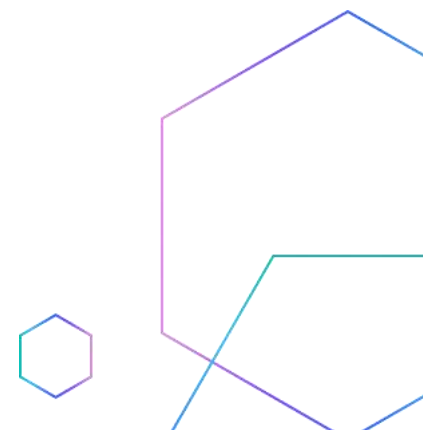
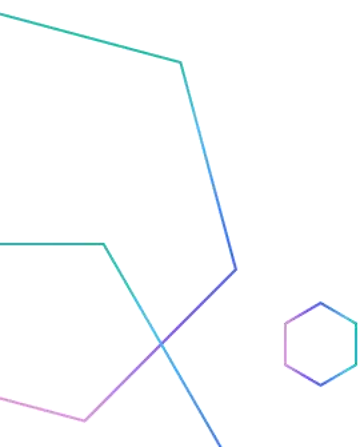
- 代码如何切分？
- 代码变动了怎么办？
- 索引和检索速度怎么保证？
- 怎么为海量的代码的 embedding 建立索引？

工作原理——Claude Context



03

Milvus落地案例



Zilliz Cloud适用业务场景



图片搜索

基于视觉或语义的图片相似性检索。



视频搜索

视频理解和基于内容的视频检索。



文本搜索

对新闻、文章、专利、文件等进行相似性搜索。



数据去重

数据聚类；文本、图像、音频、视频等数据去重。



跨模态搜索

支持灵活的多模态数据查询体验。



推荐系统

对用户和产品的丰富特征进行编码，探索他们之间的深层关联。



问答系统

个性化推荐系统和定制化搜索。



版权保护

侵权检测，跟踪并追踪数据源。



网络安全

基于 AI 的恶意软件检测和网络安全攻击检测。



AI 制药

AI 挖掘文献，快速分析分子结构。



特征处理

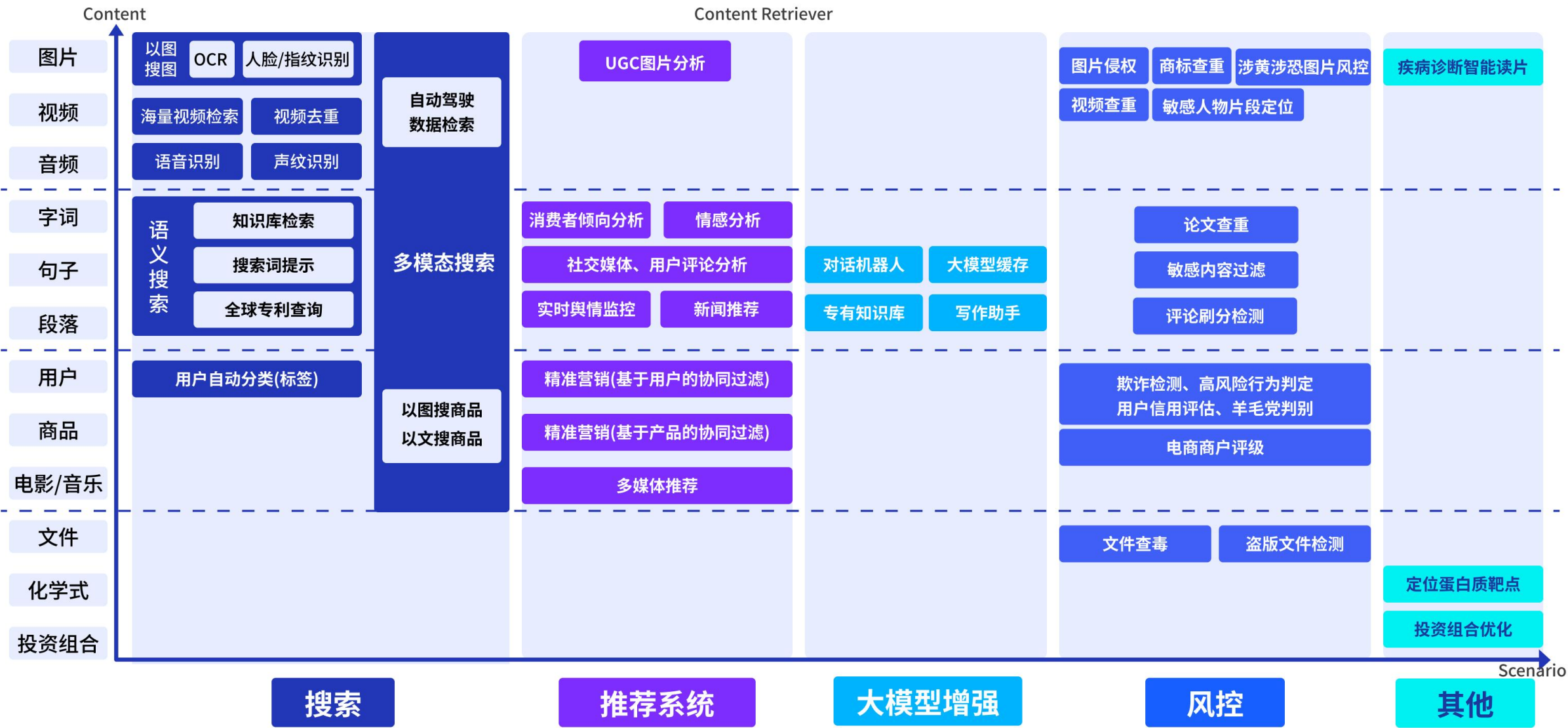
特征聚类和异常检测等。



欺诈保护

揭示用户行为和欺诈行为之间隐藏的关联性。

向量数据库场景探索



RAG应用 - Shulex使用Zilliz Cloud打造 VOC 数字化“加速器”

SHULEX 国际化VOC SaaS公司



业务价值

30% ↑

报表生成速度

50% ↓

成本

98%

召回率

90次/秒

客户问询

0

宕机

VOC评论分析

VOC评论分析



微信扫码阅读案例

文本搜索场景——VOC评论分析

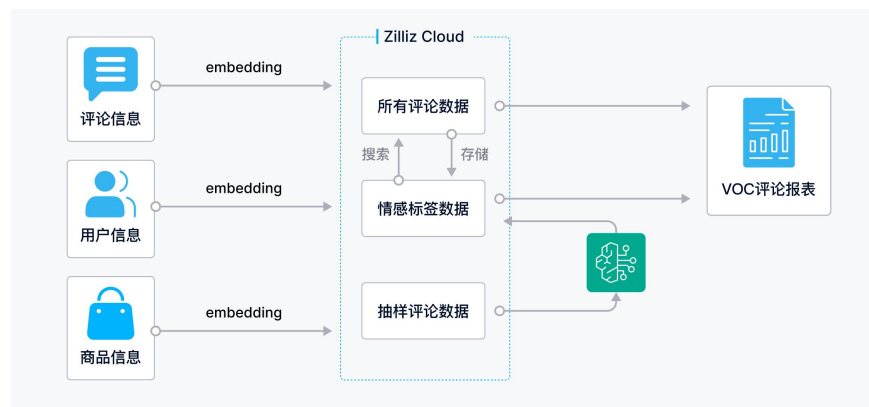


图1 | 基于 Zilliz Cloud 的 VOC 评论分析流程

大模型 RAG 应用——VOC 智能问答系统

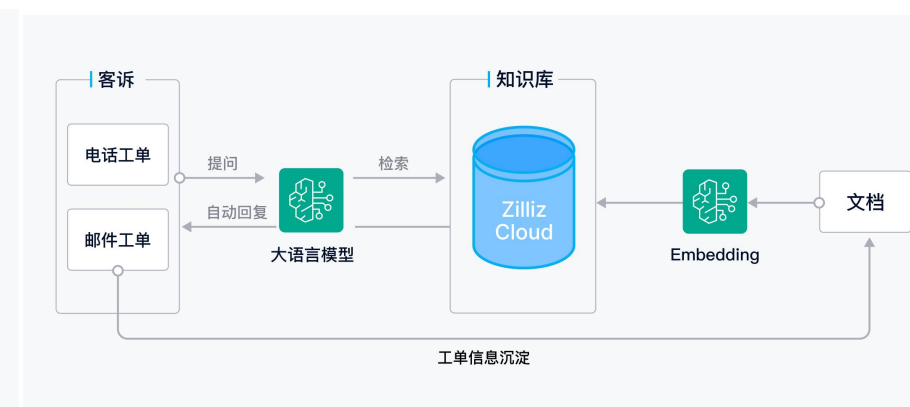


图2 | 基于 Zilliz Cloud 的 VOC 智能问答系统

某车企自动驾驶研发



长尾样本

➤ 异性车辆：

各种异形施工车辆、拖挂车、特种车辆、伸出触角的卡车铲车等

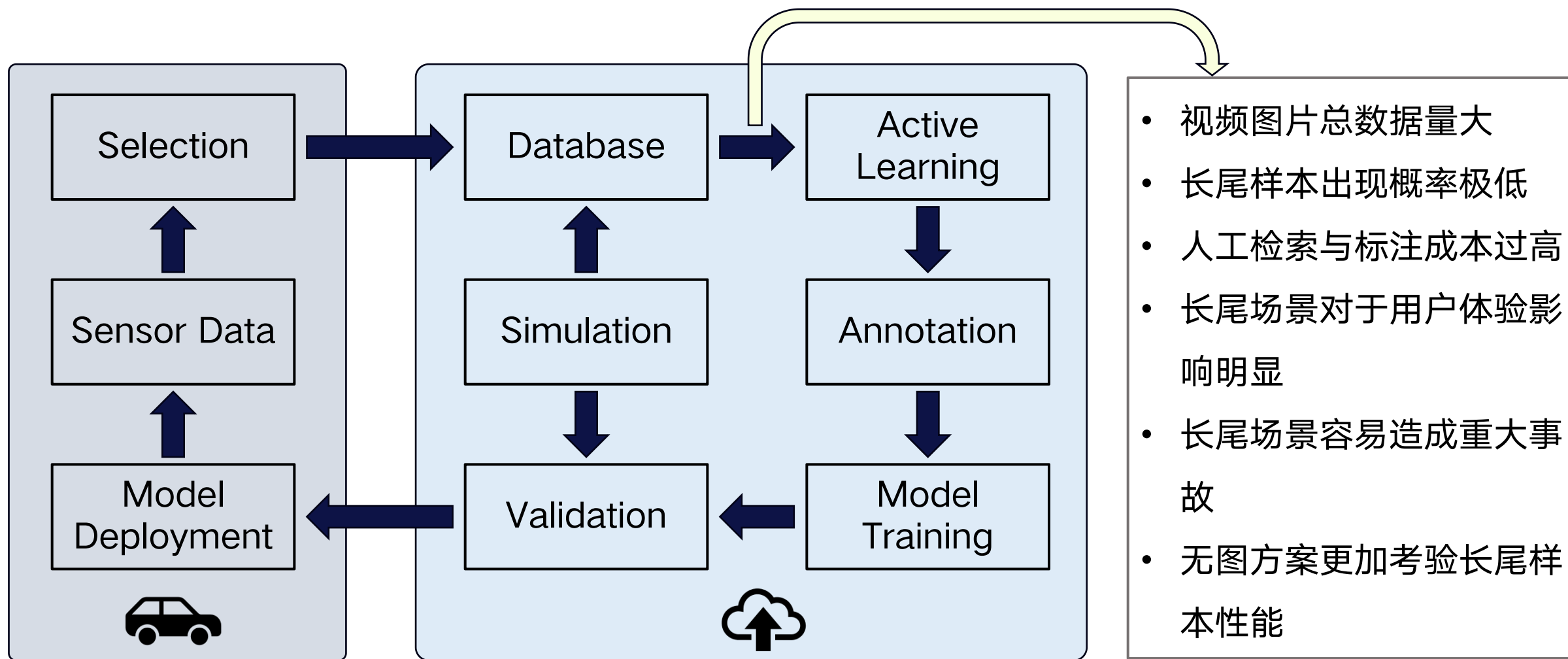
➤ 异形物体：

各种动物、低矮物体、山上落石、树木、高速公路设施部件等

➤ 道路变更：

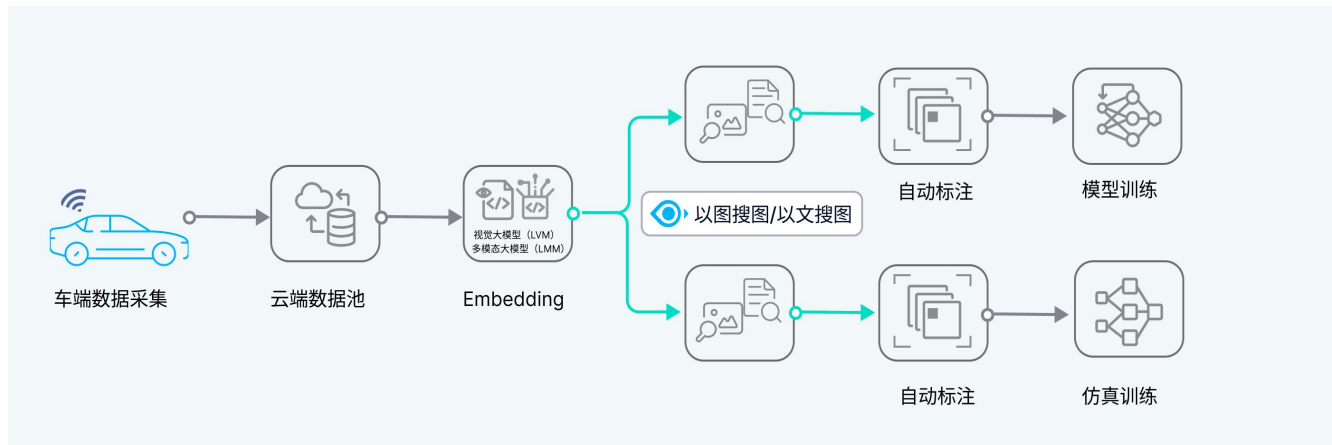
道路标线不清晰，临时施工等

长尾样本挖掘的挑战及业务价值



自动驾驶研发数据闭环

大模型+Milvus 长尾样本挖掘

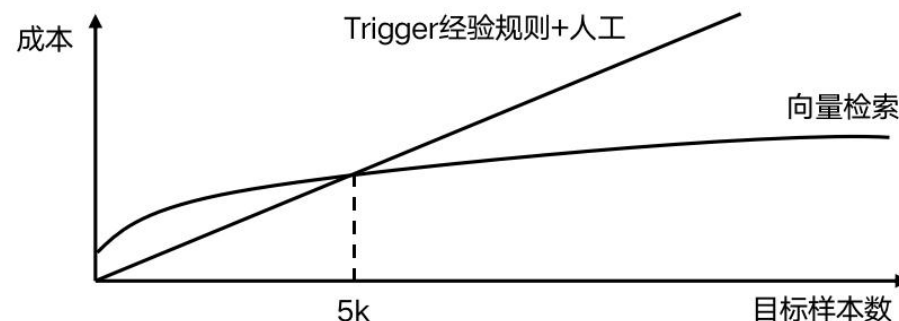


方案基本原理

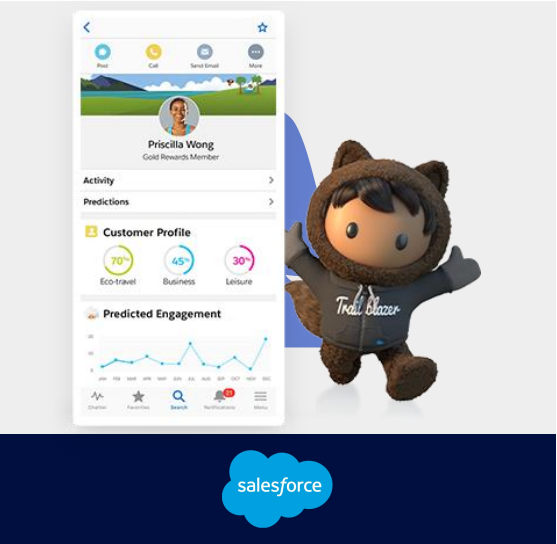
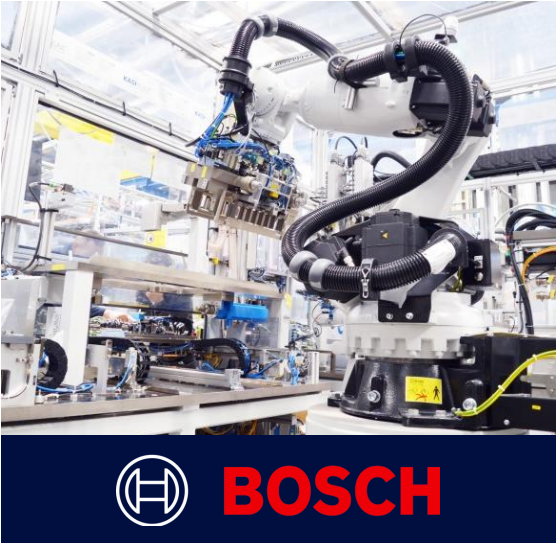
- 利用视觉大模型 LVM (Large Vision Model) 进行特征特征提取实现了大货车&异型车细分类检索能力
- 进一步利用多模态大模型 LMM (Large Multimodal Models) Embedding 模型实现任意场景文字to图片的模糊语义搜索
- 采用 Milvus 的高度横向扩展能力, 实现十亿、百亿级别向量高速检索

效果及优势

- 完成180TB视频数据入库
- 交付约120帧有效数据, 并完成人工标注
- 准确率高、样本分布均匀: 高速异形工程车0.7分以上相似图准确率>80%, 数据分布在100城市/20场景
- 成本对比: 当找图规模扩大后, 相比与人工检索方式边际成本明显降低



Industry leaders already use milvus in their apps

Use Case: Agentic RAG Vectors: 2 Billion Req'ts: 200 ms, Cost Index: DiskANN for cost savings 	Use Case: Drug Discovery Vectors: 12 Billion Req'ts: High Recall Index: IVF_BIN_FLAT 	Use Case: Image Search Vectors: 20 Billion Req'ts: High Insertion, Cost Index: Vector Lake 	Use Case: Recommender System Vectors: 10 million Req'ts: 5,000 QPS, Low latency Index: HNSW & CAGRA 
--	--	---	--

THANKS FOR WATCHING



公众号



加入交流群



<https://github.com/milvus-io/milvus>



<https://zilliz.com.cn>