基于大模型的智能应用平台在券商典型场景的探索与实践

吴克乾 | 长江证券 信息技术总部

目录

01 大模型快速的发展历程

准备期

故事的起点：chatgpt发布，国内迅速形成大模型共识

成长期

- OpenAI发布GPT4
- 百度发布文心一言1.0
- 谷歌发布chatglm
- 360发布360智脑
- 科大讯飞发布星火
- 阿里发布通义千问了手机
- Meta开源Llama2
- 百川智能开源Baichuan
- 华为发布盘古3.0
- 腾讯科技发布混元2.0
- 云从科技发布从容大模型
- 西湖心辰发布西湖大模型
- ……

数量爆发期

- OpenAI发布多模态Gpt4V
- GPT-4 Turbo发布
- 百度升级文心一言4.0
- 清华发布ChatGLM3
- 字节跳动发布大模型产品豆包
- 腾讯科技升级混元3.0
- 阿里发布Qwen-7B
- 零一万物发布Yi-34B
- Google发布多模态大模型Gemini
- OpenAI发布文生视频Sora
- 百川智能开源Baichuan3
- 智谱AI发布GLM-4
- 科大讯飞发布大模型3.5
- 商汤发布Gemini1.5
- 阿里云开源Qwen2
- 幻方开源DeepseekV2
- OpenAI发布Gpt4o
- 零一万物发布Yi-1.5
- ……

突破期

- Deepseek-R1推理大模型
- QwQ推理大模型
- Qwen3
- Manus
- MCP
- ……

发展快

效果好

投入大

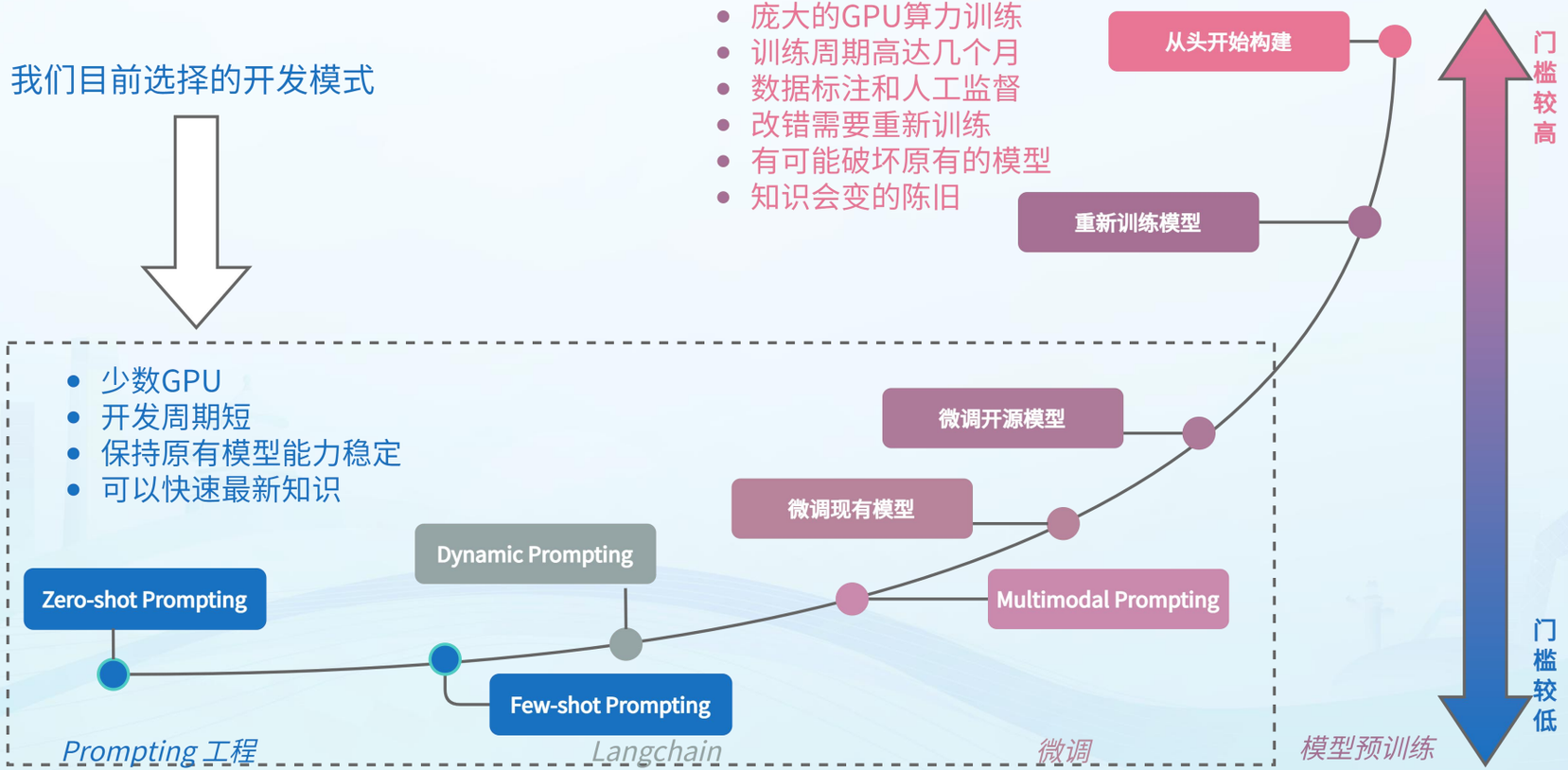
2022年12月 → 2023年2月 → 2024年3月 → 2025年1月至今 → 时间

01 主流的大模型开发模式

我们目前选择的开发模式

- 庞大的GPU算力训练
- 训练周期高达几个月
- 数据标注和人工监督
- 改错需要重新训练
- 有可能破坏原有的模型
- 知识会变的陈旧

- 少数GPU
- 开发周期短
- 保持原有模型能力稳定
- 可以快速最新知识



01 大模型开发中面临的问题

成本高

- 本地部署开源大模型需要大量算力资源
- 高端算力购买难、部署周期长、成本高

模型多

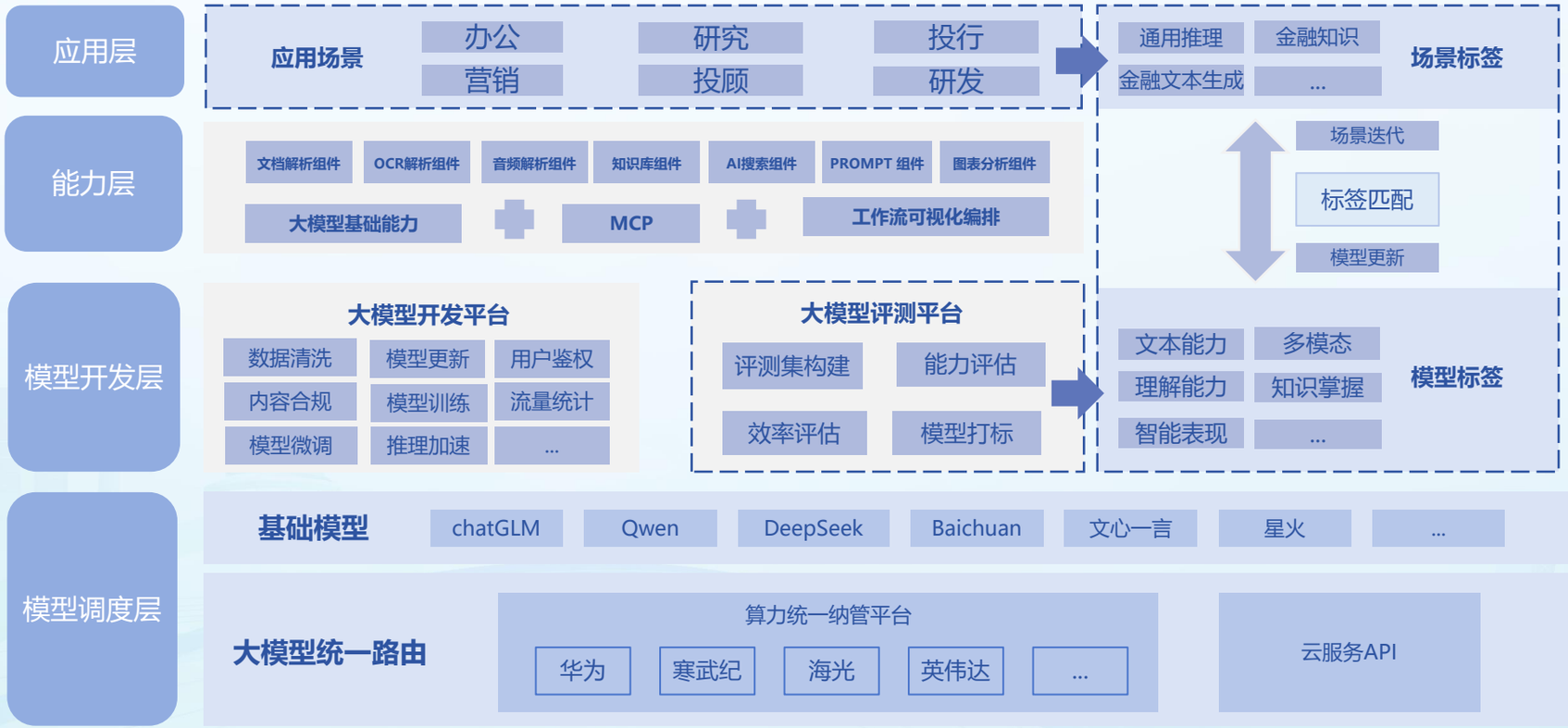
- 通用模型和垂直模型百花齐放、迭代快
- 模型真实性能和特点有待评估

业务繁

- 基础能力缺少封装，应用开发效率低
- 业务繁多，代码编写门槛高，无法快速满足需求

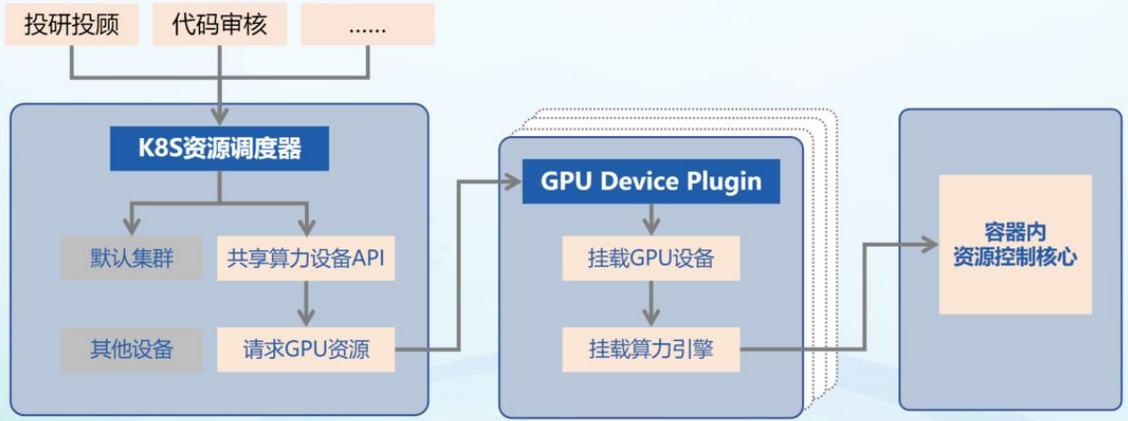
目录

02 “长江灵曦”大模型平台架构

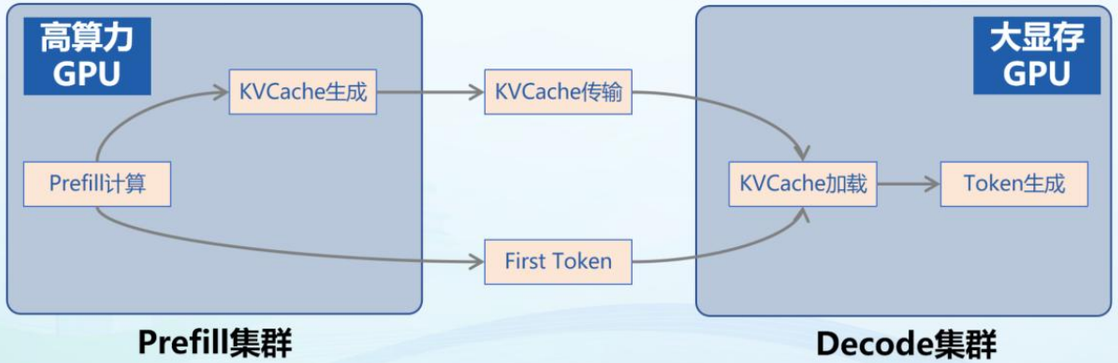


02 解决成本高——统一算力纳管平台

多源异构GPU纳管：算力资源池通过扩展Kubernetes的调度机制，实现对昇腾、寒武纪、英伟达等**异构算力资源**的**统一管理**和**灵活调度**。系统采用设备插件技术构建资源抽象层，解决了不同厂商GPU的驱动兼容和资源协同问题，支持细粒度的资源切分和动态分配，为上层应用提供标准化的资源视图。

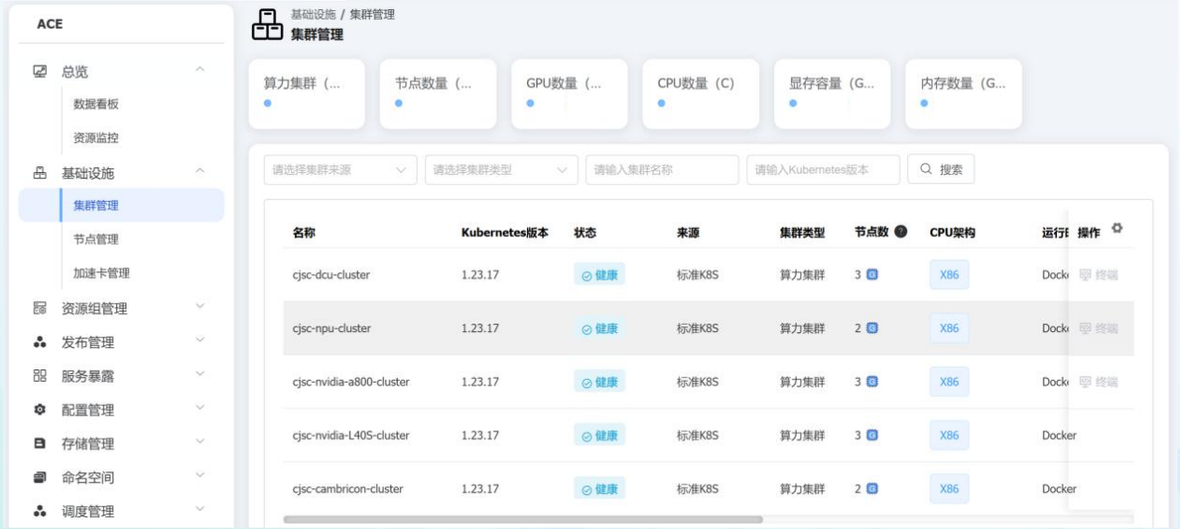


PD分离分布式推理：将大模型推理过程中**计算密集型的Prefill阶段**与**显存密集型的Decode阶段分离部署**，使高端GPU专注于复杂计算，中端GPU承担显存密集型任务，同时优化了分布式推理的KV Cache传输机制，实现按层异步传输，并通过First Token优化方案减少参数加载开销，提升了整体推理性能。



02 解决成本高——统一算力纳管平台

异构算力平台



Qwen2.5-14B-Instruct场景

传统推理并发
指标监控数据

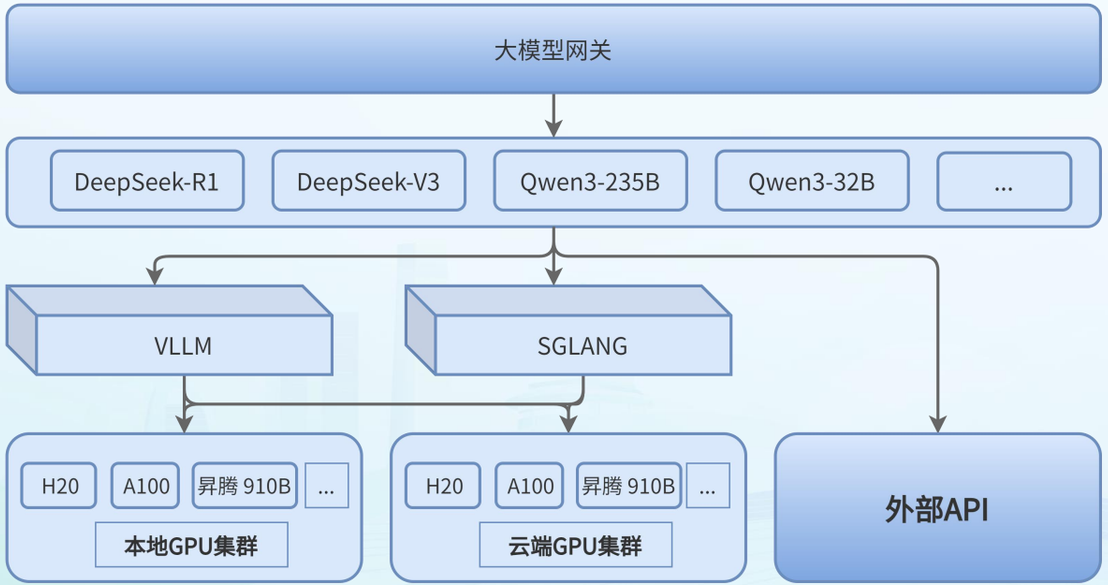
推理平均延时下降48%

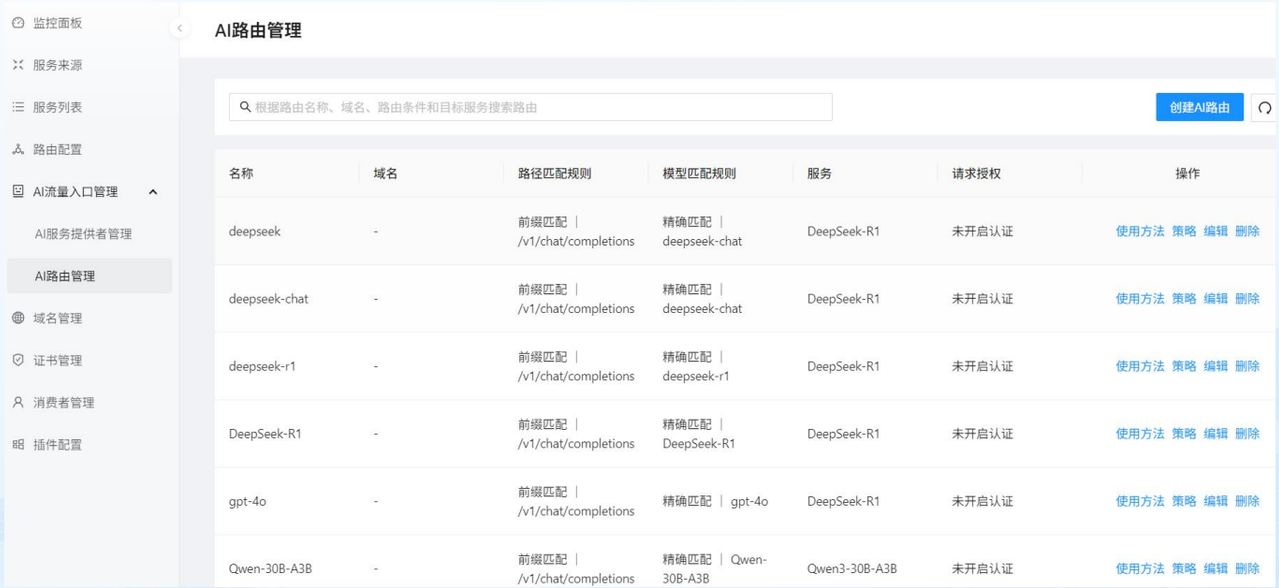
PD分离分布式推理
指标监控数据

在14B推理场景下，同样大幅降低了推理延迟，同时耗时曲线也更为平滑，证明了PD分离分布式推理在不同规模模型上的普遍适应性

02 解决成本高——大模型统一路由

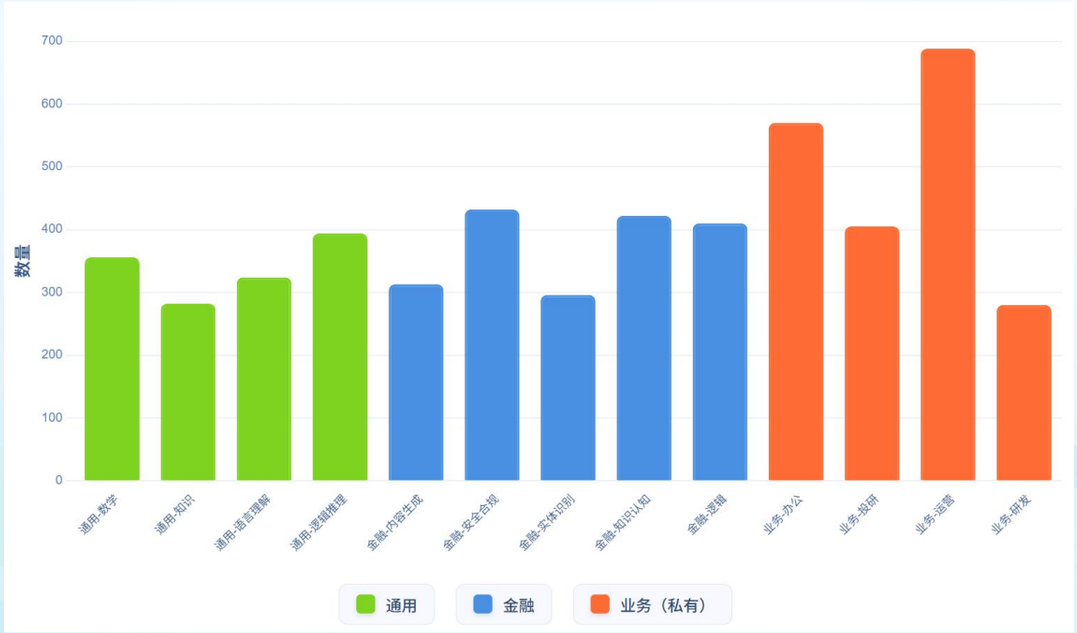
大模型路由：GPU部署大模型 结合 商用大模型API调用



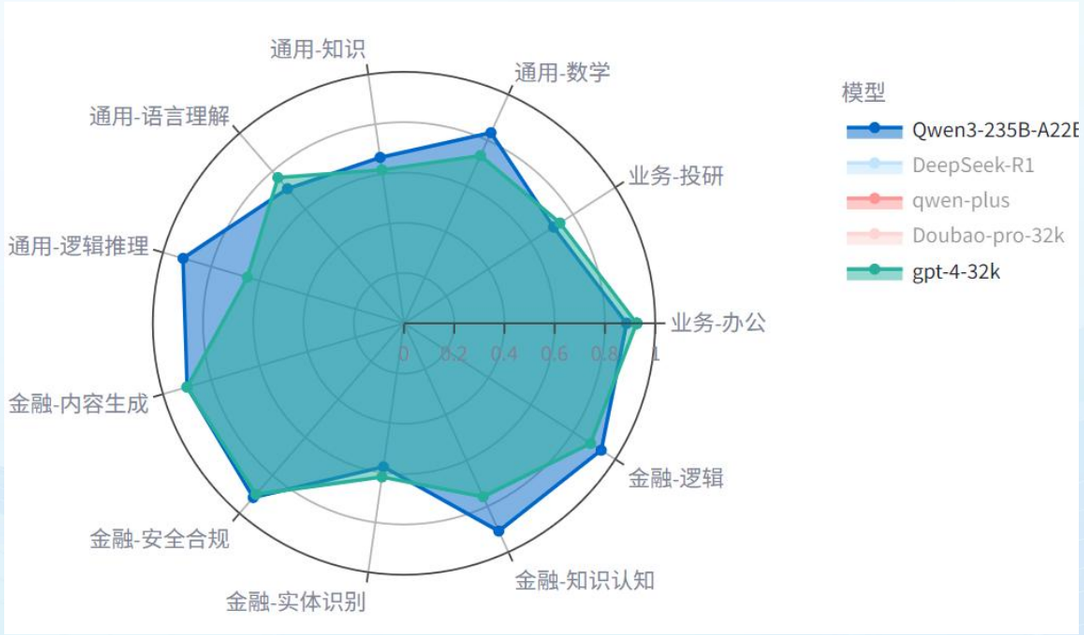


02 解决模型多——”灵曦“大模型评估平台

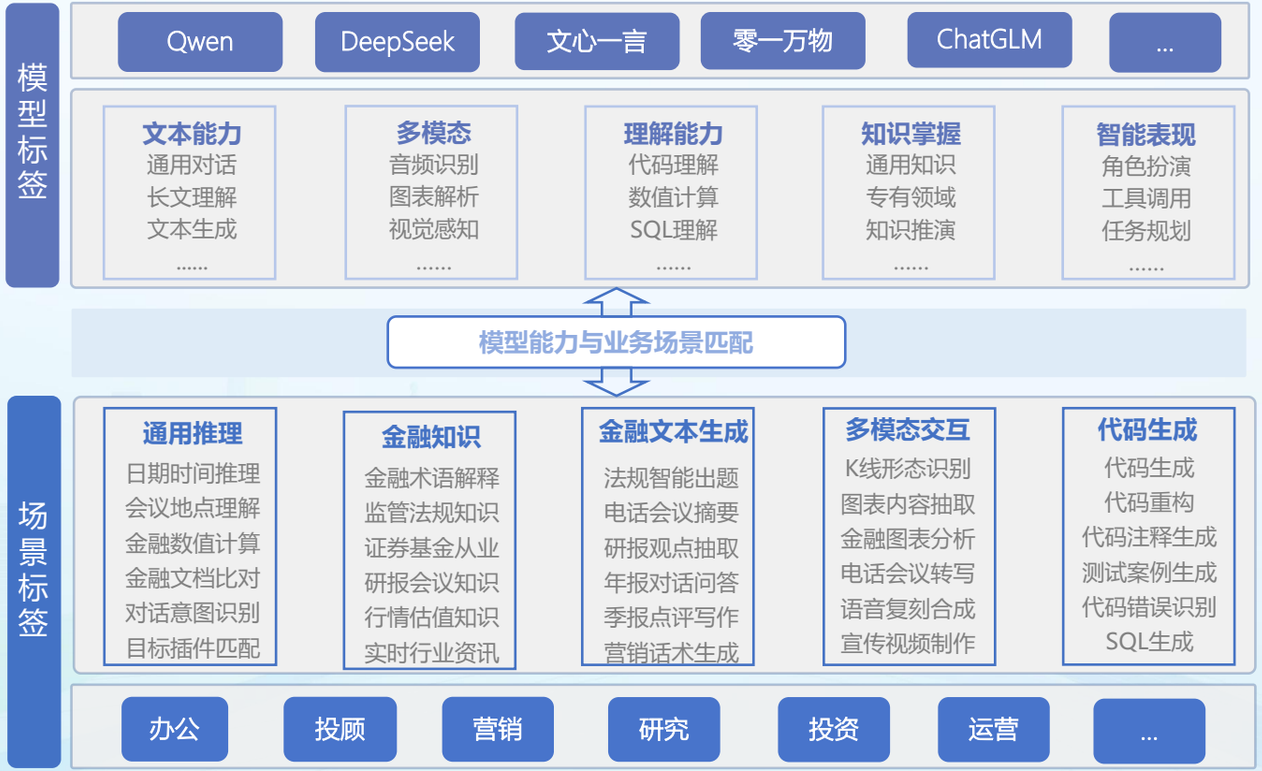
数据集分布



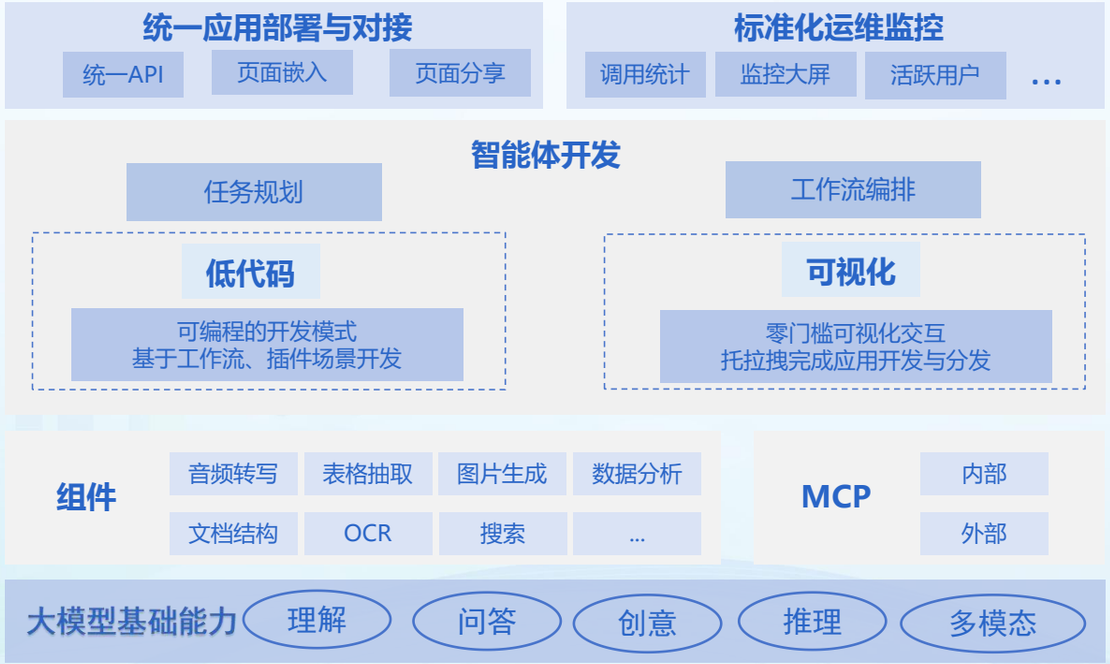
模型评估



02 解决模型多——”灵曦“大模型评估平台



02 解决业务繁——”灵曦“大模型应用开发框架



- 开发应用前&后端

基于工作流开发；
节省80%时间

- Prompt Engineering

基于可视化编排；
节省25%时间

- 应用日志与分析

实时日志与分析
节省70%时间

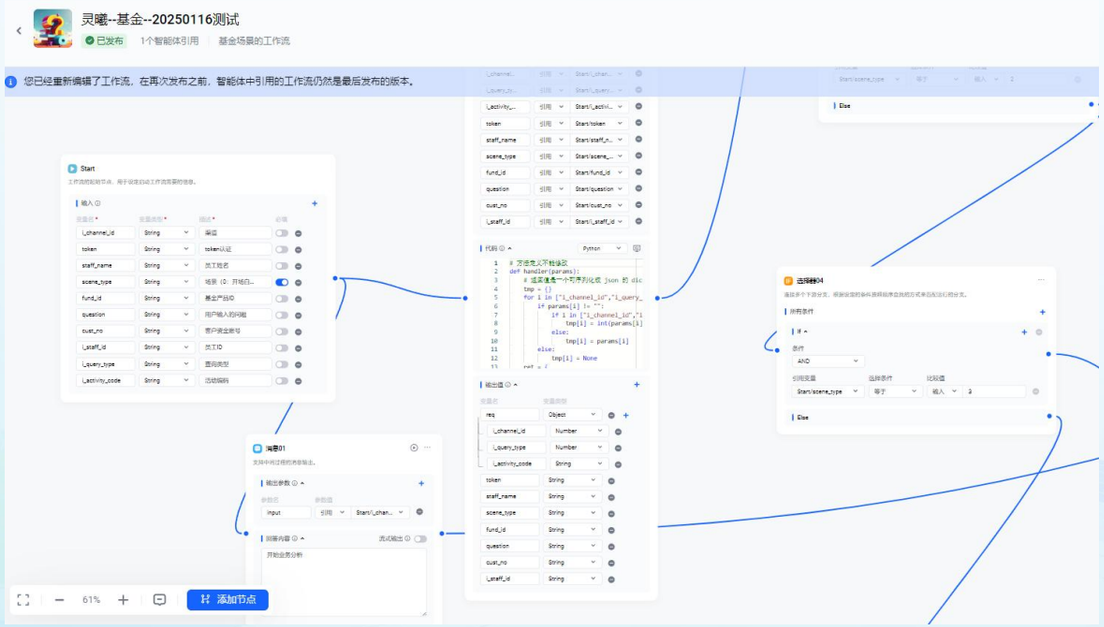
- AI 插件开发与集成

可视化工具创建
快速集成新插件
节省50%时间

02 解决业务繁——”灵曦“大模型应用开发框架

- 目前已经有400+员工参与了智能体的开发，完成了1300多个智能体的开发。

工作流编排



智能体中心

目录

03 灵曦：大模型证券行业应用

DeepSeek 通义千问 同花顺 科大讯飞 文心一言Ernie 月之暗面KIMI 字节豆包 财联社 恒生warrenQ 零一万物 百川 商汤

灵曦海量模型应用

通用接入

+ 垂类业务

灵曦@投研

- 智搜图文
- 智源对话
- 智译纪要
- 智写研报

灵曦@办公

- 对话式预定会议
- 智能纪要生成
- 思维导图
- 待办生成

灵曦@运营

- 生成相似问
- 大模型知识抽取
- 回答准确率提高20%
- 知识积累速度提高5倍

灵曦@营销

- 金融产品分析
- 基金推荐
- 营销方法生成
- 个性化营销话术

灵曦@研发

- 需求交互生成
- 代码开发测试
- 代码上线评审
- 部署检查与信息生成

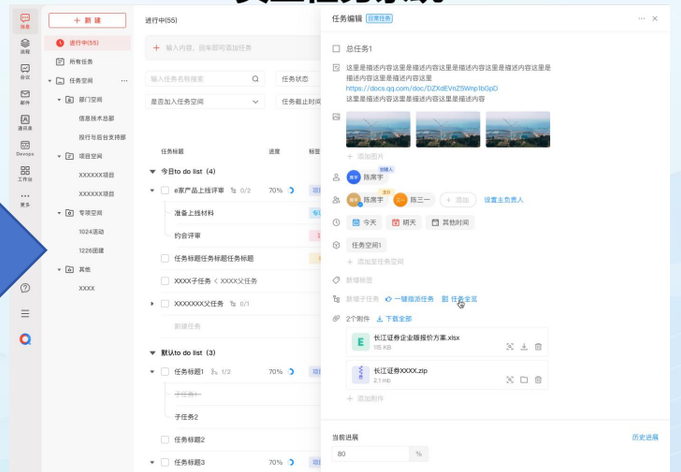
03 “灵曦“赋能办公

- 智能会议纪要已经在多个部门使用，完成**12000+**音频文件的智能分析，完成近**8800**场会议自动撰写纪要、生成重点内容摘要和思维导图

E闪记：智能会议纪要

任务提取

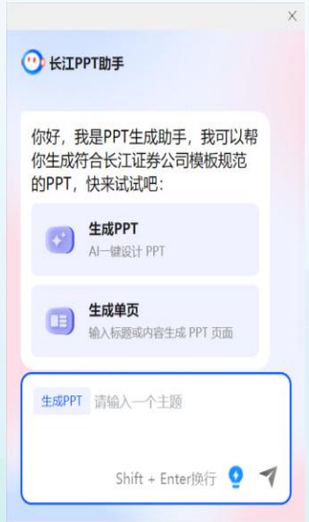
员工任务系统



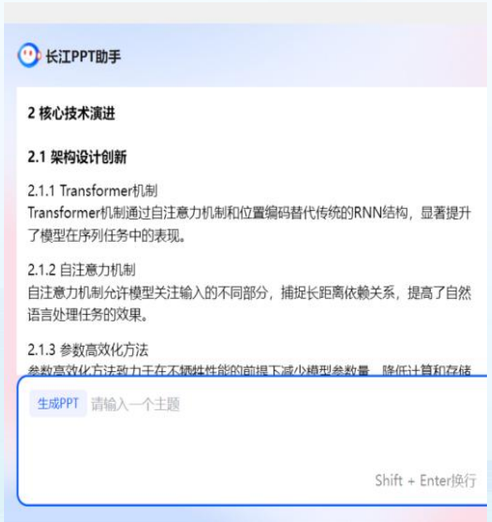
03 “灵曦“赋能办公

- WPS插件：ai-to-ppt，通过大模型生成ppt大纲和内容，并根据公司模板生成完整PPT文件。

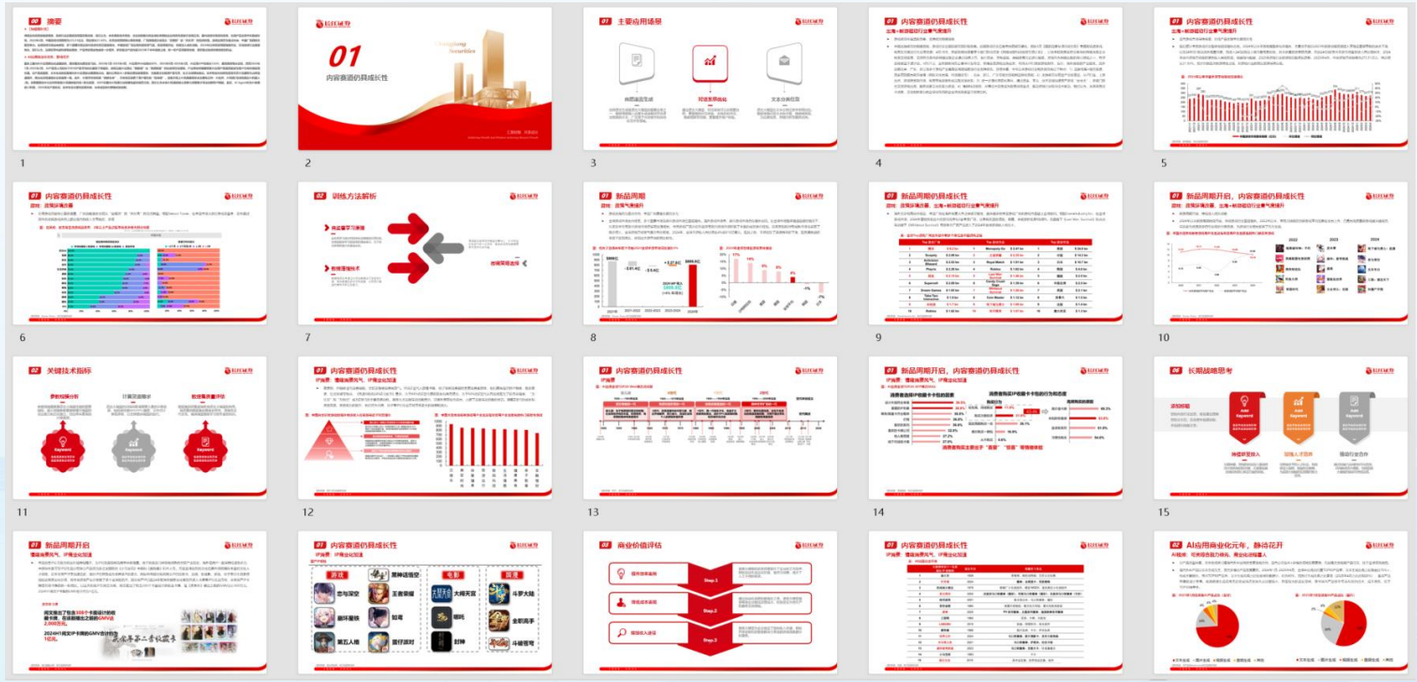
PPT主题



PPT大纲和内容



长江证券模板风格PPT

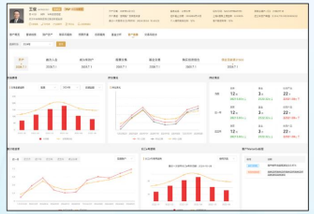


03 “灵曦“赋能营销

- AI营销服务辅助**1000**多名销售人员，覆盖公募基金、理财产品、投顾服务和数据资讯，服务百万客户，节省营销人员搜集数据和用户营销方案准备**60%**的时间。

一体化营销系统

客户画像



- 基本画像
- 自选偏好
- 投资预期
- 历史行为
- ...

产品推荐

产品画像

- 公募基金
- 理财产品
- 投顾服务
- 数据资讯
- ...

聊天辅助

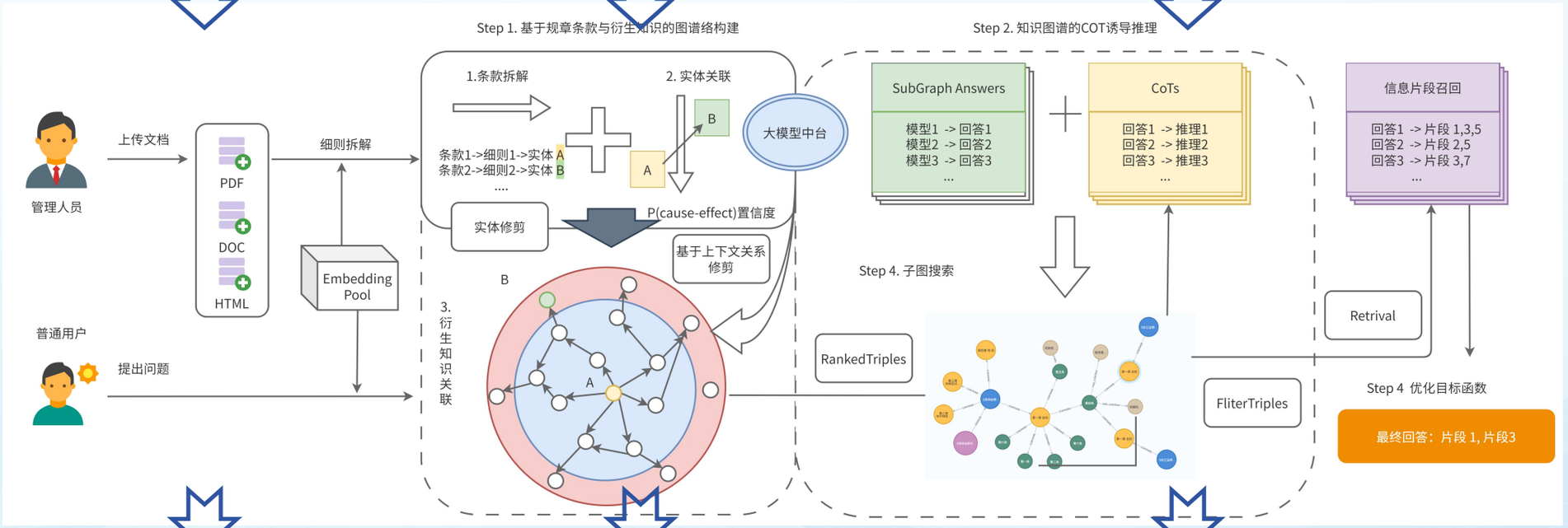
03 ”灵曦“赋能知识检索

大模型知识管理底座

文档复杂，包含图表信息？

知识非结构化，无体系？

语义检索结果不完整，效果差？



文档结构解析+多模态大模型

结构化知识抽取+图谱存储

向量检索+图谱检索+推理模型

03 “灵曦“赋能知识检索——运营业务

- 利用大模型快速生成相似问，辅助传统模型回答准确率从73%提高到90%
- 知识库积累速度对比人工提高了5倍，准确回答运营问题并快速定位答案原文
- i问爱答助手辅助运营完成了全部企业知识的解析和抽取任务，并给全体员工提供问答服务。

制度详情

制度图谱

03 “灵曦“赋能投研

搜

检索长江（或部分友商）研报和路演观点、图表；
实时分析卖方研究（研报、路演）市场动向；

读

1）大量资讯、研报、路演信息快速提取重要信息并按照特定要的要求进行总结。
2）对政策文件、行业重点事件核心观点对比，挖掘核心内容的矛盾点、提取关键要素。

写

提纲撰写内容扩写
段落润色与改写
总结与标题撰写
重点章节对比
财报对比

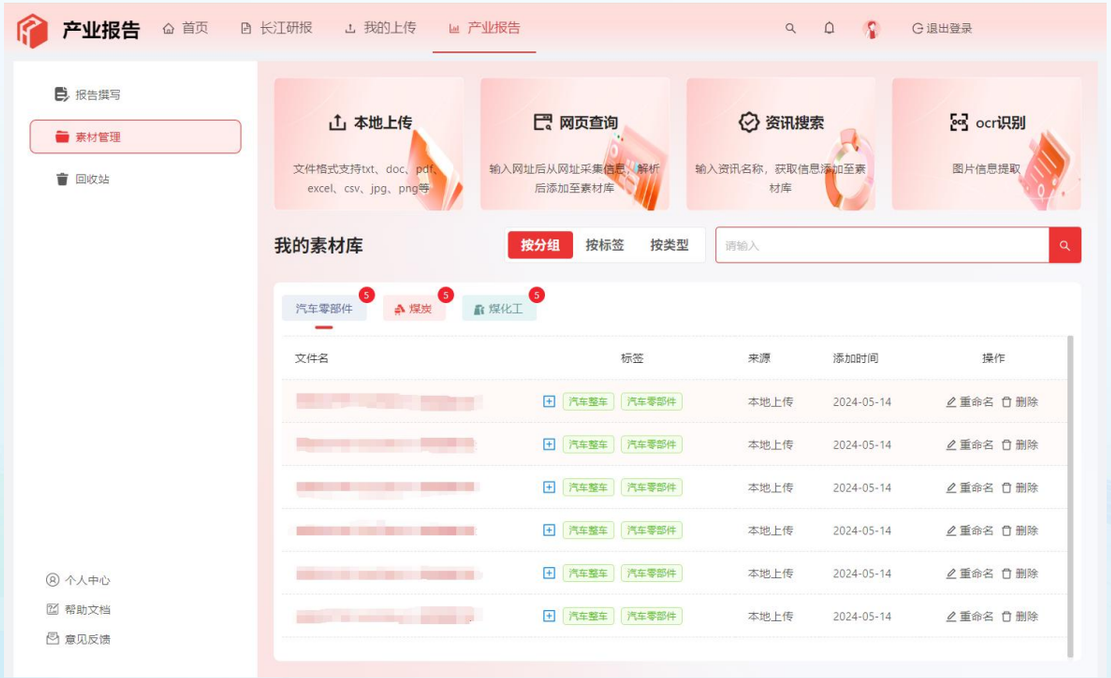
聚

电话会议、路演、调研信息快速转写
快速形成会议纪要
重点内容和思维导图

03 “灵曦“赋能投研——智写研报

- 从**30多万**份素材中整理出**4000**多份核心的素材，进一步提取出**6000+**重点事件和**24000+**有效图片，研究员检索与整理效率提高**60%**
- 提升年报点评写作效率**5**倍以上，覆盖**500+**核心股票池，覆盖研究所和机构客户**80%**以上的用户

素材管理



素材写作

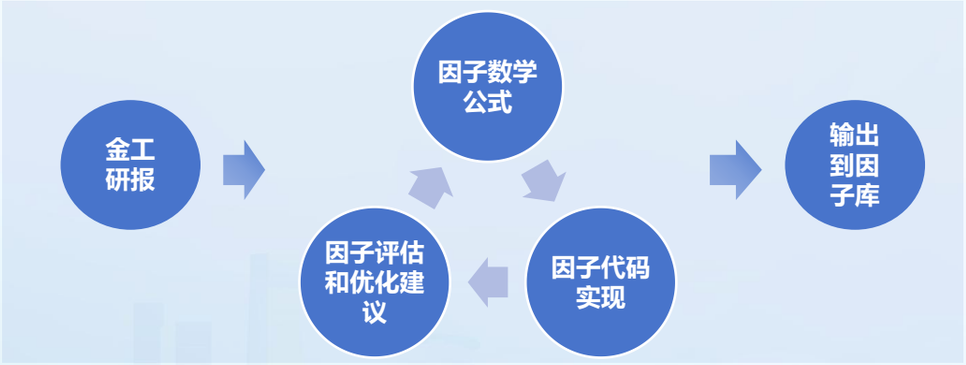
目录

04 未来展望


自主智能
学习和决策
认知智能
知识、推理、执行
感知智能
识别和处理各种感官数据，
如图像、声音
运算智能
基本的数学运算和逻辑处理
manus
The general AI agent
扣子空间
BETA
?
LangChain
Dify
Microsoft
AutoGen
PYTORCH
Google
BERT
ChatGPT
NumPy

04 未来展望

投资：多智能体协同策略构建



报告自动生成

谢谢观看