Xorbits

企业级AI和数据基础设施平台

发挥企业数据价值，让大模型触手可及

Xorbits

CONTENT 目 录

Xorbits

01

关于Xorbits

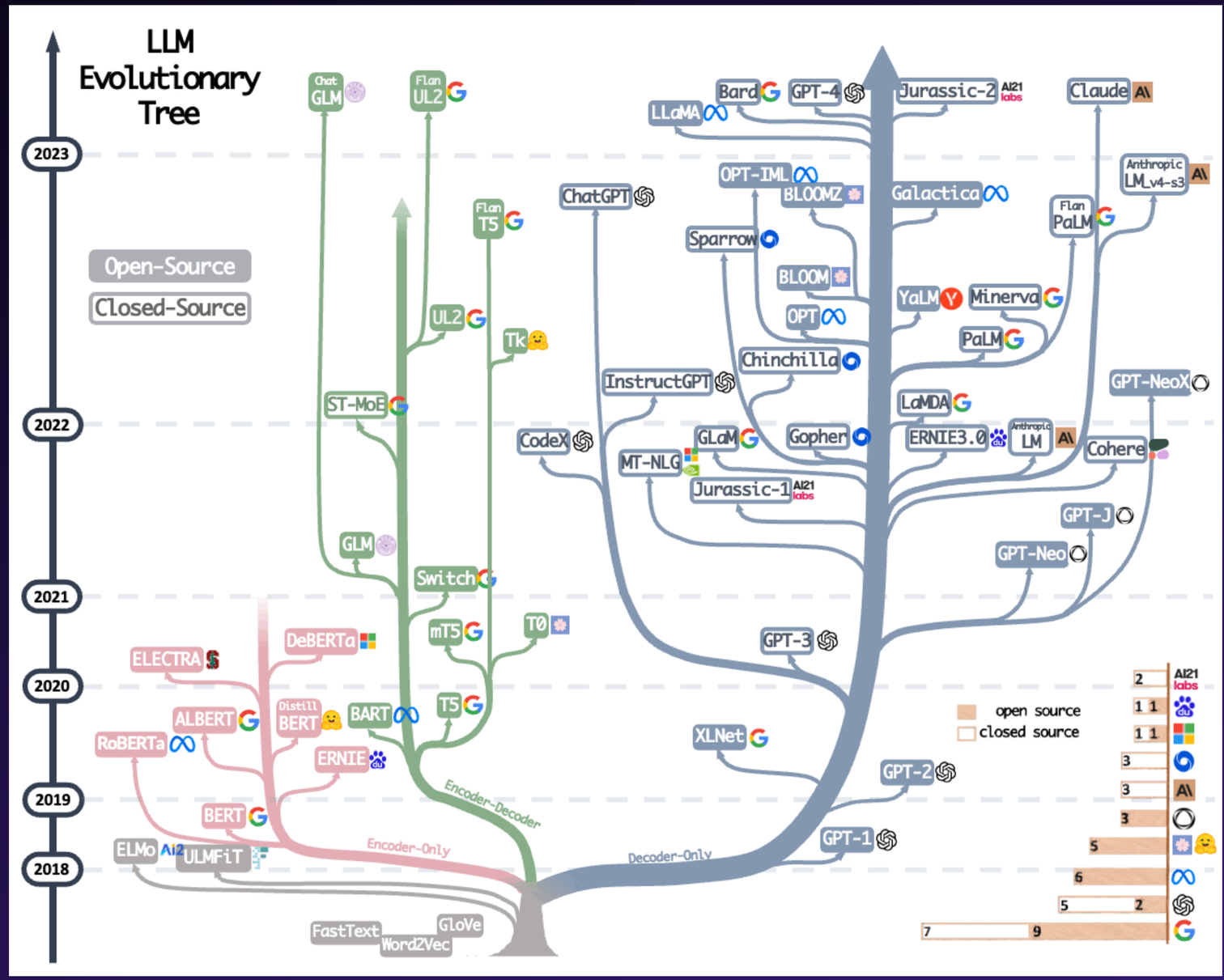
走进Xorbits

Xorbits

杭州未来速度科技有限公司（简称Xorbits）是一家专注于企业级人工智能基础设施平台公司，致力于研发能够满足客户不断增长的计算需求的解决方案。无论是大模型推理、数据处理和微调，我们的基础设施都能够提供稳定、高性能的支持，并随着您的业务规模扩大而轻松扩展。未来速度旗下的Xorbits模型平台赋能企业人工智能转型，帮助企业快速部署和接入大模型，同时，能结合企业自身的数据来提供对模型的增强。主要团队成员来自互联网大厂和世界500强企业，在分布式计算、人工智能领域有丰富技术实践经验。

LLM 正在飞速发展

- LLM 以迅猛的速度发展，几乎每天都会诞生新的开源模型
  - 国内：baichuan、ChatGLM、Qwen
  - 国外：Llama2、Falcon、Mistral



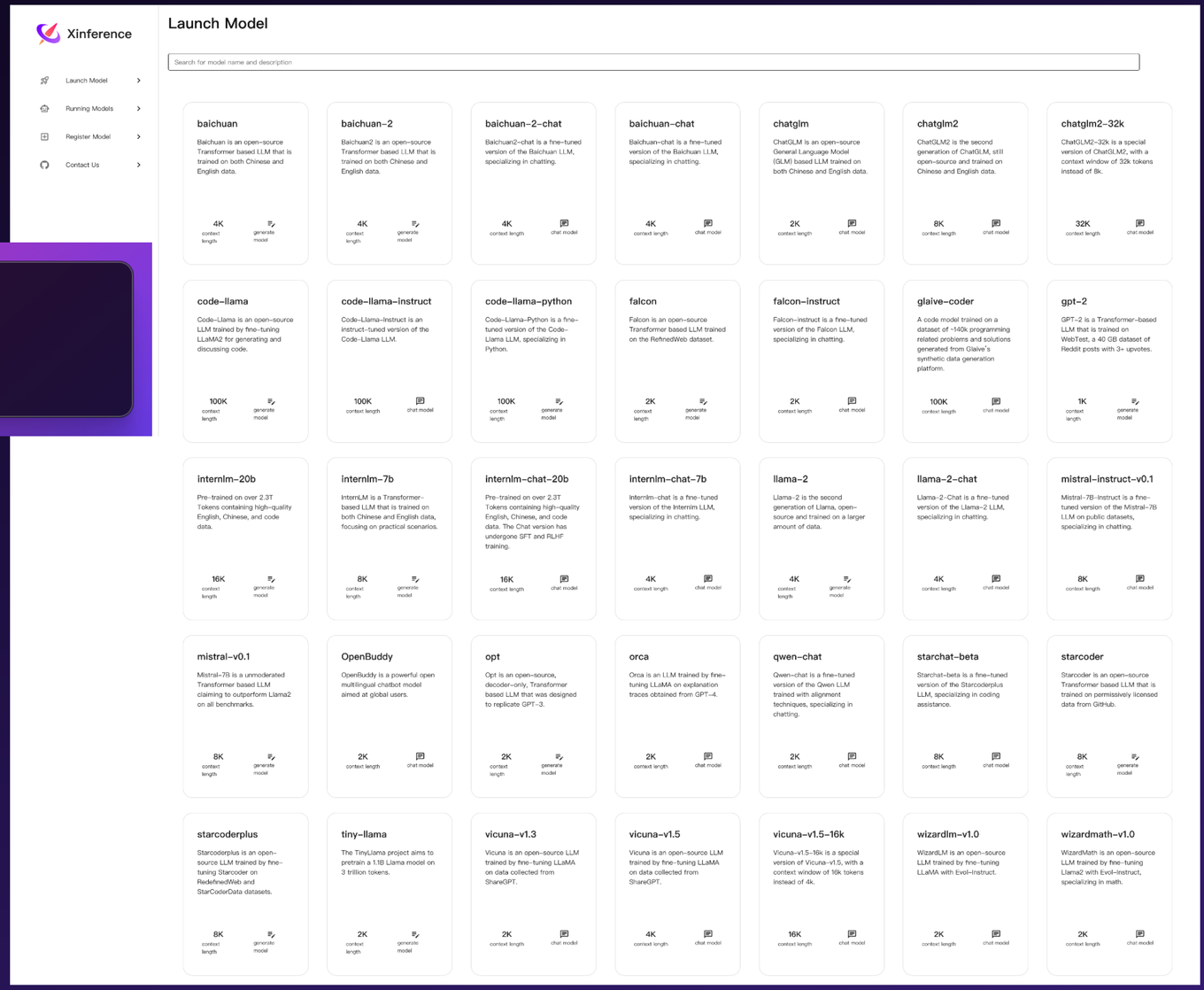
Xinference 支持所有主流开源模型

Xorbits

- OpenAI 接口兼容
- 全面的模型支持
  - 大语言模型（Large language models）
    - 文本生成模型
      - chatglm
      - baichuan
      - llama2
      - …
    - 代码生成模型
      - code-llama
      - starcoder
      - …
  - 嵌入模型（Embedding models）
    - bge-large-en
    - bge-large-zh
    - …
  - 文生图、图生图模型
- 硬件支持，CPU、GPU（Nvidia、AMD、Mac Metal、沐曦）
- 分布式支持
- 集成至 langchain、LlamaIndex、Dify

Terminal

```
$ pip install xinference
$ xinference
```



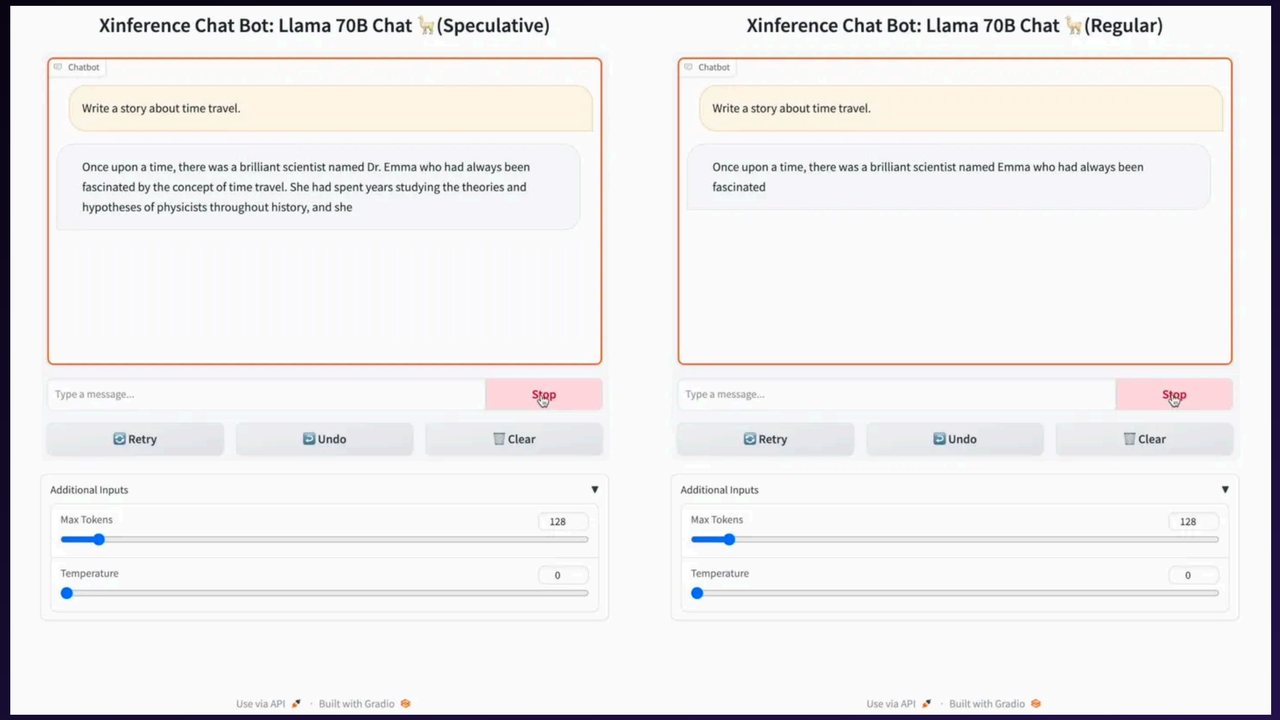
无缝衔接 OpenAI SDK

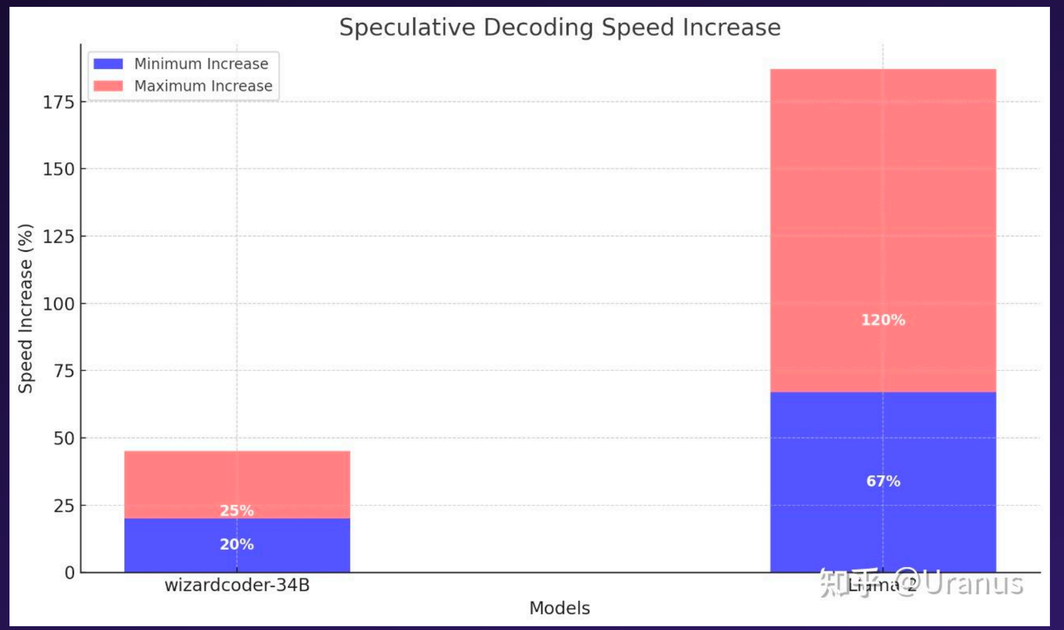
Xorbits

```python
import openai
import sys

openai.api_base = 'http://127.0.0.1:9997/v1'
openai.api_key = ''

model ='1a71f69e-3d25-11ee-a893-67f1d6f729c9'

prompt = '''
Problem:
Consider the equation of a hyperbola given by:
x^2/a^2 - y^2/b^2 = 1
where a, b > 0.

Given that the hyperbola has asymptotes y = ±(b/a)x and it intersects the line y =
2x + 3 at two distinct points, find the values of a and b.
'''
for resp in openai.Completion.create(model=model,
        prompt=prompt, max_tokens=512, stream=True):
    sys.stdout.write(resp.choices[0].text)
    sys.stdout.flush()
```

Xinference 性能优化

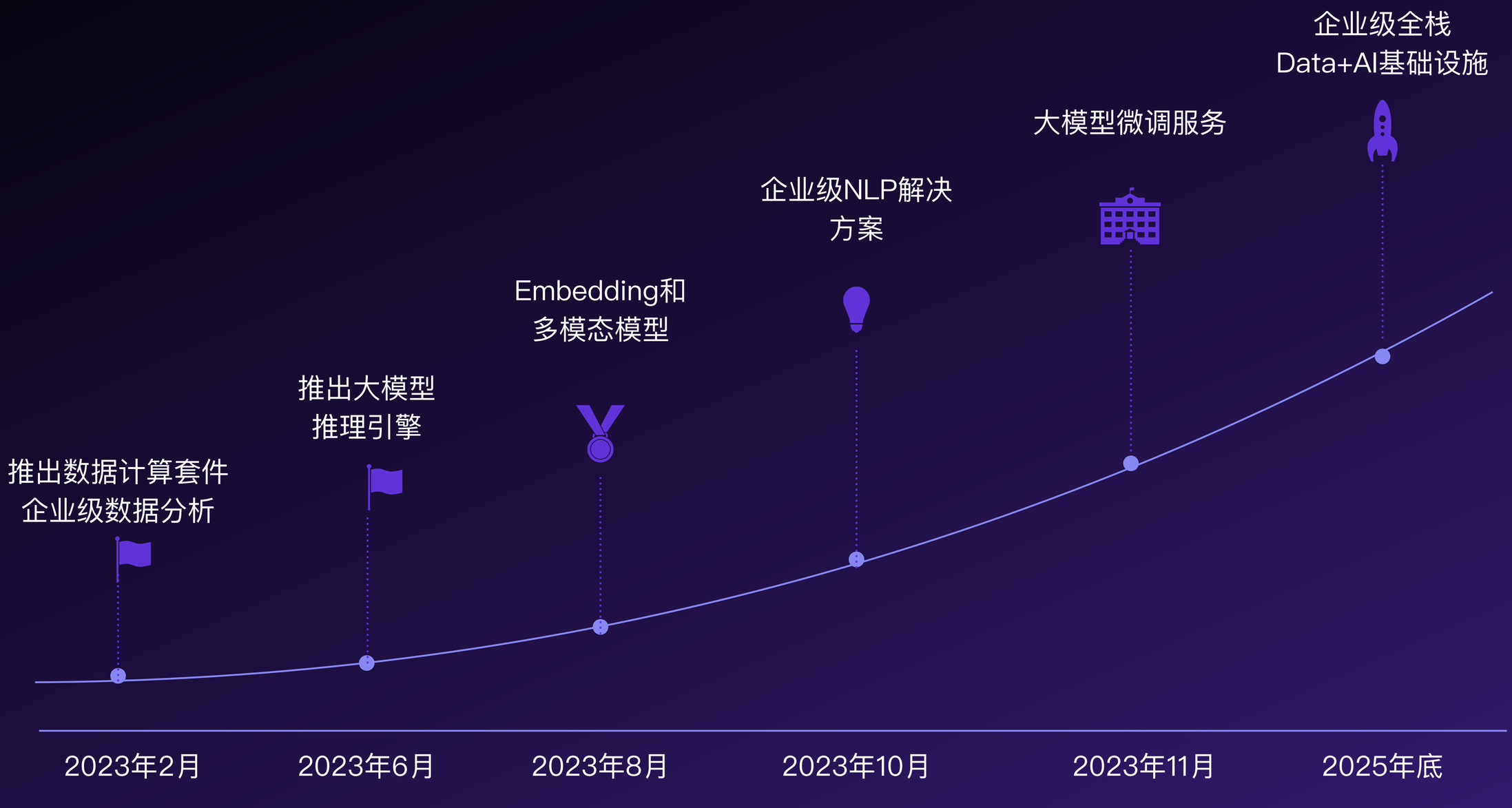
- 多后端
  - GGML
  - vllm
- 模型量化技术（GPTQ、AWQ）
- GPU 优化技术
  - Continuous Batching
  - Paged Attention
  - 张量并行
  - Flash decoding
- CPU 优化技术
  - 向量化，SIMD
  - 算子融合
  - 分布式





打造全栈基础设施平台，加速企业数据+AI一站式落地

Xorbits



构建多元化生态合作，覆盖企业数据+AI落地全链条服务

Xorbits

jina

Hugging Face

DataFocus

LangChain

Dify

百川智能 BAICHUAN AI

zilliz

aws

ModelScope

云器

LlamaIndex

ClickPaaS

KubeBlocks

LMFLOW

Chatbox

Xorbits

Xorbits产品介绍

Xorbits模型平台：一站式模型部署、推理、微调、应用解决方案

Xorbits

模型部署 – MaaS

- 轻松接入 ChatGLM、百川等 50 多种最先进的开源模型
- 支持一体机、云端部署等多种部署模式，降低运维成本
- 提供兼容OpenAI的RESTful API和Python SDK

模型推理 – Xinference

- 高效适配不同硬件，大幅提升吞吐量，降低推理延迟
- 使用vLLM和投机采样等优化技术，大幅提升推理吞吐，降低API延迟
- 原生分布式架构，可轻松水平扩展集群，支持负载均衡

模型微调 – Finetune

- 根据自己的需求和数据进行模型微调，提升模型在特点业务场景下的效果
- 可以灵活地从企业数据库中抽取数据，也可以通过API接口读取
- 在预处理数据环节对命名实体进行替换，提升模型泛化能力

AI应用

- Xorbits模型平台对接了多款开源AI应用和开发框架
- 如Dify.AI、Chatbox、LangChain、LlamaIndex和Continue.dev等
- 通过Xorbits模型平台，方便地接入开源AI应用，或构建自己的AIGC应用

帮助企业轻松部署和管理大模型，开启AIGC之旅

Xorbits

产品特性

Xorbits

01 安全

- 独立环境：提高安全性和可靠性，便于根据特定需求进行定制
- 数据隔离：确保其敏感信息不会泄露，更容易遵守数据保护规范
- 权限控制：确保只有授权的用户可以访问关键资源

02 开箱即用

- 低门槛：支持一体机和云端部署，降低运维成本和复杂度
- 可定制：基于企业自有数据进行模型微调
- 开发便利：兼容主流AI应用开发框架

03 企业级

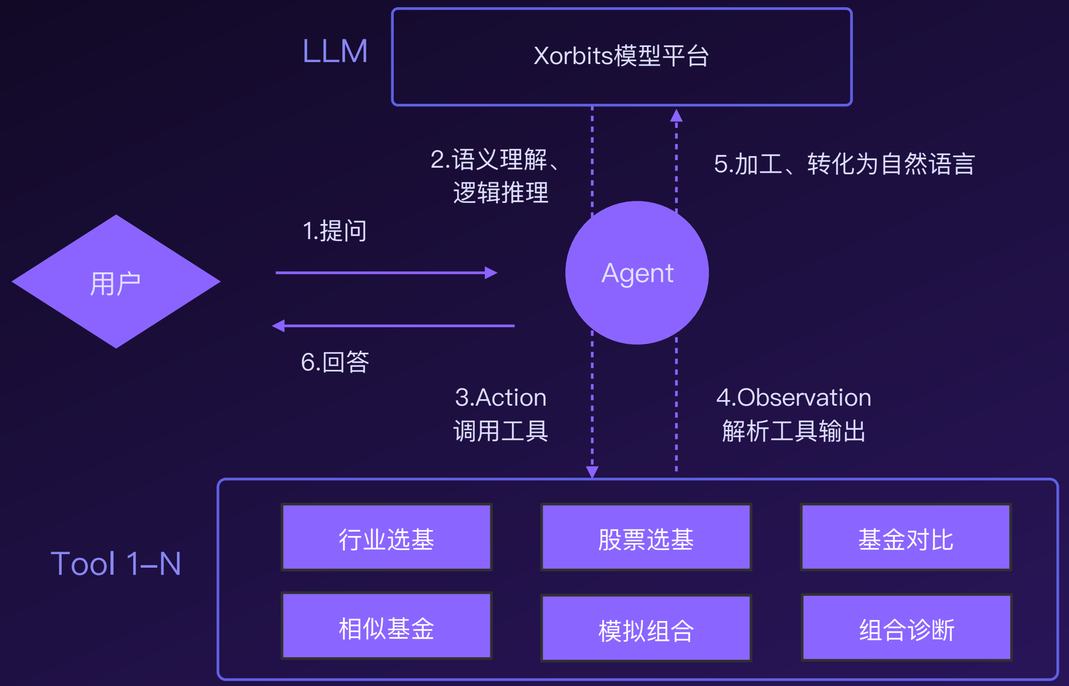
- 高性能: 高吞吐量、低延迟的模型访问，支持多客户端访问
- 灵活扩展：分布式集群，随着您业务规模的扩大轻松扩展
- 高可用：多副本、不停机升级和高可用性等功能，确保模型服务的稳定和连续

Xorbits

Xorbits最佳实践

典型案例：基金投顾Agent

通过对接丰富的基金投研API，基金投顾Agent可以通过自然语言回答用户问题，如：基金经理信息查询、对比分析、基金筛选等，是辅助一线投顾人员提升服务能力的专业AI Agent。



客户痛点

- 为符合金融行业的严格监管要求需要私有化部署模型，但过程复杂且耗时。
- 为了快速响应市场变化，需要高吞吐量和低延迟的推理服务，同时还要保证服务的稳定性。
- Agent场景对模型能力有一定要求，需要测试多种开源模型。

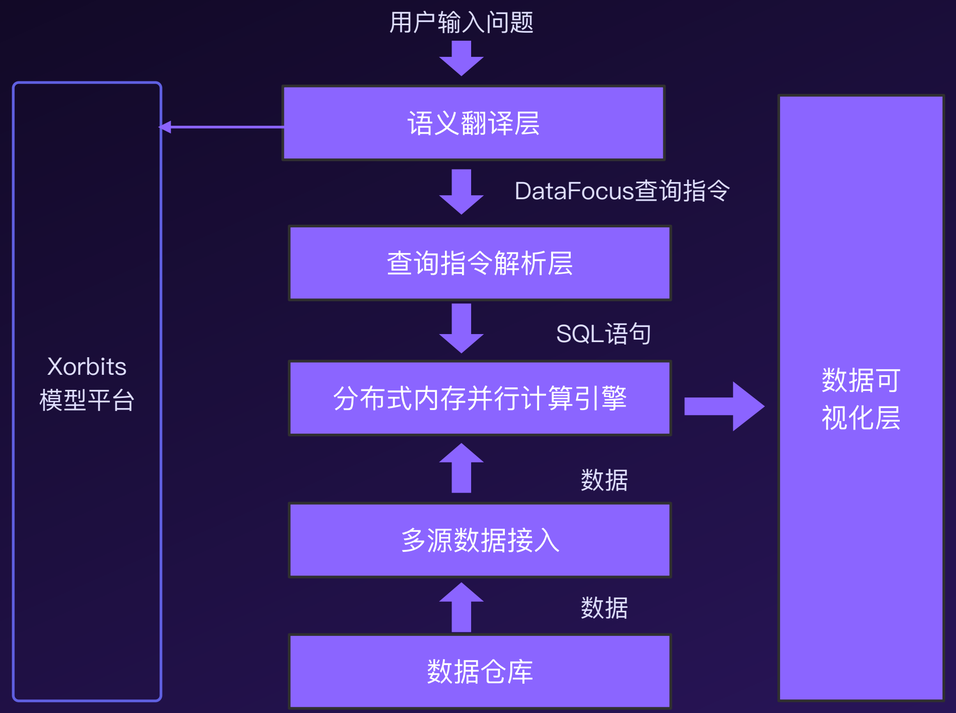
产品赋能

- 通过Prompt Engineering和对接丰富的基金投研API，基金投顾Agent可以通过自然语言回答用户问题。
- 不仅可以对基金信息进行加工和输出，还可以通过API，实现基金多指标条件筛选、基金组合构建、基金组合诊断等复杂和专业的操作。
- 为基金投顾Agent后端服务提供了高可用、低延迟的模型推理后端。

典型案例：智能BI助手

Xorbits

智能BI助手通过LLM识别业务人员的问题，理解其含义，并将问题转化为DataFocus的查询指令，通过DataFocus Search技术转化为SQL进行查询，解决了Text2SQL准确率不高的问题。在一站式的数据集成、分析、可视化和应用的基础上，实现更智能的数据分析。



客户痛点

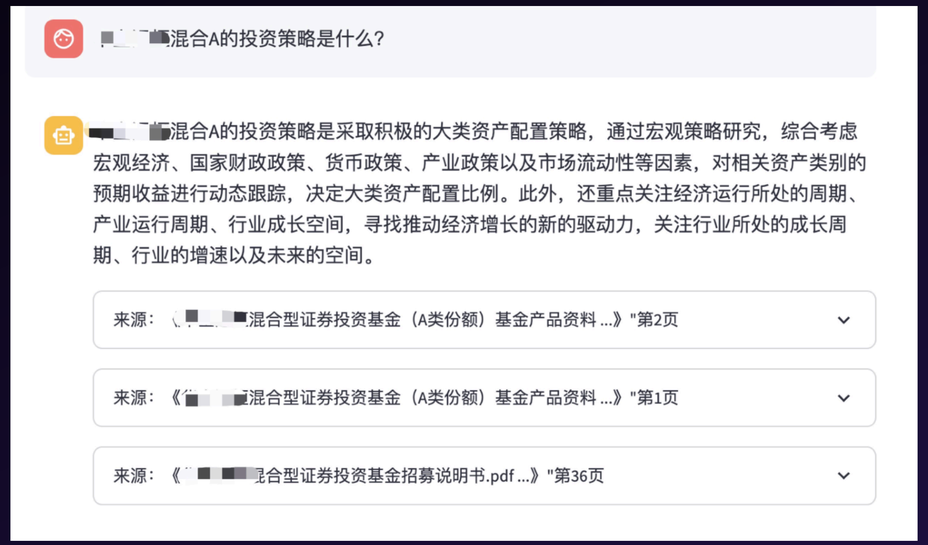
- 传统的BI解决方案和死板的报告流程将无法满足业务人员对数据的分析的即时需求
- 数据库种类繁多，尽管大部分SQL语法类似，但仍然存在细微的差别，直接用GPT来生成SQL，无法确保识别这些差异，生成的SQL语句往往不可用。
- 为了满足企业数据安全性的要求，大模型相关方案需要私有化部署模型，整个过程复杂且耗时。

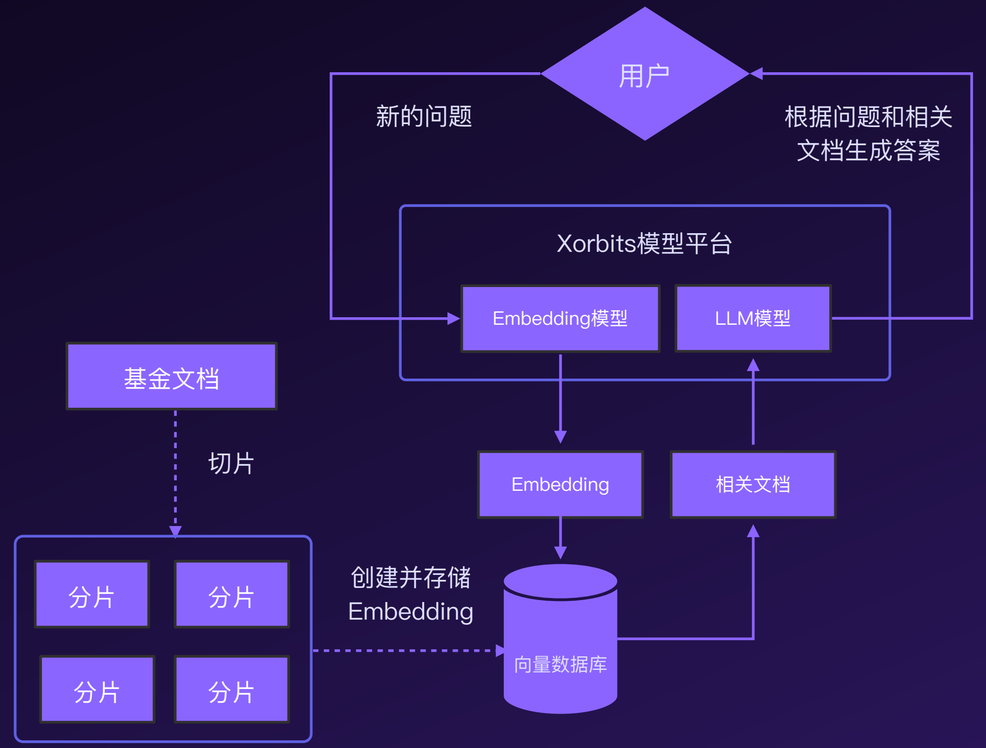
产品赋能

- 借助DataFocus Search技术，使用了可靠性更高的中间语言，解决了传统Text2SQL准确率低的问题。
- 通过Xorbits模型平台的自定义模型部署功能，可以部署用DataFocus查询指令数据微调后的版本。
- 自然语言交互方式，上手更门槛，让业务人员轻松探索和分析数据。

典型案例：基金智能客服助手

基金智能客服利用RAG技术，使用基金公告信息和基金产品文档作为知识库，回答投资者的各类疑问，比如：基金产品特性、最新公告解读、投资策略说明等，它是为基金用户提供专业知识支持的客服助手。





客户痛点

- 需要实时同步和处理基金公告和产品文档，确保信息的即时性和准确性。
- 面对高峰期的咨询请求，需要高吞吐量和低延迟的推理服务，同时还要保证服务的稳定性。
- Xorbits平台提供了多层次的数据保护措施，确保用户数据的安全性和隐私性。

产品赋能

- 智能化客服解决方案显著减少了人工成本，同时提升了客服的响应速度和服务质量。
- Xorbits不仅提供了LLM模型的部署能力，还提供了多个前沿的Embedding模型供部署和接入。
- Xorbits平台支持灵活的API和SDK，易于与客户现有系统集成，保持业务连续性。

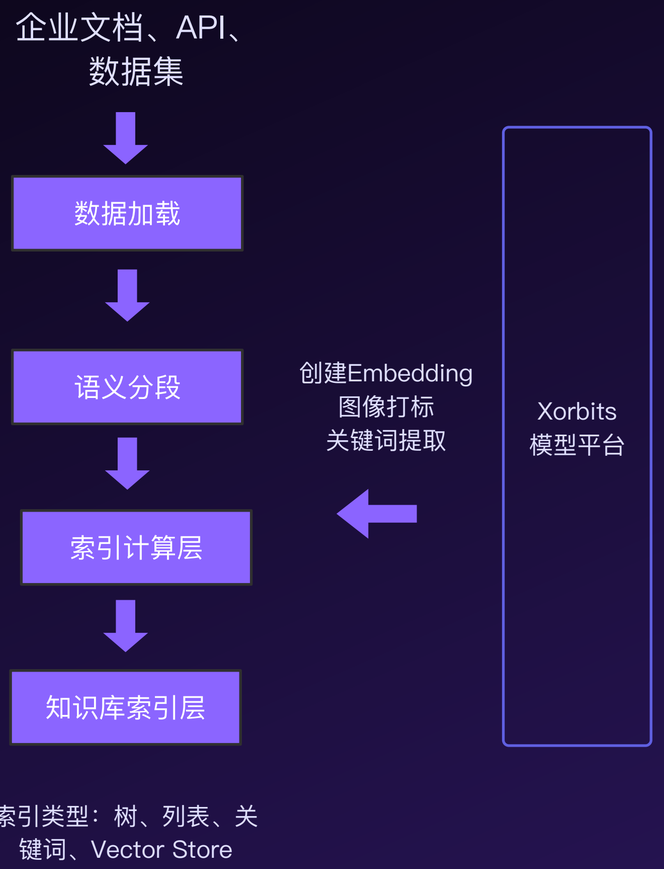
解决方案：RAG应用开发

- 搜索引擎
  - 是一种信息检索系统，用来找到存储在计算机系统上的信息。
  - 搜索引擎的目标是提高人们获取和收集信息的速度。
  - 它们通常基于关键词匹配来提供搜索结果。
- RAG (Retrieval-Augmented Generation)
  - 是一种将检索技术和语言生成技术相结合来增强生成过程的技术。
  - RAG的目标是对搜索结果进行增强。
  - 可以帮助传统搜索引擎生成更加准确、相关和多样化的信息

解决方案：RAG应用开发（续）

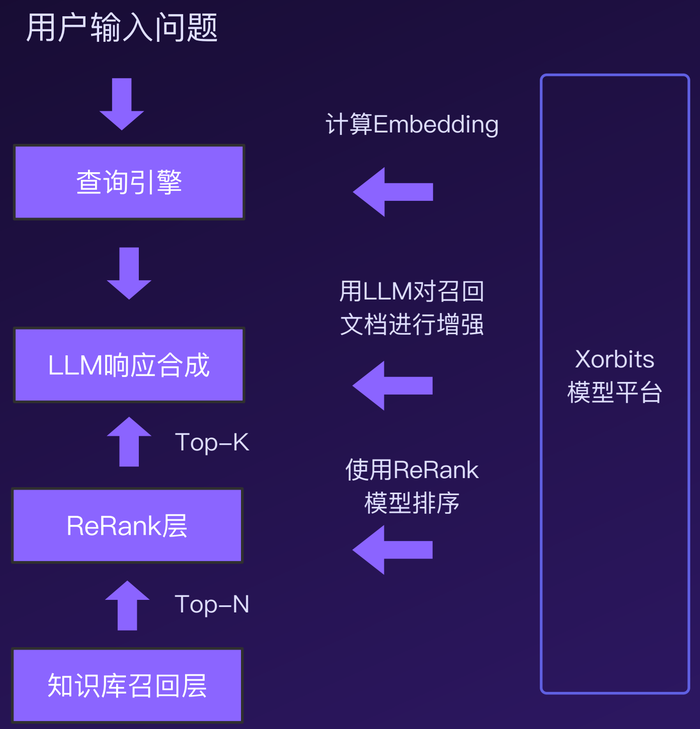
Xorbits

索引阶段



索引类型：树、列表、关键词、Vector Store

查询阶段

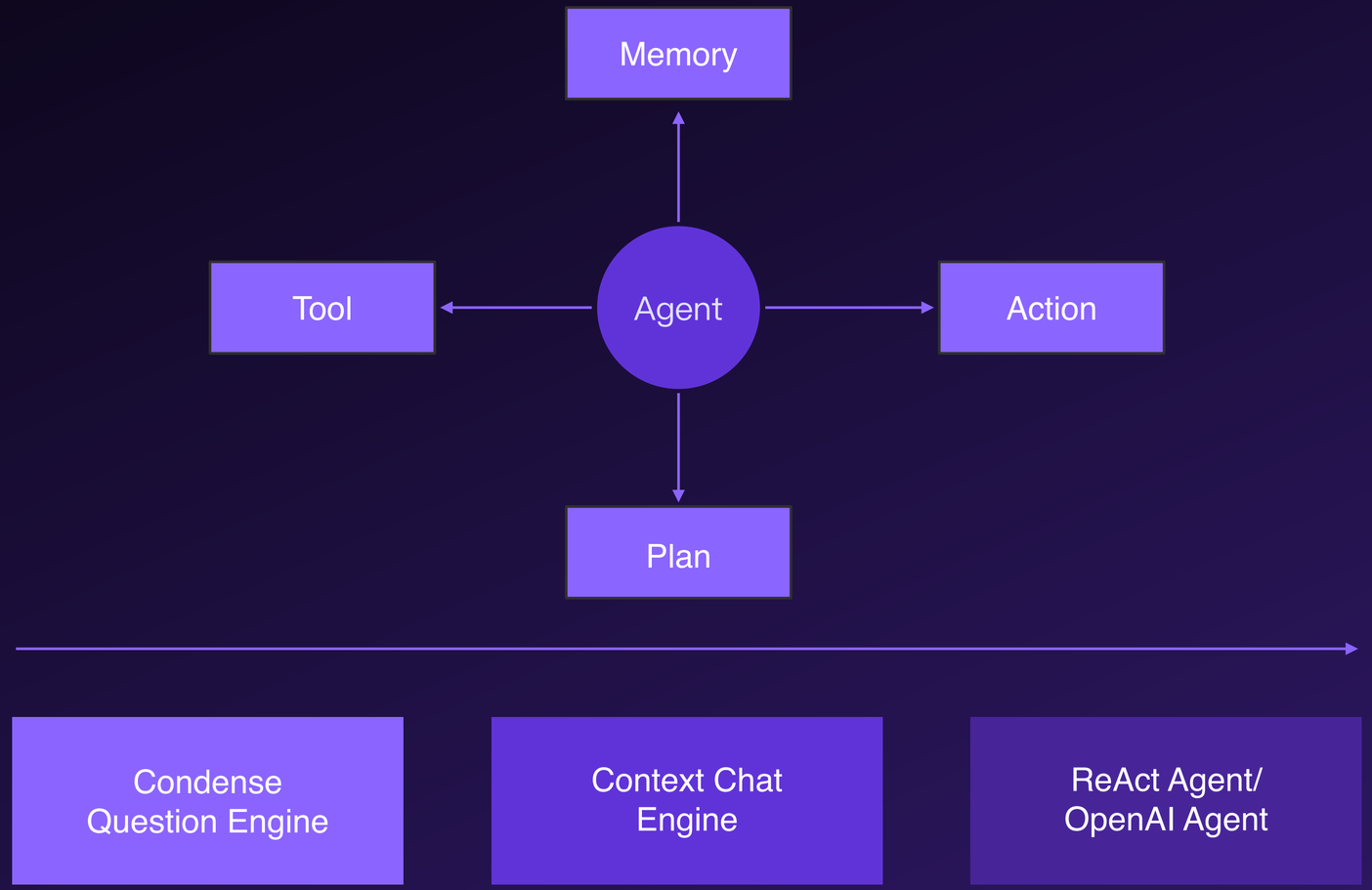


召回策略：关键词、语义、混合

解决方案：RAG应用开发（续）

Xorbits

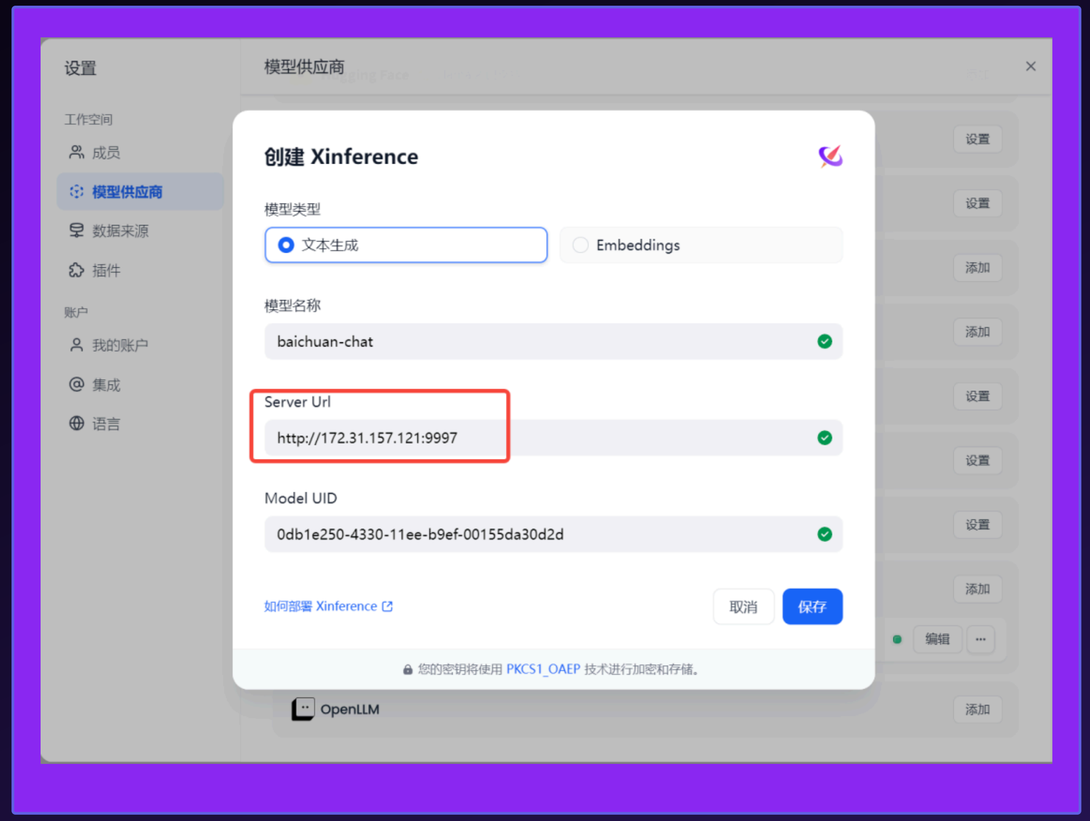
通过给Agent增加不同的能力，可以实现不同行为的聊天引擎



解决方案：RAG应用开发（续）

Xorbits

Xinference对主流的RAG开发框架生态进行了深度集成



Dify

```python
from llama_index.llms import Xinference
llm = Xinference(
  server_url=os.getenv("XINFERENCE_SERVER_ENDPOINT"),
  model_uid=os.getenv("XINFERENCE_LLM_MODEL_UID")
)
```

LlamaIndex

```python
from langchain.embeddings import XinferenceEmbeddings
from langchain.llms import Xinference

embed_model = XinferenceEmbeddings(
   server_url=os.getenv("XINFERENCE_SERVER_ENDPOINT"),
   model_uid=os.getenv("XINFERENCE_EMBEDDING_MODEL_UID")
)

llm = Xinference(
  server_url=os.getenv("XINFERENCE_SERVER_ENDPOINT"),
  model_uid=os.getenv("XINFERENCE_LLM_MODEL_UID")
)
```

LangChain

Why Xorbits

Xorbits

SLA服务保障高达 99.99%

3大核心技术
（异构计算+分布式计算+细粒度并行）

2大产品满足企业级需求
（大数据+大模型）

超1千万行代码的精华沉淀

Xorbits

发挥企业数据价值
让大模型触手可及

Twitter: https://twitter.com/xorbitsio

Website: https://xorbits.cn

Email: sales@xprobe.io