

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

1

大模型应用开发背景

2

“长江灵曦”大模型平台

3

“长江灵曦”应用场景

4

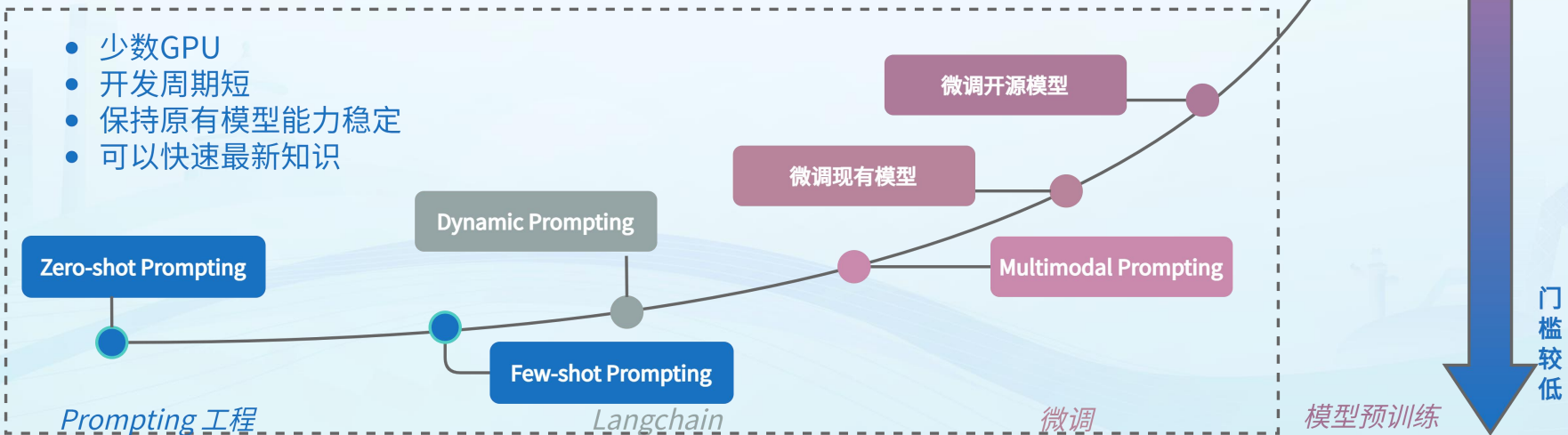
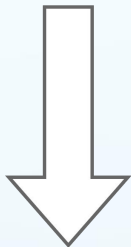
未来展望

01 大模型快速的发展历程



01 主流的大模型开发模式

我们目前选择的开发模式



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

1

大模型应用开发背景

2

“长江灵曦”大模型平台

3

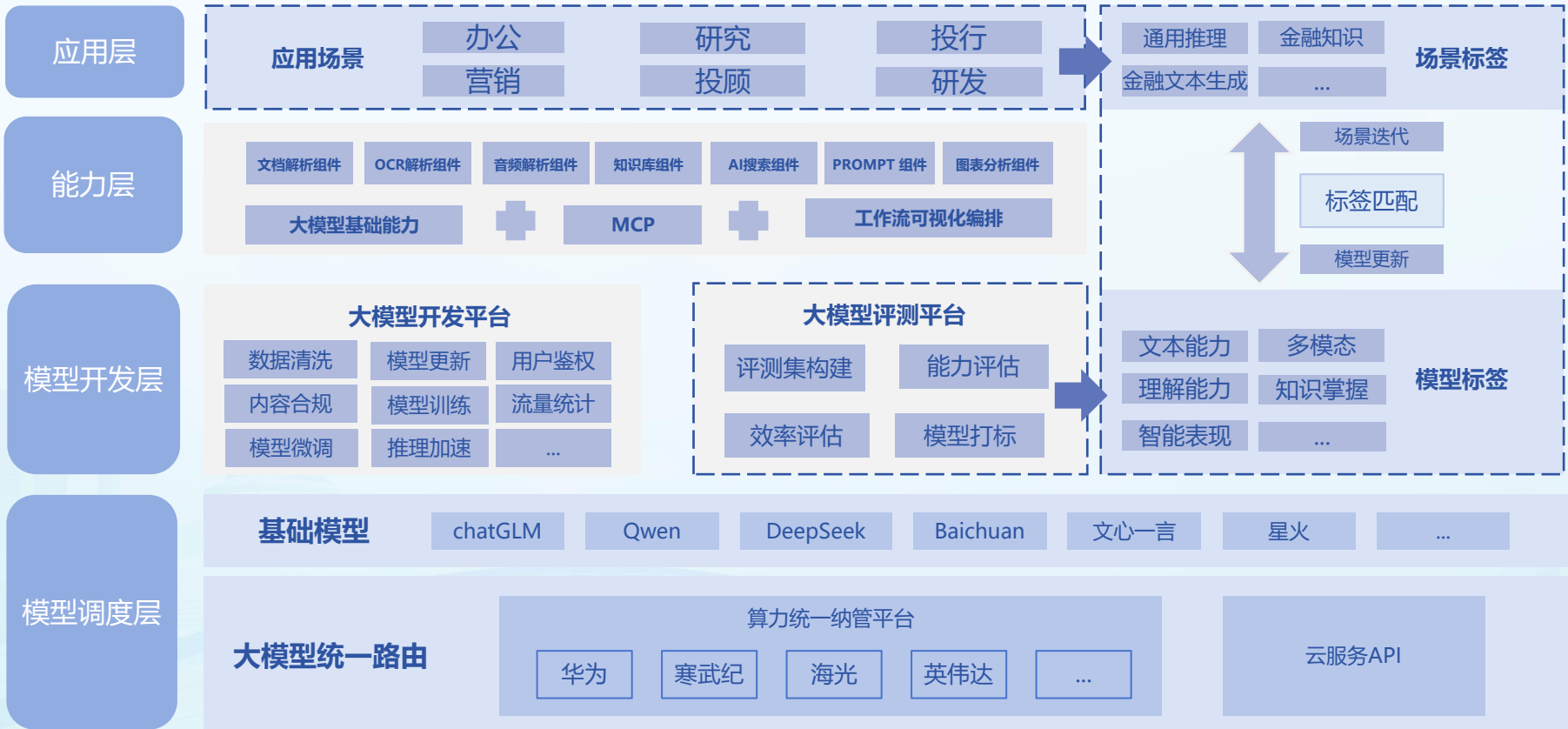
“长江灵曦”应用场景

4

未来展望

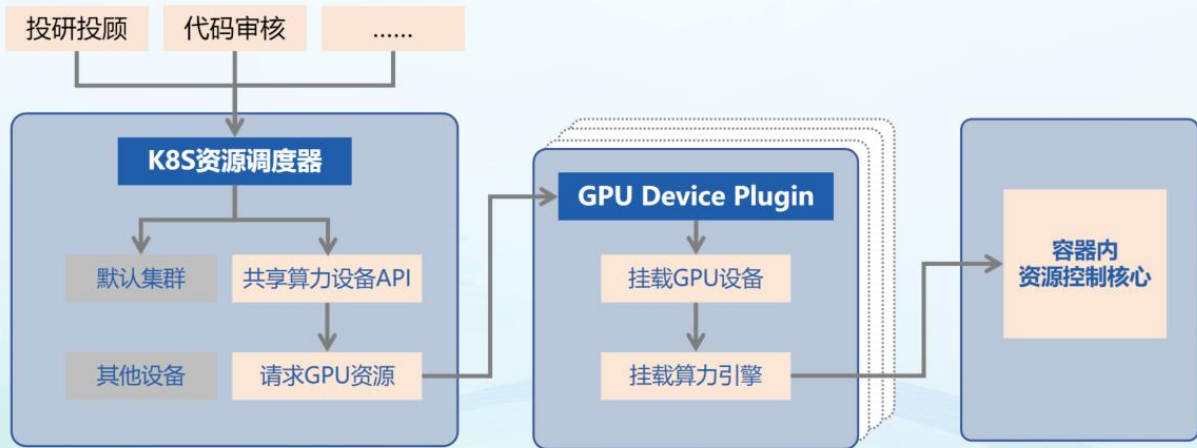
02

“长江灵曦”大模型平台架构

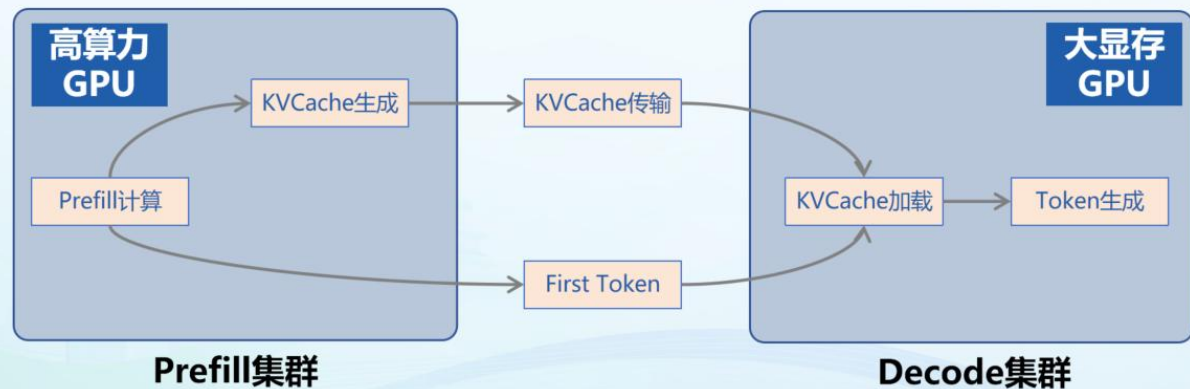


02 解决成本高——统一算力纳管平台

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。

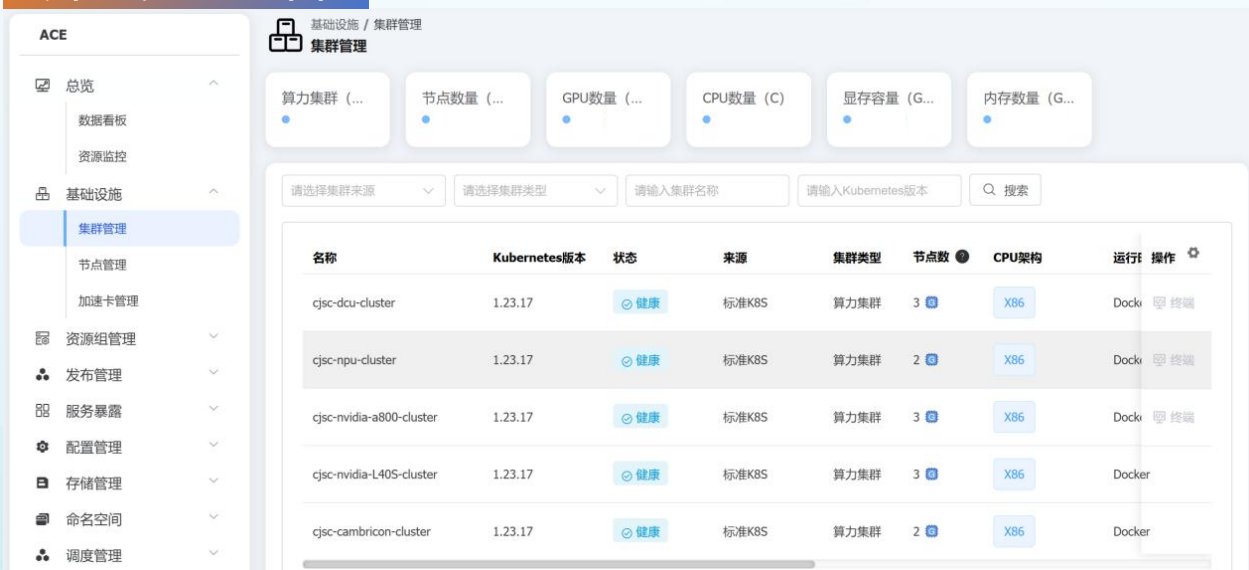


PD分离分布式推理：将大模型推理过程中**计算密集型的Prefill阶段**与**显存密集型的Decode阶段分离部署**，使高端GPU专注于复杂计算，中端GPU承担显存密集型任务，同时优化了分布式推理的KV Cache传输机制，实现按层异步传输，并通过First Token优化方案减少参数加载开销，提升了整体推理性能。



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%

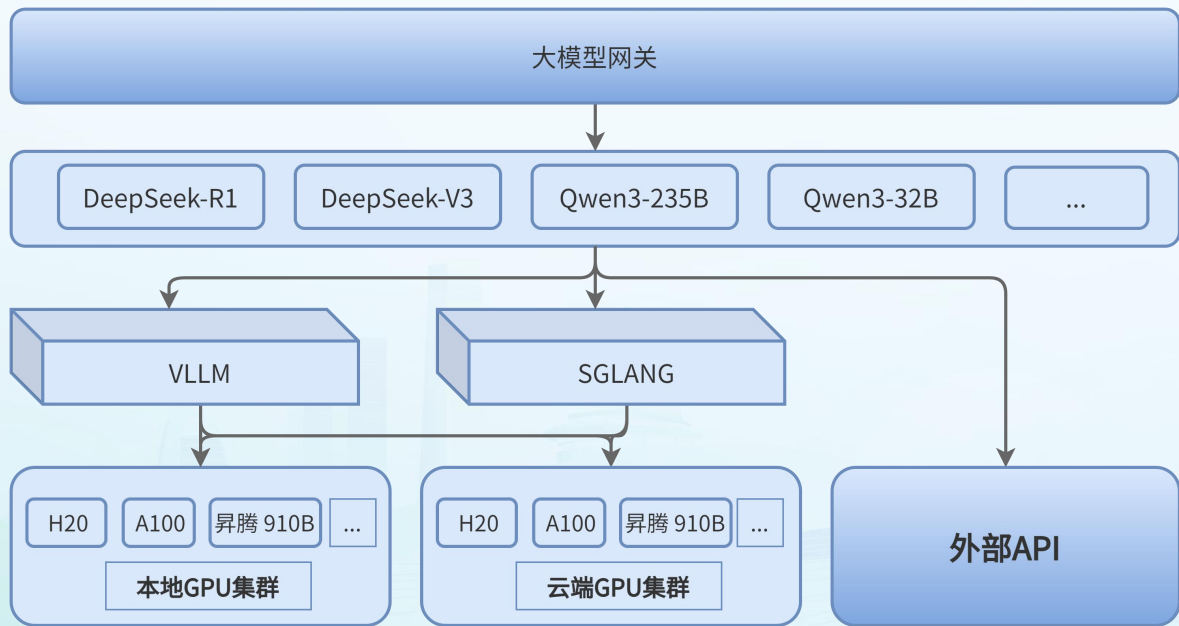


PD分离分布式推理
指标监控数据

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



AI路由管理

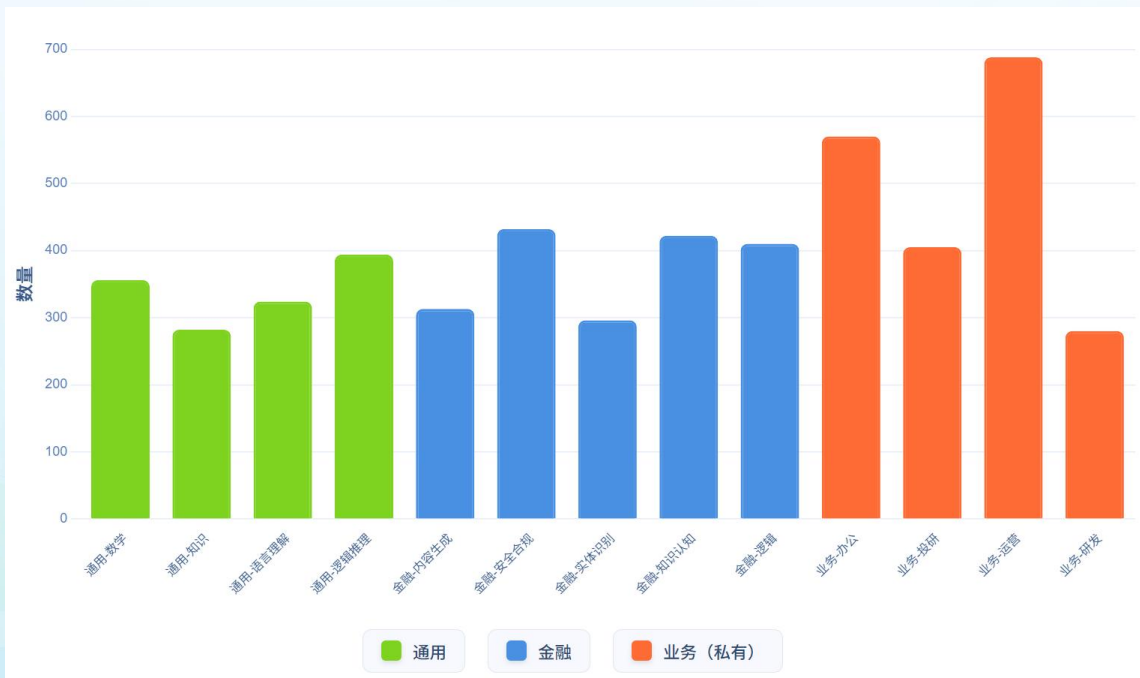
根据路由名称、域名、路由条件和目标服务搜索路由

创建AI路由

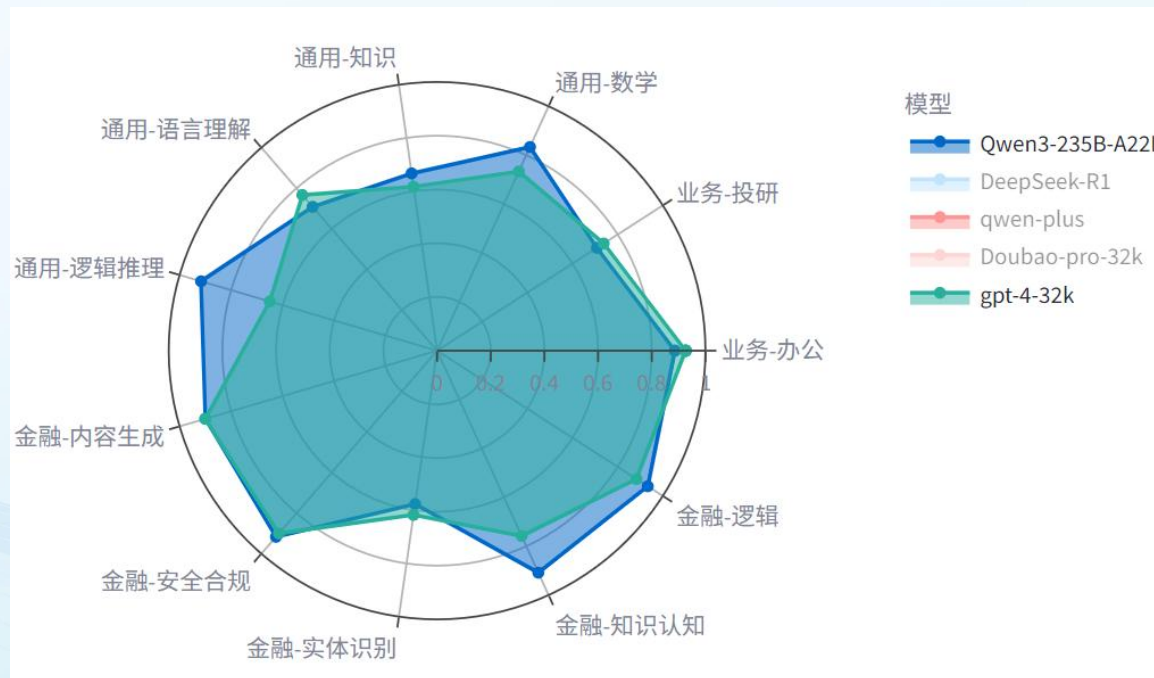
名称	域名	路径匹配规则	模型匹配规则	服务	请求授权	操作
deepseek	-	前缀匹配 /v1/chat/completions	精确匹配 deepseek-chat	DeepSeek-R1	未开启认证	使用方法 策略 编辑 删除
deepseek-chat	-	前缀匹配 /v1/chat/completions	精确匹配 deepseek-chat	DeepSeek-R1	未开启认证	使用方法 策略 编辑 删除
deepseek-r1	-	前缀匹配 /v1/chat/completions	精确匹配 deepseek-r1	DeepSeek-R1	未开启认证	使用方法 策略 编辑 删除
DeepSeek-R1	-	前缀匹配 /v1/chat/completions	精确匹配 DeepSeek-R1	DeepSeek-R1	未开启认证	使用方法 策略 编辑 删除
gpt-4o	-	前缀匹配 /v1/chat/completions	精确匹配 gpt-4o	DeepSeek-R1	未开启认证	使用方法 策略 编辑 删除
Qwen-30B-A3B	-	前缀匹配 /v1/chat/completions	精确匹配 Qwen-30B-A3B	Qwen3-30B-A3B	未开启认证	使用方法 策略 编辑 删除

02 解决模型多——“灵曦”大模型评估平台

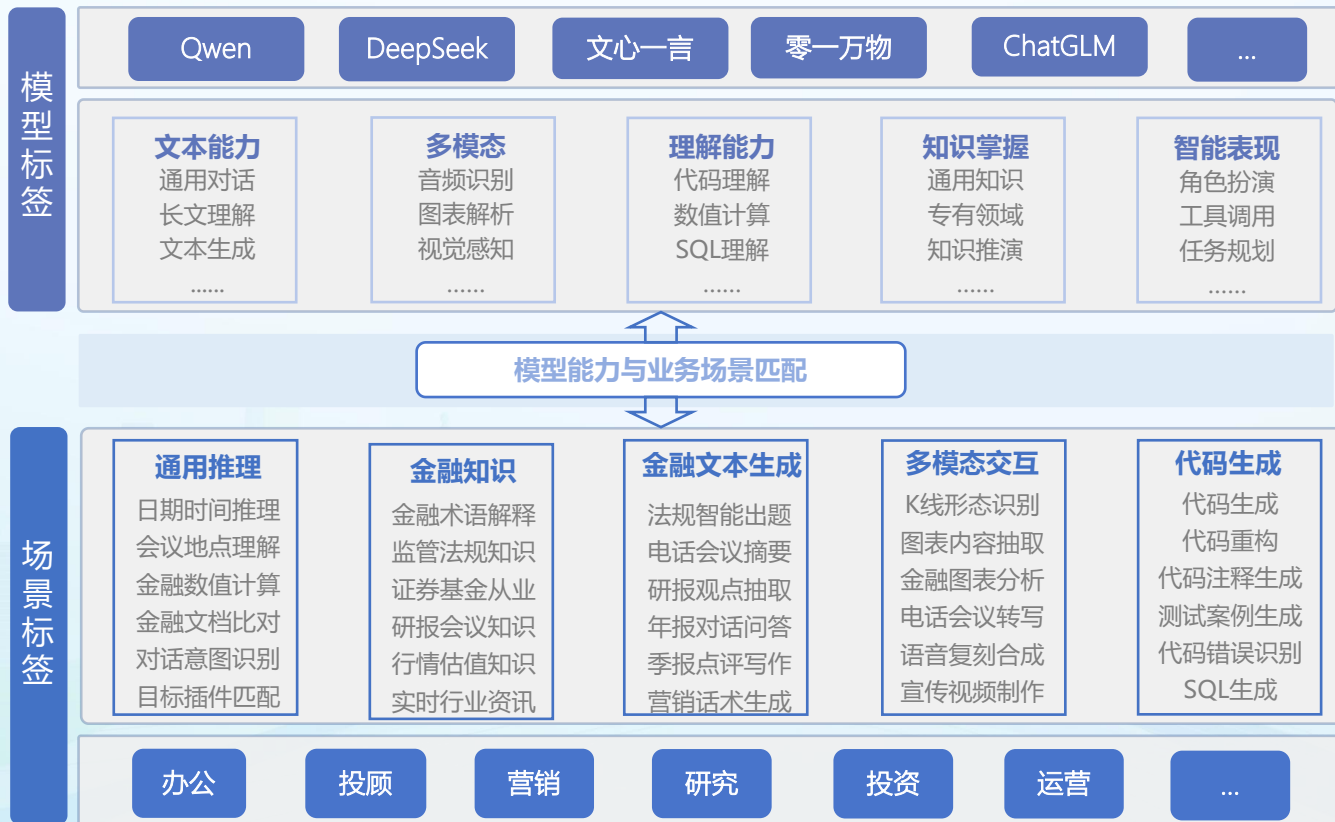
数据集分布



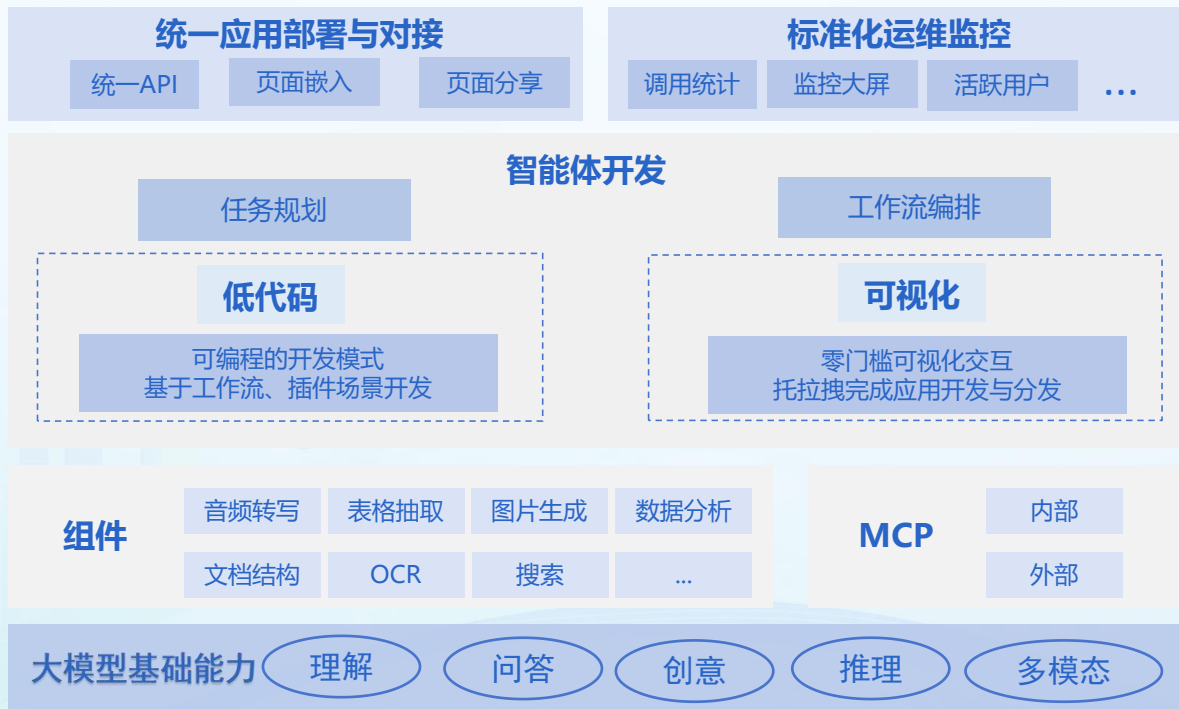
模型评估



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 workflow 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。

workflows编排

The screenshot shows a workflow editor for a task named "灵曦-基金--20250116测试". It features a central code editor with Python code for handling parameters and a right-side configuration panel for variables like token, start_name, and fund_id. A "Start" node is visible on the left, and a "结束" (End) node is on the right. A "选择节点" (Select Node) dialog is open, showing a list of nodes to be added to the workflow.

智能体中心

The screenshot displays the "智能体中心" (AI Agent Center) dashboard. It contains a grid of agent cards, each representing a different AI-powered service. Each card includes a title, a brief description, and the AI model used for its operation. The agents are categorized by status (e.g., "未发布" - Not Published, "已发布" - Published) and the specific AI model (e.g., "千问plus", "GPT4", "deepseek").

Agent Name	Description	Status	AI Model
客户近期关...	客户近期关注偏好查询	未发布	对话型
个股扫雷	个股扫雷	已发布	对话型 - GPT4
期权大赛报名查...	期权大赛报名信息查询	已发布	对话型 - 幻方deepseek
诊股	对股票进行深入分析和评估	已发布	对话型 - KIMI128K
今日热门股票	热门股票推荐选股	已发布	对话型 - 千问plus
今日机会	条件选股主题查询	已发布	对话型 - 千问plus
我的客户经理	我的客户经理	已发布	对话型 - GPT4
掌握我的客户	掌握我的客户	已发布	对话型 - 千问plus
检查手机号码激...	检查手机号码激活APP记录	已发布	对话型 - GPT4
IP归属地查询	IP归属地查询	已发布	对话型 - GPT4
E号通行证查询	E号通行证查询	已发布	对话型 - 千问plus
差标查询	差标查询	未发布	对话型
托管产品策略查...	托管产品策略查询和修改	已发布	对话型 - KIMI128K
AI早报	资讯信息获取和推送	未发布	对话型 - 千问plus
AI资讯	从网站提取资讯信息并推送，以机器人行业网站为例	已发布	流程编排型
路演会议查询	研究所路演会议信息查询	已发布	对话型 - 本地deepseek1
研报查询	研报查询	已发布	对话型 - 本地deepseek1
研报文档审核助...	帮助纠正研报文档中的错别字、符号运用以及识别敏感词，助力于全面提升文档内容质量	未发布	对话型 - 本地千问32B
AI选股	根据多种指标叠加，选股市场上满足条件的股票	未发布	对话型 - 幻方deepseek
AI资讯_工作流...	机器人爬取网站并整理资讯数据输出，以机器人行业早报为例	已发布	对话型 - 千问plus

目录

- 1 大模型应用开发背景
- 2 “长江灵曦”大模型平台
- 3 “长江灵曦”应用场景
- 4 未来展望

03 灵曦：大模型证券行业应用

DeepSeek
同花顺
文心一言Ernie
字节豆包
财联社
恒生warrenQ
百川
商汤

通义千问
科大讯飞
月之暗面KIMI

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

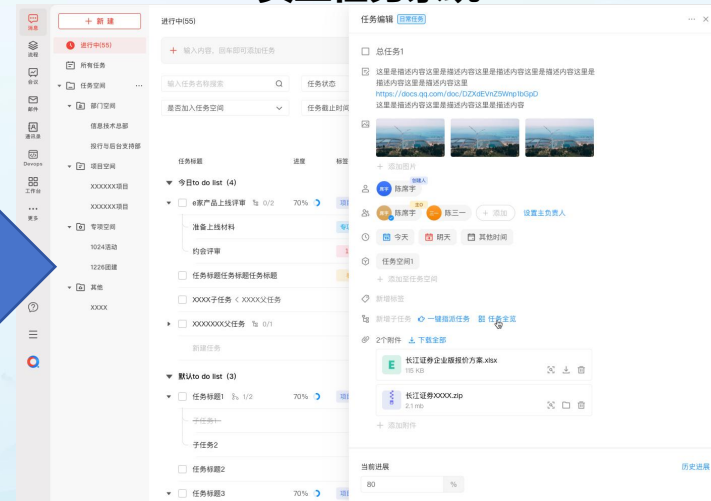
03 “灵曦”赋能办公

- 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



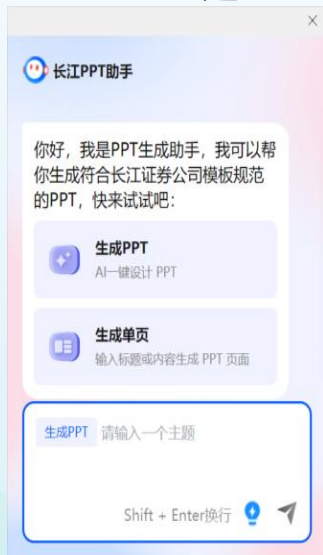
员工任务系统



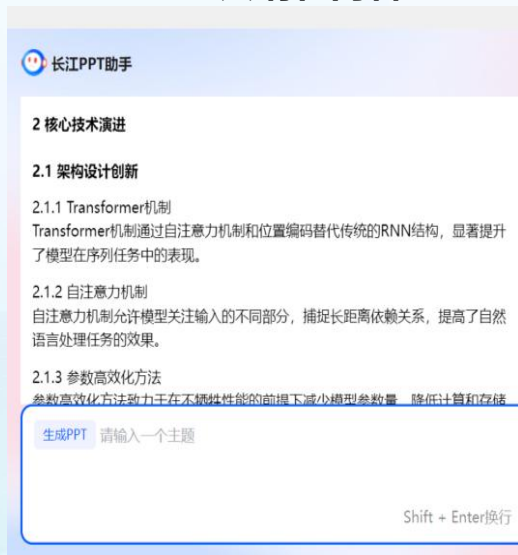
03 “灵曦”赋能办公

- 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。

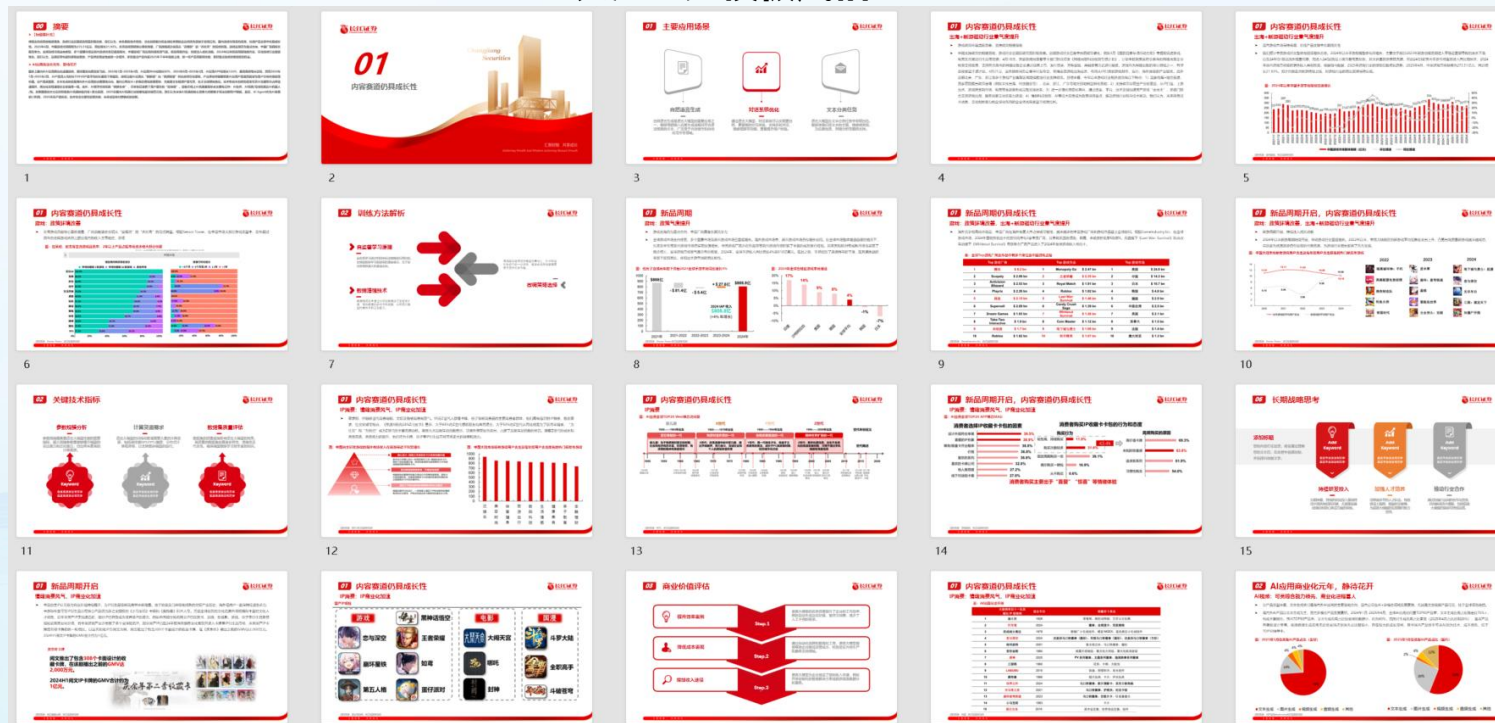
PPT主题



PPT大纲和内容



长江证券模板风格PPT

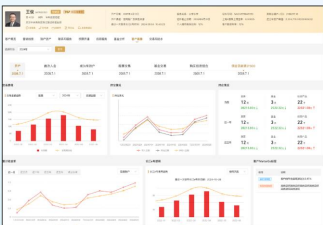


03 “灵曦”赋能营销

- 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
...

灵曦AI营销助手

营销机会 考核任务

组合策略盈利

2025年公募产品“开门红”营销活动

客户列表

- 谢鸣
- 陈芝凤
- 李庆仁**
- 高勇强
- 丁健

客户触达

发短信 打电话 写日志

尊敬的(先生/女士), 您好!

恭喜您本月收益率达到[本月收益率]%, 本年累计收益也达到了[本年收益率]%, 真是太棒了! 您的投资眼光真的非常精准, 选择跟随[组合名称]的策略真是明智之举! 组合近一月收益率达到[近一月收益率]%, 胜率高达[胜率]%, 充分体现了管理人的专业能力, 您能取得这样的成绩, 完全得益于对投资老师的信任和支持!

不知道您对目前的投资体验是否满意? 有没有什么建议或疑问需要和我交流? 我们可以根据您的风险等级[Cx]和投资目标, 进一步优化配置, 帮助您锁定收益、降低波动。

再次为您的成功点赞! 希望未来能继续陪伴您收获更多收益! 如果需要了解更多产品或服务, 随时联系我哦! 祝投资顺利, 生活愉快!

【营销话术】

尊敬的[客户姓名](先生/女士), 您好!

首先必须给您点个大大的赞! 您选择的[组合名称]组合, 近期表现真是太出色了! 本月收益率达到了[本月收益率]%, 本年收益率也有[本年收益率]%, 这充分证明了您敏锐的投资眼光和明智的决策能力! 跟着咱们投资老师的策略走, 果然收获满满, 真是太棒了!

[组合名称]自开始以来累计收益已达[累计收益]%, 近一月收益率[近一月收益率]%, 同时回撤控制在[近一月最大回撤]以内, 波动也很小, 说明咱们的投资老师不仅追求收益, 还特别注重风险控制, 这和您的风险等级[Cx]简直是完美匹配!

不知道您这次的投资体验是否满意? 有没有什么建议或需求想和我聊聊? 比如对后续市场怎么看, 或者有没有其他感兴趣的产品? 我会全力为您解答和支持!

希望咱们继续携手, 一起创造更多收益! 祝您生活愉快, 投资顺利!

您的专属助理
[您的名字]

产品名称	产品代码	客户分类
新华资管科技股票C	017800	客户具有以下特点: 激进型C1; 产品具有如下特点: 股票基金, 以上特点分
新华资管科技股票A	001230	客户具有以下特点: 激进型C1; 产品具有如下特点: 股票基金, 以上特点分
新华资管红利股票A	018800	客户具有如下特点: 稳健型C2; 产品具有如下特点: 股票基金, 以上特点分
新华资管30天滚动持有债	012648	客户具有如下特点: 稳健型C2; 产品具有如下特点: 债券基金, 以上特点分
新华资管90天滚动持有债	013537	客户具有如下特点: 稳健型C2; 产品具有如下特点: 债券基金, 以上特点分

客户信息 **产品信息** **服务记录**

近3月收益 | 近1年最大回撤 | 累计收益 | 开始时间

近一个月表现

收益率	最大回撤	年化波动	胜率

暂无数据

聊天辅助

产品推荐

产品画像



公募基金
理财产品
投顾服务
数据资讯
...

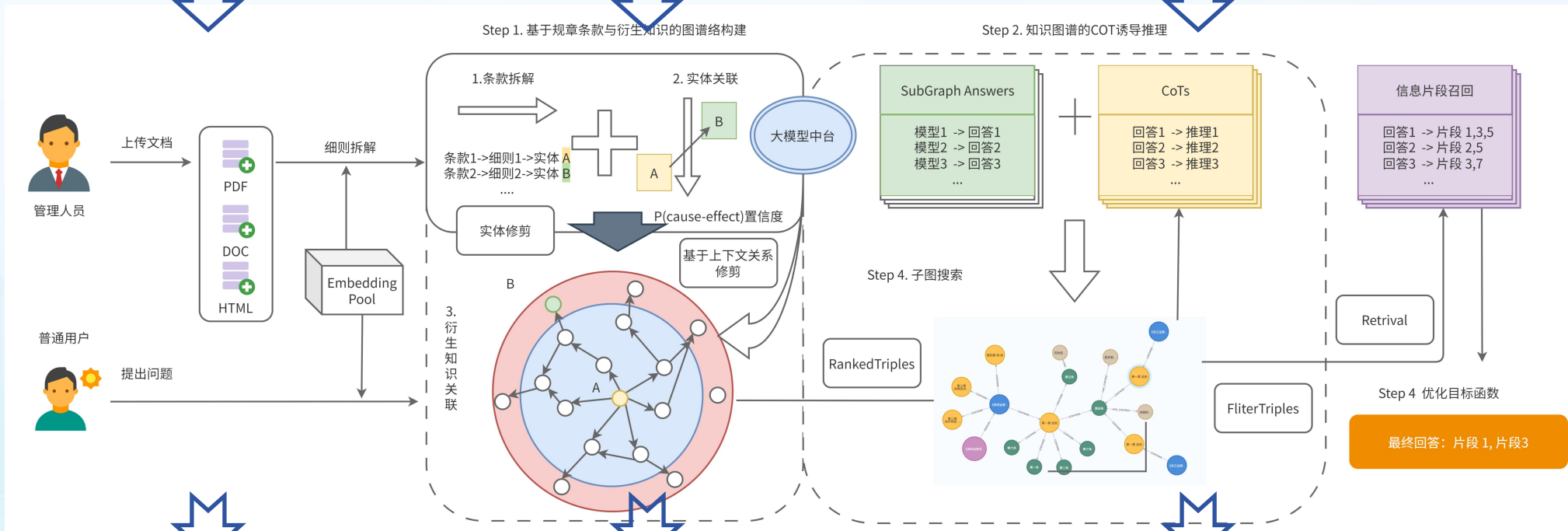
03 “灵曦”赋能知识检索

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？

大模型知识管理底座



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

行业创新业务建设管理系统

业务建设管理
审批管理
计划管理
任务管理
跟踪管理
交付文档管理
智能制度库

上海证券交易所交易规则（2020年第二次修订）

【发文字号】上证发〔2020〕17号 【发布日期】2020-03-13 【施行日期】2020-03-13
【时效性】现行有效 【效力级别】自律规则 【发文单位】上海证券交易所

制度变迁 对比

- 2023-02-17 上海证券交易所交易规则（2023年修订）
- 2020-03-13 上海证券交易所交易规则（2020年第二次修订）
- 2020-01-07 上海证券交易所交易规则（2020年修订）

正文 下载

关于修订《上海证券交易所交易规则》的通知
上证发〔2020〕17号

各市场参与者：
为贯彻落实《证券法》关于证券交易制度的相关规定和要求，经中国证监会批准，上海证券交易所（以下简称本所）对《上海证券交易所交易规则》（以下简称《交易规则》）进行了修订，现予发布，并就有关事项通知如下：
一、本次修订条款包括《交易规则》第2.3条、第5.1.1条、第7.1条、第7.4条、第7.6条、第7.8条等，主要涉及增加存托凭证作为交易品种、完善交易异常情况和重大异常波动处置规定、明确证券即时行情实时发布等内容。
二、本次修订后的《交易规则》全文重新发布（详见附件），并自2020年3月13日起实施，本所于2018年8月6日发布的《关于修订〈上海证券交易所交易规则〉的通知》（上证发〔2018〕59号）和于2020年1月7日发布的《关于修改〈上海证券交易所交易规则〉第3.1.5条的通知》（上证发〔2020〕1号）同时废止。
三、考虑到市场及会员业务技术准备情况，《交易规则》中此前暂未实施的部分内容继续暂缓实施，具体实施时间由本所另行通知。
特此通知。
上海证券交易所
二〇二〇年三月十三日

附件：《上海证券交易所交易规则（2020年第二次修订）》暂缓实施条文

上海证券交易所交易规则（2022年第二次修订）
（2006年7月1日实施，2007年4月24日根据《关于调整无价格涨跌幅限制股票申报价格范围的通知》第一次修订，2013年12月14日根据《关于修订〈上海证券交易所交易规则〉若干条款的通知》第二次修订，2013年10月18日根据《关于修订〈上海证券交易所交易规则〉及相关事项的通

关联制度

本规则引用	引用本规则	本篇废止	废止本篇
(3)	(1)	(0)	(0)
91 中华人民共和国证券法 2023-02-17 发布			
91 上海证券交易规则章程 2023-02-17 发布			
91 证券交易所管理办法 2023-02-17 发布			XXXX号

制度图谱 >>

制度图谱

制度高级搜索 × 制度快速 ×

长江证券股份有限公司债券质押式协议回购交易业务标准化细则

现行有效

查看原文 >

制度类别：内部制度
发文单位：运营管理中心
效力级别：公司具体实施细则
发布日期：202301
施行日期：202301
制度标签：运营管理中心

制度图谱 >>

制度图谱

内部制度 外部制度 配套制度

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。



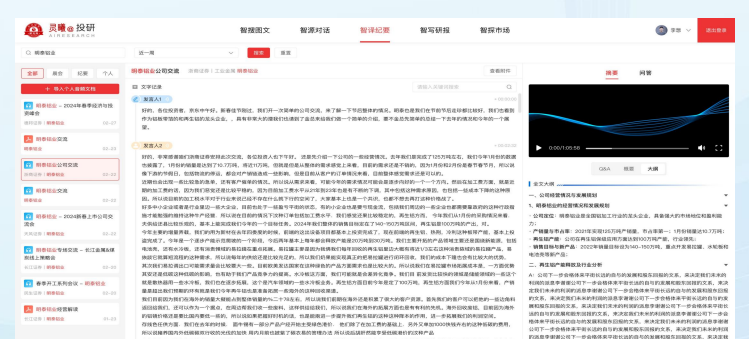
写

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

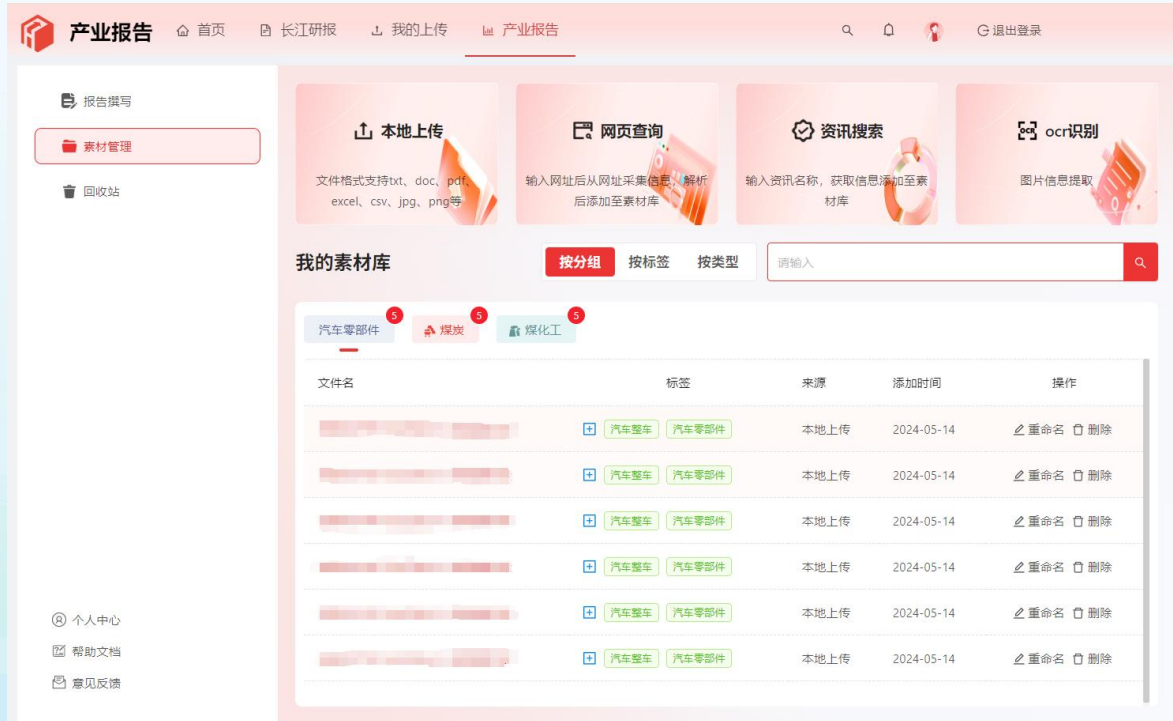
电话会议、路演、调研信息快速转写快速形成会议纪要重点内容和思维导图



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

1

大模型应用开发背景

2

“长江灵曦”大模型平台

3

“长江灵曦”应用场景

4

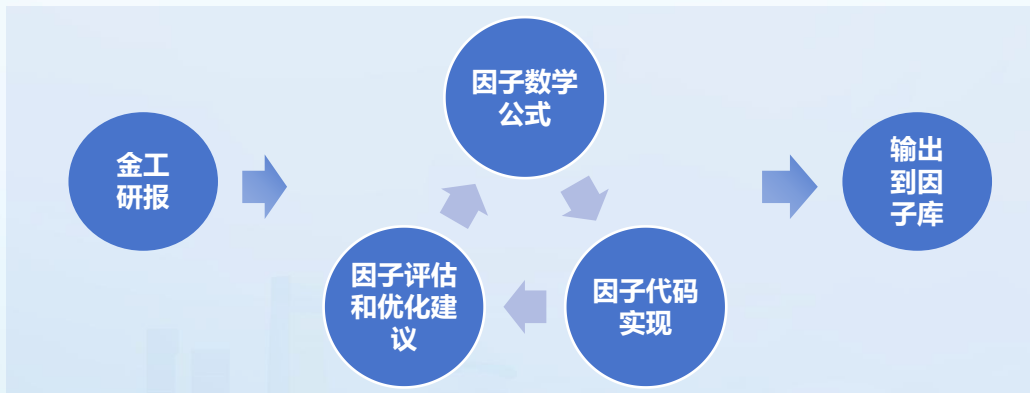
未来展望

04 未来展望



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