

工作流和智能体组成的 大模型应用的 构建、迭代和高可用

AI 进阶指南 (上) 从理论到实践

课程安排 AGENDA

07.23 深入浅出：大语言模型

张凡石 「DaoCloud 道客」高级研发工程师

07.30 基于 RAG 的 AI 应用落地

陈 佳 「DaoCloud 道客」架构师

08.06 智能体的构建、迭代和高可用

尹伯昊 猴子无限 创始人

08.13 云原生技术优化模型推理

王 璠 「DaoCloud 道客」高级研发工程师

08.20 基于分布式和容器的方式微调大模型

黄敏杰 「DaoCloud 道客」高级研发工程师

Content

目录

1. 大模型、工作流和智能体
2. 工作流和智能体的构建
3. 工作流和智能体的迭代
4. 云原生驱动工作流和智能体高可用落地

■ 十倍加速大模型应用构建、迭代和高可用

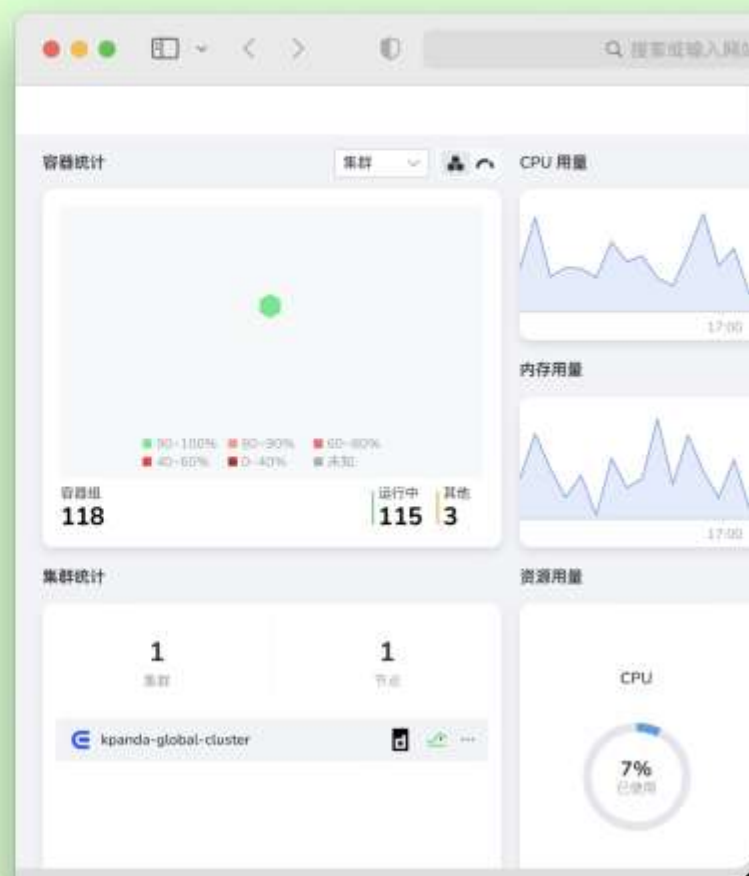
构建 云原生流程引擎



迭代 可复制专业服务



高可用 容器编排管理











Part 01

大模型、工作流和智能体

■ 大模型应用的三个范式

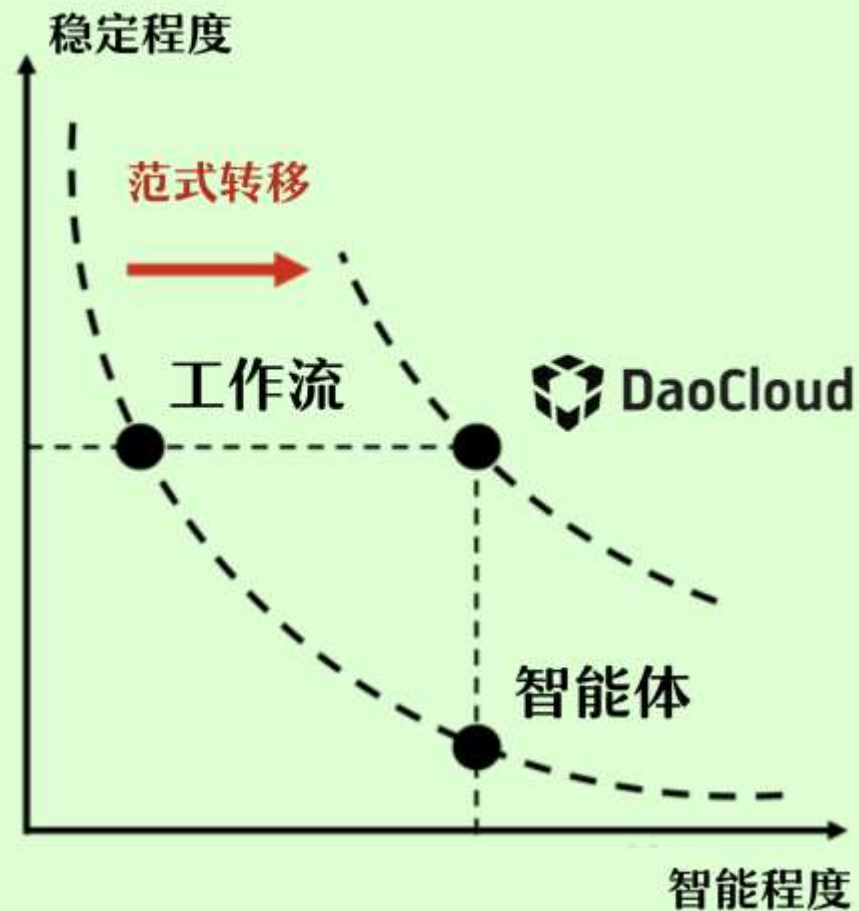
| | 主导角色 | 技术范式 | 快思考和慢思考 |
|----------------------|-----------------------|--------------------|-----------------|
| 聊天助手 Chatbot | 根据用户需求 大模型强行给出回复 | 神经网络 Since 1990 | 系统一 快思考 / 直觉 |
| 工作流 Flow | 根据预设流程 大模型和工具一视同仁 | 符号语言 Since 1940 | 系统二 慢思考 / 规划 |
| 智能体 Function Call | 根据用户需求 大模型自主决定调用工具 | 神经符号 Since 2023 | 系统二 慢思考 / 规划 |

■ 大模型应用的三个范式

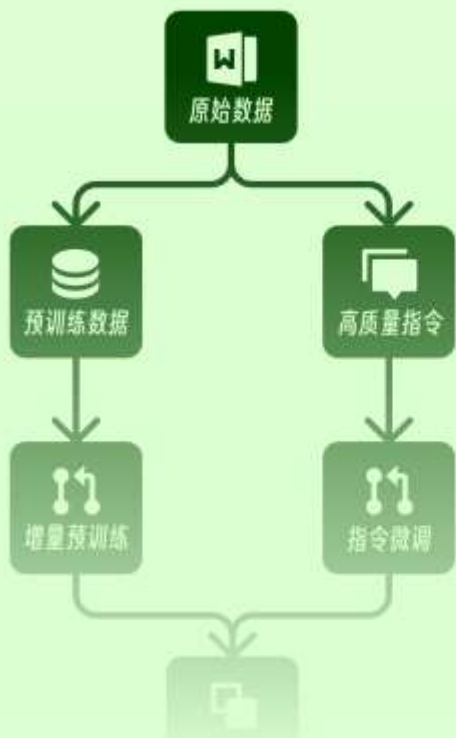
| | 稳定程度 | 智能程度 | 构建成本 |
|----------------------|--|--|--|
| 聊天助手 Chatbot |  |  |  |
| 工作流 Flow |  |  |  |
| 智能体 Function Call |  |  |  |

■ 大模型应用第四范式：工作流 + 智能体

| | 稳定程度 | 智能程度 |
|------------------------|--|--|
| 工作流 Flow |  |  |
| 智能体 Function Call |  |  |
| 工作流 + 智能体 Flow Call |  |  |



■ 工作流是大模型的天然延展



模型训练
天然是一条流程

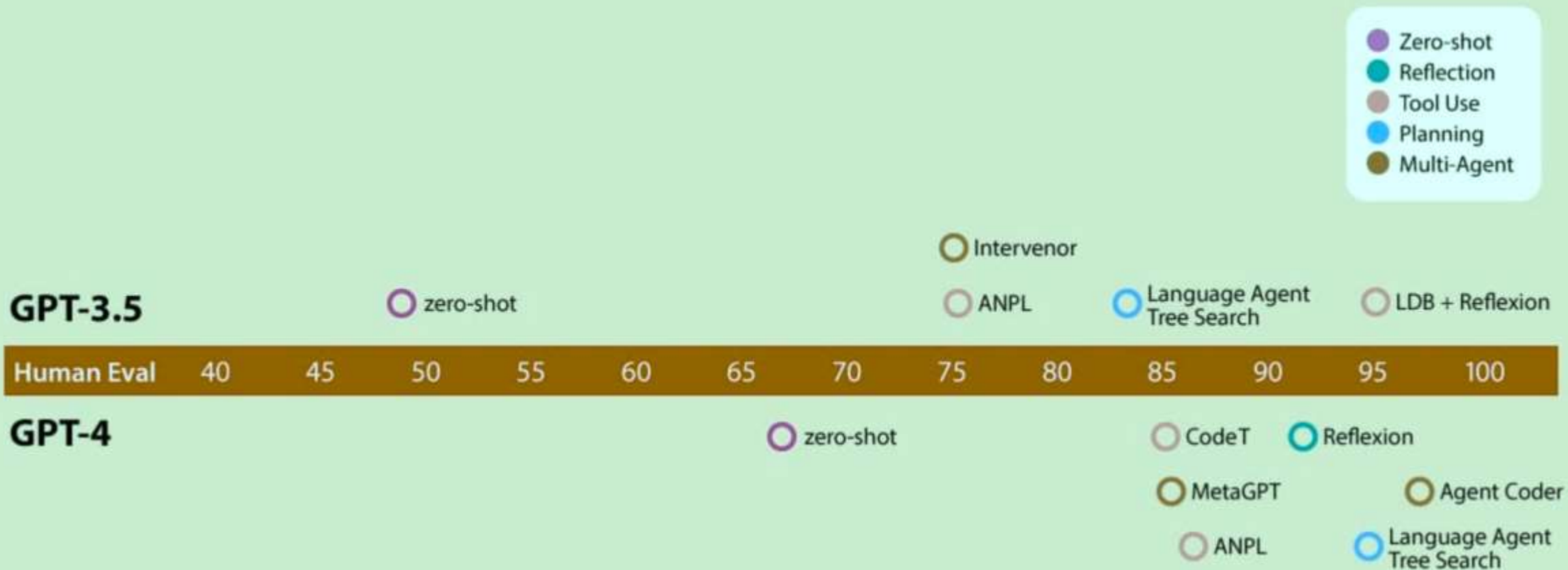


信息检索和处理
天然是一条流程



信息检索和处理
天然是一条流程

■ workflow是效果提升的核心抓手



■ workflow 是提示词工程、检索增强和智能体的基础抽象

商业前沿

OpenAI Dev Day
发布 GPTs
同时提到唯二创业公司
就包括 Zapier

大模型本身不能直接
与其他大模型和软件进行交互

Zapier 等流程引擎填补空白
是大模型必要的手和脚
让大模型能够解决更复杂的问题

场景落地

搜索增强RAG
天然就是一条流程
Langchain & LangGraph

技术前沿

多智能体是
流程的另一种表达
e.g. AutoGen

场景落地

多模态内容生成
天然就是一条流程
e.g. ComfyUI

技术前沿

自回归智能体是
流程的另一种表达
e.g. AutoGPT

Part 02

工作流和智能体的构建

■ 十倍加速大模型应用构建、迭代和高可用

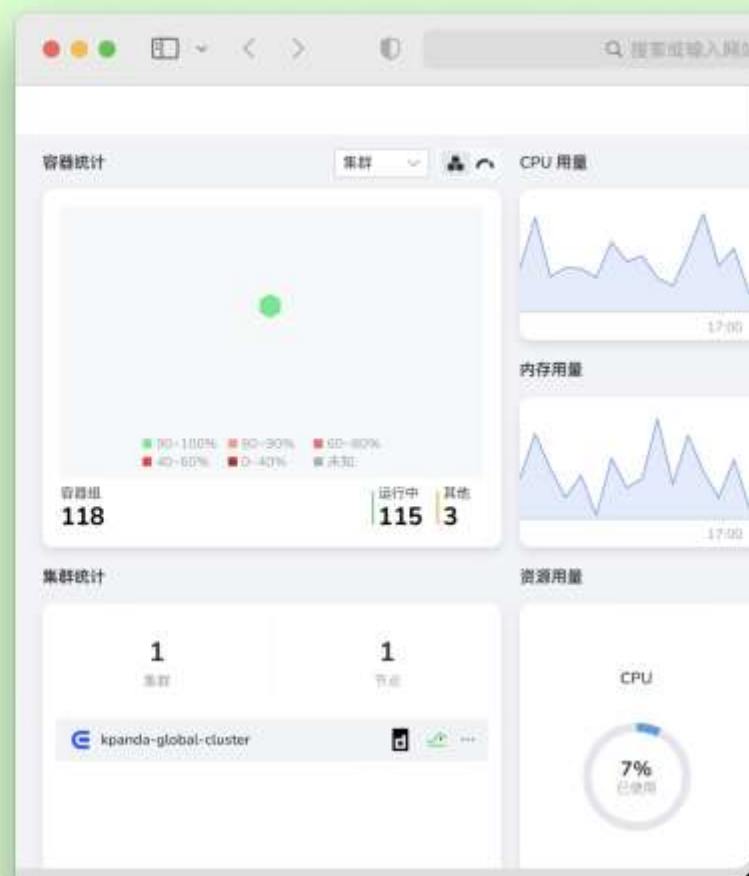
构建 云原生流程引擎



迭代 可复制专业服务



高可用 容器编排管理



■ 十倍加速大模型应用构建、迭代和高可用

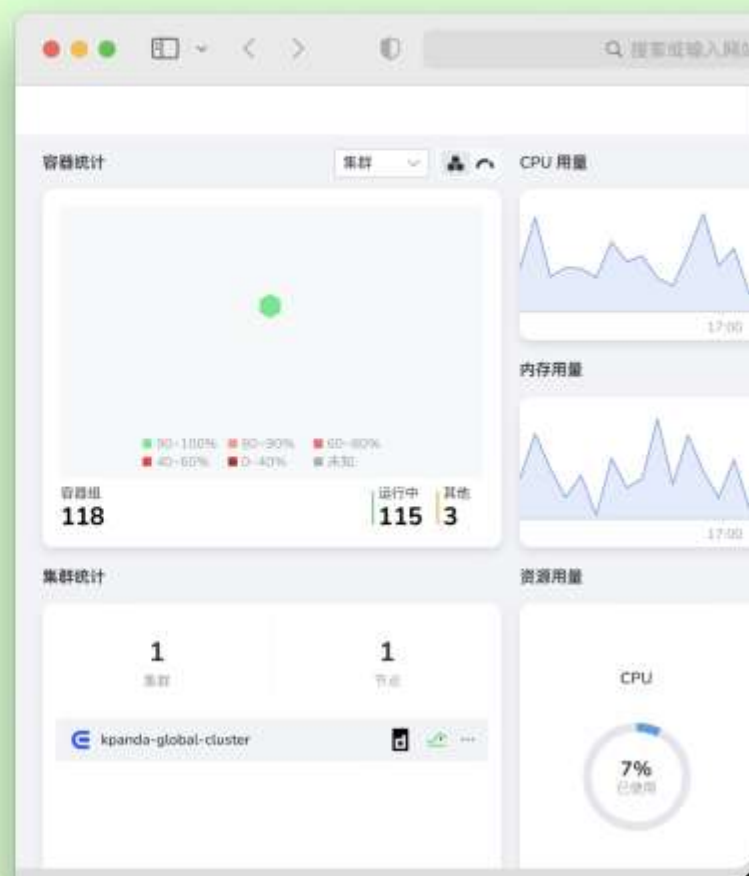
构建 云原生流程引擎



迭代 可复制专业服务



高可用 容器编排管理



我们以流程为中心寻求解决方案

工具 Tool

预先筛选和构建高可用高质量工具



流程 Flow

将工具结合业务逻辑变成可用流程



应用 APP

从流程到表单、对话、API 应用



工具 Tool

配置参数并实现工具间的参数传递



应用 APP

观测流程应用运行并进行持续迭代



■ 模型训练的背后是看不见的流程



■ 信息检索和处理的背后是看不见的流程



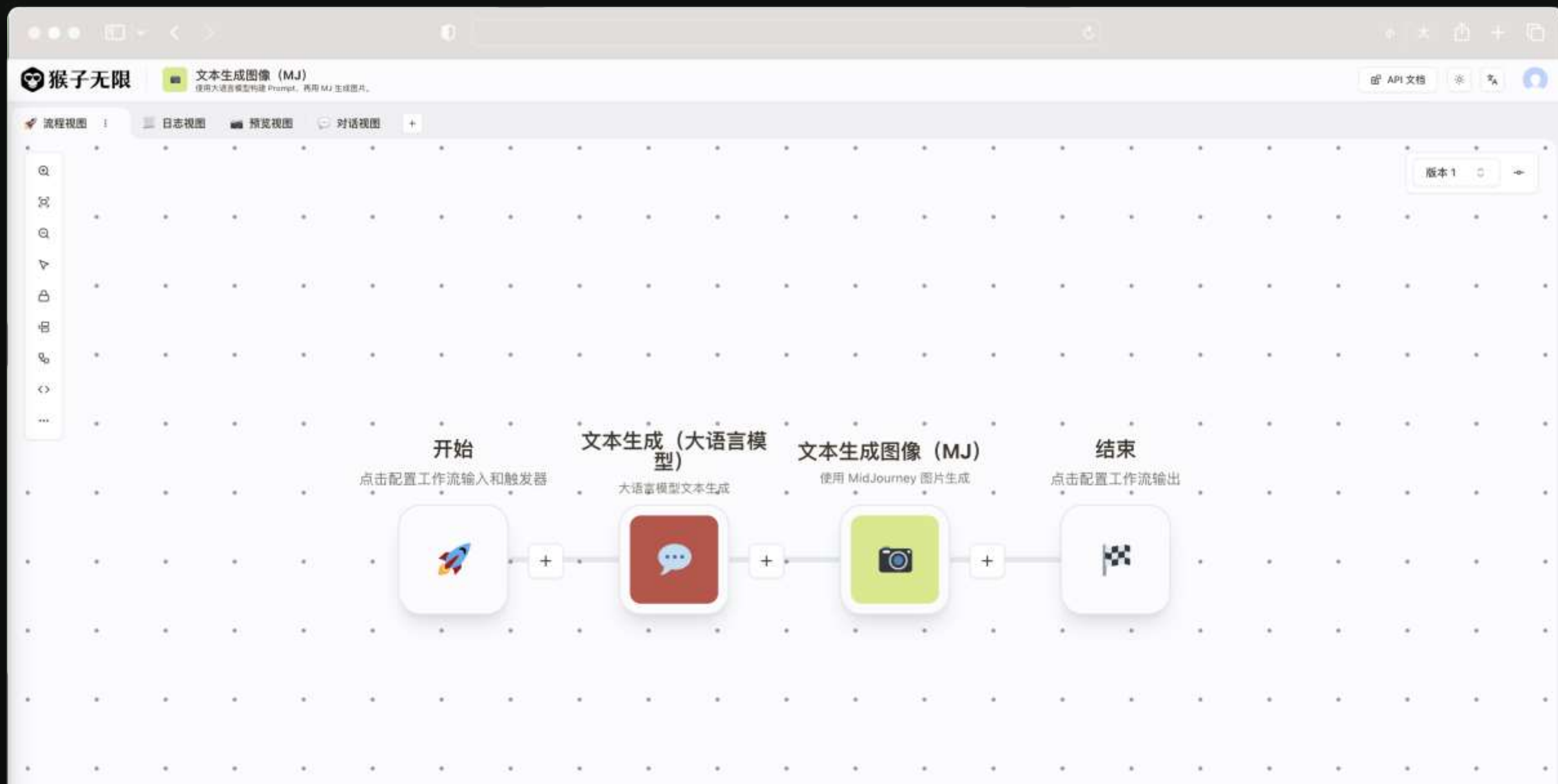
■ 信息检索和处理的背后是看不见的流程



■ 信息检索和处理的背后是看不见的流程



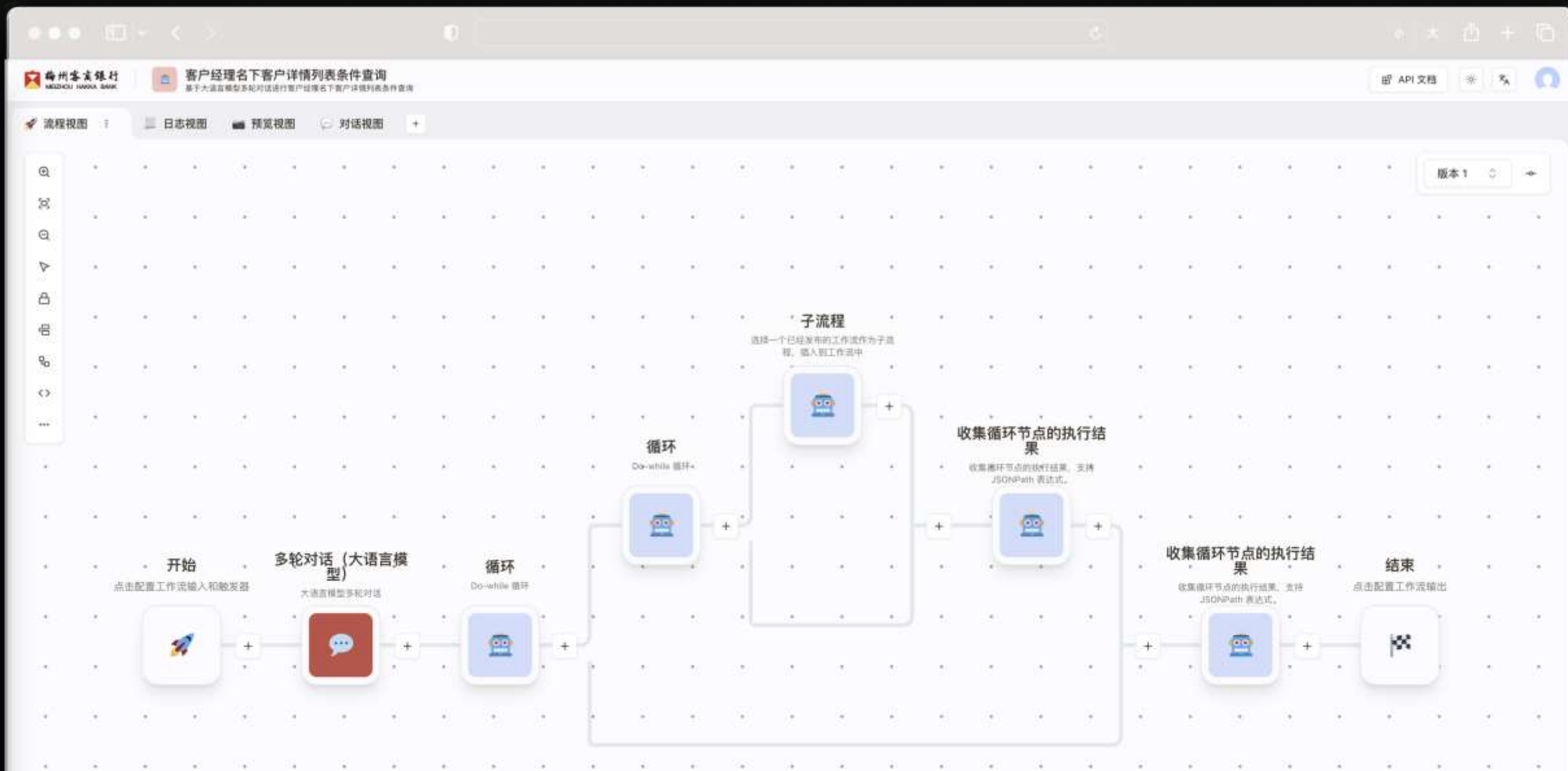
■ 流程驱动的内容创作平台



■ 流程驱动的内容创作平台



■ 流程驱动的数据查询平台



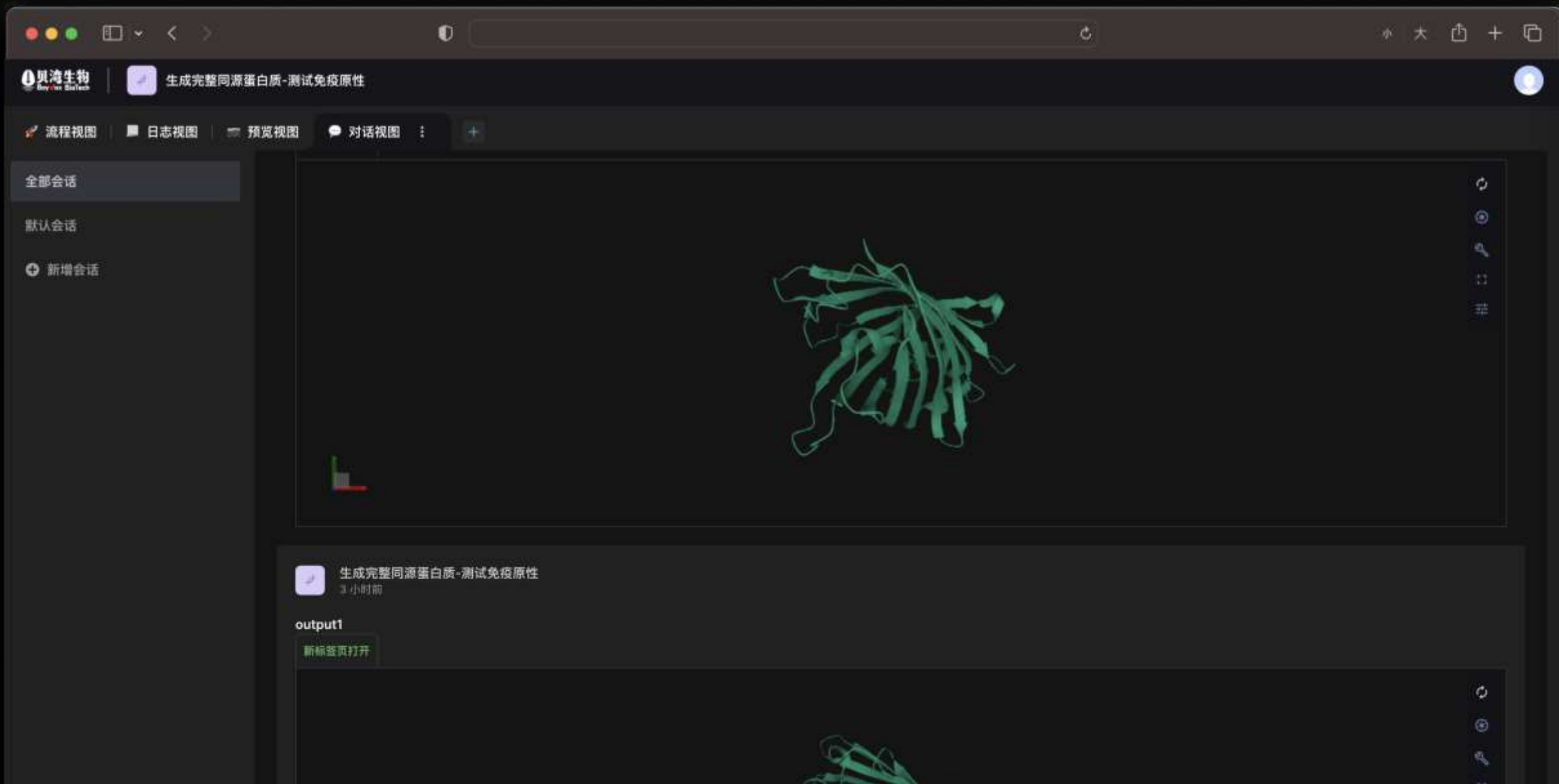
■ 流程驱动的数据查询平台



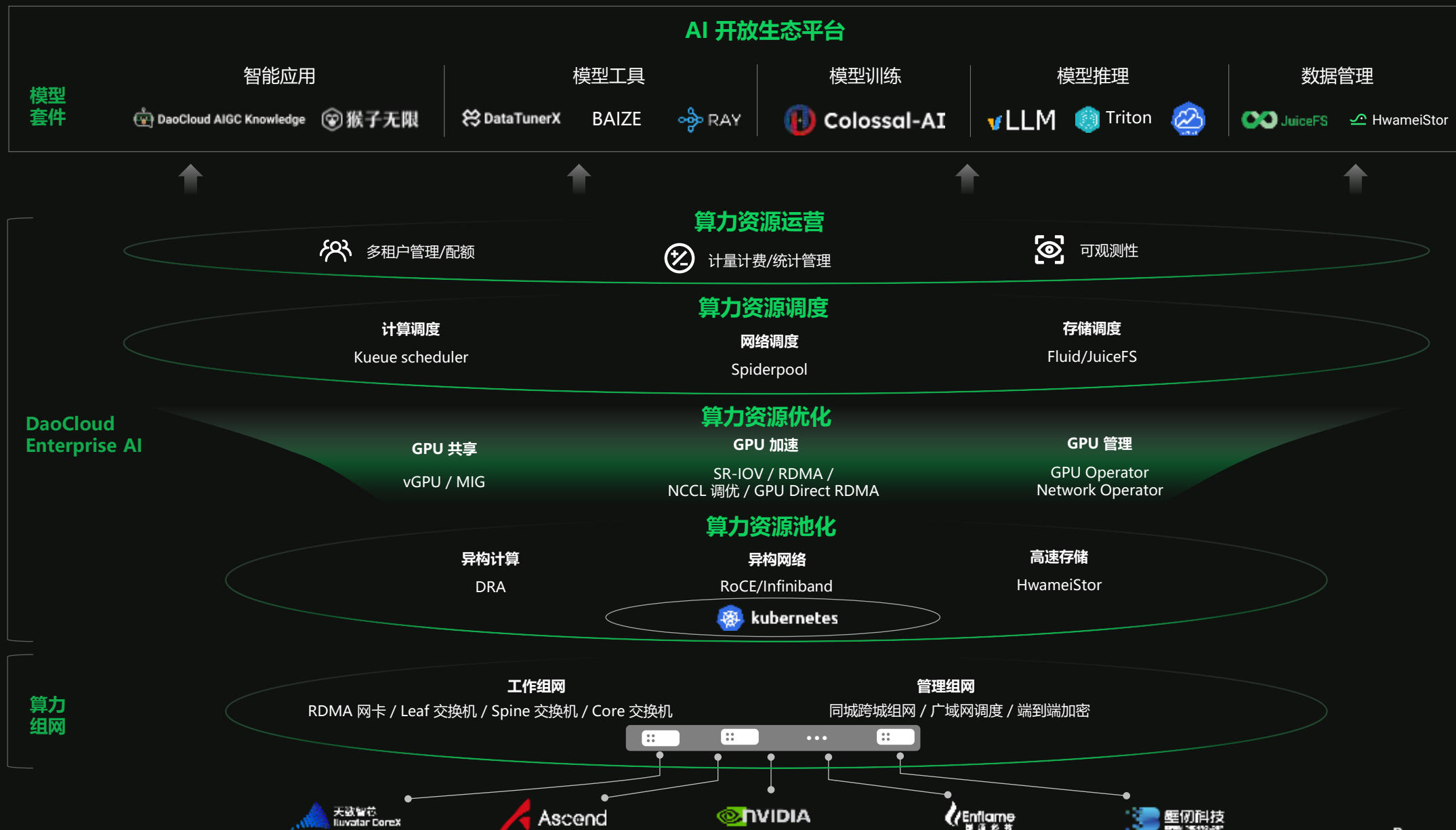
■ 流程驱动的生命科学干实验平台



■ 流程驱动的生命科学干实验平台



d.run 产品全景图



智能算力

购买算力

算力集群

可观测性

模型应用

智能问答

流程编排

模型工具

模型微调

算法开发

模型中心

管理

算力

查看购买记录

订购的 1 个算力集群正常使用中; 近2天内 1 个集群「zone-b-cluster-05」会到期。如需续期请 [联系客服](#)。

选择地点

合肥



上海



选择显卡 *



A100 80G * 1
Price: ¥875.33



0%已购买

总数: 10



A100 80G * 2
Price: ¥1230.22



0%已购买

总数: 10

购买



扫描二维码，添加我的企业微信

Part 03

workflow和智能体的迭代

■ 十倍加速大模型应用构建、迭代和高可用

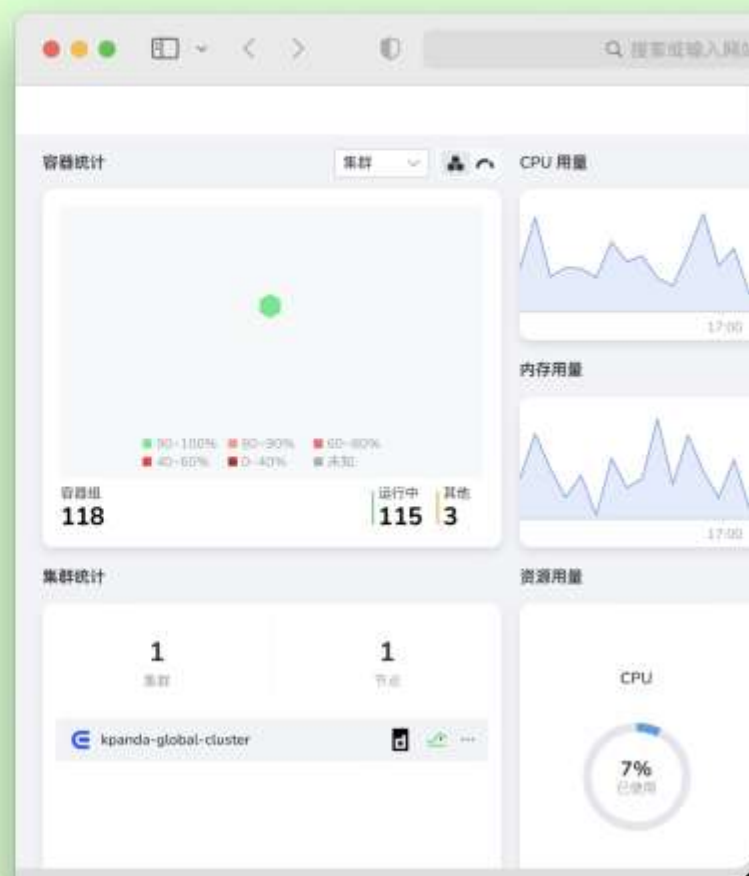
构建 云原生流程引擎



迭代 可复制专业服务



高可用 容器编排管理



■ 构建工作流和智能体迭代的通路

基于目标
拆解多条流程
完成构建

构建评价体系
评估多条流程
完成评估

基于评价结果
优化流程
持续迭代

■ 构建工作流和智能体迭代的通路：评价

已经保存到云端

分享

统计结果

详细结果

菜单

撤销

重做

格式刷

清除格式

插入

B

S

I

U

田

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

三

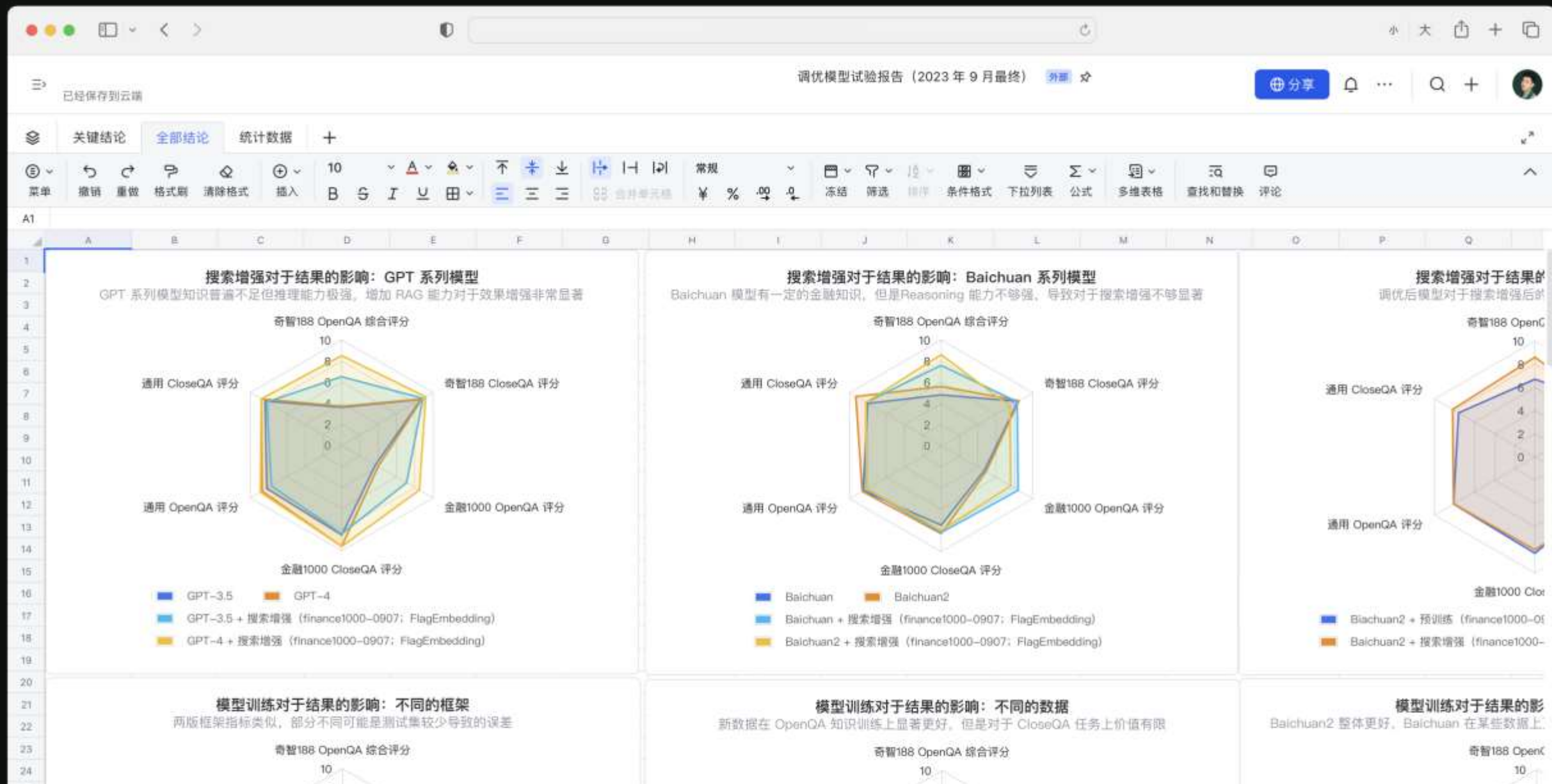
三

三</

■ 构建工作流和智能体迭代的通路：评价

| | | | | |
|--|--|--------------------|--------|-------------|
| 调优模型试验报告 (2023 年 9 月最终) 外部 | | | | |
| 最近修改: 10月8日 15:25 | | | | |
| 关键结论 全部结论 统计数据 | | | | |
| 菜单 撤销 重做 格式刷 清除格式 插入 10 常规 冻结 筛选 排序 条件格式 下拉列表 公式 多维表格 查找和替换 评论 | | | | |
| A1 | 方案唯一命名 | | | |
| | A | B | C | D |
| 1 | 方案唯一命名 | 方案类型 | 模型综合评分 | OpenQA 综合评分 |
| 2 | GPT-3.5 | 基座模型 | 6.48 | 5.47 |
| 3 | GPT-4 | 基座模型 | 6.87 | 5.79 |
| 4 | Baichuan | 基座模型 | 6.84 | 6.25 |
| 5 | Baichuan2 | 基座模型 | 7.29 | 6.59 |
| 6 | GPT-3.5 + 搜索增强 (finance1000-0907; FlagEmbedding) | 基座模型 搜索增强 | 7.58 | 7.16 |
| 7 | GPT-4 + 搜索增强 (finance1000-0907; FlagEmbedding) | 基座模型 搜索增强 | 8.83 | 8.66 |
| 8 | Baichuan + 搜索增强 (finance1000-0907; FlagEmbedding) | 基座模型 搜索增强 | 8.13 | 8.08 |
| 9 | Baichuan2 + 搜索增强 (finance1000-0907; FlagEmbedding) | 基座模型 搜索增强 | 8.04 | 8.1 |
| 10 | Biachuan + 预训练 (finance1000-0822; 第一版框架) + 指令微调 (finance+alpaca) | 基座模型 预训练 指令微调 | 6.71 | 6.31 |
| 11 | Biachuan + 预训练 (finance1000-0907; 第一版框架) + 指令微调 (finance+belle) | 基座模型 预训练 指令微调 | 7.53 | 7.56 |
| 12 | Biachuan + 预训练 (finance1000-0907; 第二版框架) + 指令微调 (finance+belle) | 基座模型 预训练 指令微调 | 7.64 | 7.45 |
| 13 | Biachuan2 + 预训练 (finance1000-0907; 第二版框架) + 指令微调 (finance+belle) | 基座模型 预训练 指令微调 | 7.58 | 7.32 |
| 14 | Biachuan + 预训练 (finance1000-0822; 第一版框架) + 指令微调 (finance+alpaca) + 搜索增强 (finan | 基座模型 搜索增强 预训练 指令微调 | 8.04 | 8.29 |
| 15 | Biachuan + 预训练 (finance1000-0907; 第一版框架) + 指令微调 (finance+belle) + 搜索增强 (financ | 基座模型 搜索增强 预训练 指令微调 | 7.98 | 8.26 |
| 16 | Biachuan + 预训练 (finance1000-0907; 第二版框架) + 指令微调 (finance+belle) + 搜索增强 (financ | 基座模型 搜索增强 预训练 指令微调 | 7.85 | 7.86 |
| 17 | Biachuan2 + 预训练 (finance1000-0907; 第二版框架) + 指令微调 (finance+belle) + 搜索增强 (finan | 基座模型 搜索增强 预训练 指令微调 | 8.14 | 8.38 |
| 18 | | | | |

■ 构建工作流和智能体迭代的通路：迭代



■ 构建能够支撑迭代的大模型原生组织



■ 构建能够支撑迭代的大模型原生组织



■ 智能体和未来的组织

未来世界的组成部分

独立行动智能体
由流程驱动
创造
社会和经济价值

未来世界的组成部分

新一代流程引擎
以大模型为核心
规模化创造
独立行动智能体

未来世界的组成部分

搭流程的智能体
用AI做AI
结合行业数据
构建新的智能体

Part 04

云原生驱动
工作流和智能体高可用落地

■ 十倍加速大模型应用构建、迭代和高可用

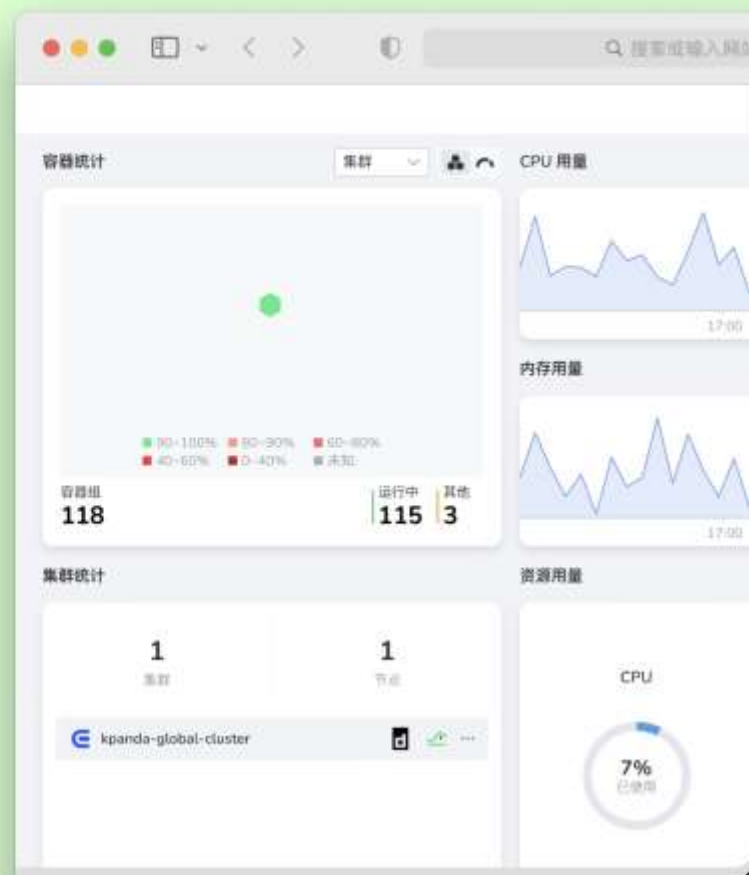
构建 云原生流程引擎



迭代 可复制专业服务



高可用 容器编排管理



■ 大模型应用的云原生技术堆栈

Docker

Kubernetes

Istio

vGPU

HPA

Helm



Docker

大模型应用中的每个工具的解耦合
和高可用

Kubernetes

大模型应用的负载均衡和主备切换



Istio

大模型应用的南北向和东西向 服务管控

大模型应用的硬件适配

芯片 FLOPS 和显存

模型参数和精度

vGPU

大模型的硬件成本控制

HPA

大模型应用的弹性伸缩

Helm

大模型应用的可迁移

■ 系统架构

前端 & 后端 & 流程编排

前端 vines-ui

后端基础服务 vines-server

流程编排 Conductor

缓存 Redis

用户数据 PostgreSQL

流程数据 Elasticsearch

对象存储 Minio

工具调用

关系数据搜索服务
vines-tool-database

文本数据搜索服务
vines-tool-text-collection

推理服务 vines-tool-llm

vLLM 多副本

关系数据 SQLite

向量数据 Elasticsearch

文件系统 FS (可选)

云原生能力

HPA
服务弹性伸缩

Istio
服务容错

vGPU
GPU 共享和超卖

Helm
集群管理和部署

Kubernetes
负载均衡和容器调度

■ 金融级高可用的智能体构建平台

容灾

中间件3副本
服务至少2副本
可用时负载均衡
不可用时主备切换

容错

降低业务间依赖
使用HTTP请求
统一降级、超时和熔断处理

容量

稳态容量
至少为 2
尖刺容量
通过HPA方式支持



扫描二维码，添加我的企业微信