

2025年6月

2025金融大模型应用与智能体建设案例集



主办单位： **金科创新社**
Fintech Innovation in China

前言

近年来，人工智能技术的迅猛发展，尤其是大语言模型（LLM）的突破性进展，正在重塑金融行业的服务生态与运营范式。从智能客服的精准响应到风险管理的实时预警，从投研分析的深度洞察到合规审计的高效执行，大模型技术已成为金融机构提升智能化水平、优化客户体验、强化风险防控的核心驱动力。

然而，大模型在金融领域的落地并非坦途。数据安全、合规监管、模型可解释性、业务适配性等挑战，使得许多机构在技术应用过程中面临“知易行难”的困境。如何借鉴行业标杆实践？如何选择适合自身业务场景的技术路径？这些问题亟需一套系统化的参考指南。

为此，金科创新社基于近两年“鑫智奖”评选积累的丰富案例，精选 50 余个来自银行、保险、证券、信托等领域的标杆实践，汇编成《2025 金融大模型应用与智能体建设案例集》。

本案例集涵盖六大核心场景：

一是智能客服与营销：展示大模型在客户服务、电话销售、数字营销等场景中的创新实践，如虚拟数字人、智能作业平台等。

二是智能风控与合规管理：聚焦大模型与知识图谱等技术融合，提升合规审查、信贷风控、司法查控等能力。

三是知识管理与智能问答：呈现金融机构如何构建智能化知识库、问答系统及创作平台，赋能内部协同与客户服务。

四是运维、安全与测试智能化：探索大模型在智能运维、安全防护、自动化测试等领域的应用，保障系统稳定与安全。

五是投顾与业务管理：涵盖智能投顾、员工协同、普惠金融等场景，体现大模型对业务管理的深度赋能。

六是创新技术与平台建设：分享金融机构在大模型平台构建、移动端架构升级、AI 中台建设等方面的前沿实践。

本案例集旨在为金融机构提供从技术选型到场景落地的全景参考，助力行业在智能化浪潮中精准锚定方向，实现技术与业务的深度融合。

Contents

目录

一、智能客服与营销

广西北部湾银行：虚拟数字人系统	01
苏商银行：大模型客服助手	03
太平洋保险：基于 AiGCP 智能生成大模型平台的智能客服应用	07
中信建投证券：基于数字金融的全场景数智化客户综合服务平台	10
中国大地保险：数字营销助理“智能小行”项目	17
浙江东方：AI Agent 营销智能体项目	20
招商银行济南分行：“营小助”智能作业平台	25
上海农商银行：零售数智化创新实验室项目	29

二、智能风控与合规管理

宁夏银行：“宁银小智”大模型在信贷风险管理中的数智化应用建设项目	32
重庆银行：基于大模型与知识图谱技术融合的数智尽调平台	36
中邮保险：基于“智能生态 + 大模型”体系下的操作风险智能联防生态搭建	41
天津银行：“AI 合规官”数智化项目	45
潍坊银行：基于大模型和 RAG 驱动的智慧合规助手	47
青岛银行：“法眼智控”网络司法查控 AI 管理体系	51

三、知识管理与智能问答

哈尔滨银行：基于大模型金融机构数智化知识管理系统创新实践	53
中邮保险：基于传统 NLP 以及大模型“双模式”的知识循环体系	56
江苏农信：基于大语言模型的运维知识管理平台	58
杭州银行：基于金融垂直大模型的商业银行制度知识库检索平台	61

北京银行：“京信妙笔”智慧创作平台	64
太平洋寿险：银保销售复盘会萃取工作台	67
太平洋寿险：智能金牌教练	70
华福证券：企业智慧问答系统	72
江南农村商业银行：“小江智脑”大模型应用平台	76
四川农商联合银行：大模型驱动的智能问数应用	79
恒丰银行：恒运 AI 答	83
财信人寿：“吉小星”AI 助手——保险全链路智能增效引擎	91
国元证券：燎元智能助手	94
中国银行江苏省分行：知行平台项目	99

四、运维、安全与测试智能化

哈尔滨银行：商业银行智能运维体系重构——AI 技术驱动的数字化转型实践	103
国泰海通证券：基于 AI Agent 的金融云平台全场景运维决策机制研究	108
北银金科：北京银行大模型安全测评平台	111
湖州银行：数据库操作安全管理创新实践	118
青岛银行：基于安全大模型 + 零信任的新一代内网访问控制体系建设	123
青岛银行：基于 Dify 的钓鱼邮件分析助手	130
镇江农商银行：基于大模型的银行网点的物理攻击行为预警探索与应用	133
中国邮政储蓄银行：AIGC 重构质量基石——智能测试自主进化实践	135
中信建投证券：AI 大模型驱动的自动化测试平台建设与应用	139
晋商银行：基于 AI 与业务组件扫描的智能化质量防护体系构建	142

五、投顾与业务管理

中信建投证券：基于 AI 大模型的多智能体技术在投顾领域的应用案例	149
中信建投证券：基于财富管理业务的员工协能全链平台	154
中信建投证券：长尾客户普惠数智金融服务平台	160



中泰证券：基于大模型的证券机构业务 AI 赋能平台	165
中原银行：大模型信贷助手	169
中国大地保险：数字审计员项目	171

六、创新技术与平台建设

山能财务有限公司：基于 DeepSeek 构建财务公司智慧金融大模型	173
平安信托：DeepTrust 大模型平台及应用项目	176
北银金科：北京银行水晶球 ChatBI	179
国元证券：面向 AI 大模型时代的新一代移动端架构——组件化与端智能深度融合实践	184
西南证券：证券公司大语言模型中台建设实践	188
人保寿险：“AI 保宝”大模型平台赋能公司数字化转型	191
大家保险集团：大家灵图影像平台	195
大家保险集团：大家保险人事办公智能化项目	198
中国大地保险：AI 中台建设项目	200

七、金融大模型解决方案选登

基于大模型构建银行统一知识库管理体系	203
领雁科技“智鉴”反洗钱智能体	208
AI 大模型在金融业务场景的应用实践	212
跬智信息 Data Agent：首个企业级自主深度推理和洞察产品在金融业的落地实践	214
云杉 DeepFlow 金融核心业务系统可观测性及运维智能体建设方案	218

广西北部湾银行：虚拟数字人系统

一、项目背景及目标

随着金融科技的持续发展及人们线上交互习惯的加深，金融服务持续向着综合化、场景化、智能化、人性化、全渠道方向转型。机器学习、大数据、人工智能等一系列新技术与业务需求深度融合，数字员工作为拟人化的新型工作人员应运而生，与自然人员工相比，具备高效率、强能力、高可控、低成本等优势，可以发挥减员增效的作用，形成数字化生产力。利用数字人技术，打造手机银行数字业务专员，助力手机银行各项业务的办理及产品介绍，相较文字客服，数字人语音交互可以提高业务和产品跳转率，数字人语音介绍产品也是适老化的重要举措，降低一线服务人员的业务压力，由数字人替代部分人工，降低成本，提高效率。

在此背景下，我行启动智能数字人交互平台项目，采购技术成熟、性能优良的智能数字人产品，提供数字人形象定制、数字人视频生成、数字人交互能力、真人接管能力、数字人后台管理等服务，满足我行数字人应用场景需求，进一步提升我行智能化服务能力。

二、创新点

1. 多语言交互能力的创新应用

广西北部湾银行虚拟数字人系统的多语言交互能力是其金融服务中的一大创新亮点。传统的跨境金融服务在语言沟通方面存在较大障碍，往往需要配备大量的多语言专业客服人员，成本高且效率低。而虚拟数字人系统通过集成的多语言处理技术，支持中文和英语等外语之间的实时交互。这种多语言交互能力极大地拓展了银行的服务范围，使国外客户能够更加便捷地获取金融服务，有效提升了跨境金融服务的可及性。通过消除语言隔阂，增强了国外客户对银行服务的认同感和信任度，为进一步深化中国与外国的金融合作奠定了良好基础。

2. 打破时空限制的全天候服务模式

虚拟数字人系统打破了传统金融服务在时间和空间上的限制，为客户提供 7×24 小时不间断的金融服务。无论客户身处世界的哪个角落，无论何时都有金融服务需求，都可以随时通过银行的线上渠道与虚拟数字人进行交互。例如客户于当地时间凌晨 3 点突然遇到对公转账的紧急问题，需要马上将一笔资金汇往合作伙伴账户。该客户可以通过北部湾银行手机银行 APP 咨询虚拟数字人，虚拟数字人迅速响应，详细解答了客户关于对公转账流程中的疑问，包括转账所需的手续、可能产生的费用、预计到账时间等，并指导客户在手机银行上完成了紧急汇款操作，避免了因时间延误可能给客户带来的损失。这种全天候的服务模式极大地提升了银行的服务效率和客户满意度，增强了银行在跨境金融市场的竞争力，让客户感受到随时随地、高效便捷的金融服务。

3. 降低运营成本与提升服务效率的双重优化

虚拟数字人系统的应用显著降低了银行的运营成本，同时提高了金融服务效率。在传统的金融服务模式中，需要大量的人工客服来处理客户咨询和业务办理，人力成本高昂，且容易出现因人工疲劳导致的服务质量下降问题。而虚拟数字人系统通过自动化和智能化的方式处理客户问题，能够同时应对大量客户的咨询请求。



在金融业务咨询高峰期，虚拟数字人能够在短时间内为多名客户提供准确的解答，而传统人工客服团队则难以在相同时间内处理如此大量的咨询。在业务办理辅助方面，虚拟数字人简化了业务流程，减少了人工操作环节。通过降低运营成本和提高服务效率，广西北部湾银行能够将更多的资源投入到产品创新和业务拓展中，进一步提升银行的核心竞争力。

ID 三、项目技术方案

在形象设计上，虚拟数字人整体造型简洁大方，身着银行职业装，展现出专业、可靠的形象。面部表情丰富自然，眼神灵动，能够根据与客户的交流情况做出微笑、思考、解答等不同表情，增强与客户的情感共鸣。

从技术架构来看，该系统运用自然语言处理（NLP）、自动语音识别（ASR）、语音合成（TTS）、大语言模型（LLM）等技术实现对客户问题的理解和分析，使虚拟数字人能够准确把握客户意图，并生成合适的回答。通过机器学习算法，虚拟数字人不断学习和优化回答策略，提高回答的准确性和质量。计算机图形学与动画技术用于构建虚拟数字人的形象和动作，通过三维建模、材质渲染等技术打造逼真的虚拟形象，使其外貌、表情和动作更加自然流畅。鉴于面向东盟的金融服务需求，系统集成多语言处理技术，支持中文、英语、越南语等多种东盟国家常用语言，通过机器翻译技术实现不同语言之间的实时转换。同时，构建金融知识图谱，将各类金融知识进行结构化整合，虚拟数字人通过知识图谱快速检索和关联相关信息，为客户提供全面、准确的金融解决方案。我行虚拟数字人系统测试环境已接入 deepseek 大模型，通过海量文本数据的深度学习，进一步加强理解、生成自然语言和推理的能力。

ID 四、运营情况

系统支持在银行官方网站、手机银行 APP、微信公众号等多个渠道接入，客户可以随时随地通过不同终端与虚拟数字人进行交互。目前支持的业务场景有开户行查询、卡片状态查询、借记卡余额查询、网点行号查询等，实现了客户咨询的智能应答，累计服务客户 12.21 万人次（25 年 1 月至 5 月底），服务量占比 39.59%。

ID 五、项目成效

经济效益：一是在业务快速发展、产品不断拓展的过程中对人力的替代、释放有明显作用，在确保客户体验不降低的同时，能够有效节约运营成本和人力成本，形成规模效益。二是加速人员转型，随着智能服务的广泛应用，将远程客服人工服务人员从重复、标准的工作中释放出来，转型从事产品推荐、复杂疑难业务处理、客户营销等高价值创效工作，进一步提升人力投入产出效益。

社会效益：通过应用智能服务，提升我行数字化应用能力，有助于提高我行在金融服务便捷性、增强客户金融服务获得感的社会影响力，进一步提高我行的企业形象。

ID 六、经验总结

广西北部湾银行虚拟数字人系统在技术集成与应用层面展现出了卓越的引领性。在广西地方银行范畴内，北部湾银行率先将多种前沿技术进行深度融合，构建起虚拟数字人这一创新服务载体。其在人工智能技术运用上颇为领先，借助自然语言处理（NLP）和大语言模型技术相结合，使虚拟数字人能够精准理解客户复杂多样的问题，无论是常规金融业务咨询，还是涉及跨境金融政策解读、东盟国家金融市场动态分析等专业性较强的内容，都能准确把握客户意图。同时，机器学习算法的运用让虚拟数字人具备自我学习和优化能力，随着与客户交互数据的不断积累，回答的准确性和质量持续提升，这一技术应用程度在同类金融机构中处于前列。

苏商银行：大模型客服助手

一、项目背景及目标

苏商银行成功应用大模型技术以提升远程银行客服中心的运营效率和服务质量。通过三大创新应用——客服知识库助手、话术推荐助手及质检助手，项目实现了机器人自助解决率的显著提高、在线客服并行会话能力的增强及服务质检的全面化和精准化。从而显著减少了客服团队的人力成本，提升了客户满意度和服务效率。

客服团队面临低效率和服务质量不一的问题，且传统客服模式已难以满足现代客户的需求。项目目标是打造“大模型客服助手”赋能远程银行客户服务业务。按照通话前、通话中、通话后的端到端业务流程闭环，形成事前知识库运营提效、事中副驾驶赋能、事后质检守护，助力客服工作模式和效率的全面升级。

二、创新点

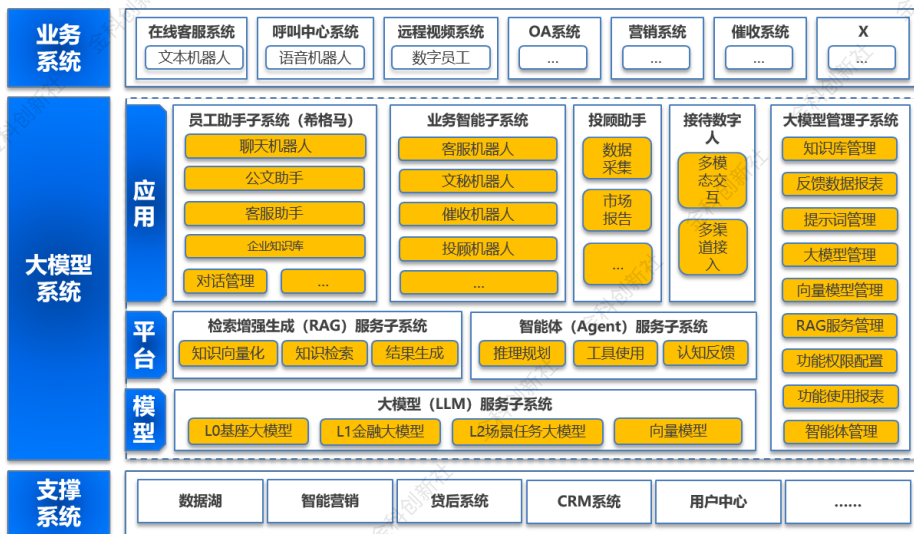
1. 客服知识库助手：传统的知识库在更新和管理上需要大量的人工投入。客服知识库助手利用大模型的自然语言理解、抽取和生成能力，实现了知识库的自动化和智能化管理，大幅度降低了人力成本和时间成本。

2. 话术推荐助手：传统的客服系统往往需要客服人员根据经验选择合适的话术。话术推荐助手通过实时理解对话情境和向量化的知识库，利用检索增强生成技术（RAG），实时推荐合适的话术，大幅提升了客服的效率和服务质量。话术推荐助手还可以根据客户的历史交互和偏好，推荐个性化的话术，增强了客户服务的个性化程度，提升了客户满意度。

3. 质检助手：传统的客服质检主要依赖人工抽查，工作量大且效率低。质检助手通过大模型提示词技术，实现了质检的自动化，大幅提高了质检的效率和准确性。传统的质检往往只关注客服的专业性，而忽视了情感交流的重要性。质检助手通过模型的情感分析能力，可以识别出客户和客服的情绪，提供更全面的质量评价。

三、项目技术方案

整体系统架构如下图所示：





大模型系统分为应用、平台和模型 3 层，主要系统的定位和功能介绍如下：

1. 员工助手子系统

定位：该子系统旨在通过对话式的交互界面和多种工作场景插件，提升员工的日常工作效率。

功能：聊天机器人、知识库问答、客服标准问生成、客服相似问生成、电销话术提炼、电催话术提炼等。

2. 业务智能子系统

定位：该子系统旨在为业务系统提供联机的智能化服务，从而提升业务系统的智能化水平。

功能：话术推荐、质检助手、电话销售对话机器人、电话催收对话机器人等。

3. 检索增强生成（RAG）服务子系统

定位：该子系统结合了向量检索和内容生成的能力，旨在为员工助手子系统和业务智能子系统提供配置化平台化的 RAG 能力。

功能：知识向量化、知识检索、结果生成等。

4. 大模型服务子系统

定位：提供基于大模型的多模态处理能力，以实现智能化的语言处理、数据分析、图像处理、视频处理。

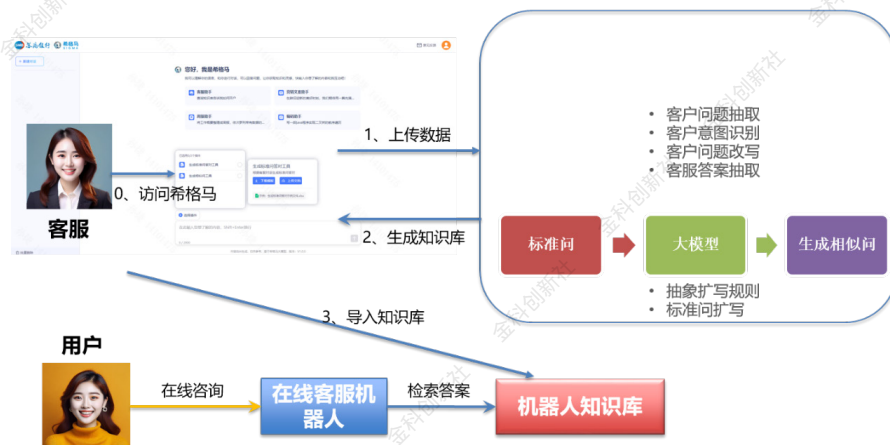
功能：大模型调用、向量模型调用。

5. 大模型管理子系统

定位：为各服务子系统提供管理和配置后台。

功能：知识库管理、反馈数据报表、提示词管理、大模型管理、向量模型管理、RAG 服务管理等。

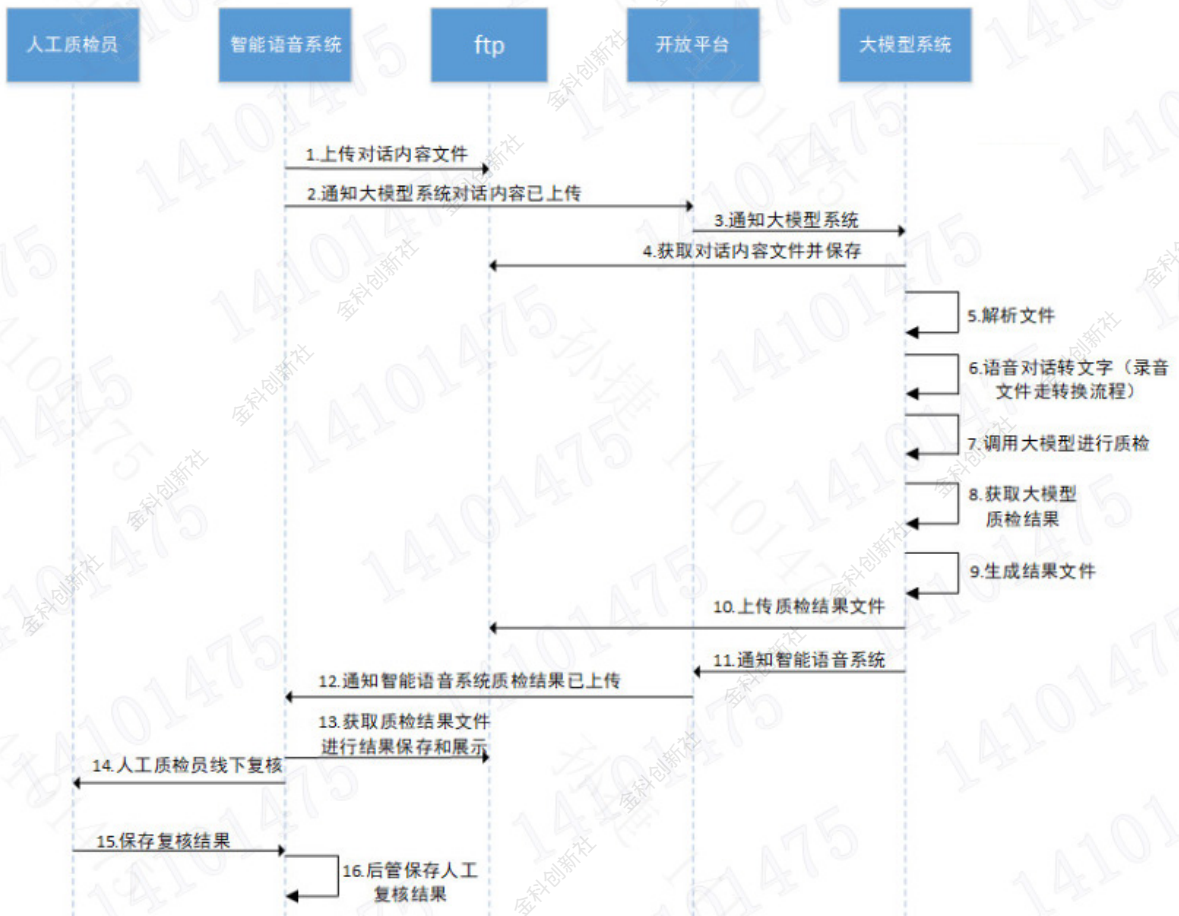
客服知识库助手业务流程如下图所示：



话术推荐助手业务流程如下图所示：



质检助手业务流程如下图所示：





四、运营情况

目前，客服知识库助手、话术推荐助手以及质检助手已全面投入使用。客服知识库助手平均每周使用 2 次，每次可生成 100 条左右标准问和 2000 条相似问。话术推荐助手每天 1000 次左右调用。质检助手已 100% 覆盖客服和电销对话质检。

五、项目成效

项目实施后，知识库助手使得机器人自助解决率从 50% 提升至 75%，话术推荐助手使并行会话数从 6 通提高至 8 通，质检准确率提升至 70%。经济效益显著，节约了大量人力成本，社会效益方面提升了客户服务体验和满意度。

六、经验总结

1. 数据与大模型技术结合：项目的成功实施依赖于大量历史服务数据的积累与整理，以及大模型技术的应用，使得知识库的覆盖和更新、话术的推荐以及服务质量的检查更加高效和精准。

2. 用户体验优先：无论是内部客服团队还是最终的客户体验，本项目都将用户体验作为核心。通过持续的用户反馈收集和优化，确保了技术应用的实用性和便捷性，进一步提高了用户的接受度和满意度。

3. 面临的挑战与对策：在话术推荐的实用性方面，初期遇到的挑战是坐席使用意愿不高。通过增强话术同步率、优化用户界面和提供定制化培训，有效提升了坐席的使用热情和话术的实用性。

本项目显示了大模型技术在提升客服中心服务效率和质量方面的巨大潜力。未来，随着技术的进一步发展和应用，期待在客户服务体验提升上实现更多创新和突破。

太平洋保险：基于 AiGCP 智能生成大模型平台的智能客服应用

一、项目背景及目标

基于 AiGCP 智能生成大模型平台的智能客服应用是基于中科汇联自主研发的 AiGCP 智能生成大模型平台基于大数据预训练、多源知识融合、持续学习等技术。在语言理解方面，这些模型可以通过预训练语言模型，自动识别语言中的词汇、语义、句法、情感等信息，并能够进行分类、命名实体识别、语义解析等任务。在此基础上搭建 AICC 统一智能交互平台，采用先进的 FreeSwitch 电话软交换解决方案，灵活的组件化模式，可以提供高度扩展、弹性、按需取用的软件服务。

随着互联网行业发展的飞速进展和社会环境的不断变化，传统智能客服在建设或者运营过程中也面临了新的机遇和挑战。当前移动互联网方兴未艾，金融科技发展迅速，互联网金融市场竞争激烈，同时消费升级趋势明显。金融行业主要包括银行、保险、证券等。金融业务通常存在客户排队时间长。业务营业时间有限、不同客户的问题重复率高、业务人员成本高等痛点。

客户呼入服务需求无一定时间规则，管理人员很难精准排班每日接线人力，时而出现排班人力冗余或无人服务的情况，线路一旦繁忙时，客户等待时间过长，且客户每次拨打呼入电话时都需要经历一系列的电话菜单选择，客户流程体验较差。

拟通过上线智能客服机器人实时识别客户需求，削减客户在线等待时间，主动实现在线报价或根据客户需求推送优质便利的自助服务，实现一步直达功能菜单。以智能客服和自助服务替代人工服务完成客户服务，提高客户体验的同时降低呼入接线人力成本。

二、项目 / 策略方案

基于 AiGCP 智能生成大模型平台的智能客服应用呼叫中心系统是一个功能强大的用于构建呼叫中心系统的产品解决方案，系统支持大容量的电话呼入 / 呼出处理、电话转接、来电弹屏、智能 IVR、智能 ACD、通话记录、呼叫坐席管理、工单管理、数据报表和绩效统计等专业呼叫中心系统功能。支持部署版、云 SaaS 版呼叫中心，通话稳定、音质清晰，多级 IVR 语音、VIP 专线等功能优势，领略一呼百应的呼叫体验。

添加工作人员及组织架构：管理员对该企业使用 aicc 的成员进行管理，账号创建支持单个添加及批量创建。管理员对企业内部员工角色及对应功能权限进行配置。系统默认了客服中心常用角色及权限，客户可根据自己实际业务定义角色权限，当一个用户出现在多个角色里时，角色权限进行累加。

知识库：知识库，对系统内部普遍问题或异样问题进行统一管理，支持音频、文字、图片、文件、等多样化知识类型，知识库分为系统知识库及自定义知识库，系统知识由平台进行导入及管理，自定义问答库则由客户自行管理，自定义知识库有自定义添加及知识学习两种途径建立。自定义知识库可以根据业务需求选择相应的系统知识进行添加。

知识学习：知识学习模块通过对未知问题、相似问题的管理，从而达到促进问答库知识优化的目的。

机器人学习：在日常工作中机器人无法回答的未知问题，已回答被踩的问题以及坐席工作中登记的问题会被统一收集到机器人学习页面，在此页面可以根据实际业务对上述几类问题做新增知识、关联相似问及加入相似问法库或忽略的操作。

机器人测试：机器人测试功能可以通过创建测试任务对该机器人所关联的自定义问答库知识进行批量检测，根据最终的检测结果对已维护知识的准确率进行查看及维护。



相似问法库：相似问法库可以对所有提问进行维护。可以直接在相似问法库添加，也可以在机器人学习、机器人测试进行维护，已被维护的相似问法可以用于机器人测试或者新增知识时直接对相似问进行勾选。

知识同步管理：支持使用插件的方式，同步获取第三方知识。依赖于插件配置，具体项目需与研发沟通讨论并定制插件同时支持插件获取知识是否需要审核入库。点击自动处理配置，根据知识同步类型自主选择。选择自动处理，处理的知识若能自动处理成功，将自动流转到已处理。自动处理失败将归纳到待处理，需管理人员手动处理。

意图库：用于对话过程中识别意图。基于识别意图给出意图标签继续引导，最终引导用户进入 ai 应用或给出推荐知识供用户选择。

AI 应用：ai 应用将具备原有多轮对话能力。同时拥有实体提取、变量应用、条件判断，以及接口调用等新增能力。

ID 三、创新点

该智能客服解决方案是基于中科汇联自主研发的 AiGCP 智能生成大模型平台基于大数据预训练、多源知识融合、持续学习等技术，现在很多 NLP 模型已经具备了超强的语言理解和生成能力。在语言理解方面，这些模型可以通过预训练语言模型，自动识别语言中的词汇、语义、句法、情感等信息，并能够进行分类、命名实体识别、语义解析等任务。在对话生成和文学创作方面，这些模型可以基于预先训练的语言模型，生成与人类类似的自然语言文本，帮助用户完成对话生成、文学创作等任务。

通过不断地吸收海量文本数据中的新知识和信息，AiGCP 智能生成大模型平台模型的效果也在不断地进化和提升。AiGCP 智能生成大模型平台技术未来可广泛应用于机器翻译、智能客服、智能写作、情感分析、舆情监测等领域，为人们的生活和工作带来了很大的便利和效益。

在智能问答的基础上搭建 AICC 统一智能交互平台，采用先进的 FreeSwitch 电话软交换解决方案，灵活的组件化模式，可以提供高度扩展、弹性、按需取用的软件服务。

平台支持电话、网页、H5、微信、企业微信、APP、智能设备等多渠道接入方式，支持公有云、私有云、混合云、本地化分布式灵活部署、云呼叫中心的资源建设方式，可以实现高效资源整合，降低部署成本，提升部署效率。

平台支持主流国产操作系统、数据库等软件环境，可以为金融行业各机构快速搭建智能呼叫中心系统、智能在线客服系统、智能工单系统、智能质检系统等全面解决方案。

ID 四、运营情况

1. 实现保单智能查询业务，投保人输入个人相关信息自助查询保单内容；
2. 实现保险报价智能查询业务，客户输入个人相关信息，查询报价相关信息；
3. 实现智能查询业务用户操作流程查询及数据分析功能。

ID 五、项目成效

智能客服应用基于 AiGCP 智能生成大模型平台在保险行业的应用将带来显著的社会效益。首先，它能够提高保险服务的效率，实现全天候在线服务，从而缩短理赔处理时间，帮助被保险人尽快解决问题。其次，智能客服应用能够提高服务质量，通过准确解答客户疑问和提供个性化的保险建议，增强客户对保险公司的信任感，促进行业的健康发展。最重要的是，智能客服应用可以普及保险知识，使得更多的人了解保险产品，增强社会对风险保障的认识，提高整个社会的风险防范意识。

引入智能客服应用基于 AiGCP 智能生成大模型平台将为保险行业带来巨大的经济效益。首先，它能够降低保险公司的运营成本，节省人力资源和时间成本，提高企业的效益。其次，智能客服应用能够提高销售转化率，通过快

速响应客户需求和提供个性化的保险方案，促进保险产品的销售，增加公司的收入。最重要的是，智能客服应用提升了保险公司的竞争力，树立了良好的品牌形象，吸引更多客户，推动整个保险行业的发展和进步。

ID 六、经验总结

基于 AiGCP 智能生成大模型平台的智能客服应用为金融行业各机构快速搭建智能呼叫中心系统、智能在线客服系统、智能工单系统、智能质检系统等全面解决方案为保险行业带来巨大社会和经济效益，通过准确解答客户疑问和提供个性化的保险建议，增强客户对保险公司的信任感，促进行业的健康发展。



中信建投证券：基于数字金融的全场景数智化客户综合服务平台

一、项目背景及目标

21 世纪以来，中国经济实现了持续的高速增长，国民财富的积累以及对资产增值的追求不断增强，进而催生了对金融机构提供高质量财富管理服务的迫切需求。据权威预测，至 2025 年，中国财富管理市场的总规模预计将突破 125 万亿元人民币，标志着其已成为金融行业未来发展的战略重点。财富管理业务在提升居民财产性收入及财富积累效率方面发挥着关键作用，并肩负着推动实现社会主义“共同富裕”目标的历史使命。

然而，近两年国内外经济形势严峻复杂，全球经济复苏进程缓慢且不稳定，国际贸易摩擦不断升级，地缘政治冲突时有发生，这些因素相互交织，给金融市场带来了巨大的不确定性。国内经济也面临着需求收缩、供给冲击、预期转弱等多重压力，经济增长速度有所放缓。在这样的宏观经济环境下，金融市场波动剧烈，证券行业首当其冲。例如 2024 年 9 月 24 日行情来袭，证券市场出现了大幅波动，证券行业受到强烈冲击，客户的投资组合价值在短时间内发生了显著变化。在这种情况下，客户的服务难度大幅增加，他们对财富管理的需求也愈发多样，不仅要求金融机构提供及时准确的市场信息和专业的投资建议，还希望得到个性化的资产配置方案和风险控制策略，以应对市场的不确定性，保障资产的安全和增值。

面对客户激增的复杂多样的需求，证券公司等金融机构普遍面临客户基数庞大与客服能力不足的挑战。具体表现在：客户服务高度依赖人工操作，客服人员的专业水平和经验参差不齐，导致服务质量难以保证，无法满足客户对高质量服务的期望；跨领域业务的综合服务能力不足，难以满足客户在不同金融产品和服务之间进行综合配置的需求；服务渠道缺乏统一规划，线上线下服务渠道分散，数据分散，缺乏集中化管理，导致客户信息不完整、不准确，影响了服务的效率和精准度。在这种情况下，仅依靠传统的人工服务模式已经无法满足市场需求，亟需依靠智能化手段来减轻人工服务压力，提升服务效率和质量，最大化提高客户转化率，增强客户粘性，提升金融机构在财富管理领域的竞争力。

随着数字化转型和人工智能技术的不断发展，中信建投证券积极探索科技赋能，以提升财富管理服务能力。团队致力于构建全场景数智化客户综合服务平台，赋能公司客服业务全链路数字化升级，为一线员工在服务客户全过程中提供强有力的技术支持。该平台通过以下三个层面实现服务全流程智能化服务：事前（服务前），利用沉淀知识帮助员工快速提升服务技能，提前触达客户了解客户意图；事中（服务中），快速响应客户需求，提升坐席服务效率；事后（服务后），跟踪分析服务全过程，为管理人员提供智能决策支持。而今，尤其是近两年生成式预训练大模型技术的快速演进，为客服业务的智能化升级带来了新的机遇和挑战。团队紧跟技术前沿，逐步探索全流程服务线全方位大模型升级，通过一系列措施满足业务部门对新兴智能化服务的需求，以此提升公司在财富管理领域的整体服务能力及市场竞争力。

二、创新点

全场景数智化客户综合服务平台整体系统框架如图 1 所示。

全场景数智化客户综合服务平台具备如下技术特点和创新优势：

1. “大模型 + 专业小模型” 打造智能客服 AI 底座



图 1 平台架构图

该平台以大模型作为核心决策和任务调度模块，集成语音识别、意图提取、知识库、智能质检等多种智能化技术，构建了综合性智能化赋能平台。利用智能化手段，促进客服业务全流程的数字化改造，叠加传统大数据技术，为客服业务提供强大的数据底座。同时，深度融合文本、语音、视频等多模态交互方式，实现全渠道客户关系触达和金融服务的事前、事中、事后的全流程赋能，形成大模型赋能的智能化金融服务生态。

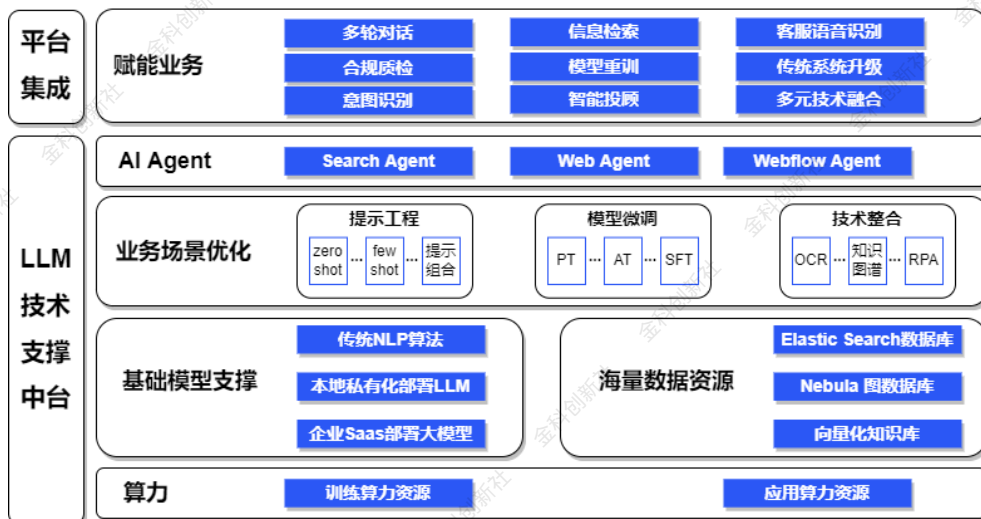


图 2 大模型系统架构图



在大模型系统架构方面，采用模块化、可扩展的整体架构设计，构建了自然语言解析、提示词设计与管理等基础功能模块。通过与传统 NLP 算法、向量知识库的结合，适应不同业务场景的需要。同时，平台通过设计封装 search 代理、web 代理等通用 agent 能力，实现了对于金融服务场景的优化适配。从而实现对于智能质检、智能投顾、多模态客服、智能检索等多个通用业务场景 Agent 构建的支持。此外，在大模型微调方面，重点完成了行业级微调语料库的构建，通过制定严格的数据标准，从多渠道收集海量数据，进行清洗和脱敏处理，确保数据高质量且安全合规。同时，对多种开源预训练模型进行评估，综合考虑其性能、适应性以及算力资源和成本投入。在技术层面，积累了全量微调以及 LoRA、提示调整、P-Tuning 等高效微调技术。目前，已针对智能客服场景构建了 10 万+ 数据集，完成通义千问、kimi、deepseek 等多种底座模型选择和部署落地，并成功将智能客服场景微调准确率提升至 90% 以上，取得了显著的成果。

2. 多种 AI 技术融合构建智能知识中台

知识图谱与自然语言处理技术构建知识中台。项目团队创新性地提出了一种结合知识图谱、大语言模型、Elastic Search 优化、向量知识库和结果重排过滤的综合搜索方案。首先，通过知识图谱和 LLM 对用户查询进行泛化，生成相关子问题，提升搜索覆盖面和精确度。其次，对 Elastic Search 进行优化，引入多字段模糊查询、过滤器机制、字段权重控制等，增强关键词匹配灵活性和准确性，提升搜索结果质量。接着，构建向量知识库，利用文本嵌入模型将口语化查询向量化，通过向量数据库实现高效存储和快速相似性检索。最后，对多源检索结果进行筛选、重排序，并结合用户上下文和反馈进行个性化调整，生成最终 Top-K 搜索结果，旨在为用户提供精准、高效且个性化的搜索体验。

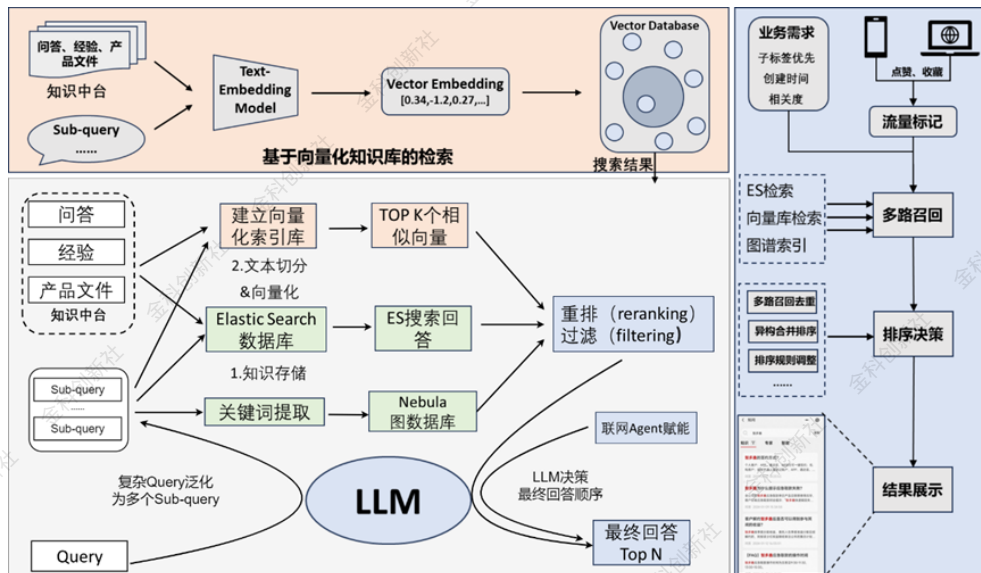


图 3 知识中台搜索总流程图

3. 基于大模型打造客服坐席智能数字化助手

落地并研发多类智能数字化助手代替员工工作。将大模型等多种 AI 技术与客户服务全流程深度融合，实现降本增效的目标。1) 智能外呼 / 电话客服系统代替人工呼叫 / 服务客户，涉及业务涵盖呼叫中心 3000+ 条常见问题以及

100+ 个复杂交互场景，并在必要时转人工进行服务；2) 智能坐席辅助全程参与员工服务，提供数字化、智能化的技术支持；3) 质检、分析系统对服务过程进行分析，帮助总结归纳经验。基于大模型自动生成服务会话小结和智能工单，并自动提取更新客户动态画像，为后续服务提供支持，全面提升员工的服务专业化水平和客户体验。4) 搭建企微群大模型机器人实现自助问答。以企微信群聊为信息通路，搭建 AI 大模型和日常业务场景的桥梁，拓展大模型在各类业务场景中的应用赋能。通过多格式文本解析、文档自动切片、大模型问句泛化、利用 Elastic Search 召回检索、索引知识重排、RAG+ 大模型问答整合、大模型意图分类、实体抽取、SQL 补全等技术，实现 IT 电商平台群智能助手、开户见证智能助手、两融券源问答助手、运营管理问答助手、两融直通车问答助手以及上市公司公告解析提取推送机器人等，赋能证金、财富委、运营管理部等多个业务部门。

4. 数智化技术全面提升客服业务运营能力

实现由“机器替代人工完成简单重复的工作”到“辅助人工完成复杂的任务”的跨越。1) 实现 7x24 小时服务，提高了客服响应速度，公司客服能力得到极大地增强，面对极端行情变化情况下的应急能力更强。2) 智能陪练、知识库、KYC 等系统使员工技能得到快速提升，新人工到岗时间大大缩短。同时培训过程中对员工的能力有了全面掌握，便于管理人员对客服业务资源的合理配置。3) 大量智能化技术应用，实现了整个客服业务的全面数字化。由初始散装的数据埋点、人工小范围检索，到集成全链路数字化看板，管理员统筹规划。客户数据需求发生变化的第一时间，随之实现数据检测、服务场景和方案的个性化匹配，运营人员能够以数字化视角来对客服过程进行监控分析。利用大模型对数据进一步挖掘，方便管理人员掌握客户的诉求、员工服务的质量等信息，进而发掘客户的潜在需求、优秀员工优秀服务范式等。

三、项目技术方案

为响应公司“致力提升客户服务能力，更好服务国家战略与实体经济，实现自身与证券行业共同的高质量发展”战略导向，项目团队深入分析客户服务全生命周期，通过业务与技术深度融合，创新构建全场景数智化客户综合服务平台。建立“AI+ 人工”协同矩阵，实现服务流程再造，针对标准化业务，部署智能客服集群，实现 7×24 小时即时响应，大幅提升服务产能。针对高净值客户与复杂业务，构建事前智“慧”，事中智“助”和事后智“学”的全流程数字化赋能体系。

1. 事前阶段：智慧赋能，精准洞察客户需求

在客户服务的事前准备阶段，通过 AI 技术与生成式预训练大模型的深度融合，为客服业务提供了强大的支持。一方面，利用大模型整合公司海量的客户信息，对客户的历史交易记录、偏好数据以及咨询记录等多维度信息进行深度挖掘与分析，生成更为精准且全面的客户画像。辅助客服人员提前洞察客户需求，进而制定出更具针对性的个性化服务方案。另一方面，智能外呼系统在大模型的改造下，实现质的飞跃。利用大模型从优秀的人工外呼服务记录中抽取优质话术与沟通技巧，提炼外呼模板。实际通话过程中，系统灵活运用生成话术，并实时分析客户反馈，动态调整外呼策略，从而显著提升客户体验与沟通效果，提高外呼的成功率与转化率，极大地节省人工时间，提升服务效率。

2. 事中阶段：智能辅助，提升服务质量和效率

在客户服务的过程中，多种人工智能技术的应用极大地提升了服务质量和效率。智能客服系统借助生成式预训



练大模型，抽取最新业务文档、行业公告等数据，实现常见问题解答（FAQ）的自动化扩充。模型通过深度学习海量业务知识，自动识别客户问题的核心意图，并精准匹配FAQ库中的回答，提升智能客服问题解决率和拦截率。同时，利用大模型根据客户问题的热度和复杂程度，自动调整回答的优先级和详细程度，确保用户获得最优质的服务体验。坐席辅助系统则利用大模型对服务流程进行预判，根据配置模板进行意图匹配，辅助坐席实现多功能服务。当客户提出问题时，大模型能够快速分析问题的性质和类别，并预测客户可能的后续需求，为坐席提供精准的建议和解决方案。此外，基于大模型实时监测坐席的服务表现，根据服务流程和话术模板，对坐席的操作进行实时指导和提醒，帮助坐席更好地应对各种复杂情况，提升服务质量和客户满意度。

3. 事后阶段：智能学习，优化服务流程和策略

在客户服务的后期阶段，对全流程进行回归分析和优化是提升服务质量的关键环节。智能质检模块在大模型技术迭代下，全面分析通话内容、服务流程、话术规范等数据，精准识别服务质量问题，并提供详细的分析报告和改进建议，进一步推动客服业务的持续优化，增强客服系统应对突发情况和异常指标的快速响应能力。同时，利用大模型对服务记录进行深度分析，提取优秀话术、抽取热点问题，并对服务流程提出优化建议。这些数据不仅便于管理层进行数据统计和监控，还为员工提供了宝贵的学习资源，帮助他们不断提升服务技能和服务质量。

通过近两年基于大语言模型技术的赋能，中信建投证券全场景数智化客户综合服务平台在事前、事中、事后三个阶段的服务能力得到了全面升级，为客户提供了更加高效、智能、贴心的财富管理和综合服务，也为公司在激烈的市场竞争中赢得了更大的优势。

四、运营情况

1. 突破呼叫中心生产力增长困局

回望过去的五年，智能客服与智能外呼两大呼叫中心核心业务齐头并进，实现了跨越式发展。公司4008-888-108客服热线客户接待量从2019年的日均500通增长至2024年的日均2000通，2024年十月全面上涨行情中表现卓越，客户接待量单日曾到达1.5万通，显著提升了客服热线承载力，打破了传统呼叫中心人工生产力弹性差的困局。另一方面，公司95587客户回访量从2019年的日均8500通增长至2024年的日均23000通，单日最高外呼量高达5.5万通。智能呼叫系统已成为呼叫中心不可或缺的核心生产力，犹如人体心脏般，为整个呼叫中心的顺畅运转提供着持续而强劲的动力支持。

2. 大力推动呼叫中心智慧建设

前端智能呼叫系统的成功实施为后续发展奠定了坚实基础，进而推动了中后台智能化水平的全面提升。通过系统性总结与创新实践，现已形成一套契合我公司客户服务需求的智能化运营体系，智能质检、智能坐席辅助等关键模块相继落地。

质检是呼叫中心质量管理体系的核心组成部分，智能质检系统的正式上线，显著提升了质检工作的整体效能，实现了质检模式的转型升级。具体而言，该系统将传统按比例抽样的质检方式升级为全覆盖模式，使质检覆盖率从系统部署前的3%全面提升至100%。质检类别也从单一质检新开户回访，接入了私募回访、日常存量回访、咨询电话等服务渠道数据，构建了更为完整的服务质量监控体系。从执行效能来看，智能质检系统每周可全量质检6万通会话，在保证质检质量的前提下，有效替代了原有质检流程中95%的人工操作环节，显著提升了质检工作的整体

效率与专业水平。

在服务质量管控方面，除事后监督机制外，更为关键的是将风险防控贯穿于服务生产全流程。坐席辅助系统的实施为服务过程的精细化管理提供了坚实保障。该系统上线后，显著降低了坐席操作风险，优化整合了 13 项跨系统查询场景，部署辅助流程及智能 FAQ 提醒功能共计 1000 余项。统计数据显示，系统已累计为人工客服提供 45.5 万次辅助支持，为回访人员提供 391 万次辅助服务。通过系统应用，单通电话平均处理时长由 9.74 分钟优化至 8.51 分钟，客户满意度指标从 97.62% 提升至 98.81%，充分体现了系统在提升服务效率与质量方面的显著成效。

ID 五、项目成效

投入产出情况：

经费实际累计投入额（万元）	976W	投入来源	自有资金
AI 替代人工年均节省资金额（万元）	3799.5W/年（12W/年单价 * 316.625 人）		

经济价值：

该平台自上线以来，项目团队构建了众多智能助手作为“虚拟员工”，由 AI 替换专职人员工作，赋能多个业务场景进而实现降本增效。

智能外呼机器人平均每年可完成约 440W 通电话，按人工每人每工作日处理 160 件回访任务计算，约节省 103 个人天工作量，有效减少了对人工客服的依赖。

智能客服机器人日常服务 2000 通来电，按人工每人每工作日处理 60 件呼入来电计算，约节省 33 个人天工作量。大模型人机会话总结助手实现海量智能电话客服服务记录复核，平均每个工作日可节省 20 个人天工作量。

上市公司公告提取助手平均每天生产有效信息 50 条，较人工查询、分析和定位工作节约工时 5h/天。消息群推送智能助手实现定时股市日历推送、特定交易所公告推送、证券交易融资“公告”推送功能，替代人工时 7h/天。合计节省 1.5 个人天工作量。

坐席辅助模块大模型做人会话小结、工单总结流转、语音转文字、录音问题更便捷锁定与回听、信息整合查询等，全链路共节省 283h/天，约节省 35.375 个人天工作量。

智能质检模块实现质检量由人工抽检 3% 提升到机器质检 100%，替代 95% 的质检工作，合计人工替代工时 990h/天，约节省 123.75 个人天工作量。

社会价值：

1. 金融科技融合

平台建设过程中，集成前沿智能化工具及多模态交互技术，是金融服务场景与科技深度融合的典型实践案例，实现了技术与业务协同助力的良好发展模式，有效推动数智化金融的建设。智能客服技术在行业中属于较为领先水平，先后跟申万宏源、国信等券商进行技术交流，无论是应用广度还是深度，均为行业提供了较好的典型实践经验。



2. 数字金融普惠服务

平台的实施应用，深度融合数字技术与金融服务，推动了数字金融的创新发展，为行业高质量发展提供了有力支撑。该平台不仅提升了金融服务的效率和质量，还通过智能化手段扩大了金融服务的覆盖范围，降低了服务门槛，让更多客户尤其是小微企业和普通投资者能够享受到便捷、高效的金融服务，有力地促进了普惠金融发展，助力国家共同富裕目标的实现。特别是在“适老化”场景中，该平台通过提供个性化的智能服务，大幅优化了老年客户的金融服务体验，有效解决了传统金融服务对老年人使用的不友好问题，从而实现了更广泛的金融普惠，实现金融服务共享。

系统在建设过程中，注重相关知识产权成果的积累，累计在《交易期刊前沿》、《金融电子化》等期刊发表论文12篇，获得国家发明专利授权3项，软件著作权9项，并荣获证券业协会优秀课题、深交所研究课题三等奖等11项行业奖项。

六、经验总结

在数字化转型浪潮下，金融机构面临客户基数庞大与服务能力不足的双重挑战。中信建投证券以科技赋能为核心，构建全场景数智化客户综合服务平台，为财富管理注入新动力。目前平台已实现客服业务全流程智能化升级，显著提升服务效能。

随着财富管理市场持续增长，客户需求日益复杂多元，对服务的及时性、精准性提出更高要求。未来，中信建投证券将紧跟业务需求，深入探索智能化场景化服务模式，优化客户体验。进一步拓展大语言模型在平台中的应用，挖掘其在更多业务场景的潜力，充分发挥智能知识中台的赋能作用，降低运营成本。同时，融合全渠道、多模态服务数据，深度挖掘非结构化文本价值，为获客、营销、合规等多领域提供支持。以数据驱动业务优化，持续改进产品和服务，提高客户满意度，助力公司实现长期发展目标，推动金融行业的高质量发展。

中国大地保险：“数字营销助理”智能小行“项目

一、项目背景及目标

在数字化浪潮席卷全球的背景下，保险行业正经历从“传统服务模式”向“智能化生态”的深刻转型。作为这一变革的引领者，大地保险推出的智能小行，通过人工智能技术深度整合业务场景，为保险公司提供端到端的智能服务解决方案，重新定义了保险行业的效率与体验边界。

近年来，保险行业面临多重挑战：客户需求日益个性化、人工服务成本居高不下、业务流程冗余低效。以车险为例，传统报价环节依赖大量人工操作，不仅耗时耗力，还容易因信息误差导致客户不满。

大地保险敏锐捕捉到这一痛点，依托 AI 大模型技术推出智能小行——一款集语音交互、图像识别、自然语言处理（NLP）与大数据分析于一体的智能服务助手，旨在通过“技术替代人力”与“流程自动化”，实现效率质量的双重提升。

为深入贯彻落实“AI 赋能”的核心理念，推动公司全面实现数字化转型，“智能小行”依托大模型平台，凭借强大语义理解能力，支持多轮对话和上下文理解智能交互系统，实现更高效、更智能的沟通，为业务人员带来全新营销体验。

二、创新点

1. 技术驱动场景落地

①多模态交互，无缝衔接用户需求

智能小行支持语音、文字、图片等多种交互方式，业务员可通过直接与 AI 对话。系统可自动识别语音、文字、图片信息，并实时生成报价方案，无需人工介入。

②非标信息标准化处理

传统保险服务中，用户提供的信息往往模糊或不完整。智能小行通过语义理解与图像识别技术，将非标准描述转化为结构化数据，精准匹配保险信息，大幅减少人工录入时间。

③多轮对话信息录入

即在多个回合的对话交流过程中，逐步引导并最终促成交易的完成。通过这种方式，销售人员能够根据与客户的沟通，逐步与 AI 对话补充完整信息，并最终实现销售订单的生成。

④动态优化与个性化推荐

基于用户对话，智能小行可动态调整方案。例如，在报价过程中，系统会根据信息推荐最优保障方案，并同步支持信息的调整，提升客户体验。

2. 从效率到效益的全面升级

①车险报价：从“人工录入”到“秒级响应”

智能小行已实现“AI 即报价”的全流程自动化。通过各机构试点数据，使用智能小行后，单笔报价处理时间从平均 1 分钟缩短至 30 秒，报价效率提升 50% 以上。



②业务运营：从“被动响应”到“主动服务”

智能小行通过接入多个智能体，聚合营销能力，主动提供包括问数，知识库等支持，提升业务人员能力，提供展业支持，为一线展业提供千人前面的业务支持。



三、项目技术方案

1. 技术架构

AI 中台（灵仙界）：集成通义平台，支持大模型训练与部署，兼容多源数据（车辆信息等）。

2. 核心模块

数据层：整合内外部数据源，建构底层信息。

算法层：应用机器学习，不断对模型进行完善。

应用层：智能营销中枢，AI 报价。

语义识别：采用大模型技术，精准解析语意，实现高效语义理解。

图像处理：运用轻量化小模型，专注于图像信息的高效识别与分析。

要素提取：自动化提取关键要素，确保信息提取的准确性和一致性。

结果整合：将语义与图像要素智能组合，生成单一、清晰的报价结果输出。

3. 业务模式

智能化报价：AI 引擎实时报价，加快营销人员展业效率。

个性化营销：千人千面的营销管理与支持，提升业务人员展业能力。

四、运营情况

智能小行作为一款高效的 AI 辅助工具，目前已经在全司各机构中广泛推广和应用。该工具通过运用先进的语义识别技术，能够精准理解和处理用户的自然语言输入，从而大幅简化了业务人员在信息录入过程中的繁琐操作。得益于此，业务人员的信息录入效率得到了显著提升，提高了约 50%。此外，智能小行还能在极短的时间内完成首次报价，将原本需要较长时间的报价过程缩短至半分钟以内，这一突破性进展极大地便利了业务人员的日常工作，使得他们能够更加高效地开展业务。

五、项目成效

智能小行通过对接多智能体的调度系统，实现了对市场营销活动的有力支撑。它能够根据不同的营销场景和需求，智能调度和协调多个 AI 智能体，从而提供更加精准和高效的营销服务，有效助力公司业绩的稳步增长。

此外，智能小行所采用的先进业务模式，不仅在实践中取得了显著成效，也为司内其他 AI 应用的发展提供了宝贵的参考和借鉴。

六、经验总结

智能小行并非终点，而是迈向“保险智能生态”的起点。随着大模型技术的突破，智能小行将进一步进化为“保险超级助手”，在健财产险，健康险等更多领域拓展更多可能性。



浙江东方: AI Agent 营销智能体项目

一、项目背景及目标

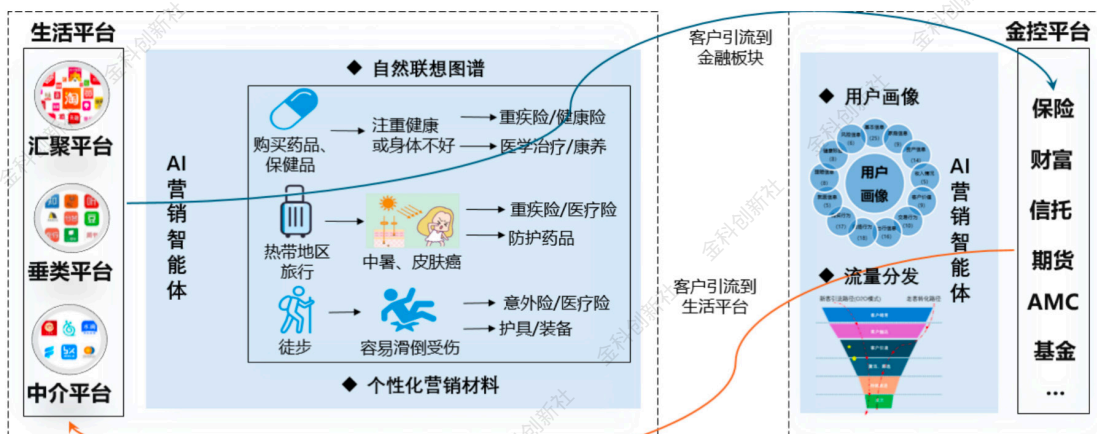
随着国家经济的发展，人民生活水平不断提高，客户的个性化程度也随之提高，不同的客户有不同的“特征”，不同的“痛点”，不同的“购买点”，千篇一律的营销话术、材料逐渐难以适配日益增大的客户个体差异。然而，为不同客户做不同方案，无异于让业务本已繁重的工作任务雪上加霜，使得业务人员陷入进退两难的境遇。浙江东方积极响应国家政策，充分发挥金控集团多元化金融服务优势，抢抓数字化发展新机遇，引入前沿新兴技术，持续开展科技创新，以科技赋能金融业务发展。

在当前经济趋势向缓的背景下，资源协同成为内生性第二增长曲线的推动力，而缺乏“金融+生活”应用场景及低效的营销工具是阻碍资源协同的重要因素，主要困难和问题是：一是缺乏“金融+生活”应用场景：“低频”的金融产品服务，缺少以客户为中心的“高频”生活场景；产品推荐缺乏自然联想和生活场景转化，缺乏多业态的产品组合和黏性高的增值服务。二是传统营销效率低、个性化程度弱、精准营销成本高：金融企业获客渠道单一，获客成本高、转化率低；千人千面的精准营销消耗大量企业资源；缺少整合多业态的客户营销平台等。

因此，针对行业业务营销难的问题，提出创新解决方案，打造 AI Agent 营销智能体，借助人工智能 AIGC 智能体实现高效、精准的营销。基本思路 and 核心优势：一是借助 AI Agent 营销智能体，通过自然联想图谱和生活情景转化，以客户为中心，结合客户的不同特征为客户精准推荐保险产品及增值服务，从低频的金融产品服务转向高频的生活+金融场景化服务，实现子公司间精准营销；二是利用科技的手段对客户特征、痛点、购买点进行分析，使用人工智能大模型自动为目标客户匹配最合适的产品，同时自动生成适配的营销话术、材料。AI Agent 营销智能体助力构建现代化产业发展新动能，塑造企业新优势，帮助业务部门提高客户黏性和营销效率，降低客户流失率，实现利润增长，助力公司第二增长曲线。

二、项目主要内容

AI Agent 营销智能体是助力企业进行产品和服务的整合营销的平台。该平台根据客户画像推荐匹配的产品和服务，并依托先进的 AIGC 多模态大模型，深度融合自然语言处理、图像识别与合成等前沿技术，生成个性化营销话术及海报，辅助销售团队的线上线下载客。



AI Agent 营销智能体的基础功能 - 精准营销：营销人员收集客户的基础信息、各方面特征、个性化需求，将这些信息数据输入 AI Agent 营销智能体，智能体根据客户信息，结合公司自有的企业知识库内的数据，自动为该客户匹配最合适的产品，以及该产品的营销话术、宣传海报，且材料内容满足客户的特征及需求点，帮助营销人员实现高效精准营销。

AI Agent 营销智能体的拓展功能 - 智能引流：智能体对接第三方平台，当汇聚平台、垂类平台、中介平台等第三方平台的用户登录平台时，平台会将用户的脱敏信息发送给智能体，智能体根据收到的用户信息自动匹配与该用户最契合的产品，生成该产品的广告海报，并将该广告海报传回给第三方平台，在平台上进行展示以吸引用户点击，在用户点击后跳转至产品的详细介绍页面，实现智能引流。

三、项目创新亮点及实施方案

项目创新亮点

1. 借力人工智能 AIGC，实现高效精准营销

传统的营销以产品特征及客户群体特征为基础，考虑到时间成本、经济成本等因素，传统营销一般不会关注客户作为一个独立个体的个性化需求。借助人工智能 AIGC，智能体实现高效、精准的营销。

(1) 高效。智能体可以根据客户的信息自动生成营销策划材料，每 10 秒可以为一位客户生成一套包含推荐产品、营销话术、营销海报的完整的方案，将营销人员从单一、重复的工作流程中释放，降低人力成本的同时也提升了业务流程的流畅性和准确性，在运营成本与业务优化双方面实现技术与业务的提升；

(2) 精准。智能体推荐的产品高度适配客户的情况，营销话术、营销海报能够根据客户的信息展开自然联想，比如客户长期在户外的烈日下工作，智能体会意识到客户有罹患皮肤癌的风险，因此会为客户推荐医疗险，并在营销话术中提示用户注意做好防晒措施。

以下是智能体根据三位客户的不同特征为客户精准推荐产品，并生成高度匹配的营销话术及海报的案例。

智能体识别信息：

姓名：王xx
年龄：30-40
性别：女
职业：医生
年收入：50-100万元
教育程度：博士
婚姻情况：已婚
子女情况：一女
健康状况：好
爱好：羽毛球
推荐产品类型：保险

推理

智能体输出信息：

推荐产品：终身寿险（分红型）

营销话术：王女士您好，我推荐您购买一款分红型终身寿险。这种保险产品不仅能够为您提供长期的保障，还能在您的投资中获得分红回报。作为一名医生，您的收入稳定且您可能想要一个稳健的投资计划，这款保险产品正是为您量身定制的。而且，作为一位母亲、妻子，您也能为您的家人提供额外的经济保障。我可以为您提供更多详细信息以及定制化的保险方案，以满足您的特定需求。



(从上述营销话术中可以看出，智能体为客户推荐了一款分红型的终身寿险，高度贴合客户的个性化需求，同时从客户的职业及收入信息入手，“作为一名医生，您的收入稳定且您可能想要一个稳健的投资计划”，以及从客户的家庭及婚姻信息入手，“作为一位母亲、妻子，您也能为您的家人提供额外的经济保障”，拉近与客户的心理距离，突出对客户需求的认识与理解。从上述营销海报中可以看出，智能体生成的海报中一家人坐在一起，完美贴合客户的子女情况“一女”。同时，副标题“医术高超，事业辉煌，财富增长，终身保障”紧贴客户的医生职业与博士学位，肯定了客户高级知识分子的身份与事业上的成功，拉近了与客户的心理距离)



智能体识别信息：

姓名：吴xx
年龄：45
性别：女
职业：企业高管
年收入：百万+
教育程度：硕士
婚姻情况：已婚
子女情况：一女
健康状况：良好
爱好：健身、阅读、护肤
推荐产品类型：增值服务

推理

智能体输出信息：

推荐增值服务：赴韩抗衰老之旅

营销话术：尊敬的吴女士，您好！作为一位事业有成、家庭幸福的企业高管，您不仅在职场上展现着非凡的领导力与智慧，生活中更是以优雅与健康的姿态引领风尚。我们深知，在忙碌的工作与家庭生活中，保持青春活力与最佳状态对您而言至关重要。鉴于此，我们特别为您精心推荐了“赴韩抗衰老之旅”这一尊享增值服务。韩国作为全球美容抗衰老技术的领先者，汇聚了最尖端的医疗科技与自然美容理念，能够为您量身打造个性化、安全高效的抗衰方案。在繁忙之余，您能够置身于韩国的美丽风景中，享受由顶尖专家团队带来的定制级抗衰老服务，从内而外焕发新生，都能让您在保持高效工作状态的同时，也拥有令人羡慕的年轻态与好气色。



(从上述营销话术中可以看出，智能体根据客户的年龄和爱好判断出客户对医美可能感兴趣，为客户推荐赴韩抗衰老增值服务。同时从客户的职业信息和家庭信息入手，“作为一位事业有成、家庭幸福的企业高管，您不仅在职场上展现着非凡的领导力与智慧，生活中更是以优雅与健康的姿态引领风尚”，拉近与客户的心理距离，肯定客户的事业和个人生活状态，突出对客户需求的认识与理解。从上述营销海报中可以看出，宣传语“韩国抗衰，焕新容颜；护肤有道，岁月无痕”紧密贴合客户的年龄及需求爱好，拉近了与客户的心理距离)

智能体识别信息：

姓名：张xx
年龄：55
性别：男
职业：企业主
年收入：300万+
教育程度：学士
婚姻情况：已婚
子女情况：一儿
健康状况：一般
爱好：饮酒、钓鱼
推荐产品类型：基金

推理

智能体输出信息：

推荐产品：养老年金保险

营销话术：张先生您好，作为一位成功的企业主，您在事业上已取得令人瞩目的成就，这充分彰显了您卓越的投资眼光与财富管理的能力。在追求财富稳健增长的同时，我们也深知您对于资产配置多元化的重视。因此，我特别向您推荐我们的私募基金产品。我们的私募基金，专为像您这样高净值、有远见的投资者量身定制。它不仅能够帮助您实现财富的稳健增长，更通过专业团队的市场分析与精准投资，有效分散风险，捕捉市场中的高潜力机会。无论是新兴产业的布局，还是传统行业的精选项目，我们都力求为您带来稳健而可观的回报。选择我们，意味着您将获得一站式、全方位的财富管理服务，这样，您便能更加专注于您热爱的事业与家庭，享受更加丰富多彩的生活。



(从上述营销话术中可以看出，客户经营企业，资产较多，能承受一定风险，但对资金灵活性有要求，因而给客户推荐私募基金，提高收益，保证资金灵活性。从上述营销海报中可以看出，智能体生成的海报硬币成阶梯式形状堆叠，暗示财富逐步增长。同时，宣传语“智投高远，钓金满载；基业长青，跑赢人生”与客户的兴趣爱好和执业紧密贴合，拉近了与客户的心理距离)

2. 对接各类平台，实现智能客户引流

传统的广告投放是较为常见的获客渠道，但是其成本高，精准度差，转化率低。经调研，例如某保及某市民卡基本对所有客户都会推送同样的、无差异的保险广告，精准度相对较弱。

为了弥补这些不足，扩大优质获客渠道，智能体可扩展对接汇聚平台、垂类平台、中介平台，根据平台细分类型及平台用户信息分析塑造用户画像，为用户匹配最适合的金融产品，生成个性化营销广告，并将该营销广告实时传递到这些平台上，吸引用户点击，点击后将其引流至对应的金融产品详情页面，实现智能引流。相比于传统的广告投放，AI Agent 营销智能体的智能引流广告客户转化率提升近 30%。



个性化广告



项目实施方案

项目第一阶段——东方金控集团子公司推广，目的是充分运用金融科技，在精准营销、场景金融、数字化营销等方面积极探索，赋能东方金融子公司的业务第二增长曲线，提升服务经济高质量发展能力。

项目第二阶段——国资委体系内各级公司推广，目的是通过科技创新助力企业营销，推进业务发展，为提升商贸流通、金融服务、生命健康等板块产业的核心竞争力、实现高质量发展提供支撑。

项目第三阶段——外部推广，目的主要是承担起社会责任，为企业通过科技创新实现业务增长提供良好的学习经验和借鉴模板，同时通过释放企业数据的潜能，促进数据资源的高效利用与共享，推动数字经济和智能化升级，增强国家数据治理和数据经济建设，赋能中国经济社会的高质量发展，实现更广泛的社会效益。

AI Agent 营销智能体采用 AIGC 多模态大模型，对接业内知名的多个大模型，基于自然语言处理（NLP）、图像识别与合成技术，实现文本、图像等多模态内容的生成与交互，项目于 2024 年初在东方金控集团子公司推广试点。为后续进一步向外推广该项目，梳理项目价值主张、关键业务、客户关系、生态伙伴、成本结构、盈利模式、推广方案、社会效益 8 个关键要素。

四、项目成效及推广应用

项目成效

1. 经济社会效益

浙江东方金控集团 AI Agent 营销智能体项目通过 AIGC 大模型引擎和模型训练有机结合与实践，利用自然联想图谱和生活情景转化，从“低频”的金融产品服务转向“高频”的生活 + 金融场景化服务，实现子公司间个性化销售精准营销，提高金融子公司营销效率和效果，在保险公司取得了显著成效，助力子公司实现收入 260 万元，通过低营销成本实现了较高收入，实现销售杠杆倍数 47 倍，协同销售收入贡献率 16.36%。

通过该场景应用，促进员工高质量稳定就业，在提供发展平台、改善工作环境等取得了显著效果，在金融行业大环境收缩的趋势下，帮助遇到较大经营压力的子公司稳定员工就业率，提高销售人员提高服务质量，增加销售收入，增强员工稳定性。

2. 经营质效和市场竞争能力等方面产生的影响

AI Agent 营销智能体作为金控集团的创新产物，在浙江东方金控集团及其子公司企业核心业务和市场竞争能力等方面，提高了客户黏性，降低客户流失率，通过子公司间交叉销售协同，享受更加多样化的增值服务，通过多元化服务组合客户日益增长的个性化需求，提高了客户黏性，使客户对东方金控集团品牌产生更深层次的认同与依赖，降低子公司客户流失率。

3. 赋能产业链生态建设和协同效果

浙江东方金控作为一家金控集团，AI Agent 营销智能体在赋能产业链生态建设和协同方面，能更好地满足客户多元化金融服务需求，做好数字金融为财富管理、保险等多家金融服务子公司企业服务赋能新质生产力，以新质生产力点燃高质量发展的新引擎。

通过 AI Agent 营销智能体，能够更准确地理解客户需求和偏好，提供个性化的产品和服务推荐，从而提升客户体验；可以减少传统营销方式中的资源浪费，如减少纸质宣传材料的使用，降低能源消耗和碳排放；AI Agent 营销智能体是金控集团在技术创新方面的积极探索，有助于提升企业的竞争力和影响力，有助于推动金融行业的数字化转型和产业升级，为整个行业树立技术创新的典范；有助于提升企业的营销效率和客户转化率，进而促进企业的经



济增长和盈利能力提升，随着企业经济实力的增强，能够更多地投入到社会公益事业中，为社会提供更多的就业机会和公共服务，推动社会的可持续发展。

4. 知识产权及荣誉成果

AI Agent 营销智能体项目于 2024 年荣获首届中国互联网创新大赛人工智能赛道优秀奖；荣获第五届长三角金融科技创新与应用全球大赛“金融机构最佳实践奖”；荣获第五届金融服务创新与客户经营峰会卓越金融科技场景应用创新奖；荣获省国贸集团创新大赛决赛三等奖，第二届省国贸集团青年创新创效大赛新风奖。同时，AI Agent 营销智能体需基于一系列数据支撑，其智能数据分析平台已获国家计算机软件著作权登记证书。

市场定位与推广应用

AI Agent 营销智能体作为大模型与营销场景结合的产物，其市场定位在于利用先进的人工智能技术，对客户特征、痛点、购买点进行深入分析，并自动为目标客户匹配最合适的产品，同时生成适配的营销话术和材料。这一市场定位满足了当前企业对精准营销、高效营销的需求，特别是在大数据和人工智能技术日益成熟的背景下，AI 智能营销的市场需求持续增长。

随着市场竞争的加剧和消费者需求的多样化，企业越来越需要一种能够快速响应市场变化、精准定位目标客户、并提供个性化营销方案的工具。AI Agent 营销智能体正是这样一种工具，它能够帮助企业实现营销流程的自动化、智能化，提高营销效率和效果。

AI Agent 营销智能体项目未来推广应用规划。一是向下子公司推广，项目可推广至下属各级金融子公司，子公司结合业务场景进行适配性改造。二是向国资委体系内各级公司推广，目的是通过科技创新助力企业营销，推进业务发展，为提升商贸流通、金融服务、生命健康等板块产业的核心竞争力、实现高质量发展提供支撑。三是向外推广，项目可进一步拓展至各类有产品营销需求的金融机构，整体市场存量中，结合现有 5000 家金融企业，市场规模可达约 25~50 亿。

五、项目意义与价值

作为国资委体系内数字化创新的积极实践产物，AI Agent 营销智能体将通过科技创新助力营销，推进业务发展，为提升国资委体系内商贸流通、金融服务、生命健康等板块产业的竞争力、实现高质量发展提供支撑。同时，AI Agent 营销智能体将承担起社会责任，为企业通过科技创新实现业务增长提供良好的学习经验和借鉴模板；通过释放企业数据的潜能，促进数据资源的高效利用与共享，推动数字经济和智能化升级，增强国家数据治理和数据经济建设，赋能中国经济社会的高质量发展，实现更广泛的社会效益。

招商银行济南分行：“营小助”智能作业平台

一、项目背景及目标

随着金融科技的迅猛发展，银行业对公客户的函证业务、取现业务、转账业务、开户业务等虽已逐步实现线上化，但仍面临每年数以百万计的柜面交易需求。同时，虽然 OCR 技术在运营作业中的应用已相当成熟，但 OCR 模型的性能更多依赖于训练数据的质量和数量，数据量不足或数据质量差可能导致模型泛化能力差。为突破这一瓶颈，济南和海口分行大胆尝试引入 AI 多模态识别技术，结合 OCR 与 AI 多模态两种先进算法进行双重验证，显著降低了识别错误率。通过与总行的密切沟通，各团队深入探讨并制定了基于字段重要性的风险评估机制，确保识别结果的安全性和可靠性。这一创新举措标志着总行、济南及海口分行在对公业务领域实现了 AI 智能录入的重大突破。

该项目以“多维运营”为牵引，以“1671 工程”为方法，共同探索拓展 AI 智能作业新赛道，打造“深度融合、人机协同、智能生态”为特征的 AI 智能录审“新质生产力”，打造数智驱动新模式、开启智慧运营新动能，实现了颠覆式的革新，为打造全行数智运营核动力贡献力量。

二、创新点

1. OCR+AI 多模态技术融合应用实现精准审核：结合 OCR 与 AI 多模态两种先进算法进行双重验证，降低识别错误率。通过同一单据不同字段信息（如大小写金额等）之间的逻辑审核，不同单据同一字段信息的交叉审核，通过智能信息录入进化为智能信息审核，进一步赋能小桂圆高效、精准完成审核工作，实现了智能化技术与实际业务需求的紧密结合。这种模式的创新不仅为 AI 赋能业务开辟了新的路径，也为其他业务场景的智能化改造提供了借鉴。

2. 海螺 RPA 助力业务流程自动化：通过 RPA 影像嵌入式应用，及最新的海螺智能作业 SDK 能力，该系统实现在不改变运营核心系统现有流程的情况下自动带入信息。通过录审数据的实时流转与处理，减少人工干预。

3. 系统扩展性和可维护性：该系统采用模块化设计，支持系统的扩展和维护，适应业务发展的需求；建立反馈机制，通过持续收集用户反馈和识别错误，不断优化模型和算法，提高系统的准确性和稳定性。

三、项目技术方案

业务模式

打通模式：多团队深度融合，探索录入新模式。总、分行共同讨论分析运用 OCR、AI 多模态两种不同算法进行双重识别、交叉验证提升识别效率的可行性；研究探讨基于字段重要性的风险性，评估字段是否可以实现科技替代录入，确保识别结果的安全性和可靠性。

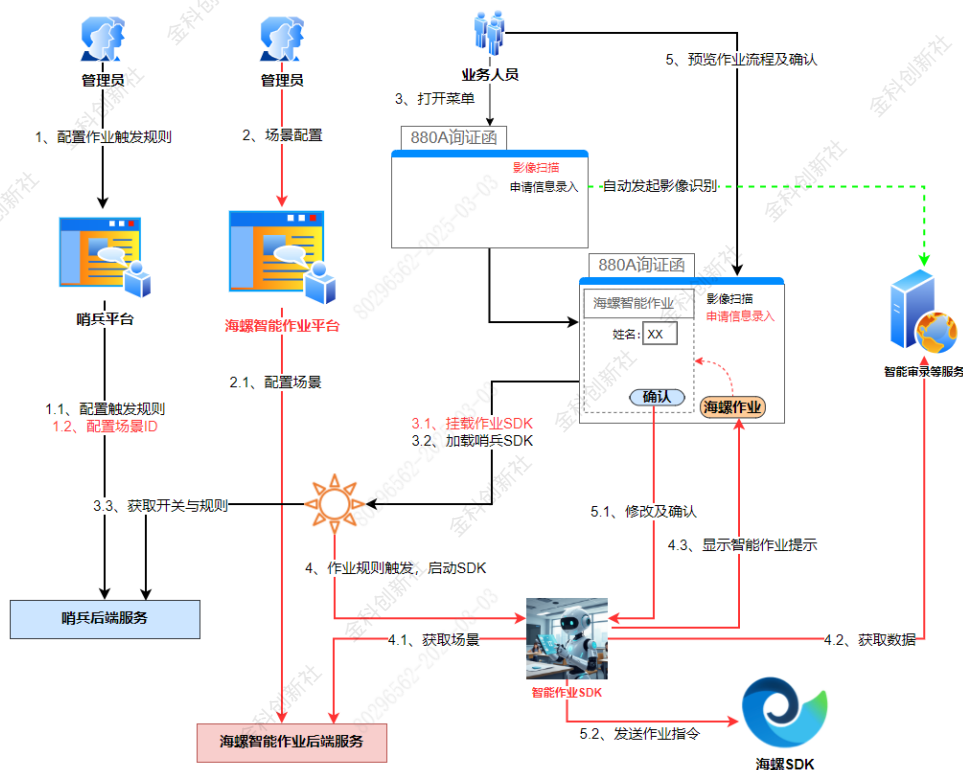
打通流程：多规则深度融合，开启智慧新流程。在确定录入新模式之后，如何成功落地应用成为摆在眼前的新专题。总行、济南、海口分行运用“PETSC”方法，选定函证业务作为 MVP 试点业务，横向从产品、事件、任务、步骤等分析确定智能录审的节点；纵向从业务种类、单据组合、凭证种类、字段信息确定识别应用规则。通过横向、纵向的规则综合推演，最终确定“先扫描、后识别、再应用”的方式，并通过业务和技术层面的双重评估，确保了流程的科学性和可行性。

打通模式：多业务深度融合，开启智能新模式。在成功推演评估 MVP 业务流程的可行性后，总行、济南、海口进一步提炼 AI 智能录审的公共要素，打通跨业务种类运用 AI 智能录入的可行性评估方法，萃取总结形成可广泛适



用的通用方法。形成通用方法后，根据业务量、单据标准化程度等要素，对全行柜面对公业务进行评估论证，最终确定在函证业务先行先试后，同步扩展至对公取现、转账等各类业务，持续扩展 AI 智能录审的业务种类，持续升级 AI 智能录审的智能程度，推动银行业务向更加智能、更加高效的方向持续迈进。

流程介绍



1. 影像扫描，纸质信息实时数据化。经办人员扫描单据影像到运营核心系统，触发“营小助”智能作业平台（询证函模块）自动调用影像系统接口获取扫描上传的影像，随即展开“询证函智能录审之旅”。

2. 多技术融合，审核信息执行智能化。“营小助”智能作业平台（询证函模块）运用行内智能录审平台多模态大模型+视觉云平台OCR的技术融合手段，同时识别并提取询证函的关键录入要素，如询证函的函证范围、截止日期、联系人、联系电话、函证范围、会计师事务所等信息。同时，在识别页面支持人工勾选AI大模型识别结果、OCR识别结果或手工录入（不使用任一识别结果），便于后续根据不同字段使用识别信息、手工录入信息分析字段识别的准确率、错误率等，进一步评估字段自动化录入的可行性。

3. 多平台联动，业务信息录入自动化。“营小助”AI智能数字运营平台（询证函模块）调用影像系统接口获取实时数据，通过RPA影像嵌入式应用、海螺智能作业SDK能力，将OCR及大模型智能录审结果通过运营核心系统接口自动填充至相应位置，实现在不改变运营核心系统现有流程的情况下自动带入信息。通过录审数据的实时流转与处理，减少人工干预，确保数据的准确性和一致性，提高了业务办理的效率，大大减少了人工录入的时间和错误率。“营小助”智能作业平台（询证函模块）通过打通影像系统、运营核心系统、RPA、智能录审平台等联接，实现了AI大模型能力与运营核心系统流程的无缝对接，全面提升了AI技术在运营核心系统实际业务中的未来应用潜力，为未来的运营核心系统全面智能化打下了坚实的基础。

四、运营情况

自第一版本 5 月 8 日上线至今的函证试点应用阶段，AI 多模态技术与招行运营核心系统进行了深度整合，验证了 AI 技术在询证函处理中的卓越效果。

1. 推广应用情况

“营小助”智能作业平台自上线以来，已在济南分行、海口分行函证业务场景中得到了广泛应用和推广，平台的自动化录入和审核功能显著缩短了业务处理时间，提升了整体业务效率。柜面人员和后台运营人员对平台的使用体验总体满意，认为其操作便捷，效率提升显著。随着版本迭代和功能优化，计划进一步扩大应用范围至其他分行和分支机构。

2. 系统运行情况

“营小助”智能作业平台在系统运行方面表现稳定，但也仍存在一些需要优化的地方。

运行稳定性：平台整体运行稳定，未出现重大故障或中断，确保了业务的连续性和可靠性。偶发性故障较少，且能够快速恢复，对业务影响较小。

性能表现：平台的 OCR 和大模型技术处理速度较快，能够满足高并发业务需求。系统响应时间较短，用户体验良好。

安全性：平台采用数据加密和访问控制技术，确保敏感数据的安全性，符合金融行业的数据安全标准。

合规性：平台的数据处理流程符合相关法律法规，未发生数据泄露或安全事件。

五、项目成效

经济效益

1. 减少人工成本、提高工作效率：通过多技术融合（如 RPA、OCR、多模态大模型）实现询证函信息的自动录入，减少了人工操作的需求，降低了人力成本。传统询证函处理需要大量人工操作，每份询证函的录入和审核时间约为 10 分钟。通过智能录入系统，处理时间缩短至 5 分钟以内。基于目前每年约 50 万份的全行纸质询证函业务量估算，未来预计可在 60%-70% 的询证函业务接单环节中，运用“营小助”智能作业平台（询证函模块）的 AI 智能录审能力实现人工录入替代。在运行效果良好后，预计可替代 30%-50% 的人工操作，年节约 FTE 约 27（FTE 计算： $5\text{min} \times 500000 / 60 / 6 / 250 \approx 27.8$ ），释放百万分钟的人工操作资源，显著提升业务处理效率和准确性。

2. 提升业务处理能力：系统可以同时处理大量询证函，提升了银行的业务处理能力，满足了高峰时段的业务需求。

社会效益

1. 提升行业标准：通过引入先进的技术手段，该系统推动了行业向更高标准的数字化、智能化方向发展。

(1) 推动行业数字化转型：传统询证函处理依赖人工操作，效率低且易出错。智能录入系统通过 RPA、OCR 等技术实现了流程自动化，推动了行业从“人工操作”向“智能处理”的转变，提升了行业的整体数字化水平。

(2) 提高规范化水平：通过标准化的字段配置和流程管理，系统确保了询证函信息的准确性和一致性，减少了人为错误，提升了行业服务的规范化水平。

2. 改善用户体验：“营小助”智能作业平台（询证函模块）提供了友好的用户界面和实时预览功能，使



用户能够更便捷地完成询证函的录入和提交。用户通过系统提交询证函后，信息自动录入和审核，操作流程更加简便，处理时间从传统的数小时缩短至几分钟，极大提升了用户的满意度。

六、经验总结

项目建设经验总结

1. 需求分析与规划

明确需求：在项目启动初期，分行进行了全面的需求分析，与用户深入沟通，明确了系统的核心功能需求，包括自动化的询证函录入、数据校验、流程管理等。

制定计划：根据需求分析结果，制定了详细的项目计划，包括时间表、资源分配和里程碑设定，确保项目有序推进。

2. 系统设计

模块化设计：采用模块化设计，将系统分为数据采集、校验审核、自动录入、流程管理等模块，便于后期维护和扩展。

技术选型：充分评估行内各技术平台及工具的优势特点，选择适合的 OCR、RPA 等技术，确保系统高效稳定运行。通过对比测试，选择了性能最优的技术方案。

3. 开发与测试

敏捷开发：采用敏捷开发模式，分阶段开发和测试，及时发现问题并修复，确保系统质量。

全面测试：在开发过程中，进行了单元测试、集成测试和用户验收测试，确保系统功能完善、运行稳定。

4. 上线与部署

平稳过渡，持续优化：优先济南、海口分行作为试点，试运行一段时间后，充分暴露系统问题、收集反馈意见，持续优化系统功能和性能，提升用户体验，再逐步推广至全行。

推广经验总结

用户培训与支持：该系统一经上线，便为用户提供详细的操作培训，帮助用户快速掌握系统功能和使用方法。同时建立了完善的售后服务体系，及时解决用户在使用过程中遇到的问题，提升用户满意度。

总结与展望

通过“营小助”智能作业平台（询证函模块）的建设和推广，我们积累了丰富的经验，成功提升了行业的数字化水平和用户的工作效率。未来，我们将继续优化系统功能，拓展应用场景，根据业务量、单据标准化程度等要素，对全行柜面对公业务进行评估论证，在函证业务先行先试后，逐步扩展至对公取现、转账等各类业务，持续扩展 AI 智能录审的业务种类，持续升级 AI 智能录审的智能程度，推动银行业务向更加智能、更加高效的方向持续迈进。

上海农商银行：零售数智化创新实验室项目

一、项目背景及目标

背景

AI 技术的兴起为银行业带来全新发展机遇。以大模型为代表的新一代 AI 技术，在自然语言处理、数据分析等领域展现强大潜力，可助力银行实现智能客服、风险评估等业务的智能化升级，成为提升银行竞争力、驱动新质生产力的关键。同时，银行同业积极布局 AI，竞争压力持续升级，已有 20 余家银行部署应用 DeepSeek 大模型，在智能客服、风险管理等多领域取得成效，人工智能能力已成为银行竞争核心要素。此外，政策也明确指引银行深化 AI 技术融合应用，加快零售业务人工智能能力建设，是顺应政策、实现“弯道超车”的关键。

目前，上海农商银行零售业务在人工智能应用上已取得一定成果。决策类模型应用初具规模，在产品推荐、信用评级等场景发挥重要作用，通过数据分析助力精准营销与风险管理，但传统决策模型在复杂场景和大规模数据处理上存在局限，难以满足业务与客户期望。为此，行内建立零售数智化创新实验室探索 AI 应用，打造零售业务人工智能能力，实现业务转型升级与可持续发展，提升核心竞争力。

目标

1. 提升零售金融业务智能化服务体验

以客户为中心，利用 AI 技术深度洞察客户需求，提供个性化、智能化的金融产品与服务，提升客户体验与满意度；优化业务流程，提高运营效率与风险管控能力，提升对营销和管理人员的服务支撑体验，实现业务模式的创新与转型。

2. 助力零售业务价值创造和竞争力提升

通过 AI 技术的应用，推动银行从传统金融服务向数字化、智能化金融服务转变，挖掘业务潜在价值，增强在金融市场中的竞争力与影响力，适应金融行业数字化发展趋势。

3. 构建零售智能化共享生态圈

通过 AI 技术的广泛应用与拓展，连接客户、合作伙伴、监管机构等各方，构建一个开放、共享、协同的智能化金融生态系统，实现金融服务与社会经济生活的深度融合，更好地服务实体经济，创造更大的社会价值。

二、创新点

(1) 敏捷创新机制：突破传统银行技术应用的线性流程，通过设立专门实验室，构建“探索—试点—快速采购”的敏捷闭环，压缩新技术从概念验证到业务落地的周期，快速响应市场变化与客户需求。

(2) 场景驱动式采购：改变传统以技术或供应商为中心的采购模式，围绕零售业务场景痛点（如智能营销、员工培训等）探索 AI 应用，以试点效果为采购决策核心依据，确保技术投入精准匹配业务价值。

(3) 数据验证决策：依托实验室试点阶段积累的真实业务数据（如转化率提升、成本降低比例）量化评估项目成效，使采购决策从经验驱动转向数据驱动，提升资源投入的科学性和成功率。

(4) 生态协同创新：通过实验室试点筛选优质项目，吸引外部技术供应商、科技企业参与合作，构建银行主导、多方协同的 AI 创新生态，加速技术迭代与资源整合。



(5) 风险可控的创新探索：以小规模试点验证技术可行性和业务适配性，降低大规模技术投入风险，同时为后续采购积累实践经验，形成可复制、可推广的数智化转型路径。

ID 三、项目技术方案

实验室的技术方案围绕三大核心策略构建，形成“安全隔离、敏捷验证、生态协同”的创新体系。首先，在环境部署上，实验室采用测试环境与生产环境深度隔离的架构设计。通过物理网络隔离与数据单向传输机制，将实验室测试环境与银行核心生产系统完全分隔，避免创新探索对现有业务的干扰。同时，配置独立的数据沙箱系统，对生产数据进行脱敏、加密及特征提取处理后同步至实验室，既保障数据可用性，又确保敏感信息安全。采用容器化技术与微服务架构，实现计算资源的动态分配与弹性扩展，支持多项目并行测试，配合版本控制对模型、代码及数据进行全生命周期管理，为技术验证提供稳定、安全的运行空间。

其次，在项目推进机制上，实验室推行免费试点与效果驱动的采购模式。与科技公司合作初期，采用免费试用方式引入 AI 技术方案，在零售业务的具体场景中开展小规模试点。试点过程中，通过多维度指标体系量化评估项目成效，如 AI 外呼的客户接通率、营销转化率，AI 客服的问题解决率、客户满意度等。对于试点效果显著、能够切实提升业务效率、优化客户体验或降低运营成本的项目，实验室考虑采购，推动技术成果向生产环境迁移。这种“先验证、后投入”的模式，既能降低技术应用风险，又能确保资源精准投向高价值项目。

最后，在生态合作层面，实验室秉持联合共创的开放理念。银行依托自身在零售业务领域的深厚积累，提供真实业务场景、历史数据资源及行业经验，为技术研发明确方向；科技公司则发挥其在大模型开发、算法优化、系统搭建等方面的专业优势，双方通过成立联合创新小组、定期技术研讨会等形式，共同探索 AI 技术与银行零售业务的深度融合。这种资源互补、协同创新的模式，不仅加速了技术成果的转化效率，更有助于打造符合银行业务特性的定制化 AI 解决方案，为零售业务数智化转型注入持续动力。

ID 四、运营情况

零售数智化创新实验室自成立以来，构建起“技术探索 - 场景验证 - 成果转化”的全链条运营体系。在技术合作层面，实验室与商汤、蚂蚁数科等头部科技企业建立深度合作，围绕 AI 问数、客户洞察等前沿领域开展联合攻关，目前已储备 4 项创新技术方案，形成差异化竞争优势。场景探索方面，通过跨部门协作机制，联动零售业务、金融科技等部门挖掘潜在应用场景，累计完成 12 个业务场景的可行性评估，其中 4 个创新项目选入试点，创新项目中 AI 教练已完成试点，且由于分支行参与度高，目前已正式合同采购；AI 外呼项目自 3 月中旬开始对客户使用，截止 5 月底已触达 13 万客户；AI 客服项目和 AI 问数仍处于部署中。

系统运行方面，实验室采用“双环境隔离”架构，将测试环境与生产系统完全分离，依托容器化部署与微服务架构，实现资源动态调度与弹性扩展。同时，建立标准化项目管理流程，通过子项目小组责任制与双周汇报机制，确保项目进度可控、风险可管。

ID 五、项目成效

零售数智化实验室的高效运营为银行带来显著的经济效益与社会效益。经济效益上，通过技术试点与成果转化，成功优化 2 项核心业务流程，推动运营成本降低和业务处理效率提升。在社会效益方面，显著提升了银行的品牌形象和社会影响力。通过引入先进的数智化技术，银行的服务质量和客户体验得到大幅改善。同时，实验室的运营也为金融科技领域培养和储备了一批专业人才，通过项目实践和技术交流，员工的创新能力和技术水平得到显著提升，为银行的可持续发展奠定了人才基础。

零售数智化创新实验室不同子项目的具体赋能方向各有侧重，下面分子项目分别说明：

(1) AI 教练项目，由 AI 扮演不同场景下的各类客户与员工进行情景对练，超级教练即刻指导建议，以“学”、“练”、“战”相结合的方式，滚动培养条线员工专业素养和实战能力，试点 2 个月期间，全行参与项目培训人次达 3070 次，覆盖 524 位理财经理。分支行培训参与率达 100%，理财经理线上考试通过率超 90%，试点效果显著，分支行参与度较高，故已签署采购协议。

(2) AI 外呼项目，打造行内暂缺的人机协呼模式，通过 AI 主动对客户进行个性化的外呼营销，帮助直营经理扩大连接客户能力，提升长尾客户的触达率和转化率。项目 3 月初开始对客，根据试运行数据，人机协呼模式为客户经理初筛了意向客户，企微转化率由纯人工外呼的 14.47% 提升至人机协呼的 21.49%，后续行内考虑采用人机协呼模式，节约人力成本。

(3) AI 客服项目，客户点击手机银行 PUSH，直接跳转微信与 AI 客服对话，AI 客服识别客户的语音或文字后，解析客户需求，并自动作答，对意向客户可无感转接人工坐席。项目结合信用卡账单分期场景目前处于项目测试中。

(4) AI 问数项目，革新银行零售业务数据应用模式，用户无需掌握复杂的数据处理技能，仅以自然语言提问，系统就能依托大语言模型与业务数据，快速输出精准答案。面对业务数据波动，深度剖析业务涨跌根源，并结合历史数据与行业趋势，从产品优化、渠道调整、精准营销等方向，提供切实可行的业务发展建议，为银行高效决策、提升竞争力提供有力支撑。项目目前处于场景讨论和数据治理中。

六、经验总结

零售数智化创新实验室项目在建设及推广过程中积累了丰富经验。在项目建设方面，首先构建了跨部门协同机制，由零售、科技等多部门组建联合团队，打破信息壁垒，确保业务需求与技术方案深度融合。其次，采用“双环境隔离”架构，将测试环境与生产系统严格分离，依托容器化与微服务技术实现资源灵活调度，保障实验安全稳定运行。同时，建立标准化项目管理流程，通过子项目小组责任制和定期汇报机制，对项目进度、风险进行动态管控。

在推广经验上，坚持“试点先行、以点带面”原则，优先选择业务痛点突出、技术可行性高的场景开展试点，通过小规模验证降低推广风险。注重与头部科技企业建立战略合作，引入前沿技术的同时，结合银行实际需求进行本地化改造。此外，强化知识沉淀与经验复用，将试点过程中的技术文档、业务案例等进行系统化整理，形成可复制的标准化方案，加速创新成果向全行推广，有效推动零售业务数智化转型进程。



宁夏银行：“宁银小智”大模型在信贷风险管理中的数智化应用建设项目

一、项目背景及目标

近年来，随着生成算法、预训练模型、多模态数据分析等 AI 技术的聚集融合，AIGC 的实践效用迎来了行业级大爆发。通用大模型技术的成熟推动了新一轮行业生产力变革，在投入提升与政策扶植的双重作用下，以大模型技术为底座、结合专业化金融能力的金融大模型产品应运而生。大模型技术在金融领域的发展呈现出蓬勃之势，为金融机构的数智化转型带来了前所未有的机遇。金融领域也凭借海量数据资源、复杂业务场景及高数字化基础，成为了大模型技术落地的先行者。

尤其是随着 2025 年初国产大模型 DeepSeek 的爆火，使得当前大模型行业发展已真正进入深水区。DeepSeek 在显著提升大模型训练效率的同时，极大程度的降低了推理成本，在提升用户体验的基础上为辅助分析及决策场景打开了空间，使得大模型产品在推进金融产业升级和在金融场景中落地变得更加容易且有效。

为加速实现宁夏银行数智化转型需求，构建基于信贷业务风控场景的大模型深度应用能力，宁夏银行建设了专属信贷管理领域的首个 AI 智能助手“宁银小智”。“宁银小智”大模型是将 DeepSeek R1 满血版大模型作为底座，引入安硕信息自主研发的金融垂直领域大模型“千寻”，深度适配宁夏银行信贷业务场景及信贷系统应用功能量身打造而成。

“宁银小智”大模型在信贷系统中的集成部署，不仅能够支持业务人员在信贷系统中灵活调用对话页面，快速获取企业信用信息、生成企业全景画像，还能够通过深度逻辑思考进行股权结构、科创能力、财务情况、风险归因、行业风险等深度专项分析，实现客情全方位挖掘，繁杂信息高效处理，报告撰写效率全面提升。

二、创新点

灵活调用、智能答疑

“宁银小智”大模型已与宁夏银行行内信贷管理系统无缝融合，信贷用户可在信贷系统主页直接唤醒“宁银小智”获得服务支持。通过“通用问答、灵感、历史记录”等功能，实现对话页面快捷调用、关联问题智能联想、复杂问题一问即答。

企业画像、一键生成

“宁银小智”大模型以宁夏银行行内信贷系统现有“智胜大数据”为基础，内置“风险归因、财务分析、科创分析、涉诉分析”等智能组件，实现企业经营、历史沿革、重大事项等多维度的企业信用信息基础问答，可一键生成客户立体全景画像，帮助信贷用户快速洞察客情。

深度分析、智能求索

“宁银小智”大模型依托自身强大的深度思考及逻辑推理能力，基于企业经营、信用、财务及行业等 12 大

领域的模型化问答知识库，实现金融信息解读、行业政策研判、风险因子挖掘等多方面的分析能力输出并智能生成专业分析内容，辅助业务人员高效处理繁杂信息，提升报告撰写效率。主要赋能内容包括：

1. 企业信用问答

“宁银小智”大模型的知识范围涵盖企业概况、经营行为、监管负面等维度，通过与企业风险归因分析模型、科创分析模型结合，“宁银小智”大模型不仅可以通过逻辑思考，回答企业工商、诉讼、舆情、经营等方面问题，还能够对企业风险态势、科创能力进行量化分析和可视化呈现，深度刻画企业信用全景画像。

2. 行业产业问答

“宁银小智”大模型将行业前沿研究与产业集群数据纳入底层知识库，用户可以针对目标行业的发展概况、竞争格局等内容展开提问，“宁银小智”大模型能够提供可溯源的分析文本及相关图表，同时，还能够实现目标区域的产业深度分析，智能生成产业集群图谱，精准推荐产业重点企业，协助信贷人员快速洞察行业、产业发展脉搏，把握商业先机。

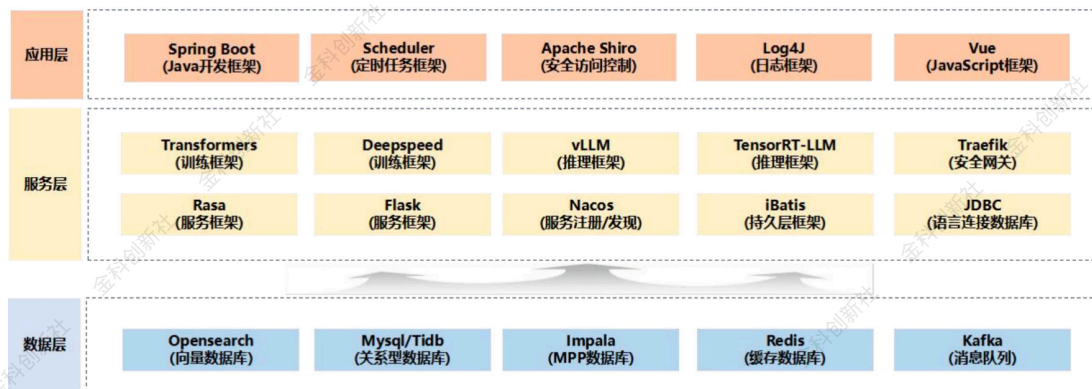
3. 财务分析问答

“宁银小智”大模型学习了公开市场企业和行业财务行为表现相关知识，内容涵盖企业财报、指标、绩效标准值等价值化信息，能够帮助信贷用户解答企业偿债能力水平、盈利能力分析、尽调审计要点、整体财务评价等相关问题。

三、项目技术方案

本项目于 2024 年 10 月启动后，历经了 6 个月的需求分析、开发实施、联调部署和充分测试，最终于 2025 年 4 月成功在宁夏银行信贷管理系统中集成部署并成功上线，并面向全行客户经理、授信审批人员等信贷系统用户开展功能普及、推广及具体应用。同时，为了使“宁银小智”的功能更加“易用”且贴合真实用户体验，项目团队不断接受用户使用反馈，不断开展大模型功能调优，截止 2025 年 5 月末，已完成了 2 个版本的功能迭代。

项目关键技术如下图所示：



技术架构图

1. 产品应用层

采用 Spring Boot(Spring Cloud 微服务方式部署) 实现 + Vue 实现，完全容器化部署。服务端利用 Spring Boot 简化了分布式系统基础设施的开发，如服务发现注册、配置中心、消息总线、负载均衡、断路器、数据监控等，都可以用 Spring Boot 的开发风格做到一键启动和采用 jar 包直接运行部署。



2. 服务层中的模型服务

采用基于主流基座大模型 + RAG 架构，搭建企业级问答服务框架，包括知识库管理、Agent 管理、Memory 管理、Action 管理等模块，支撑意图对齐、上下文理解、文案生成等对话服务；接口应用服务通过 Flask 框架实现，任务规划服务编排各个接口服务，调用不同 Agent 服务获取检索知识库，结合专家知识，输出问答结果。

3. 大模型 RAG 服务

RAG 前沿技术（如向量检索、文档检索、知识切片等），结合金融领域专家知识，支撑结构化、非结构化、多模态各类型数据的知识召回；数据分析、统计类 Agent 高度集成我行及实施厂商的模型资产和场景大模型（如归因分析、财务分析等），产出垂直领域业务数据特征描述、专家分析意见等结果。

四、运营情况

“宁银小智”大模型目前已全面推广应用至宁夏银行所有信贷业务场景中，为近千位客户经理及授信审批、风险管理等部门人员提供了“客情问题智能联想、复杂问题一问即答、金融信息深度求索及风险因子深度挖掘”等智能化分析能力输出，辅助业务人员快速洞察客情，高效处理繁杂信息，提升报告撰写效率，实现了信贷业务全流程管理的数字化、智能化。

五、项目成效

1. 助力智能营销

通过分析海量企业外部数据，实现企业分析、商机挖掘、圈脉触达、科创评价、政策解读、话术推荐等能力，支持快速生成企业多维营销画像，不仅能够帮助客户经理挖掘潜在客户，还能够根据实际场景生成个性化营销指引，大幅提升客户触达效率。

2. 辅助智能尽调

“宁银小智”大模型通过逻辑思考及推理能力，分析输出企业概况、经营行为、财务分析、运营能力、监管负面等维度问题分析结果，并以文字、表格、图形等方式全方位展示，尽显尽调场景智能化，辅助提升尽调效率及质量。

3. 驱动领域风控

“宁银小智”大模型基于深度理解企业行政、司法、税务、舆情等风险数据以及监管制度、违约案例等内容，在风险筛查模块，实现了企业名单批量风险扫描及分值排名，辅助信贷人员清晰观测企业风险指数及风险动态变化，综合评估企业风险及波动情况，深度赋能信贷风控全流程。

4. 深化市场洞察

“宁银小智”大模型将行业前沿研究与产业集群数据纳入底层知识库，信贷用户可以针对目标行业的发展概

况、竞争格局等内容展开提问获取可溯源的分析文本及相关图表，辅助市场研判。

5. 提高工作效率

“宁银小智”大模型目前已无缝集成至宁夏银行行内信贷管理系统中，并在全行所有信贷业务场景中全面推广应用。为近千位客户经理及授信审批、风险管理等部门人员提供了“客情问题智能联想、复杂问题一问即答、金融信息深度求索及风险因子深度挖掘”等智能化分析能力输出，辅助业务人员快速洞察客情，高效处理繁杂信息，提升报告撰写效率，实现了信贷业务全流程管理的数字化、智能化。

六、经验总结

中小银行的业务场景具有“小而散”的特点。聚焦中小客群，客户需求差异大，传统管理模式高度依赖人力经验。但当前，外部竞争压力与内部效率瓶颈已不容回避，大型银行下沉区域县域、净息差收窄、客户对数字化服务期待攀升，这些挑战倒逼中小银行必须将智能化提升至战略高度。

AI 技术的平民化，为中小银行提供了“以小博大”的可能。DeepSeek 出现后，中小银行可以在组织架构和应用推广重视的基础上考虑爆点应用场景，持续以点带面推动大模型智能化应用，但要避开“伪智能化”陷阱，因为效果导向比技术堆砌更加重要。中小银行的智能化绝非单纯的技术采购，而是一场“顶层主导、全员参与”的变革。

大模型能发挥预期作用，离不开“实质落地”：一是要避免“重展示轻实用”。大模型若无法嵌入实际业务应用场景，反而浪费资源；二是要避免“重采购轻运营”。引入智能工具并实施部署后，要不断开展大模型运营监测，深度结合应用场景，持续开展参数调优及模型训练，否则大模型将沦为“空壳”。

基于项目实践经验，宁夏银行深耕信贷管理业务场景，并紧密结合用户需求持续开展系统功能优化。未来，宁夏银行将紧跟人工智能技术发展，不断挖掘“宁银小智”大模型应用能力，深度赋能信贷全流程、多场景领域。将在“调查报告智能撰写、客户风险智能筛查、审查审批智能辅助、行业产业智能洞察”等业务场景中不断深化应用，并联动行内信贷系统、押品系统以及营销类系统等，高效响应多场景需求，实现大模型的高效部署、创新应用。



重庆银行：基于大模型与知识图谱技术融合的数智尽调平台

一、项目背景及目标

随着信息技术的飞速发展，金融行业正经历着深刻的变革，银行传统的业务模式面临着前所未有的挑战。银行授信业务的尽职调查作为信贷业务开展的核心环节，对于评估企业风险、保障资金安全以及制定合理的金融服务策略起着至关重要的作用。然而，传统的人工尽调方式在数据获取、处理效率、风险识别等方面存在着诸多不足，已难以满足现代银行业务快速发展的需求。近年来，生成式人工智能的迅速发展正在重塑银行未来业态，银行业已开启数字化转型“加速跑”。重庆银行将人工智能技术作为核心驱动力，全面深化战略布局，以“技术筑基-场景破局-生态共赢”为战略框架，从场景、业务、数据、算力、安全五个维度进行能力提升，全方位推进“技术创新+场景赋能”双轮驱动模式。

我行数智尽调平台以公司业务尽职调查作为大模型应用“小切口”，构建“数据整合-智能分析-决策支持”三层能力体系，实现尽调流程数字化升级，为一线提效减负，旨在促进金融产品和服务创新，提升服务实体经济质效，以金融创新驱动高质量发展。

二、项目创新点

1. 多渠道数据整合

集成来自政府公开渠道、信用评级机构、行业协会和社会公众媒体等多个外部数据源，主要为税务信息、发票数据、征信数据等授权数据，工商信息、司法信息、知识产权、招中标信息、舆情信息等公开数据；充分与行内数据进行整合，主要为信贷业务数据、行为数据、交易数据、财务数据等；确保信息的全面性和实时性。在数据整合过程中，对不同来源的数据进行清洗、标准化处理，消除数据冗余、重复和错误，确保数据质量。通过高性能的数据处理引擎，支持并行计算和分布式数据处理，能够同时分析多条数据流，显著提升数据处理的速度，使得客户经理可以在最短时间内获得分析结果。

2. 数据分析与客户画像

在信息整合和加工共享的基础上，对客户信息和数据进行结构化处理与特征指标加工等深度挖掘与处理，依据行内风险偏好，设置百余个风险指标，对客户数据进行全面细致的扫描，用于负面事件筛查。一旦发现潜在风险点，系统会迅速输出详细的筛查结果，包括风险类型、风险程度以及可能产生的影响等信息。同时，基于专业的风险评估模型和丰富的业务经验，系统还会给出针对性的行动建议，如进一步核实某项信息、调整授信额度或采取相应的风险防范措施等，真正实现了风险管理的前置。在风险筛查的同时，借助大数据技术，系统将为客户生成全方位 360 画像。这一画像不仅仅局限于客户的基本信息，如企业名称、注册地址、法定代表人等，还深入挖掘客户的关联信息，包括股权关系、上下游供应链关系、担保关系等，清晰呈现企业在市场网络中的位置和影响力。对于财务风险，通过对财务数据的深度分析和模型预测，评估企业的财务健康状况，预测潜在的财务危机。在经营模式方面，结合企业的业务流程、市场定位、销售渠道等信息，剖析其经营模式的优势与潜在风险。此外，

利用自然语言处理技术对海量的行业舆情数据进行分析，及时掌握行业动态、竞争对手情况以及企业自身的口碑评价，为银行提供全面、立体的客户视角。

3. 量化模型支撑智能决策

探索以“数据+算法”驱动，结合大数据、机器学习和决策引擎等前沿技术，应用于精准营销和风控决策等场景。为了更精准地衡量各因素对企业风险的影响程度，系统借助数据分析技术，基于行内风险偏好，结合企业财务指标、经营绩效评估、信用历史与行为分析、行业与市场环境分析等多维画像，将各个指标的权重和影响力进行量化。在此基础上，系统精心构建了四大决策模型，分别为产业分析模型、客群挖掘模型、量化评估模型和企业综合评分模型。

产业分析模型通过对大量行业数据的深度挖掘和分析，把握不同产业的发展脉络、周期特点以及竞争态势。它能够预测产业的未来走向，帮助银行提前布局，优化信贷资源在不同产业间的配置，避免因产业系统性风险导致的信贷损失。客群挖掘模型利用机器学习算法，对海量客户数据进行聚类分析和特征提取，它能够发现不同客户群体的潜在需求、行为模式和风险特征，帮助银行精准定位目标客户群体，实现差异化营销策略。量化评估模型运用复杂的数学模型和统计方法，对企业的各项风险因素进行量化计算，它综合考虑企业的财务风险、经营风险、市场风险等多个方面，为企业的风险状况提供一个精确的量化评估结果。企业综合评分模型则是在前述三个模型的基础上，对企业的综合实力和信用状况进行全面评价。它将各个维度的评估结果进行整合，通过历史数据的训练，不断优化模型的参数和算法，提高模型的精准度和业务可解释性。经过反复训练和验证，该模型能够精准地计算出企业的综合能力评分，这一评分全面反映了企业的综合实力和信用状况。客户经理可以依据这一评分，快速、准确地了解企业的基本情况，为其提供精准的企业能力评价，从而在信贷业务中做出更明智的决策，如确定合理的授信额度、贷款期限和利率等。

4. 基于 AI 自动生成尽调报告

通过梳理行内不同行业、不同产品的业务尽职调查模板，形成预设的报告模板，基于自然语言处理技术，将尽调报告自动生成结构清晰、逻辑连贯的报告框架和内容。报告涵盖了企业经营分析、行业分析、财务分析、风险缓释措施等多个方面，支持将报告输出为 Word 格式，方便用户进行保存和编辑。此外，还可以根据数据情况实现报告的实时更新，确保数据的时效性和准确性，满足银行对于尽调报告的全面要求。

三、项目技术方案

系统架构

为提升信贷报告撰写质效，拟在银行尽调场景中应用大模型技术，实现报告模版理解、自动化文档解析、信贷报告生成等功能。需建立以下大模型能力：一是模版理解，解析各产品、各行业尽调模版样式，理解模版填写要求；二是信息提取，多模态识别原始资料，对多种信息源进行识别、清洗、提取；三是数据分析，根据已有数据进行风险分析、进行数据分析及对比、趋势预测等；四是报告生成，将各类信息结构化，并填写到模版对应部分；五是报告复核，对报告内容进行审核及溯源。

基于以上功能要求，构建了“数据-模型-应用-安全”四层协同架构体系。底层数据层，整合财务数据、交易数据等行内数据和工商、司法、舆情等外部数据以及非结构化文档，通过 ETL 管道实现多源异构数据的清洗和标准化。模型层，运用 DeepSeek-V3-0324 满血版和 Qwen-QwQ-32B 等开源大模型，并基于 vLLM 分布式部署，在开源大



模型应用平台进行场景开发。融合规则引擎和知识图谱构建“生成+推理”双驱动架构，支撑文档解析、风险关联分析等核心能力。应用层，通过 API 接口调用，在数智尽调平台集成大模型应用平台功能，将大模型能力具像化为自动化报告生成、风险可视化分析等模块，嵌入现有尽调工作流程，实现人机协同。安全层贯穿全域，采用动态脱敏、联邦学习技术保障数据隐私，实现全流程风险可控，形成从数据治理到业务价值转化的闭环技术支撑体系。

技术实现特点

1. 分布式数据湖结构：数据融合与价值挖掘的基石

分布式数据湖架构在数智尽调平台中处于关键地位，它搭建起跨平台、多模态的数据治理中枢，为全域数据融合与价值挖掘创造了条件。通过引入 CDC (Change Data Capture) 技术和高可用部署体系，平台能够实时同步行内业务数据，确保数据的时效性与准确性，让银行在尽调时获取的信息始终贴近企业运营实际情况。

从数据存储层面看，分布式数据湖采用混合存储策略。对于结构化数据，如企业交易流水，平台运用列式压缩存储和增量更新技术，既减少了存储空间占用，又能快速处理新增数据，提升数据查询和分析效率。对于非结构化数据，像抵押物扫描件、尽调录音等，平台实施对象存储并进行元数据标签化处理，借助 NLP 技术将尽调笔记转化为可检索的知识节点，便于后续深度挖掘和利用。

凭借统一的数据目录服务和细粒度权限控制机制，分布式数据湖不仅满足了银行合规审计的要求，还为智能信贷决策提供了坚实的数据支撑。在实时风险预警场景中，系统能依据整合后的数据，迅速发现企业异常交易行为；在关联交易穿透分析时，可清晰呈现复杂的企业关联关系，助力银行全面把控风险。

2. NLP 技术：从文本到洞察的桥梁

自然语言处理 (NLP) 技术赋予数智尽调平台理解和处理自然语言的能力，实现从形式化文本分析到语义深层次理解的跨越。在词法分析环节，平台借助 jieba 工具进行分词、词性标注和命名实体识别，精准提取文本中的关键信息。例如，从企业新闻报道中识别出企业名称、相关人物、事件等实体。句法解析阶段，利用隐马尔可夫 HMM 模型等传统算法和 CKY 算法，平台对句子结构进行拆解分析，理解文本的语法结构和语义关系，为后续的语义理解和逻辑推理奠定了基础。通过预训练语言模型 Masked Language Modeling 和逻辑推理框架算法 Neural Theorem Prover，平台实现动态语义理解和篇章结构推理，能够深入解读文本背后的含义和意图。

在数智尽调中，NLP 技术与语音识别、语音合成、视觉问答等多模态技术协同工作，对企业司法、舆情等非结构化数据进行情感解析时，NLP 技术可判断出文本的情感倾向，帮助银行发现潜在风险点。若大量舆情信息对企业持负面评价，可能暗示企业存在声誉风险，提醒银行在尽调时重点关注。

3. 知识图谱技术：构建企业关系全景图

知识图谱以图形化的方式呈现企业实体、事件及其之间的关系，为数智尽调平台提供了一种全新的数据分析和理解方式。我行建立的图分析平台，当前采用 ArangoDB 数据库，兼有 key/value 键/值对、graph 图和 document 文档数据模型，基于其本地集成的多模型特性，搭建高性能程序，支持灵活的可视化效果配置，包括实体、边的可视化效果展示，可支持 K 层展开、最短路径、全路径，也可通过自定义图查询语句进行图查询的 GQL 功能。

借助图分析平台的标准化、自动化、灵活化，搭建风控应用场景，充分发挥知识图谱的价值。基于知识图谱技术，平台深度挖掘企业全方位的关联关系，包括股权、人事、担保、资金交易等多种关系类型，并穿透式地分析实际控制人、一致行动人等关键关系。有助于全面了解企业的背景和潜在风险。在评估企业信用风险时，若发现企业与高风险企业存在紧密的股权关联或频繁的资金往来，可提前做好风险防范措施。

通过聚类、分类等手段，平台对全量企业客户进行分群和群体特征分析，挖掘企业客户群体风险形态，包括

集团派系识别及风险洞察、担保圈链识别及风险洞察、黑名单关联族谱挖掘及风险洞察、信贷资金用途监测、企业风险传导监测等。在担保圈链识别及风险洞察场景中，平台能清晰展示担保圈内企业的相互关系，预测风险在圈内的传导路径，辅助评估关联授信整体风险，提前制定风险应对策略。

4. 人工智能技术：智能决策与风险预判的引擎

人工智能技术贯穿于数智尽调平台的各个环节，为智能决策和风险预判提供了强大动力。基于多线性回归算法和半监督机器学习等技术，平台综合分析企业财务状况、宏观经济指标、行业舆情等多维数据，结合业务专家经验，量化不同行业景气度指数，推导产业链上下游行业风险传导的相关性。这使银行能够提前预判行业风险，调整信贷策略，优化资源配置。

在财务分析方面，人工智能技术通过分析企业财务报表勾稽关系、数据变化情况，并与大量行外财务数据样本交叉验证。并创新性地采用动机分析识别方法，在基于财报完整性、平衡性及波动性分析等自下向上的传统方法基础上，从企业造假动机出发自上向下建立指标联动，构建财务分析模型。该模型能有效识别企业财务造假行为，提高风险评估的准确性。

四、运营情况

数智尽调平台上线以来，取得了显著的成效。已覆盖全行最高额授信、普惠科企贷等多个对公产品，服务全行多家分支机构，助力对公业务实现上亿元的信贷投放。

在优化信贷业务流程方面，数智尽调功能通过同步业务流程、线上智能生成尽调报告，解决了以往尽职调查靠手工耗时费力等痛点。智能生成尽调报告上百份，平均报告自动化完成率高达 60%，为客户经理提效减负，为评审与贷后管理提质增效，同时大幅提升了企业融资效率和客户体验。

在提升尽调质量方面，通过对企业所在行业的发展情况、企业资质、财务情况的智能分析，辅助客户经理提供更加全面和准确的企业尽调报告，为后续的审批决策打下坚实的基础，提升评审人员贷中审核效率，提升信贷风险管理质量。

在支持新兴产业发展方面，数智尽调平台辅助我行将信贷资金投向新兴产业，从金融供给侧有效助力地方重点产业高质量发展。例如，在支持某新兴科技企业的过程中，平台通过对企业的技术创新能力、市场竞争力、财务状况等进行全面评估，为银行提供了准确的风险评估和授信建议，帮助企业获得了所需的信贷资金，推动了企业的快速发展。

五、项目成效

1. 经济效益

一是减少数据处理时间。传统尽调模式下，数据处理工作繁琐且耗时，业务人员需要花费大量时间收集、整理和分析数据。数智尽调平台通过高性能的数据处理引擎，实现了数据的自动化采集、清洗和分析，大大缩短了数据处理时间。平台能够同时分析多条数据流，快速生成数据分析结果，使业务人员能够在最短时间内获取全面、准确的企业信息，为决策提供有力支持。例如，在进行一笔大额信贷业务的尽职调查时，传统方式可能需要数天甚至数周的时间来收集和分析数据，而数智尽调平台可以在几个小时内完成数据处理工作，大大提高了工作效率。

二是提高风险识别能力。数智尽调平台借助先进的技术手段，能够从海量数据中精准识别风险点，为业务人员提供全面、深入的风险分析报告。知识图谱技术和人工智能技术的应用，使平台能够挖掘出企业之间隐藏的关联关系和潜在风险，提高风险识别的准确性和及时性。业务人员可以根据平台提供的风险提示，提前制定风险防范措施，



降低信贷风险。例如，在识别企业的欺诈风险时，平台可以通过分析企业的交易流水、关联方关系、财务数据等多维度信息，利用机器学习算法构建欺诈风险模型，及时发现异常交易和潜在欺诈行为，为银行资金安全提供保障。

三是简化报告撰写流程。传统的尽调报告撰写工作需要业务人员耗费大量时间和精力，且报告质量容易受到个人经验和能力的影响。数智尽调平台通过自然语言处理技术，自动生成结构清晰、逻辑连贯的尽调报告。平台根据预设的报告模板，将分析结果转化为规范的报告内容，涵盖企业经营情况分析、所在行业分析、财务分析、风险缓释措施等多个方面，并支持将报告输出为 Word 格式，方便用户进行保存和编辑。业务人员只需对报告进行审核和补充，即可完成报告撰写工作，大大简化了报告撰写流程，提高了工作效率和报告质量。

2. 社会效益

一是推动实体经济发展，通过构建本平台，促使银行不断创新金融产品和服务模式，开发出更符合市场需求的个性化产品使得金融服务精准支持实体经济发展的重点领域和薄弱环节，为实体经济发展提质增效。二是提高社会资源分配效率，通过海量数据分析，精准评估企业价值和风险，使银行的资金能够更精准地流向优质企业和项目。有助于提高社会资金的使用效率，避免资金错配，促进资源向更有潜力和价值的领域流动。

六、经验总结

1. 项目建设经验

从项目建设流程来看，可在需求顶层设计、技术架构选型、数据治理三个方面总结经验。在需求设计时，应当重点关注痛点精准识别，通过深度调研一线业务人员需求，明确尽调场景核心痛点，包括多元数据整合困难、报告撰写耗时、风险识别遗漏等问题。在项目管理方面，应注意分阶段规划目标，形成“短期试点 - 中期扩展 - 长期深化”的实施路径。在技术架构和选型方面，可采用“大模型 + 小模型”的组合方案，根据尽调模块特征，选择能力适配的模型。在数据治理及安全保障方面，根据调查报告撰写要素构建结构化知识库，同时通过联邦学习、动态脱敏等技术确保数据不出行，保障数据安全。

2. 推广经验

在应用推广方面，可构建场景验证和敏捷迭代的推广策略。场景化试点方面，在产品选择上，可优先选择标准化程度高的业务场景进行试点，如小微企业贷款。在试点机构选择上，可优先选择业务量大、具有一定数字化转型思维的机构展开试点，收集足量样本，为后续优化迭代做准备。在敏捷迭代方面，构建快速反馈优化机制。以两周为周期收集用户反馈，持续迭代系统功能，保证系统正常运行。

中邮保险：基于“智能生态 + 大模型”体系下的操作风险智能联防生态搭建

一、项目背景及目标

近年来，国家金融监管总局持续强化操作风险管理要求，明确提出金融机构需深化风险工具联动应用，实现“风险识别 - 评估 - 监测 - 整改”全链条闭环管理。《银行保险机构操作风险管理办法》强调，应通过整合 RCSA（风险与控制自评）、KRI（关键风险指标）、LDC（损失数据收集）三大工具，构建多维风险视图，提升风险防控的协同性与前瞻性。

中邮保险基于监管导向与行业趋势，针对传统风险管理中存在的孤岛、基层触达弱、响应滞后等痛点，启动“智能生态 + 大模型”数字化闭环体系建设。以智能风险管理（IRC）系统为核心载体，响应监管对“工具联动、穿透管理”的要求，率先探索“AI 大模型 + 风险工具联动”模式，着力解决三大核心问题。一是风险工具割裂，RCSA、KRI、LDC 独立运行，数据无法交叉验证，风险信号碎片化；二是基层风险盲区，总省两级监控难以覆盖地市、县区及网点末梢，风险识别颗粒度不足；三是处置效率低下，人工依赖度高，风险预警与整改脱节，难以满足监管“T+1”时效要求。

二、创新点

本成果通过“技术重构 + 管理革新”双轮驱动，攻克保险业数据孤岛、基层触达滞后、动态响应薄弱三大痛点，以“大模型 + 智能生态”为核心，构建 RCSA-KRI-LDC 立体化联动模型，实现风险防控能力跃升与行业范式升级，兼具创新性、实用性和可推广性。

1. 技术创新

首创“智能生态 + 大模型”协同模式。突破传统风控技术工具化局限，首次将 AI 大模型与风险监测、预警、评估等系统深度融合，构建覆盖预警、评级、偿付能力管理等全场景智能生态体系，实现风险管理从“被动响应”向“多维前置管控”跃迁。研发保险风控领域专用生成式 AI 技术，搭建风险管理知识库，实现海量数据秒级精准查询，突破传统人工查询效率瓶颈，显著提升数据应用价值。

2. 管理机制创新

T+1 动态风控网与五级监测体系。创新构建覆盖“总 - 省 - 地市 - 市县 - 网点”五级监测网络，实现 36 个 KRI 指标 T+1 时效预警全国覆盖，结合数据标签化与可视化技术，为基层提供实时风险画像，将监管“早识别、早预警”要求转化为量化操作标准，推动风险治理颗粒度下沉至业务末梢。

3. 应用效能提升

通过构建 RCSA-KRI-LDC 立体化联动模型，通过预设阈值与事件关联规则，实现数据交叉验证与风险信号智能

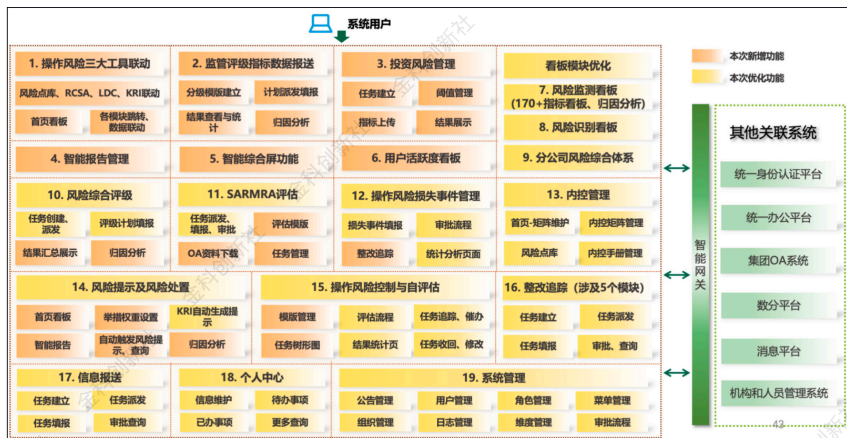


推送，满足监管对工具协同的要求；依托动态风险指标体系，将 KRI 监测延伸至网点层级，确保五级机构风险偏好一致性。集成 AI 大模型与专用知识库，实现风险数据秒级定位、归因解析及处置建议生成，推动风险处置从“人工研判”向“智能决策”升级；搭建“总-省-地市-市县-网点”五级穿透式监测网络，风险识别颗粒度细化至末梢流程，提前拦截潜在风险。通过联动看板实时展示工具交叉验证结果，支持风险点分类管理与流程地图一键生成，实现可视化穿透分析；建立“识别-评估-预警-整改-报告”全流程闭环，整改响应时效压缩至 24 小时内，推动风险管理从“分散式事后处置”向“集约化前置管控”转型，对标行业领先实践。

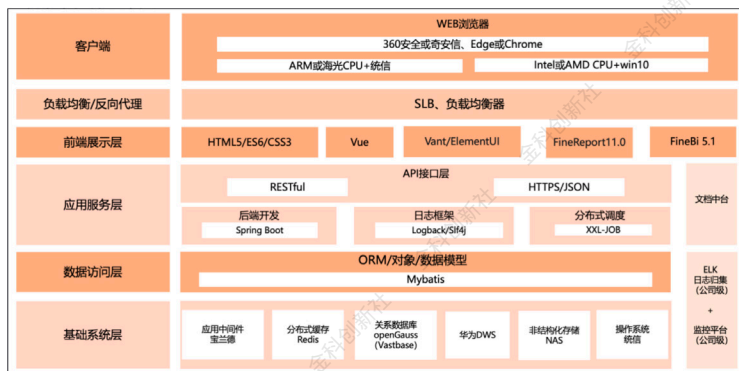
三、项目技术方案

智能风险管理（IRC）系统采用分层解耦架构设计理念，构建起功能架构-技术架构-部署架构的三维协同体系。在功能架构层面，通过业务模块化设计与标准化接口封装，实现风险识别-评估-监测-整改的全流程闭环管理；技术架构依托微服务框架与分布式计算引擎，保障实时决策引擎的高并发处理能力；部署架构采用容器化编排与自动化运维平台，支持混合云环境下的弹性伸缩与灰度发布。通过架构间的分层治理与松耦合集成，最终达成应用层人机交互敏捷高效、服务层业务逻辑精准执行、基础设施层资源调度稳健可靠的立体化建设目标，为金融机构实现风险防控数字化转型提供全栈式技术底座。

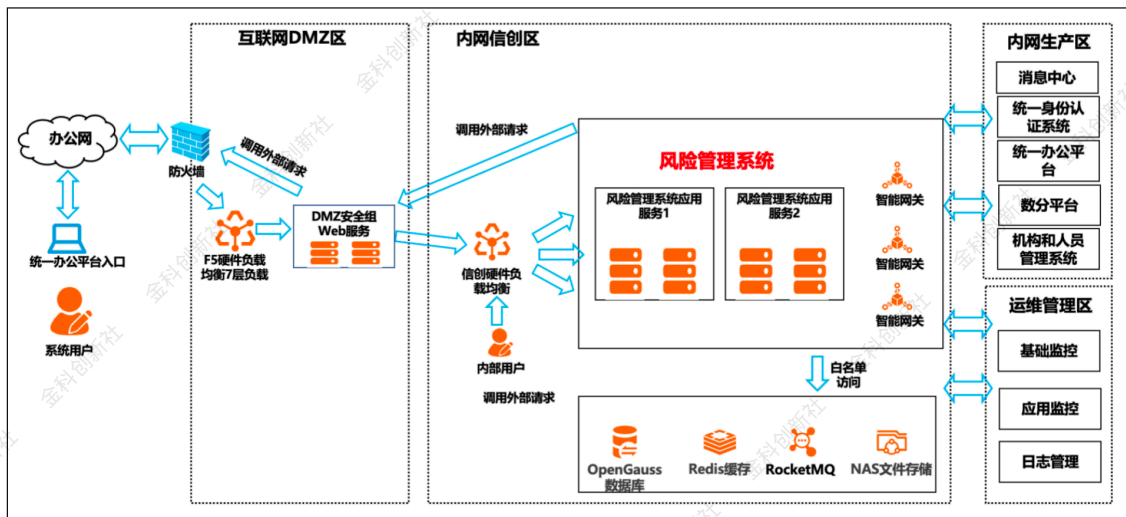
1. 功能架构



2. 技术架构



3. 部署架构



四、运营情况

系统已推广面向中邮保险总省两级近 3000 人使用，并进行系统功能使用培训。对用户权限进行分级管理，不同管理层次用户对应不同查看与使用权限，在确保系统数据安全前提下，实现风险管理信息精准推送与分级赋能，使得各层级能够迅速、有效的接收到最全面的风险信息并及时处置 / 缓释风险。

构建五级监测体系，KRI 指标 T+1 预警覆盖全国网点，风险识别效率提升 200%；LDC 事件全流程线上化，智能表单替代人工填报，数据收集效率提升 150%；RCSA 评估自动关联风险点与责任部门，整改任务一键生成，操作风险评估效率提升 120%。通过标签化、可视化与自动化联动，实现“监测 - 处置 - 优化”闭环提速，推动管理下沉与业务高效运转。通过数据穿透、智能预警与动态决策，推动风险管理从“被动响应”向“主动防控”转型。

五、项目成效

1. 全流程数字化闭环：系统性重塑行业风控标准

通过自主研发深度融合风险管理工作需求，打造了覆盖风险识别、评估、预警、处置、整改的完整数字化闭环。该体系突破性地解决了保险业长期存在的“线下手工操作 - 线上碎片化管理”割裂问题，实现风险事件上报时效从 72 小时压缩至 12 小时。特别是在操作风险应对领域，系统通过规则模型的实时查询，填补了系统性线上风控的空白。完成风险管理“大模型 + 智能生态”的协同模式落地，实现对企业发展过程中面临的风险隐患进行多维前置管控。

2. 数据穿透可视化：重构风险管理空间维度

构建了行业“总 - 省 - 地市 - 市县 - 网点”五级穿透式监测网络，突破传统机构层级的平面化监控模式，将风险定位精度提升至网节点级颗粒度。配套开发的“机构风险全景图”等可视化看板，运用动态流数据处理技术，实现各机构层级、上万个网点的实时风险态势投射，支持管理者在短时间内完成区域性风险点的有效定位，推动行业



风险治理进入“细胞级”管理时代，进一步提升风险管理智能支撑水平。监测数据时效实现 T+1，网省级 KRI 指标达 36 个，赋能省分机构实现风险“早识别、早干预”。

3. 生态化平台集成：打造智能风控超级枢纽

成功实现与整合风险预警（KRI）、操作风险与控制自评估（RCSA）、操作风险管控（LDC）等 12 个核心子系统的深度整合，构建了日均处理量达百万条的智能风控中枢。率先探索 AI 生成式大模型在风险垂直领域的应用，突破传统风控数据孤岛瓶颈，该项技术在行业处于先进水平，以“数据直达、智能直达”的实践成果，研发的“风控认知智能引擎”实现语义解析准确率达 90% 以上，风险决策响应速度提升至秒级，标志着保险业正式进入风险认知智能决策阶段。

六、经验总结

中邮保险“智能生态 + 大模型”操作风险联防项目的成功实践，为金融业风险管理数字化转型提供了可复用的方法论：一是以“监管导向 + 痛点攻坚”双轮驱动，精准定位数据孤岛、基层触达弱、响应滞后三大行业痛点，通过技术重构与管理革新协同发力，构建覆盖“总 - 省 - 地市 - 市县 - 网点”五级穿透式风控网络，将监管要求转化为量化操作标准；二是“分阶段 + 模块化”实施策略，采用瀑布模型与敏捷开发融合的交付模式，通过需求优先级矩阵与 DevOps 持续集成，严控技术债务的同时实现核心功能快速上线；三是“用户中心 + 生态协同”设计理念，以基层需求为导向，通过 AI 大模型与 RCSA-KRI-LDC 联动模型，实现风险数据秒级定位，推动风险处置从人工研判向智能决策升级；四是“数据直达 + 分级赋能”运营机制，构建五级权限管理体系与可视化穿透看板，通过标签化、自动化技术实现风险信息精准推送，保障总省至末梢机构风险偏好一致性。项目验证了“技术 + 管理 + 生态”三位一体的风控范式升级路径，为行业提供了从工具协同、流程重塑到智能跃迁的完整解决方案。

天津银行：“AI 合规官”数智化项目

一、项目背景及目标

十八届三中全会以来，党中央在以习近平同志为核心的党中央团结带领全党全军全国各族人民，持续从社会主义市场经济体制、经济高质量发展体制、创新体制机制、经济治理体系等多方面通过建立和健全制度体系、进一步深化各项改革措施。为此，针对金融行业，党中央和国家监管机构也在各类管理措施中，明确加强了银行金融机构对内控合规制度管理体系的要求：“及时、动态地将监管规定转化为内部规章制度”、“要把合规性审核作为制定或修订内部重要制度的必经程序”、“要强化管理制度化、制度流程化、流程信息化的内控理念，将各项业务制度的合规管理要求嵌入业务流程中”。2024 年的政府工作报告中也明确提出“深化大数据、人工智能等研发应用，开展‘人工智能+’行动”。国家金融监督管理总局令 2024 年第 7 号公布了《金融机构合规管理办法》，将于 2025 年 3 月 1 日正式实施。此举标志着我国金融业合规管理迈入了一个新的阶段，具有里程碑意义。

面对党中央、监管机构持续加强针对金融机构的风险整治和管理力度，银行机构在合规管理方面面临巨大挑战，合规成本投入不断增加。当前信息时代正在加快进入银行合规内控的发展阶段，伴随着以大数据、人工智能(AI)、ChatGPT 等为代表的信息技术的高速发展，为合规内控管理提供了新的手段与方法，人工智能技术上的突破层出不穷，逐渐深入地赋能千行百业，推动人工智能要素已成为新质生产力的典型代表。

在此背景下，天津银行通过引入和合信诺“AI 合规官”领先解决方案，打造横向到边、纵向到底的数智化监管合规管理体系，将合规基因注入天津银行发展决策、业务经营的全过程、全领域。

例如，在外规监管制度跟踪过程中，通过“AI 合规官”可以实现监管法规政策条款的自动化拆解，完成对制度层级以及条款层级的自动化打标签，从而实现监管政策解读以及要求从文本内容向条款化、结构化转化的过程。此外，可按照最新的监管法规政策要求，自动化沉淀一套监管的红线合规要求，突破了以前需要人工整理监管要求的做法，为后续开展监管检查提供了精准的相关检查线索，大幅提升了监管合规管理效率。

此外，天津银行“AI 合规官”解决方案，搭载了全球领先 copilot 智能合规 AI 助手，可提供智能合规问答、法规政策智能问答、制度规范性和一致性智能审查、多知识库智能问答、长报告 / 法规政策快速解读等诸多实用和创新功能，大幅提升合规管理工作的效率、降低运营成本，真正达成降本增效的业务目标和数智化监管合规的高质量发展目标。

二、创新点

1. 全栈支持信创，架构先进、可拓展性、兼容性好、适用性、稳定性、安全性强；
2. 国内首个基于 DeepSeek 等大模型技术的创新合规管理工具，助力天津银行实现合规管理效率的大幅提升；
3. 方案对齐监管规则和要求，稀缺性价值高，以数智驱动管理思路，形成超越流程管理的领先数智化监管合规管理；
4. 用户交互体验更友好，利用智能化技术手段大幅降低了用户使用的难度，让系统更好用，用户愿意用；
5. 为一线用户减负赋能，提供系列 AI 合规助手和工具，实现智能化辅助支持高效工作。

三、项目技术方案

“AI 合规官”的设计和规划与天津银行战略重点、技术升级和合规管理流程紧密联系，通过以 DeepSeek 大模型一体机为核心底座、软硬件的深度融合与优化，旨在打造以“监管法规”为核心，以“数智化”为手段、以畅通“制



度传导”为主要目的“合规内控生态圈”，通过引入大语言模型及 AI 技术，对接内外部知识库、数据库，实现自然语言交互实时问答及回复监管合规领域相关问题等，具备高效性、准确性、个性化、安全性以及低成本特点。

ID 四、运营情况

该项目上线后在我行得到全面推广应用，覆盖天津银行各相关业务部门。运行期间系统稳定，未出现重大故障，为用户提供高效、可靠的监管政策法规咨询和分析服务，有效支持业务开展，得到了全行用户广泛认可。对于银行业内其他正在寻求转型或准备转型的机构具有很大的启发和借鉴意义。

ID 五、项目成效

这一创新举措体现了数智化合规管理对业务发展的全面支持，通过迭代传统合规管理手段和做法，构建科学、智能、有效的合规管理架构和 AI 工具，实时跟踪监管政策变化，并赋能一线业务，有效提升数智化水平及工作效率、提升合规风险监测、预警及治理能力，确保天津银行合规管理始终符合最新监管要求，加快实现从“被动监管遵循”向“主动合规治理”的转变，推动天津银行法治建设和高质量发展。

ID 六、经验总结

通过对天津银行“AI 合规官”项目的跟踪、复盘和总结，有如下经验在后续建设中为同业机构建议参考借鉴：

1. 加强企业合规管理，提高依法合规经营水平不是单一搭建数字化、智能化信息系统和工具这么简单。它是一项企业级工作，需要整体布局规划，包括业务流程的优化、业务系统的管控改造、全面的数据治理。高层真正的重视是成功关键。
2. 数智化合规管理项目不能烟囱式建设，一定要打通底层数据，这样才能够做到口径统一，数出一致，也为未来灵活扩展和进一步深化在 AI+ 法律、AI+ 合规、AI+ 内控、AI+ 风险、AI+ 审计等高价值及复杂场景的应用奠定基础。
3. 合规治理数智化不是一蹴而就的，它是个持续工程，需要长期坚持不懈。坚持以用提质，数据的质量是在高频的使用中得到提升的。

在数智化金融监管趋势下，天津银行将持续加强合规管理体系和能力建设，不断提升监管合规管理质量和效率，深化锻造数智化能力，赋能公司业务稳定、健康和高质量发展，为金融强国建设贡献天津力量。

潍坊银行：基于大模型和 RAG 驱动的智慧合规助手

一、项目背景及目标

在金融监管政策持续加码、合规要求日益精细化的背景下，监管机构明确将数字化转型作为提升合规管理质效的核心路径。国家金融监督管理总局 2025 年 3 月施行的《金融机构合规管理办法》强调“充分运用数字化、智能化手段提升合规管理的实时性与精准性”，人民银行《金融科技发展规划（2022-2025 年）》明确提出“探索人工智能、大数据、区块链等技术在风险防控、合规管理中的深度应用，构建智能化风控模型与动态合规监测体系”。面对银行业制度库中海量制度条文，传统检索因效率低、时效性差、解读碎片化等问题，已难以满足“实时响应、精准识别”的要求。在此政策框架下，建设智能化合规工具成为金融机构应对监管趋严与业务敏捷发展双重挑战的必然选择。

随着潍坊银行数字化业务场景的持续拓展，高频次的监管政策迭代与复杂化的业务合规需求，对传统合规服务模式形成显著挑战。业务部门对制度查询的实时性、多维度关联解读需求日益迫切，而传统查询模式效率、人力瓶颈逐步凸显。在此背景下，潍坊银行聚焦合规管理数字化升级，依托大模型技术的语义理解能力与 RAG 技术的知识检索优势，打造“智慧合规助手”，构建“检索 - 生成 - 校验”一体化智能引擎。通过精准匹配制度条文与智能化解读输出，在提升合规响应效率的同时强化风险管控能力，推动合规服务向精准化、标准化、可溯化的新模式升级。

二、创新点

在本次基于大模型和 RAG 驱动的智慧合规助手建设中，我们通过创新的方式实现了合规服务的智能化升级。以下为具体创新点描述：

1. 技术融合创新

基于行内大模型，深度融合 RAG 技术实现制度库的动态检索与智能问答，有效突破传统检索模式效率瓶颈。通过自建政策知识库的动态维护机制，实时整合最新制度文本，采用“向量检索 + 语义增强”双引擎技术，将非结构化制度文本转化为高精度语义向量，在问答过程中实时注入最新制度内容，既保留了大模型的自然语言理解优势，又通过结构化知识锚定有效抑制模型幻觉问题。系统特别强化了多轮对话与上下文关联分析能力，可精准解析复杂业务场景中的隐含合规诉求，例如跨章节制度关联、模糊条款解释等场景，确保输出的每个合规结论都严格遵循现行有效制度规范。

2. 交互优化与效能监测

系统提供多维度交互分析看板，支持实时查看问答日志、标注关键会话案例，并内置性能监测模块统计日活用户量、平均响应时长等核心指标。管理人员可通过可视化界面分析高频咨询话题分布，识别政策盲区或表述模糊条款，手动调整模型检索权重或补充标注数据。通过会话日志记录高频咨询要点，为人工优化模型提供数据支撑。针对复杂政策咨询场景，系统提供人工标注工具，支持对模型输出的条款匹配度进行分级标记，沉淀的标注数据集将用于定向优化检索模型效果。



3. 智能辅助与精准推荐

系统创新设计政策关联推荐功能，在回答用户提问时自动呈现相关度较高的关联条款，推荐依据来源于历史会话的共现分析和政策文本的语义关联。对于高频咨询的共性政策问题，支持管理人员手动配置标准解读模板，当检测到同类问题时优先调用预设解读方案，提升响应一致性与准确性。通过分析用户历史咨询记录和岗位特征，主动推送关联度高的制度条款和典型案例，实现合规服务从“人找知识”向“知识找人”的范式转变。

三、项目技术方案

该项目上线后在我行得到全面推广应用，覆盖天津银行各相关业务部门。运行期间系统稳定，未出现重大故障，为用户提供高效、可靠的监管政策法规咨询和分析服务，有效支持业务开展，得到了全行用户广泛认可。对于银行业内其他正在寻求转型或准备转型的机构具有很大的启发和借鉴意义。

1. 项目规划

智慧合规助手建设遵循“技术筑基 - 流程优化 - 服务升级”的推进路径，分阶段完善合规智能服务能力。

一阶段完成基础能力搭建，自建政策知识库，部署 RAG 引擎实现政策文本向量化存储与基础问答功能，开发会话日志记录系统与基础性能看板（日活用户、响应时长等指标），在业务部门开展试点并建立人工标注优化流程。

二阶段重点优化服务流程，扩展多轮对话理解能力，开发高频问题统计模块与人工标注工具，支持管理人员根据日志分析调整检索权重，在试点部门实现制度匹配准确率量化评估，建立人工标注数据反哺模型优化的标准流程。

三阶段推进服务闭环建设，基于历史会话日志构建常见问题知识沉淀库，开发简单版条款关联推荐功能，完善覆盖检索准确性、响应速度、人工干预率的多维度评估体系，形成“问题记录 - 标注优化 - 模型迭代”的可持续改进机制。

2. 建设方案

2.1 平台业务架构

潍坊银行智慧合规助手是以大模型与 RAG 技术为核心，结合银行业合规管理需求构建的智能问答服务平台。平台集成制度动态检索、多轮对话理解、人工标注优化、效能监测等核心模块，自建知识库精准检索关联制度条文，支持政策文本向量化存储与语义检索，确保合规回答的准确性与时效性。基于“大模型 + 人工校验”的双重机制，建立覆盖问答全流程的闭环管理，包括会话日志记录、高频问题分析、人工标注反馈等环节，形成“检索 - 响应 - 优化”的可持续改进体系，助力合规服务从传统人工查询向智能辅助决策转型，提升合规响应效率与风险防控能力。

2.2 平台技术架构

智慧合规助手采用轻量化微服务架构，实现大模型推理与 RAG 引擎的融合部署，集成政策文本向量化引擎、多轮对话管理模块、人工标注工具及效能分析看板。内置会话日志系统记录用户咨询详情与模型响应数据，支持人工标注关键案例并反哺检索模型优化。质量管理方面，构建覆盖制度匹配准确率、响应时效、人工干预率的多维度评估指标，结合可视化看板动态监测日活用户量、会话峰值等运行状态。同时，与行内日志管理系统、用户权限管理体系无缝集成，确保服务合规性与数据安全性，为全行提供高效、可控的智能合规支持。

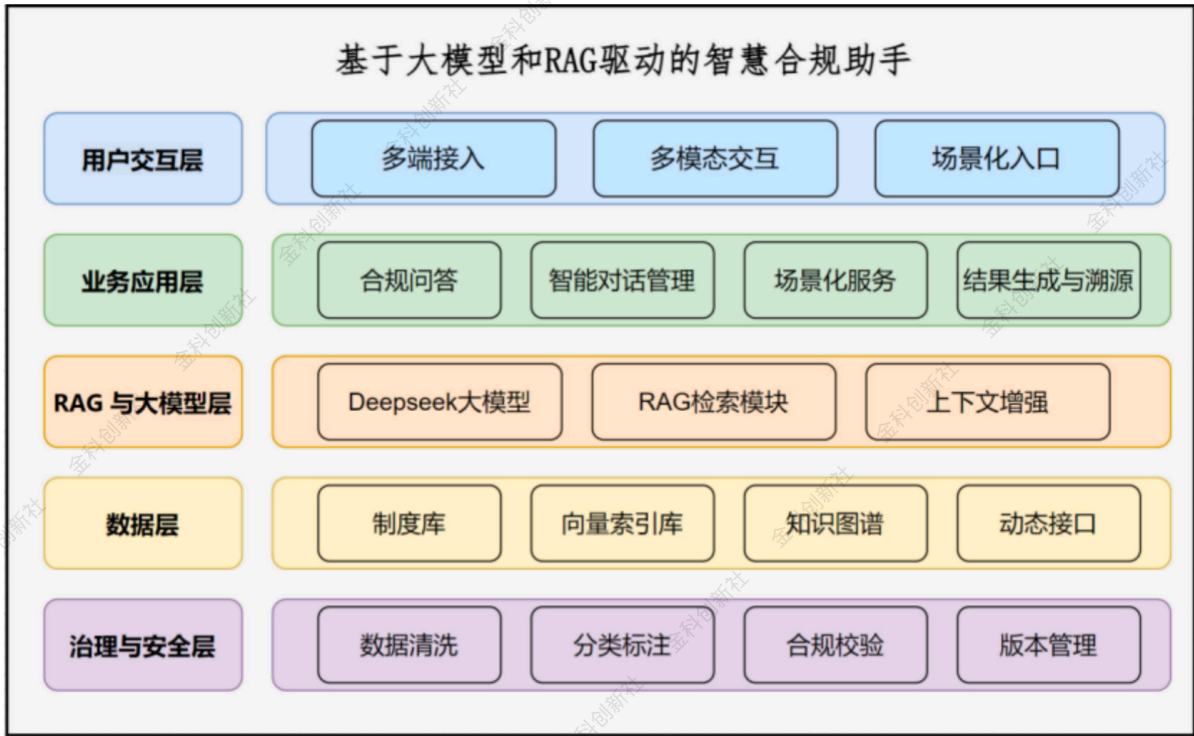


图 1. 平台应用架构

四、运营情况

通过建设智慧合规助手，将制度检索、智能问答与服务优化流程深度融合，实现合规咨询服务的标准化管理与智能化升级。目前，该平台已覆盖多个主要业务条线的日常合规需求，支持复杂场景下的多轮对话交互与政策精准匹配，显著缩短人工查询制度的时间成本。通过持续优化语义理解算法与人工标注反馈机制，有效降低条款误读风险，提升合规应答的准确性与权威性。系统运行期间积累的对话日志与用户行为数据，为识别制度盲区、优化服务流程提供实证依据，逐步形成“智能响应-人工校验-知识迭代”的良性循环机制，为全行合规管理数字化转型奠定实践基础。

五、项目成效

1. 经济效益

(1) 提升合规服务效率。通过智能问答与精准检索技术，显著缩短人工查询制度的时间成本，减少重复性工作投入，加速业务决策流程。

(2) 优化人力资源配置。自动化问答服务与高频问题沉淀功能，释放合规人员基础查询压力，使其聚焦于复杂场景分析与风险研判，提升专业价值产出。

(3) 增强业务敏捷性。通过快速响应制度咨询需求，支持业务部门及时把握合规边界，缩短新产品、新业务的合规评估周期，助力业务创新落地。



2. 社会效益

(1) 推动合规管理数字化转型。构建智能化合规服务模式，为银行业探索人机协同的合规管理体系提供实践范例，促进行业合规服务标准化进程。

(2) 强化风险防控能力。通过精准解读与动态更新的制度服务，提升全行合规执行力，增强金融业务规范性，助力维护区域金融稳定。

(3) 培养“业务+技术”复合人才。系统建设与优化过程中，推动合规团队掌握智能工具应用与数据分析能力，培育兼具制度解读与技术思维的复合型人才梯队。

六、经验总结

1. 制度为本

项目启动阶段需立足合规管理本质，优先建立“制度解读标准化”体系。通过全面梳理行内制度库结构与业务场景关联性，明确智能问答服务边界与责任归属，制定标注标准、响应分级、人工校验等全流程规范。重点固化制度检索优先级规则与模糊条款处置机制，确保智能应答既符合技术逻辑，又严格遵循行内合规管理要求，为服务可控性奠定基础。

2. 渐进拓展

推广策略需兼顾服务价值与实施可行性，分层次推进能力覆盖：优先选择业务影响面广、技术实现复杂度低的场景建立服务标杆，通过标杆案例沉淀标准化实施模板；随后延伸至低频但规则明确的场景，依托既有经验加速复制；对跨系统协同、高复杂度场景，采用“业务专家+技术团队”联合攻坚模式，逐步突破技术瓶颈；针对投入产出比低的边缘场景，建立动态评估机制，聚焦资源优先保障核心服务能力迭代。

3. 数据驱动

构建覆盖问答全链路的量化评估体系，持续追踪制度匹配准确率、响应时效、人工干预率等核心指标，结合满意度反馈与标注数据分析服务短板。通过高频问题聚类识别制度盲区，利用误判案例逆向优化语义检索模型，形成“数据采集-问题归因-策略优化”的闭环管理。定期输出服务效能报告，为制度库更新、功能迭代提供决策依据，推动智能合规服务从经验驱动向数据驱动的精细化运营转型。

青岛银行：“法眼智控” 网络司法查控 AI 管理体系

一、项目背景及目标

近年来，网络查控业务量迅猛增长，传统人工审核模式在处理效率与操作风险防控方面已难以满足实际需求。同时，业务的快速发展也对事后质检覆盖度和实时风险干预能力提出了更高要求。为应对挑战，青岛银行立足“数智化运营”体系建设，依托人工智能技术，打造集实时监测、智能审核、精准质检于一体的网络司法查控 AI 管理体系。项目旨在通过科技赋能，全面提升司法查控业务的处理效能与风险防控水平，构建“覆盖全流程、管控全链条、响应全时段”的新型业务模式，为有权机关高效、合规开展司法查控工作提供坚实服务支撑。

二、创新点

风控模式革新：建立“法眼洞察”实时异常监测机制，覆盖 40 余类异常场景，7×24 小时实时监测系统运行与业务流程，精准识别异常情况并主动预警，借助短信、系统通知等多渠道，确保风险信息第一时间触达责任人，大幅提升应急响应效率，推动风控模式向“主动干预”转变。

审核技术突破：“智脑会审”依托“智能 OCR+DeepSeek”双引擎，实现从精准文书识别、智能信息提取、深度内容分析到报文自动校验的全链条闭环处理，全面革新文书审核模式，审核效率提升 97%，人工干预需求大幅降低。

质检体系升级：全面整合多源数据，借助自动化质检程序对业务进行“全控严检”，将传统人工抽检模式升级为 100% 全量自动化检测，结合“机审+人核”双轨机制，质检效率提升 93%，升级重构了网络司法查控业务处理的事后监督模式。

三、项目技术方案

“法眼洞察”智能监控机制：为全面提升网络司法查控业务的安全性与稳定性，搭建“系统+业务”双维度监控网络。系统层面，“法眼洞察”对各接入渠道的服务状态、接口响应及数据传输等关键环节进行实时监测，精准识别异常，保障平台稳定运行；业务层面，深度融合司法查控业务处理流程，动态追踪关键节点，如账户冻结失败、报文反馈失败等异常情况，实现自动识别、智能预警，并通过短信、系统通知等多渠道高效传递风险信息，显著提升应急响应效率，实现风险“自动识别-主动预警-快速处置”闭环，全面筑牢司法查控业务安全防线。

“智脑会审”AI 审核功能：司法文书审核是金融机构协助有权机关开展司法查控业务处理流程中的关键环节，青岛银行依托“智能 OCR+DeepSeek”双引擎，构建起一套高效、精准、智能的司法文书自动审核体系。借助智能 OCR 技术实现对各类司法文书进行精准识别和内容提取，并利用 Deepseek 自然语言处理和深度语义分析能力，智能校验文书信息的准确性与合规性，形成从精准文书识别、智能信息提取、深度内容分析到报文自动校验的全链条闭环处理机制，提升司法文书处理的智能化水平与风控能力。

“全控严检”业务质检机制：青岛银行全面整合有权查控平台、企业知识平台及核心业务系统等多源异构数据，依托自主研发自动化质检程序，构建“全控严检”高效智能的校验机制。通过深度融合业务规则与监管要求，实现对全量网络司法控制类业务处置结果与反馈信息的自动核验，针对程序标记的异常数据，采用“机器初筛+人工复核”机制，全面重构网络司法查控业务的质检体系，提升质检覆盖面与效率，强化业务数据的完整性、准确性与合规性管理。



四、运营情况

“法眼智控”网络司法查控 AI 管理体系上线以来已取得显著成效。在异常检测方面，已覆盖接口异常、控制处理报错等 40 余类场景，贯穿业务接收、解析、处理及上报各环节，全面嵌入系统运行与业务处理全流程；在 AI 文书审核方面，月均处理量超 2 万笔，审核准确率达 97.8%，人工审核量下降 85%。在智能质检方面，质检覆盖率由原 10% 提升至 100%，质检效率提升 93%，大幅提升了业务处理的高效性与合规性，为网络司法查控业务稳健运行提供了有力保障。

五、项目成效

项目实施后，青岛银行在网络司法查控领域的经济效益与社会效益显著。在经济效益方面，通过构建实时异常检测机制，系统响应与处置能力显著提升，有效防范因处理滞后带来的经济损失；同时，依托智能化体系建设，业务运营效率持续优化，月均处理量突破万笔，审核准确率达 97.8%，节约人力成本约 35 人。在社会效益方面，“三位一体”智能支撑体系的建立，显著缩短了业务响应周期，全面提升了服务专业度与业务响应效能。项目打造了智能、高效、专业的新型业务模式，为有权机关开展司法查控工作提供了坚实保障，树立了金融科技赋能司法实践的标杆，彰显了青岛银行在金融科技创新方面的示范引领作用。

六、经验总结

当前，金融科技已成为驱动金融创新的核心引擎。本项目是青岛银行首次将 DeepSeek 大模型应用于业务流程中，通过深入调研智能化应用场景，精准识别业务痛点，依托 AI 技术与业务规则的深度融合，成功构建了“事前—事中—事后”全周期风控闭环。项目建设过程中，“业技术”敏捷小组高效协同，在场景梳理、规则制定及系统优化等关键环节发挥了重要作用，显著提升了我行在网络司法查控业务领域的处理能力与响应效率。项目的成功落地，标志着人工智能技术在金融业务流程中的深度应用，充分体现了技术赋能业务的战略价值。同时，跨部门协作机制凝聚了集体智慧，形成了可复制、可推广的“青岛银行方案”，为行业智能化转型提供了有益借鉴。

哈尔滨银行：基于大模型金融机构数智化知识管理系统创新实践

一、项目背景及目标

随着金融行业数字化转型加速，各机构面临海量非结构化文档治理难题。据行业调研显示，头部金融机构年均产生技术文档超 50 万份，但知识利用率不足 35%。传统基于关键词检索方式的知识管理系统存在如下痛点，语义理解能力薄弱导致召回率低下，知识更新滞后造成维护成本高企，敏感数据外泄风险制约智能化进程。本项目创新性地提出“分布式大模型 + 智能知识管理”双轮驱动方案，通过本地化部署满足金融级安全要求，采用 RAG 技术实现动态知识更新，结合深度语义理解突破复杂场景应用瓶颈。技术架构设计重点攻克多源异构文档处理、检索性能优化、基础环境性能及稳定性等核心问题，为行业知识管理范式转型提供可复用的解决方案。

二、创新点

技术创新维度

1. 采用“大模型分布式推理 + 知识管理流水线”融合架构，通过 vLLM 框架实现百亿参数模型在国产化硬件环境的高效集群部署，推理资源利用率达 78%。
2. 研发面向金融文档的动态分块算法，通过对 RagFlow 二次开发改造，使之深度融合 MinerU 智能解析引擎与动态分块策略，构建了面向复杂文档的智能分块体系。针对非常规复杂格式文档的特点，引入 LlamaIndex 框架实现定制化的文档分析和分块，提升分块的合理性及效率。
3. 构建基于多阶段重排序的混合检索体系，通过混合检索双路召回与深度精排模型组合，实现金融领域查询 MRR 指标行业领先。
4. 构建幻觉抑制体系。一方面，为控制生成答案的质量，配置拒绝回答阈值。当生成答案的可信度低于设定阈值时，系统拒绝输出答案，避免生成低质量或不准确的内容。另一方面，采用检索置信度评分、LLM 事实核查双重验证、提示器控制等机制，相较于传统 RAG 方案，幻觉率降低了 42%，有效提升了生成内容的质量。

应用创新维度

1. 打造“知识即服务”新模式，通过 API 网关将各类知识库能力注入 OA、ITIL 等多个业务系统，统一接入与标准化接口，降低平台集成复杂度。
2. 建设智能监报告警系统，基于故障事件智能关联处理方案，提升 ECC 故障排查处置效率。
3. 构建批量作业调度事件智能化发现与处置能力，提升批量作业调度任务故障排查处置效率。
4. 实现了文档数据的统一管理，该平台对文档的整个生命周期进行了系统化管理，不仅提升了文档的组织性和可访问性，还确保了信息的时效性和准确性，提升各类文档在存储、检索方面的便利性。系统拒绝输出答案，避免生成低质量或不准确的内容。另一方面，采用检索置信度评分、LLM 事实核查双重验证、提示器控制等机制，相较于传统 RAG 方案，幻觉率降低了 42%，有效提升了生成内容的质量。

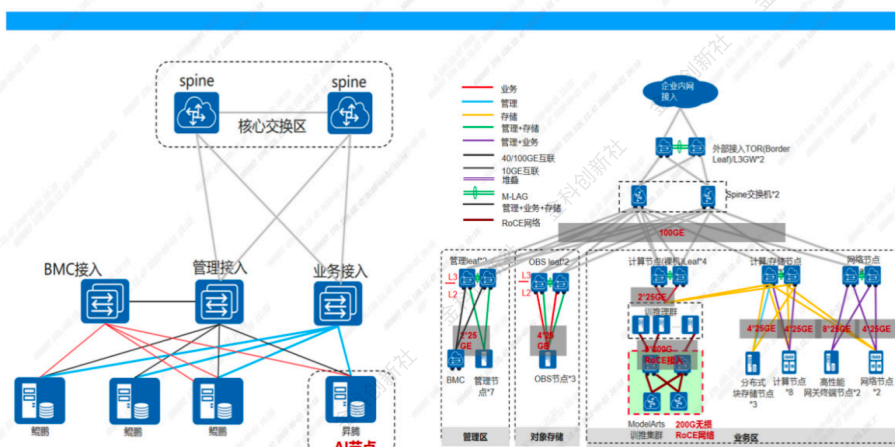


三、项目技术方案

1. 系统架构设计

系统采用分层解耦的微服务架构，构建“数据层-算法层-服务层”三级技术栈。底层基于国产华为昇腾 910B 的模型推理服务集群，依托 vLLM 框架实现 Qwen2.5-32B 大模型的分布式部署，通过张量并行与流水线并行技术，将模型推理延迟控制在 800ms 以内。中台层集成经过二次开发后的 RagFlow 知识库引擎，强化了文档解析能力，可支持超过 20 种格式的文件，涵盖 PDF、Word、Excel、PPT、图片以及影印件等金融文档的自动化处理，日均处理量达 2000+ 文档。服务层通过 Dify 平台实现业务流程可视化编排，构建查询理解-向量检索-重排序-结果生成的完整服务链路。

大模型配置及组网方案



2. 文档解析技术

经实际测试表明，RagFlow 自带的原生 DeepDoc 智能解析引擎在处理多层次嵌套的图文排版及复杂表格结构的 PDF 文档时，存在解析精度不足的问题，难以满足金融行业对非结构化数据的深度语义识别需求，基于此，亟需一种更优质的文档解析算法进行替代。经过对市面上常见开源文档解析引擎调研、分析、论证，最终确定了 RagFlow 集成 MinerU 的文档解析管理方案。RagFlow 通过深度融合 MinerU 智能解析引擎与动态分块策略，构建了面向复杂文档的智能分块体系，在提升知识处理效率的同时保障了语义完整性。系统依托 MinerU 组件的多模态解析能力，结合 OCR 光学识别与版面分析技术，精准解构 PDF、扫描件等非结构化文档的图文布局，实现对表格、公式、多栏排版等复杂元素的上下文感知。在此基础上的动态分块机制采用双层驱动架构，预设模板库内置学术论文、法律合同、财务报表等 20 余种垂直领域的分块规则，通过语义边界检测自动匹配文档类型。用户自定义模块则支持块大小、重叠比例、特殊标记保留等参数灵活配置，辅以基于 Transformer 的语义完整性评估模型，动态优化分块颗粒度。

3. 检索优化体系

检索通过运用多路召回技术，结合关键词、语义以及知识图谱检索，实现全方位检索优化。同时，引入 Raptor

分层摘要技术，采用树状分层摘要架构，从块级总结逐步构建到全局索引，显著提升检索效率与准确性。此外，通过调整 Rerank 模型权重，对检索结果的排序进行优化，从而提高检索结果的准确性和相关性。最终由 Qwen2.5 大模型进行生成式结果优化，采用指令微调模板确保输出符合金融文本规范。

4. 支持批量知识加载与知识库灵活挂载

在知识批量加载能力方面，RagFlow 通过 RESTful API 体系提供了工业级数据管道支持。其核心知识摄取接口采用分布式任务队列架构，支持 JSON/XML 标准格式的多模态数据流式上传，单次 API 调用可承载百万级文档片段的批量注入。在知识库动态挂载层面，RagFlow 基于 API Token 的弹性知识路由体系。每个知识库在创建时自动生成具有细粒度权限控制的访问令牌，支持通过 HTTP Header 进行动态绑定。这种设计使得单个推理接口可以实时切换多个知识源，为多租户和各类切换测试场景提供了原生支持。

四、运营情况

数智化知识管理系统严格遵循业务场景需求导向原则，在需求调研、方案设计、系统开发及测试验证等全生命周期阶段与业务部门保持深度协同，通过精准对接业务流程痛点与知识应用场景实现技术方案的精准落地。自今年 4 月正式上线以来，系统运行稳定，日均知识检索及调用量约 1200 笔次，业务连续性保障有效，获得了行内用户对知识获取效率提升的积极评价，同时通过建立常态化需求反馈机制持续收集具有建设性的功能优化建议，为后续敏捷迭代及服务能力升级提供了重要支撑。

五、项目成效

项目实施后，行内知识管理效能实现质的飞跃。在运营层面，技术文档处理效率提升 4 倍，知识库更新周期从周级压缩至小时级。ECC 故障及批量作业调度任务故障排查处置效率显著提升，处置准确率从 85% 提升至 95%，人为操作失误导致的事故减少 80%。业务端数据显示，通过集中管理和快速检索各类知识文档，员工可以快速获取所需信息，减少了资料收集时间；通过对内部经验和知识的有效管理，避免信息孤岛，提升行内的知识管理能力，促进知识共享，推动了全行业务效率的显著提升和管理水平的全面升级。此外，系统建设过程中培养多名复合型 AI 人才，也为金融行业智能化转型树立标杆。

六、经验总结

项目的成功实施，使每个员工，都能生产知识，获取经验；每个管理者，都能统计知识水平，掌握组织成长；每个组织，都能沉淀知识资产，提升协作效率。本项目成功验证了“大模型 + RAG”技术在金融领域的落地可行性，尽管大模型参数规模不断增长，但 RAG 在成本、速度和准确性上仍具优势，尤其适用于海量数据和垂直领域，具有不可替代性。三个核心经验值得借鉴：第一，知识管理体系建设需遵循“数据治理先行，智能应用渐进”原则，建立完善的文档质量管理体系；第二，检索效果优化需要构建“召回 - 排序 - 生成”的全链路评测机制，特别是在金融专业术语处理上需设计领域适配方案；第三，大模型落地需与实际业务场景结合，确保技术能更好的为业务服务。

未来规划沿着三个方向深化，首先扩展多模态处理能力，实现图表、录音、视频等非文本知识的统一管理；其次探索小样本持续学习机制，构建自演进的知识服务体系；最后在推动国产化生态建设方面，完成从芯片层到应用层的全栈技术验证。



中邮保险：基于传统 NLP 以及大模型“双模式”的知识循环体系

一、项目背景及目标

传统保险营销人员常面临专业知识更新快、客户咨询量大且多样化、个性化推荐难度大等困境，智能营销助手系统，采用前沿人工智能技术进行搭建，具备 7*24 小时智能问答、保险知识管理和智能推荐的能力，能有效减轻营销人员负担，同时也能帮助营销人员更精准地触达目标客户，助力营销人员拓展业务提升服务质量和效率，增强市场竞争力，实现业务的持续增长，为保险营销和企业发展注入新的活力。

销售人员获取知识需要层层问询相关人员，耗时长、效率低且信息损耗较大，无法及时向销售人员反馈正确的信息。在销售人员面对面与客户销售时，无法第一时间给与支持，仅能依赖销售人员个人知识与技能。系统上线后，销售人员可随时随地自行询问，大幅提高知识获取效率，成为展业出单的得力助手。

二、创新点

智能营销助手系统在技术体系建设方面，由传统业务驱动的定制建设模式转变为 AI 平台化的服务能力建设模式。建立从开发到运行、到运维监控的软件全生命周期管理和治理能力，实现技术 + 管理双驱动，并广泛应用分布式数据库、缓存、消息中间件等技术，满足多种营销场景，同时具备弹性部署，快速扩容的能力，实现对未来业务高速增长下的持续性能保障。另外，研发管理上实现了迭代升级可控的项目实施过程。该平台的建设为我司营销活动开展及业务增长注入新动力，向智能化发展迈出坚实一步。

三、项目技术方案

1. 基础设施层

- 使用 Kubernetes 作为容器编排平台，用于管理和运行应用程序的容器化实例。
- 在 Kubernetes 集群上配置自动伸缩、高可用性、负载均衡和服务发现等功能。

2. 微服务架构

- 使用 Spring Cloud 提供的组件，如 Ribbon（客户端负载均衡）、Hystrix（熔断器）、Feign（声明式 REST 客户端）等来构建微服务架构。
- 将应用程序拆分成独立的微服务单元，每个服务单元具有特定的业务功能，并运行在独立的容器中。
- 使用 Nacos 来管理应用程序的配置信息，实现配置的统一管理和动态更新。

3. 持续集成与持续部署

- 使用 git 工具进行代码的持续迭代和管理，使用发布脚本进行分发部署。

4. 日志与监控

- 使用 ELK+Skywalking+Elasticsearch 收集和分析应用程序的日志。
- 使用天牛平台（基于 Kubernetes）对服务和组件进行管理和监控。

5. 安全与权限管理

- 使用 Kubernetes 的角色 based 访问控制（RBAC）功能，对集群和资源进行访问控制和权限管理。
- 在应用程序层面，使用 Spring Security 来实现身份验证和授权，保护微服务的安全性。

6. 基础设施监控和故障恢复

- 使用天牛平台（基于 Kubernetes），对集群和容器的运行状况进行监控和警报。
- 实施自动扩展，根据应用程序的负载和指标进行动态伸缩。

除了开放实践中描述的底层技术栈外，还是使用了 ASR、TTS、NLP、RAG 等技术组件和方案。

相关技术与项目业务目标完成智能问答相吻合。

本次是创新性使用人工智能技术为业务赋能。

四、运营情况

自上线以来用户覆盖 4000 余人，完成超 40 余万次的问答交互。

五、项目成效

智能营销助手是为公司和渠道营销人员打造的学知识、提能力的专属智能展业工具，在其现有全面产品信息、专业智能问答的基础上，本次通过引入并应用 DeepSeek 模型以及通义千问模型，并创新性地使用“双引擎”模式，使用传统 NLP 模型与大模型相结合，实现公司级知识库内循环，完成整体知识闭环流程，在节约算力、提升效能等方面将得到进一步增强。

六、经验总结

智能营销助手深度融合 DeepSeek、Qwen 等大模型，通过多模态数据处理与多任务协同技术，显著提升分析精度与运行效能。该系统具备三大核心能力：一是智能推理引擎，突破传统数据分析边界，实现用户画像、消费行为轨迹、风险偏好多维度数据的融合解析，深度洞察用户需求，实现“一人一策”的精准保障方案定制，推动营销人员的服务从标准化向个性化跃迁。二是内置智能知识库，集成动态进化的企业级知识图谱，构建包含产品特色、公司内产品分析的立体化知识矩阵。通过自然语言处理技术实现复杂业务场景的智能解析，为营销人员提供实时决策支持，赋能从业人员实现从经验驱动向数据驱动的专业提升。三是依托大模型的算力优化技术，系统在保持低资源消耗的同时，处理复杂需求的响应速度大幅提升，缩减人工处理问题时间，提供更流畅的交互体验，进一步提升用户体验和满意度，使其聚焦高价值客户深度解析，实现营销能力和服务质效双升级。



江苏农信：基于大语言模型的运维知识管理平台

一、项目背景及目标

江苏农信历来注重科技创新，积极采用分布式、大数据、人工智能、云计算、AIOps（智能运维）等新兴技术，推动IT运维数字化转型，实现系统运维自动化和智能化，提升运维产能和效率，从而更好地保障金融科技系统安全、稳定、持续运行，为客户提供高质量的金融业务服务。

随着 ChatGPT 4.0 的发布与流行，AI 大模型技术呈现快速发展趋势，国家宏观政策也鼓励金融机构进一步投入科技创新，以数智化转型增加金融服务的科技含量。面对新一轮人工智能规模化应用浪潮，江苏农信积极开展生成式 AI 大模型技术的钻研和探索，以其作为新质生产力建设的发力点之一。其中，信息科技部系统运维团队聚焦研发基于 AI 大模型技术的运维知识管理平台——“运维小助手”，结合实际工作经验，将传统知识库与 AI 大模型技术有机融合，实现不同角色的运维人员在多场景中对运维知识、信息的快速查询与运用，辅助运维人员高效利用相关运维知识，使其能够快速作出精准的运维决策，进一步提升运维工作效能。

江苏农信科技系统承载全省农商行万亿级金融资产规模的系统运转，日均处理近亿笔日交易，保障业务的安全性、稳定性与可靠性是一项艰巨而复杂的任务。面对日益增长的海量复杂交易，运维人员缺少快速定位问题、智能辅助分析、自动化归纳总结并形成可复制经验的管理工具，运维管理工作具体面临以下几方面挑战：

1. 数据分析与理解能力：随着技术的不断迭代更新和 IT 架构的日益复杂化，IT 运维人员每日需处理的信息量呈指数级增长。他们需要不断学习和适应新技术、新架构，以应对这一挑战。如何快速而高效地分析海量的文本和数据，准确理解并妥善处理这些信息，成为了运维人员面临的一大难题。

2. 系统故障问题定位：针对复杂的 IT 系统，故障预测和诊断工作尤为关键且艰巨。运维人员需深度分析系统日志、性能数据等，以提前洞察潜在故障，迅速定位问题所在，并及时采取解决措施。这一过程中，他们需要凭借丰富的经验和敏锐的洞察力，快速思考，面临的压力巨大。

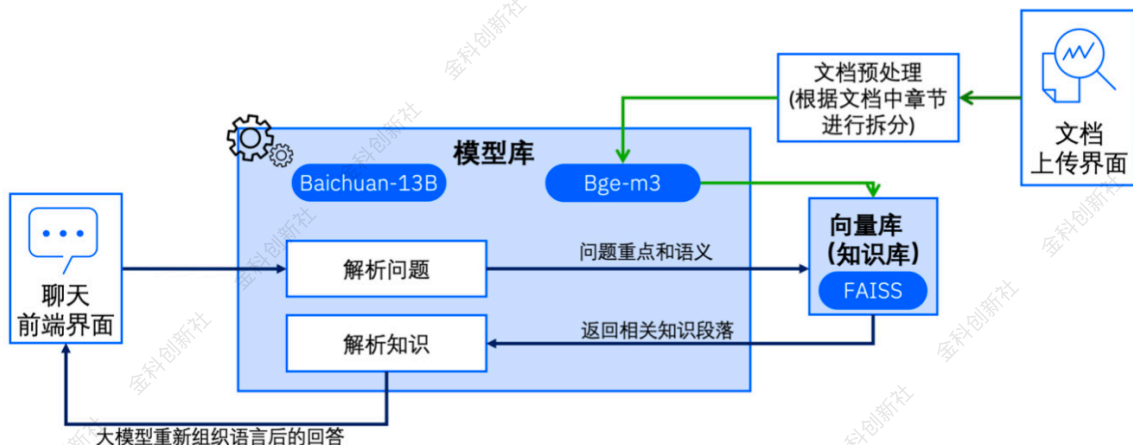
3. 自动化分析与总结：运维工作涉及大量重复性任务，如监控系统状态、记录异常事件、分析性能指标等。因此，撰写事件报告、日报等文档也是运维工作的重要组成部分。运维人员渴望拥有一个智能工具，能够协助他们完成文档和报告的撰写工作，减少手动输入和整理数据的时间，从而加快工作进度，提升整体工作效率。

4. 知识与流程标准化：在 IT 运维领域，经验和知识的积累与传承至关重要。将各个运维人员的宝贵经验和知识转化为标准化的流程和最佳实践，不仅有助于减轻运维人员的日常工作负担，还能帮助新入职的员工更快地融入团队，掌握工作技能。通过有效的知识管理，可以确保运维团队的知识体系得以持续更新和完善，为银行业务的稳健发展提供有力支持。

二、实践目标和方案

此次实践，目标是在国产的海光 DCU 上，建设基于 AI 大语言模型的知识问答系统。借助大语言模型本身的学习能力，在对运维文档进行学习后，回答运维人员提出的问题，从而辅助运维人员快速找到相关的运维信息。让大语言模型从大量未标记的文档中学习知识，通过理解这些文档，大语言模型可以捕捉到其中的关键信息、模型和规律。一旦模型学习到了足够的知识，它就能够对运维人员提出的问题进行理解和回答。这样，运维小助手作为一个强大的辅助工具，能够帮助运维人员便捷地获取所需的运维信息，而无需手动查阅文档，从而提高运维人员的工作效率。

1. 总体架构



“运维小助手”的逻辑架构如上图，包括文档预处理、文档向量化、问题解析、问答知识段落解析、答案生成、前端界面等功能模块。

为了兼顾系统在专业性知识的准确性和模型的训练效率，采用将知识外挂作为大语言模型的记忆体的策略，将知识通过 BGE-M3 这一 Embedding 模型向量化后存储在可修改的向量数据库中。我们选用 LangChain 作为基础开发框架来搭建和管理知识库问答系统，不仅为用户提供了便捷的对话交互方式，也为后续的功能扩展和优化提供了良好的基础。而集成的大语言模型 BAICHUAN-13B，则承担了系统的核心功能，负责理解用户提出的问题和知识库中的内容，并重新组织语言生成相应的答案返回给用户。

2. 知识库的建立

运维文档的切分是构建知识库问答系统中至关重要的一环。我们需要将文档切分成意思完整且独立的知识段落，作为知识库中的最小单位，用于与用户的问题进行匹配。文档的切分直接影响到问答助手对知识的理解和系统的命中率。

在切分器的选择上，考虑到运维文档中一个章节内上下文的强关联以及对顺序的要求，我们发现 LangChain 中的通用的切分器如 Recursive Character Text Splitter 等，所使用的切分依据并不适合运维小助手知识库。在这种情况下，我们自开发了切分器，根据文档中自身的章节结构进行切分，保留了文档内部的逻辑关系和语义连贯性。

在拆分后，我们引入了每个知识段落所在文档的标题作为额外信息，这一策略能在检索过程中能够更加准确地定位到与用户查询相关的知识点，提升命中率，帮助系统更好的捕捉文档的语义信息，使得系统在进行答案检索时能够更加准确地匹配用户的查询。

在使用自开发的拆分器并引入标题作为辅助信息后，测试集问题在知识库中的命中率从 60.5% 提升到了 77.3%。

3. 文档内容向量化

在切分文档后，章节内容会被通过 Embedding 模型向量化，将它们编码为一个统一的向量空间中的向量。我们在对 Embedding 模型选型时首先尝试了 BGE-LARGE-ZH-V1.5，但由于其支持的 token 数量较低，无法支撑部分较长的运维文档的章节，因此我们最后选用了 BGE-M3。其有能力处理较多的 token，token 数最多可达 8192，这足以覆盖主要运维文档的章节长度，从而确保在知识库中，每个知识点都能包括更为详尽和完整的信息。



为了提升模型的表现,我们收集了运维专家针对知识库中的运维文档内容提出的问题,以及在文档中对应的答案。这些问答对作为小样本被用于微调 Embedding 模型,以确保模型能够更好地理解和捕捉到关键信息,并能够在向量空间中更准确地匹配相应的知识点,从而成功提高了模型在运维文档知识库中的命中率和检索性能。

同时,我们将微调后的模型和基础模型进行了合并。通常对基础 BGE 模型微调可以提高其在特定目标任务上的性能,但这可能会导致模型在目标领域意外的通用能力严重退化。通过合并微调后的模型和基础模型,可以显著增强特定任务的性能,同时保持其在其他任务中的有效性,为未来拓展其他运维场景铺平道路。

4. 大语言模型的微调

我们采用 Prompt Engineering 来改善大语言模型的表现。Prompt Engineering 的核心是在提供尽可能多的上下文信息的同时,通过提供少量示例来更好地让大模型了解当前的任务。大语言模型本身具备一定的关键词提取能力和语义理解能力,可以根据向量化知识库中的内容进行重新理解、提取内容,并组织语言进行回答,但对 prompt 的用词比较敏感,有时候改变一个词甚至动词都会导致完全不同的回应。通过多轮 prompt 的迭代,我们发现大语言模型可以更准确地根据知识库中提取的相关内容生成相应的回答。

考虑到运维工作的严谨性,为了保证运维小助手的可靠性和安全性,我们也特别调整了 prompt 的设计,加入例如“根据已知信息”、“不允许在答案中添加编造成分”等限制,以确保大语言模型严格遵循文档所包含的信息来回答用户的问答,避免模型产生 hallucinate (虚假的信息)。同时,在回答问题时,运维小助手会明确告诉用户答案出自哪个文档的哪个章节,并展示该章节的内容。这样既保证了答案的真实有效性,又方便用户快速了解答案的上下文。如果用户提出的问题无法在现有文档中找到答案,运维小助手会明确回答用户无法找到该问题的答案,不会虚构或杜撰任何内容,防止了因为误导而可能导致的运维事故的发生。

5. 总结

从技术角度来看,江苏农信运维小助手已经端到端走通,在业务专家准备问答测试集中,回答的命中率超过了 85%,完成了基于国产 GPU 的大语言模型知识库问答系统的概念验证,但后续仍需要进一步迭代,持续更新和完善知识库内容,并不断优化模型的表现。

一方面,文档质量是运维小助手系统性能的关键因素。为了确保信息的及时性、准确性和全面性,知识库需要根据行内的运维工作要求持续完善和更新,以提高知识库的质量和覆盖范围,这可能涉及到自动化的文档更新和生产数据读取的机制的引入。

另一方面,模型本身的能力也至关重要,因此建立健康的模型生命周期是必不可少的。在用户使用过程中,不断收集用户反馈,并依据反馈数据定期对模型进行微调和性能评估,以确保模型始终能够保持高效、准确和可靠。

杭州银行：基于金融垂直大模型的商业银行制度知识库检索平台

一、项目背景及目标

银行规章制度是银行正常运作的重要基础。随着金融科技的迅猛发展，银行业务日益复杂，法律法规、内部政策、合规要求也在不断更新。银行及金融机构的合规部门、审计部门等，都需要快速检索和获取相关的制度和政策信息，同时也面临来自监管机构的严格合规要求，对于能够实时更新法规和政策信息，确保合规性、高效的信息检索，提高工作效率，降低人为错误的制度管理和检索系统的需求日益增加。

在整个金融监管加强、金融合规严控的政策背景下，银行需要建立健全合规管理体系，确保遵循相关法律法规。随着信息技术的发展，尤其是大数据和人工智能应用如大语言模型的出现，为银行制度知识库检索系统的建设提供了技术支持。越来越多的金融机构开始重视信息化建设，尤其是在合规和风险管理领域，推动了市场对制度管理和检索系统的需求。在这个大背景下，杭州银行围绕“二二五五”战略和“数智赋能”建设，银行制度知识库检索项目开始受到重视并立项。

对于我行的制度知识库检索项目，有四大项目目标：

1. 系统目标：建立一个高效、智能、可扩展的银行制度检索系统，提升信息检索效率；
2. 业务目标：降低各大业务机构及部门在日常运营中对于制度知识的检索成本，减少因信息不对称、不全面带来的合规风险；
3. 技术目标：实现高可用性、高安全性和高扩展性的检索系统和文档管理中台，满足不断变化的政策及用户需求；
4. 用户目标：提升用户的工作效率和决策能力，提供优质的用户体验。

二、项目 / 策略方案

整个项目策划包括以下几点：项目目标、需求分析、项目范围。

项目目标

1. 建立一套完备的银行制度知识库，提供便捷的检索服务；
2. 实现高效的信息管理与知识共享，支持银行内部及外部的政策法规合规。

需求分析

1. 用户需求：明确用户（合规部门、审计部门、其他工作人员等）的检索需求和使用场景；
2. 数据源分析：确认知识库所需的数据源，包括法规条款文件、外部政策文件、行业指南以及操作手册等。

项目范围

包括知识库的构建、数据处理与管理、检索系统和管理平台的开发、用户界面的设计等。



整个项目实施包括以下几个阶段：数据处理与整合、系统设计、开发与测试、上线与培训。

数据处理与整合

1. 根据各部门提供的行内制度、法规、政策及相关文档，进行分部门分类 (excel/doc/pdf/markdown/html/ofd 等) 的标准化及定制化处理；
2. 建立数据更新机制，确保知识库信息的时效性。

系统设计

1. 确定系统的整体架构，包括前端用户界面、后端数据存储与处理模块；
2. 设计端到端的算法模型，建立向量存储知识库，支持语义搜索，实体硬过滤，大模型生成式对话以及其他多模态识别等功能。

开发与测试

1. 前后端开发：用户友好的文档中台界面开发，确保易用性。建立数据库，实现数据存储和文档管理模块；
2. 算法模型开发：包括知识图谱的构建、图谱链路的检索、query 拓展、文档 OCR 标注与上下文语义的切分、多重检索链路召回逻辑、重排模型优化、多模态大小模型协同优化等；
3. 测试：进行功能性测试、性能测试和安全性测试。

上线与培训

1. 系统上线，通过行内微课进行线上 + 线下的初步推广；
2. 针对业务运营管理员进行培训，确保用户能够有效使用系统。

ID 三、创新点

该项目的创新点主要体现在以下四个方面：

①采用大模型 +RAG 技术：本项目引入了先进的垂直金融零售领域基座模型——金磐大模型与 RAG(Retrieval-Augmented Generation 检索增强生成) 框架，实现了端到端的模型开发与优化，包括文档切分、query 改写、搜索匹配、向量库存储等，有效缓解了大模型在生成过程中可能出现的幻觉现象，确保生成内容的准确性和相关性，从而大大提高了模型的可靠性。

②采用知识图谱技术：银行制度知识需要跨部门跨知识库的解决方案，现有的 RAG 方法往往依赖简单的数据结构，限制了它们理解基于实体间复杂关系信息的能力。并且这些系统往往缺乏必要的上下文知识，无法在不同实体及其相互关系之间保持回答的连贯性。我们基于 GraphRag 的思想构建了杭银制度图谱，将图结构纳入文本索引和信息检索中，并在图构建上做了剪枝优化，以及图问答的意图识别和多实体路径搜索，保证了对多跳复杂问题更有效的推理和召回。

③采用多模态大小模型协同技术：银行业的制度文档类型不光格式多样，内容也很复杂，通常将文本、表格和图片组合在一起，以传达丰富的想法和见解。虽然 RAG 擅长理解长文本，但难以有效的集成和理解多模态内容。我们将主流的多模态大模型和小模型（包括 Clip、Sam 等）相结合，能够更高效准确的分类和理解流程图、扫描发票、大型表格等模态，实现精准问答。

④文档数据管理中台：本项目开发了一款专为开发与运维人员设计的文档数据管理中台，该平台对文档的整个生命周期进行了系统化管理，不仅提升了文档的组织性和可访问性，还确保了信息的及时更新和准确性。此外，中台还可以为各个部门和业务条线量身定制专属的多层次文档集群，以满足不同部门个性化需求，根据自身的业务特点和操作流程，快速获取所需信息，提升各类文档在存储、检索方面的便利性。

ID 四、运营情况

在推广方面，采取了多种策略以确保制度知识库检索系统的广泛应用。首先，线上录制系统使用的视频课程通过公司的内部渠道推广至全行，使用户能够便利的学习相关知识。此外，我们还组织线下培训合成，以便于面对面的交流和互动，帮助用户更好的理解和应用系统。为了增强使用文档的可见性，我们在工作沟通 APP 中发布公告，鼓励用户积极查阅和使用相关文档，从而提高整体使用效率。

在用户反馈方面，为了及时收集用户的意见和建议，我们设置了“点踩”按钮，方便用户对内容进行快速反馈，这一功能不仅能帮助我们了解哪些部分存在问题，还能引导我们进行改进。同时，我们对用户的 query 进行算法建模分析，以识别常见需求和潜在问题，从中提取有价值的信息，帮助我们更好的理解用户需求。

在系统运行方面，将用户活跃量作为重要指标进行分析，评估系统的使用效果和用户参与度。同时，我们还开发了一套问题回溯机制，确保能够快速定位为问题，保证系统稳定运行。

ID 五、项目成效

银行制度知识库检索系统为银行内部各部门（如合规部门、审计部门等）以及行内其他人员提供服务，带来的成效具体体现在以下几个方面：

1. 提升工作效率：通过集中管理和快速检索各类制度和政策文件，员工可以迅速获取所需信息，减少了查找资料所耗费的时间。这种效率的提升将直接影响工作进度，使得各部门能够更快的完成任务，提高整体生产力。
2. 促进知识共享与传承：制度知识库能够有效记录和分享内部经验和知识，避免信息孤岛现象，提升行内的知识管理能力。这不仅促进了部门间的协作，也能帮助新员工快速上手，从而节省培训成本。
3. 降低合规风险：合规部门能够更有效的访问最新的法律法规和内部政策，从而降低因信息滞后导致的合规风险。及时了解和遵循相关规定，可以减少因违规操作而造成的潜在罚款和法律责任。

ID 六、经验总结

项目经验总结主要体现在以下几个方面：

首先，深入了解项目需求相当重要。在项目启动阶段，与相关方进行充分沟通，明确期望和需求，制定科学合理的开发计划。这一过程可以有效减少重复开发或废弃功能的发生，从而避免开发资源的浪费，提高项目的整体效率。

其次，在技术调研方面，对开源产品进行全面测试，充分了解这些产品的适用性和稳定性，避免后期出现不适用的情况，导致额外的迁移成本和资源浪费。

最后，在推广方面，将项目嵌入用户最常见的应用程序中，能够有效降低用户的使用门槛。这种策略不仅能够提高用户的接受度，还能促进更广泛的使用，从而提升项目的价值。通过优化用户体验，确保用户能够无缝的访问新功能，可以最大限度的发挥项目的潜力，实现预期目标。



北京银行：“京信妙笔”智慧创作平台

一、项目背景及目标

银行内部工作涉及大量的文书报告写作工作，既包括通知公告、述职报告、心得体会等较为通用的行政和个人文书，也包括贷款报告、内审报告、合规文件等业务性很强的专用文书，利用人工智能技术，尤其是大模型技术，进行大规模资料阅读分析和写作，可以支撑银行内各个部门各个岗位的员工，具有较高的业务价值。

京信妙笔智能报告平台利用大模型、文档解析、文档检索等人工智能技术，能够自动化处理信审过程中的数据，快速生成高质量的信审报告。平台不仅支持多源数据接入，还具备强大的分析能力和可视化展示功能，为北京银行的信审业务提供了强有力支持。

二、项目 / 策略方案

项目充分利用生成式大模型及大模型写作工具的能力，基于银行对写作数据需求、模板需求、审核需求，在通用写作产品的基础上，增加多源数据对接、数据理解分析、模板写作、溯源审核等功能，形成银行可信写作平台。

业务模式

围绕贷款审查业务和日常办公两个场景，开发了信贷报告生成和办公文稿两大系列十余种文档写作功能。包括：

对公初审报告：基于支行尽调报告、行内系统数据，通过大模型分析和智能生成能力，提炼企业信息、贷款申请信息、股东信息、财务数据、抵押信息等关键数据，按照分行对公初审报告模板，一键生成分行对公初审报告。

个贷初审报告：基于支行尽调报告数据，通过大模型分析和智能生成能力，提炼借款人基本情况、用款企业情况、抵押房产情况等关键数据，按照分行个贷初审报告模板，一键生成分行个贷初审报告。

初审报告摘要：基于总行初审报告，通过大模型能力，提炼企业信息、报审方案、财务数据、抵押信息等关键数据，按照总行摘要模版，一键生成初审报告摘要。

总行通知：通过内置提示模板对用户输入进行引导，通过更改标题、时间、对象等信息，快速生成对应类型的通知内容，提高工作效率和内容质量。

PPT 大纲：利用大模型来组织和整理信息，明确展示的重点和逻辑流程，为后续的内容创作提供清晰的框架。提高制作效率、确保内容完整性和逻辑性。

心得体会：基于预设的框架，心得体会可以在短时间内自动生成，有效节约时间成本。通过变换模版中的关键词可生成多样化的内容，以增加文本的丰富性。

新闻稿：通过大模型生成的新闻稿可以保证品牌信息的一致性、降低语言错误的可能性。有助于提升品牌稳定性。

工作汇报：快速和高效的工作汇报生成可以更频繁地更新项目状态和业务洞察，为管理层提供及时的信息，帮助他们做出更为数据驱动的决策。

工作进展汇报：利用先进的大模型技术，生成高质量、风格一致的工作进展汇报文档。有效减少人工编辑时间，确保信息的准确性和时效性，同时强化品牌形象的稳定性与专业性。

工作计划撰写：基于大模型技术的智能工具，帮助用户快速、准确地制定符合要求、结构清晰、目标明确的工作计划。提升工作效率与规划质量。

技术架构

1. 基于 yayi 大模型以及多种类型报告模板，通过客户经理文档上传、多源数据库接入进行数据汇聚；构建文档解析、向量化等模型对数据处理；根据不同报告模板构建知识体系，分解子问题并生成报告；最后，利用大模型润色初稿，生成定制化报告，支持在线编辑与 Word 导出，实现高效、精准、自动化的报告生成流程。
2. 数据库采用的是 Mysql，虚拟机部署，一主一从；Redis 采用哨兵的架构模式，测试环境采用虚拟机部署，生产环境采用业务系统使用虚拟机部署，大模型使用实体机部署。
3. 后端应用基于标准的 J2EE 架构、MVC 模式，服务层依赖 SpringBoot、Mybatis 实现核心组件的支撑；前端 WEB 应用采用 Vue 框架进行系统搭建，使用 Element-UI 组件进行页面组件的设计。
4. 算法服务基于 Python 语言开发，采用 Flask 框架搭建 RESTful API 服务，算法模块依赖 transformers、PyTorch 等库实现核心功能。

ID 三、创新点

1. 项目基于大模型的可信公文创作技术，以 Transformer 为核心的生成式人工智能技术始终面临幻觉问题，这给 aigc 技术在工业界（尤其是看重准确性、可信性的金融领域）的应用带来了限制。本项目研发了可信公文创作技术，通过将智能生成、检索生成、模型自省、专家框架、智能溯源等技术的落地实现和综合运用，解决了 aigc 模型在具体应用中不可解释、不可追溯的黑箱问题，为大模型在金融创作场景的落地开辟了一条道路。
2. 基于大模型的文档解析技术实现 PDF、WORD、TXT 等多种格式，尽调报告、初审报告等多种类型的文档解析能力，针对不同格式、不同类型的文件，采用不同的文档解析模块，构造对应知识树，满足公文创作场景下的文档处理需求。本项目采用基于大模型的文档解析技术，WPS 文档解析模块基于文档解析引擎提取文字内容。PDF 解析模块将结合文档版面分析、表格识别、段落文字块合并等多项技术，以准确划分标题、文本段落、图片、表格等内容信息，从而实现对文档段落结构的精确恢复，最大程度地保留文档的语义信息；并针对不同类型构造对应知识体系，最大程度保留了文档的知识信息。
3. 兼容适配多模板报告创作，大模型根据不同的报告模板自动解析报告目录，拆解问题，基于知识体系明确各部分内容的取数来源及撰写要求，依据撰写要求及取数来源，从知识库中查询数据。基于大模型 prompt 工程，利于大模型生成对应的输出内容。并基于索引服务，明确生成内容的数据来源，从而实现结果溯源和数据稽核。这样一套完善的报告生成流程可以帮助系统更加高效地生成各类报告，确保报告内容的准确性和可靠性。

ID 四、运营情况

目前京信妙笔智能报告平台已在北京银行全面上线，支持总行初审、分行初审岗位业务人员进行信贷初审报告的编辑审核工作，相关岗位人员涉及 10000 人左右，帮助业务人员从原先最快 3 天完成一份完整报告到 5 分钟生成、半天审核完成一份初审报告，提升初审效率 80% 以上。

ID 五、项目成效

京信妙笔智能报告平台可轻松解决信贷审批过程中，海量非标数据处理难、数据信源不统一的问题。贷前审核过程中涉及到大量企业信息，完全依靠人工进行核实和梳理需要耗费大量时间和精力，效率较低，且容易发生人为错误和疏漏。且信贷报告中使用的企业信源自访谈、底稿、研究报告甚至互联网信息，各个信息源的可信度不尽相同，根据不同信息得出的观点也大相径庭，审核者难以一一查证来源，确认可靠性。项目主要实现四项经济效益：



1. 提高工作效率

自动化生成：京信妙笔智能报告平台能够自动化完成报告的生成过程，减少了人工编写报告的时间和精力。传统报告编写需要人工收集数据、分析数据、撰写报告等多个步骤，而平台则可以在短时间内完成这些工作，大大提高了工作效率。

快速响应：在需要快速生成报告的场景下，如总行通知、工作进展汇报等，京信妙笔智能报告平台能够迅速生成报告，确保信息的及时传递和决策的快速制定。

2. 提升报告质量

精准分析：京信妙笔智能报告平台通过大模型及文档解析、向量化等技术，能够对数据进行深度挖掘和分析，生成更加精准、科学的报告内容。

丰富多样：京信妙笔智能报告平台能够生成包括文本、表格、图像等多种形式的报告内容，使报告更加直观、易懂。

3. 降低成本

减少人力成本：京信妙笔智能报告平台的自动化生成功能减少了人工编写报告的需求，从而降低了行内的人力成本。行员可以将更多的人力资源投入到其他更有价值的工作中去。

降低错误率：人工编写报告容易出现错误和疏漏，而京信妙笔智能报告平台则能够通过算法和模型来减少这些错误和疏漏的发生，提高报告的准确性和可靠性。

4. 促进创新

新模式探索：京信妙笔智能报告平台为北京银行提供了全新的报告生成模式，有助于北京银行探索新的工作方式和管理模式。

激发创意：京信妙笔智能报告平台通过自动化生成报告，使行内员工能够有更多的时间和精力去关注创新工作，激发创意和灵感。

六、经验总结

京信妙笔智慧创作平台基于北京银行实际业务需求，利用大模型技术开发了贷款审查和日常办公的智能写作工具，打通了银行内部的信贷、风控系统与外部数据库，实现了报告生成的自动化和溯源功能。通过配置不同业务模板，项目有效提高了信审报告生成和审核的效率，简化了人工修改与审核流程，显著提升了审批质效。在推广过程中，项目结合银行业务场景不断优化，得到了广泛好评，成功推动了银行数字化转型和智能风控体系的建设。

太平洋寿险：银保销售复盘会萃取工作台

一、项目背景及目标

随着保险行业竞争加剧，银保渠道作为核心业务增长点，亟需通过精细化运营提升销售效率。中国太保寿险银保渠道每周需召开 600 余场销售复盘会议，传统管理模式面临多重挑战：一是会议录音转写错误率高，非结构化内容导致管理层难以快速获取有效信息；二是人工回听录音耗时耗力，单次会议分析需消耗管理人员 1 小时以上；三是优秀销售经验依赖业务专家主观筛选，案例覆盖范围有限；四是总部对一线经营动态感知滞后，难以及时响应市场变化。

在此背景下，公司启动会经智能化改造工程，明确三大核心目标：一是通过自动化质检与摘要生成，实现会议内容结构化分析；二是构建多维度数据分析模型，精准洞察一线共性问题 and 市场动态；三是建立动态知识库，推动优秀案例全域共享，赋能销售队伍能力提升。

二、项目 / 策略方案

本项目在技术创新与业务模式重构上实现双重突破：

1. 技术融合创新

自主研发的行业专有大模型融合保险专业词库与历史业务数据，通过切片预处理、多轮微调和后置校验机制，攻克保险术语识别、多角色发言区分等技术难点。例如，针对会议中常见的“增额终身寿”“个养产品”等专业词汇，优化 ASR 模型后识别准确率提升至 95%；通过长文本切片技术，将单次会议 1 小时的录音分割为 15 个逻辑段落，显著提升大模型上下文理解能力。

2. 业务模式重构

首创“质检 - 萃取 - 复制”三位一体管理体系。质检模块从流程合规性、内容深度、情绪识别等维度量化会议质量，自动生成评分报告；萃取模块通过双重分析机制，先提取单场会议热词与客户案例，再聚合全域数据形成共性洞见；沉淀模块构建动态知识库，将优秀案例直接推送至外勤展业 APP，形成“分析 - 改进 - 复制”的业务闭环。

3. 管理模式升级

引入恶意行为识别与互动性分析技术，量化评估会议内容与参与度。例如，结合语义分析判断讨论焦点，帮助管理层识别会议异常行为，及时过程管理，确保总部要求贯彻落实。

三、项目技术方案

1. 数据层

构建多源数据处理平台，整合语音、文本及业务系统数据。通过优化 ASR 模型，保险专业术语识别准确率



从 82% 提升至 95%；建立噪声过滤机制，有效消除会议背景音干扰。

2. 模型层

针对长文本处理难题，开发上下文关联算法，确保跨段落语义连贯性。

3. 应用层

设计可视化管理看板，支持多维度数据钻取。管理人员可通过热词词云、趋势图谱等工具快速定位问题；一线销售通过外勤 APP 查看标杆案例，平均学习时长从 30 分钟缩短至 8 分钟。

技术实施中攻克三大难点：一是通过角色分离算法，准确区分保险顾问、银行客户经理等发言主体；二是设计异常结果自动重试机制，当模型输出偏离阈值时，系统自动触发二次分析；三是建立人工复核通道，每月抽样 5% 的会议结果进行校验，持续优化模型参数。

四、运营情况

效率提升

管理人员日均处理会议数量从 8 场增至 20 场，信息提取效率提升 60%；人工回听工作量减少 80%，每年释放 4000 人天用于高价值分析。会议流程合规率从 73% 提升至 96%，恶意踩点行为减少 82%。

知识沉淀

1. 萃取热点话题，判断政策是否直达。例如，银保业务首季红期间紧盯“重客储备”进度，紧跟“分红险产品及个养产品”推动，从萃取平台可以看到当期开会的热议话题与总部策略方针方向一致，说明政策层层宣贯，落地执行情况较好。

2. 萃取热议问题，推动赋能问题解决。萃取平台还专设热议问题模块，将银保业务专家的经营理念借助大模型技术结构化呈现，实时识别提炼开会中提及的网点经营、客户经营的难点与痛点，比如队伍在“识客户”环节提及较多的问题是客户邀约存在困难、客户对保险产品有疑虑等，业务专家或各级管理人员就可以根据现存问题，有针对性地提出解决方案，赋能一线经营。

3. 萃取经营标杆，看见优秀，学习优秀。萃取平台还能提炼出在开会过程中提及的成交客户案例进行总结分析，比如企业家客群主要通过购买增额终身寿产品，实现家企资产的隔离；而购买个人养老金产品鑫福享的客户主要担心利率下行，看重长期锁定利率。对于萃取的成交案例，银保全员可见，方便队伍看见优秀，学习优秀。

五、项目成效

1. 直接成本节约

自项目创立以来，人工质检复核的工作量大幅减少。据统计，每位经营管理日均成本约为 3000 元，通过优化流程，每年可节省超过 1200 万元的人力成本，极大提升了管理效能。

2. 业务增长贡献

通过大模型技术聚合销售话题和客户经营案例，形成标杆案例库，促进优秀经验的复制和共享。这不仅提高了质检模型的智能化水平，优化了语音转文本的准确性，还实现了会议流程的自动化判断。此外，平台通过结构化提炼会议内容，快速捕捉关键信息，洞察共性话题，并提炼优秀案例，为其他会议提供参考。

六、经验总结

1. 业科深度融合

建立“业务需求池 - 技术开发池”双轨机制，业务专家全程参与模型训练与结果校验，确保功能贴合实际场景。

2. 敏捷迭代体系

构建“数据采集 - 模型训练 - 人工校验”闭环，通过每两周优化会议持续提升系统性能。

3. 知识资产化运营

将萃取成果系统化编入公司级知识图谱，支持多场景复用。



太平洋寿险：智能金牌教练

一、项目背景及目标

太保寿险已有的训练系统已实现基础的学练考评和基础管理内容建设。但仍存在能力评估粗放，缺乏数据化评估工具，难以精准定位代理人能力短板；训练效率低下：标准化课程无法满足个性化需求，难实现因材施教；管理穿透不足：总部机构难以快速知晓一线实际培训情况，无法做到过程和行为管控。

在此背景下，进行该项目建设，明确项目三大核心目标：一、构建代理人能力评估体系，实现从“经验驱动”到“数据驱动”的转型；二、打造智能化训战闭环，缩短培训周期并提升实战转化率；三、建立总部到一线的全流程穿透式管理，推动业务策略精准落地。

二、创新点

1. 技术融合创新

能力画像引擎：基于 Transformer 架构的大模型，融合知识图谱与历史业务数据，构建代理人多维能力标签（如产品知识、客户洞察、异议处理）；

虚拟陪练引擎：基于通义千问大模型与语音交互技术，模拟 20+ 实战场景（如客户异议处理、产品讲解），根据场景进行多类角色扮演，角色扮演一致率达到 80%+；

动态推荐算法：自动化采集学习行为数据，结合当前代理人所处销售阶段，优化个性化学习路径。

2. 业务模式重构

训战一体闭环：首创“测 - 学 - 练 - 考 - 评 - 管”六阶段模型，将传统单向培训升级为“诊断 - 训练 - 验证”动态循环；

知识资产化运营：通过 AI 萃取优秀代理人销售话术案例，构建动态知识库。

3. 管理模式升级

全链路穿透式管理：总部可实时追踪 38 家机构训练计划执行情况、全国 10 万 + 代理人的训练进度与效果，通过数据看板识别低效机构并定向干预，确保要求落实到位。

三、项目技术方案

1. 数据层：构建多源数据处理平台。融合代理人客经、培训、授课等多维度数据。

2. 模型层：针对角色扮演、智能点评，打造大模型能力 + 小模型能力，确保模型评价时准确理解文本内容。

3. 应用层：设计智能金牌教练队伍端和管理端应用，实现训练场景管理角色和用户角色即时联动。

技术实施中攻克三大难点：一是通过角色扮演算法，确保模型扮演的角色能根据代理人描述内容进行灵活响应及答复，同时当代理人描述内容偏离主题时，进行及时的场景拉回和提示；二是设计内容生成复核机制，从原

始素材到培训素材的生产过程中，支持生成后人工复核校验；三是建立人工复核通道，每月抽样对练内容及评价结果进行校验，持续优化模型参数。

四、运营情况

通过智能金牌教练，将原本需要逐层级反馈的机构训练情况由周级缩短为小时级，同时实现对代理人能力情况的实时观测和分析。月均节约主管辅导时长约 5 小时，内容准备效率提升 50% 以上。

智能金牌教练不仅提升了内勤训练人员准备训练素材的效能，也为公司训练管理工作注入智能动力。同时，其成功应用也为其他业务场景提供了宝贵借鉴，推动公司整体决策的科学化进程。

五、项目成效

1. 经济效益

自项目创立以来，根据测算，通过逐步线上化训练，每年预计可节省超过 2000 万元的人力成本，极大提升了训练效能。管理效能提升 30%+。

2. 社会价值

响应监管关于推动深化人身保险行业个人营销体制改革要求，提升保险销售人员专业化水平。做好队伍的全流程数字化管理，增强保险销售人员的合规意识、专业水平和服务能力，提升客户满意度。

六、经验总结

1. 业科深度融合

建立“业务需求池 - 技术开发池”双轨协同机制，由业务专家主导提炼代理人能力评估标准、实战场景库等核心需求，技术团队同步构建大模型训练框架与评估算法。业务专家全程参与模型训练与结果校验，确保能力画像与真实展业场景高度契合。

2. 数智化训战 SOP 体系

构建“数据采集 - 模型训练 - 人工校验”闭环，后续可以复制到内勤培训、讲师培训等多类场景。

3. 知识资产化运营

将萃取成果系统化编入公司级知识中心，支持多场景复用。



华福证券：企业智慧问答系统

一、项目背景及目标

在当今数字化浪潮席卷各行业的背景下，金融领域正面临着前所未有的变革机遇与挑战。中央及福建省委金融工作会议明确提出，要加快建设金融强国，充分把握数字技术革命带来的契机，大力推动金融与数字技术的深度融合与创新，全方位加快监管数字化智能化转型进程。

在此政策引领下，我司积极响应，致力于深入挖掘数字化转型的多元场景，从规章制度的建设以及日常运用实践为切入点，依托大模型技术开展深入研究探索，旨在打造一个能够显著提升业务知识和规章制度使用效率与传导效果的企业智慧问答系统。企业智慧问答系统共分为智慧制度与智慧财务两大模块，为全司员工营造便捷获取财务费控报销相关信息和规章制度知识的环境，精准识别员工提问意图并高效匹配相应规章条款，赋能一线员工专业素质提升。

长期以来公司业务知识和制度规章繁多、分布零散、难以查找，同样的，企业内部的财务板块也面临着亟待解决的难题。财务人员每日深陷大量与费用报销相关的咨询工作泥潭，各类问题高度相似且重复性极强，致使人力成本不断攀升，工作效率却受限难以提升。尽管企业内部积累了海量的业务和管理数据，并且已配套建设多个功能模块助力数据应用，但由于不同系统间存在差异，业务流程复杂度高，用户在准确获取所需数据时困难重重，往往需要对系统和业务有深入且全面的了解才能完成数据提取任务，这无疑进一步抬高了数据利用的门槛与难度，严重制约了企业运营效率的提升。

鉴于上述复杂严峻的形势，本企业智慧问答系统项目应运而生。项目旨在凭借先进的人工智能技术，充分挖掘大语言模型的强大潜能，对企业内部的财务咨询与数据获取流程进行全面优化升级，从而提升整体运营效率与服务质量，为企业在数字经济时代的持续稳健发展提供坚实有力的支撑。

二、创新点

企业智慧问答系统共分为智慧制度与智慧财务两大模块：

1. 智慧制度

- “数据驱动”：创新性提出多维度评估规章制度和业务知识，对制度和知识按公司治理要求进行科学分类，开展存量制度数据治理、建立增量制度管理机制，标本兼治，有效归集知识库。
- 智能知识梳理，按照一定规则帮助业务人员高效梳理规章制度和业务知识，快速形成知识库。
- 大模型人机交互，利用大模型技术优化人机交互体验，降低业务人员获取规章制度和业务知识时的个人知识积累门槛。

2. 智慧财务

(1) 极简架构设计

- 摒弃冗余组件：完全舍弃向量数据库、向量模型、多路召回策略等复杂技术栈，无需混合检索（如全文检索+向量召回）和资源隔离运维。

- 去人工化设计：无需人工定制数据切片策略、存储加工策略或召回策略，避免因复杂系统带来的开发维护成本。

(2) 数据处理与算力优化

- 动态信息处理：通过动态文件过滤和窗口分区（而非固定 chunk 切割），避免数据切片导致的信息丢失和排序干扰。

- 上下文完整性：利用大模型直接处理原始文件（非碎片化文本），保留完整语义关联，消除 chunk 同质化问题。
- 暴力美学计算：基于大模型支持超长上下文（百 K token 输入），通过窗口分区管理（文件区、片段区、系统缓冲区），最大化利用 GPU 算力。

- 性能与成本优化：极限缓存命中（缓存与 miss token 比例 2:1）降低推理成本；仅需个位数 token 输出，规避解码瓶颈，扩展性极强。

(3) 工程创新与优化

- 全流程大模型驱动：从文件筛选到答案生成，全程依赖 LLM 自主决策（非人工规则），减少人为干预导致的性能天花板。

- 实时性与灵活性：动态响应 query 需求，无需全量更新向量库（对比一代 RAG 依赖批量更新），适应实时场景。
- LLM-Native 设计：完全围绕大模型能力构建（非依赖传统 NLP 组件），通过算法创新（如 128k 窗口分区）实现效果跃升。

- 无预处理的端到端流程：无需复杂数据预加工，直接处理原始代码 / 文档，避免预处理导致的信息扭曲。

三、项目技术方案

智慧制度

1. 文档预处理阶段

文档筛选：对华福内规库文件进行全面梳理，依据业务相关性、法规效力等标准，剔除过期、失效及与核心业务关联度低的非必要文件，筛选出有效的内规文件集合。

文本切分：将筛选后的有效内规文件进行切分，按照合理的语义单元（如按章节、条款等）将其分解为多个文本片段，便于后续处理和检索。

向量编码与存储：调用文本向量化模型，将切分后的文本片段编码为高维语义向量，捕捉文本深层次的语义信息，并将这些语义向量存储于高性能的向量数据库中，为后续的语义匹配奠定基础。

2. 大模型微调阶段

模型选择：在众多预训练大模型中，经过充分评估和对比，选定与华福业务领域契合度高、具备较强语言理解和生成能力的基础大模型作为初始模型。

数据准备：收集整理华福内规文档以及相关的高质量问答对，问答对涵盖内规的核心内容、常见业务场景中的合规问题及解答等，作为模型微调的数据集。

微调训练：利用准备好的数据集对选定的大模型进行监督微调，通过调整模型参数，使模型学习到华福内规领域的专业语言风格、知识要点和问答逻辑，从而提高模型对内规相关问题的回答准确性和专业性。

3. 在线问答阶段

问题处理与语义匹配：当用户提出内规相关问题时，系统首先利用语义模型将问题文本转换为语义向量。同时，



结合经典的 BM25 算法，综合计算问题语义向量与内规文本片段语义向量的相似度，快速、精准地从向量数据库中检索出若干个与问题语义最接近的内规文本片段。

提示词构建与回答生成：将匹配到的相关内规章节文本片段与用户问题进行整合，精心构建为适合大模型输入的提示词。将该提示词输入微调后的大模型，大模型依据其强大的语言生成能力，基于提示词中的信息和已学习到的内规知识，生成针对用户问题的详细、准确、专业的回答内容，并将回答反馈给用户。

4. 持续优化阶段

Agent 服务架构：系统内置 Agent 服务，负责统筹协调用户请求中涉及的各项具体任务。Agent 服务可灵活调用多种专业工具，包括语义理解工具用于精准解析用户问题语义；数据召回工具高效检索内规知识库；数据重排工具对检索结果进行优化排序；API 调用工具整合公司内部其他相关系统数据；以及复杂计算工具处理涉及合规指标等的复杂计算任务。

幻觉问题解决与优化：借助 Agent 的精准召回能力，能够在生成回答前，对相关信息进行严格筛选和验证，有效减少大模型可能出现的幻觉问题，确保回答内容的准确性和可靠性。同时，基于用户与系统的交互数据，包括用户反馈、新出现的内规问题及解答等，Agent 不断对系统进行自我优化和迭代，持续提升系统性能和回答质量。

智慧财务

用户请求触发检索流程，系统智能调用三个模型实现高效筛选。

1. 初始阶段简单检索模型依据段落特征快速初步筛选，初步定位关键信息；
2. 小参数量大模型 (Qwen-2.5-14B-Instruct, 参数量 140 亿) 对全文进行初筛，凭借其轻量级优势快速过滤无关内容；
3. 同类型小参数量大模型接力进行全文二筛，进一步优化筛选结果，精准定位关键信息；
4. 将筛选后的文档传递给大参数量大模型 (Qwen-2.5-72B-Instruct, 参数量 720 亿)，由其深度处理并生成精准回答，整个流程智能高效，确保高质量的检索与回答生成。

四、运营情况

自智慧问答系统上线至 2025 年 5 月底，访问人数达 1400+，使用次数达 6300 人次。

同期团队针对 AI 技术进行推广，同时推进企业内部智慧问答系统的宣讲。2024 年 11 月，AI 内训团队组织开展分公司应用技巧线上培训，培训覆盖 27 家分公司。2025 年 3 月，AI 内训团队在全国 27 家分公司，总部 21 个部门开启轮回培训宣讲。内部建立答疑群，及时在群内推送应用更新优化情况。

五、项目成效

直接经济效益

1. 大幅降低人工工作量：AI 机器人可替代部分基础咨询工作，减轻人工处理工作量，可更好的服务业务。
2. 提高处理效率：机器人可 24/7 实时响应，缩短用户等待时间，提升整体效率。
3. 减少错误率：自动化回答可避免人工解答的疏漏，降低因错误信息导致的重复沟通和标准出入。

间接经济效益

1. 提升员工满意度：快速、准确的解答能提高内部用户（员工）的体验，减少因报错问题导致的负面情绪。
2. 优化财务团队工作重心：释放员工精力，使其专注于高价值任务（如数据分析、流程优化），提升整体财务效率。
3. 增强数据洞察能力：通过机器人收集的高频问题，为后续流程优化提供数据支持。

技术效益

1. 推动企业智能化转型：落地 AI 技术应用，积累经验，为未来更多场景的 AI 部署奠定基础。
2. 提升系统集成能力：机器人可与现有系统（如 ERP、OA）对接，强化企业技术生态的协同性。

战略效益

1. 支持规模化发展：随着公司扩张，机器人可无缝应对咨询量增长，避免人力资源的线性增加。
2. 前瞻性布局：抢占 AI 在财务领域的应用先机，为未来智慧财务（如自动化审核、预测分析）积累经验。

六、经验总结

技术架构与业务需求深度融合，保障项目实效性

1. 对接业务需求：深入剖析企业知识管理与财务咨询等业务流程中的痛点与难点，将其拆解为多个具体的应用场景。通过打造智能问答核心引擎、多源知识整合、精准答案推送等关键模块，确保系统功能紧密贴合业务实际需求。
2. 分步推进项目实施：运用“需求调研—模型选型—系统开发—测试优化—推广应用”的逐步推进模式，加速项目从概念设计到全面落地的进程，确保各阶段成果都能有效承接并推动后续工作开展。

智能技术优化提升，构建长效运营机制

1. 智能问答性能优化：建立系统性能监测体系，实时追踪问答准确率、响应时长、用户满意度等关键指标，依据数据反馈对模型参数、知识库结构、问答逻辑进行持续优化，实现问答质量与效率的螺旋式提升。
2. 全员赋能知识共享：公司内部成立跨部门项目推广小组，采用线上线下联动的方式，为 27 家分公司及总部 21 个部门提供培训，着重讲解智能问答系统的应用场景拓展、知识更新维护以及复杂问题处理技巧，打造企业内部知识共享与智能协作的良性生态。



江南农村商业银行：“小江智脑”大模型应用平台

一、项目背景及目标

随着人工智能技术的快速发展，发展数字经济和数字金融已成为推动银行业数字化转型、拥抱新一轮科技革命的重要途径。利用大模型技术可以提升银行服务的智能化水平，降低人力成本，增强客户体验，提高内部运营效率。因此，江南农村商业银行持续开展新技术研究，跟进大模型发展趋势，立足本行自主研发大模型应用平台，以实现技术自主可控，并推动金融服务的创新，加快数字化转型步伐，提升竞争力。主要实现目标有：提升数据知识利用率、提升办公效率、降低运营成本、提升客户满意度。

二、创新点

1. 全栈信创生态体系的突破性构建

江南农村商业银行信创大模型应用平台，依托千亿级开源大模型，深度整合国产软硬件资源，包括向量数据库、麒麟操作系统、达梦数据库及信创 GPU、CPU 等，实现技术自主可控与安全，为金融信息安全提供坚实后盾。

2. AI 能力场景化多覆盖

平台打造“1+3+N 大模型应用体系”，采用了国产化显卡 DCU 与英伟达 GPU 的异构算力架构，构建了稳定的基础计算能力。在此基础上应用场景集中在办公、营销、运营管理、信贷及研发等五大关键业务领域，建立一套全面的员工办公辅助系统，紧密契合一线员工的具体工作需求，达到了成本优化与效率提升的双重目的。

3. 研发效能革命性提升的工程实践

内置辅助代码插件，支持多种编程语言、多种编辑器，具备多项提升开发效率的创新力。支持基于上下文的实时代码续写，显著提升编码速度；可将自然语言需求转换为可执行代码，降低开发门槛；并能自动生成高质量的单元测试案例，提升代码可靠性。通过智能分析代码结构，提供优化建议与注释生成，增强代码可读性与维护性。同时，其具备强大的代码解释与智能问答能力，能够快速解答复杂问题，并在出现异常报错时，实现快速定位与修复建议推送，全面提升软件研发效率与质量。

4. 融合跨模态协同训练的多模态 OCR 大模型

江南农村商业银行利用多模态大模型应用技术，集成了文本、图像等多源数据，实现了高效的信息提取，大幅提升了 OCR 识别的效率和准确性。

5. 强语义理解能力驱动的智能文本机器人

江南农村商业银行在构建智能问答过程，通过构建合理的提示词 prompt 不断生成和修正答案，并利用外挂知

识库的方式，将大模型与本行特色业务逻辑相结合，提高了答案生成的准确性。这体现了 RAG 技术的核心思想，即结合外部知识库来增强大模型的生成能力。文本机器人展现了出色的语义理解和上下文捕捉能力。

三、项目技术方案

系统架构

“小江智脑”大模型应用平台可以划分为多个层次，每个层次都承担着不同的功能和任务，共同协作以实现高效大模型功能应用。采用分层架构设计，可作为大模型架构参考，遵循分层架构的设计原则：

1. 基础资源层：提供通用大模型应用平台运行时候的资源支持，包括各种 CPU、GPU、内存、硬盘和操作系统资源。
2. 基座模型层：提供主流开源大模型支持，包括 Qwen2.0-72B、Qwen2.0-7B、Stable Diffusion3、GLM-4V-9B 等。这些组件共同构成了应用平台的核心引擎。
3. AI 能力层：AI 涵盖了大模型平台的核心能力部分，为应用场景提供大模型能力。
4. 应用层：使用 AI 能力赋能各种业务场景。大模型的应用主要集中在办公、营销、运营管理、信贷及研发等关键业务领域，其目标是建立一套全面的员工办公辅助系统，紧密契合一线员工的具体工作需求，从而达到成本优化与效率提升的双重目的。

项目技术

1. 基于 RAG 技术的智能办公助手研发与应用

江南农村商业银行在 AI 办公助手研发中，创新性采用检索增强生成（RAG）技术架构，构建“知识检索 + 内容生成”双引擎驱动模式。通过将行内业务知识库、制度文件、操作流程等结构化数据与大模型能力深度融合，实现三大突破：

业务知识调用：系统可从外挂知识库中检索政策、业务规则及历史案例，突破传统大模型知识固化局限，确保生成内容紧跟业务需求；

可信决策支持：所有生成答案均附带知识来源标注，形成可追溯的决策链路，有效降低信息偏差风险，助力打造“可解释、可审计”的智能办公体系；

该实践不仅解决了传统 AI 在金融场景中知识滞后、领域泛化能力弱等痛点，更通过 RAG 技术实现了大模型通用能力与银行业务专精能力的有机融合，为金融机构智能化转型提供了可复用的技术范式。

2. 构建场景化 Prompt 工程体系

江南农村商业银行系统性构建了面向金融场景的 Prompt 工程体系，通过多维度技术手段提升大模型对银行业务的理解与执行能力。具体包括：

场景化提示词设计：基于财务支出等高频业务场景，设计包含业务规则的结构化提示模板，引导大模型生成详细的符合要求的输出；

上下文增强机制：在客服问答中注入历史交互记录等上下文信息，提升应答个性化水平；

领域知识注入：将银行热词等专有知识嵌入 Prompt，强化大模型对专业概念的精准理解；

动态策略优化：持续迭代提示词模板，结合用户反馈数据动态调整，使大模型输出与业务需求保持同步演进。

该实践显著提升了大模型在行内知识库检索等任务中的准确率，这一工程体系为金融领域大模型场景化落地提供了可复用的方法论支撑。



3.AI Agent 智能体矩阵

江南农村商业银行基于大模型 AI Agent 技术，深度融合自然语言处理、机器学习与多模态交互能力：

多模态交互引擎：通过语音语义一体化建模，支持自然语言、语音、图像、等跨模态输入解析，构建日程助手、AI 识图助手、AI 绘图助手等场景化智能体，实现 " 语音指令 + 视觉反馈 " 的沉浸式交互体验；

智能协同进化：通过用户反馈数据持续优化任务执行策略；

AI Agent 智能体场景化应用成效包括：领导日程助手实现会议安排、语音发布工作日志等事务的自动化处理，节省了业务人员事务性工作时间；AI 识图助手、AI 绘图助手支持图片理解、生成图片等任务，应用于客户经理上门打卡照片审核、绘制宣传海报等多模态场景；语音转文字助手可识别语音内容，转换成文字，应用于客户经理上门拜访客户记录，客户画像生成等场景。

该技术体系不仅突破了传统单点 AI 工具的功能局限，更通过智能体间的协同联动，为金融领域智能化交互提供了可拓展的技术底座。

重点解决问题

提升内部知识管理效率：通过智能办公 AI 助手，快速检索内部知识资产，提高决策质量；

优化客户服务体验：利用文本机器人提供 24 小时智能客服，提高问题回复率和准确性；

提高软件开发效率：通过辅助代码编程工具，提升编码效率和代码质量；

增强业务系统智能化：通过多模态大模型和专用小模型，优化业务办理流程。

四、运营情况

江南农村商业银行自主研发的“小江智脑”大模型应用平台自上线以来运行稳定，已成功应用于智能客服、信贷审批、风险控制等多个业务场景。整体运营效果良好，为银行数字化转型提供了有力支撑。

五、项目成效

智能办公 AI 助手：服务超过 36.3 万人次，显著提升了员工的工作效率；

文本机器人：工作效率提高 16 倍，客户问题的有效回复率提升约 30%；

辅助代码编程工具：采纳行数超过 24.6 万行，开发人员工作效率平均提升 15%；

业务系统智能化：预估单笔交易节省 2 分钟，全年节约工时高达 40 万小时。

六、经验总结

江南农村商业银行的“小江智脑”大模型应用平台展示了如何通过技术创新来提升金融服务的智能化水平。通过自主研发的“小江智脑”平台，江南农村商业银行不仅实现了技术的自主可控，还成功地将大模型技术应用于多个业务场景，显著提高了工作效率和客户服务体验。未来仍需不断探索新技术融合，秉持创新精神，不断探索尝试新技术，持续进行模型的优化和更新，采用迁移学习等技术，以适应不同业务场景的需求。持续升级优化平台，以更好地满足客户日益多样化、个性化的需求。

四川农商联合银行：大模型驱动的智能问数应用

一、项目背景及目标

1. 项目背景

随着国家数字经济战略的深入推进和人工智能技术的飞速发展，金融行业的数字化转型已进入深水区。在这一浪潮中，数据被提升到前所未有的战略高度，成为驱动业务创新和精细化运营的核心生产要素。我行积极响应国家关于强化科技创新、推动金融高质量发展的号召，深刻认识到数据资产化、知识化、智慧化的重要性。正如《新一代数据架构》所倡导的，数据架构需向以知识和智慧为方向的更高价值密度演进。

在数字经济战略驱动下，金融业正加速向智能化转型。数据作为核心生产要素的战略地位日益凸显，但传统数据服务模式面临多重结构性挑战：一是系统架构之困—经多年信息化建设，银行普遍存在多套异构业务系统，数据整合效率低下，需求交付周期难以匹配业务敏捷化要求；二是人机协同之困—业务人员普遍受限于数据库等专业技术，长期依赖 IT 部门支持，导致数据价值释放滞后于业务创新节奏；三是安全发展之困—多数 AI 技术链依赖外部厂商，存在核心技术受制于人和数据安全隐忧。正如《新一代数据架构》所指出，数据服务需向知识化、智慧化方向演进，方能实现价值密度跃升。

作为地方性金融机构，我行已完成数据治理体系的基础建设，构建了统一数据仓库、实时计算平台、数据中台等核心能力。然而，实践中仍存在四大痛点：

数据“可用”与“易用”的鸿沟：尽管我行已建成强大的数据中台和各类数据服务，但业务人员在实际用数过程中，仍高度依赖 IT 部门进行数据提取、报表开发，响应周期长，难以满足日益敏捷化、场景化的业务需求。

业务需求与技术实现的错配：业务人员难以用精确的技术语言描述其数据需求，而 IT 人员对业务场景的理解深度有限，导致沟通成本高，数据成果与实际业务目标可能存在偏差。

数据价值释放不充分：海量数据中蕴藏的洞察和知识未能得到充分、高效的挖掘和利用，业务决策的智能化水平有待提升。一线业务人员渴望更直接、更便捷的方式洞察数据，赋能日常工作。

对自主可控技术的要求：在核心技术领域，特别是人工智能和大数据处理方面，积极采用并发展国产化技术，确保金融系统的安全稳定与长远发展。

在此背景下，大语言模型（LLM）技术的突破为解决上述痛点提供了革命性的路径。本项目通过构建基于大模型的智能问数，将复杂数据分析能力赋予业务人员，旨在打通数据服务的“最后一公里”，实现“人人用数据、时时用数据、处处用数据”的愿景，真正做到“用数赋智赋能业务发展”。

2. 项目目标

本项目以我行现有成熟的数据服务体系为坚实基础，通过引入先进的大语言模型技术，构建“大模型驱动的智能问数应用”。其核心目标在于通过技术革新，重塑我行业务人员与数据交互的方式，全面提升数据要素的价值创造能力，为我行高质量发展注入强劲的数智动能。具体目标归纳如下：

一是实现业务用数的范式革新与极致提效。本应用旨在赋予业务人员通过自然语言便捷、高效地进行自主数据查询与即时分析的能力。这将极大降低用数门槛，使一线人员能够独立、快速地响应业务变化，显著缩短数据获取与分析周期，支持敏捷决策。长远来看，便捷的用数工具将积极培养业务人员的数据思维和分析能力，营造“人人



用数据、时时用数据、处处用数据”的浓厚数字化氛围，从而加速我行整体数字化转型进程，全面提升全行数字化作业水平与运营效能。

二是深化数据智慧洞察与知识赋能决策。该应用致力于推动数据从原始记录向结构化信息、业务知识乃至决策智慧的逐级转化与价值沉淀。通过大模型的理解与分析能力，辅助业务人员更深层次、更精准地洞察数据背后隐藏的规律、市场趋势与潜在业务机遇，从而显著提升业务决策的科学性、前瞻性与精准度。

二、创新点

本项目在深入研判金融行业数字化转型趋势、充分借鉴业界先进经验的基础上，紧密结合我行多年信息化建设所积累的坚实成果与独特优势，进行了一系列具有实践性的关键创新。主要体现在以下三个方面：

（一）架构创新与自主可控 构建基于“Text2API”的国产化智能问数新范式，兼顾敏捷与稳健。

本项目在自然语言到数据的转换路径选择上，独辟蹊径。业界虽有直接“Text-to-SQL”的技术方案，但该方案在处理涉及复杂多表关联、深层嵌套查询以及特定业务逻辑封装时，大模型往往难以生成完全准确且高效的 SQL，尤其在我行业务系统繁多、数据模型关联复杂的背景下，错误 SQL 的风险与后期维护成本极高。为此，我行创新性地确立了以成熟、稳定、安全的数据服务 API 作为大模型统一调用“后端”的核心架构。这一“Text2API”模式，既充分复用了我行多年数据中台建设在数据标准化、服务化封装、业务逻辑沉淀及统一出口管控等方面的宝贵成果，又通过大模型赋予了这些既有服务自然语言交互的“前端”智慧能力。此举大幅降低了业务人员直接操作底层数据库的潜在风险，巧妙规避了复杂 SQL 自动生成的业界难题，显著简化了大模型理解和处理复杂业务逻辑的难度，实现了“前沿 AI 技术”与“稳健企业级架构”的完美融合与高效协同。

（二）融合赋能与安全保障 深度践行“数据智理”与“用数赋智”，并以金融级安全体系保驾护航。

在“数据智理”层面，规划通过结合大模型强大的自然语言理解与上下文感知能力，帮助业务用户更透彻地理解数据来源、加工过程与业务内涵，从而建立对数据的信任感，提升数据应用质量。在“用数赋智”层面，该应用直接赋能各级业务人员进行自助式、探索式的数据获取与即时洞察，使其能更便捷地将数据转化为驱动日常工作、优化业务流程、创新产品服务的智慧与动能，有力推动全行数据驱动型业务模式的构建与深化。

（三）敏捷迭代与前瞻进化 坚持业务价值驱动的敏捷开发。

本项目自始至终秉持以业务价值为导向的敏捷开发与迭代理念。从项目启动初期，便请核心业务部门深度参与需求梳理、场景共创和原型验证。通过“快速上线、小步快跑、持续反馈”的迭代模式，项目团队能够迅速捕捉用户在实际使用中的痛点与期望，并据此不断调优模型交互策略、完善应用功能设计、拓展高价值应用场景。这种紧密围绕业务需求、快速形成“应用 - 反馈 - 优化 - 再应用”的闭环机制，确保了技术创新与业务价值实现的紧密贴合和持续交付。

三、项目技术方案

项目规划

我行大模型应用自 2024 年 1 月正式启动，为确保“大模型驱动的智能问数应用”的高效构建与成功落地，已系统地规划了清晰的建设路径。整体建设将遵循“夯实基座、敏捷迭代、场景牵引、价值驱动”的核心策略，通过分阶段、递进式的方式稳步推进，确保每一阶段成果可验证、风险可管控，并能快速响应业务需求变化，持续创造并交付业务价值。

第一阶段核心目标在于奠定坚实的技术与应用基础，并初步验证应用核心价值。将全面完成超百卡国产算力集

群的优化配置，并搭建稳定高效的大模型推理环境。在此基础上，针对金融领域特性及智能问数核心场景，审慎进行大模型的选型与应用适配。通过 Prompt 工程设计、结合行内业务术语库与数据元数据信息，并优化与大模型的交互逻辑，重点提升大模型对银行业务语境、数据结构及用户真实意图的精准理解与响应能力，确保其能有效服务于下游任务。同时，构建“Text2API”核心转换引擎，打通大模型与本行现有数据服务总线，实现自然语言到 API 调用的精准、高效转换，首期将覆盖行内高频使用及核心业务领域的的数据服务接口。

第二阶段致力于横向拓展服务范围并纵向深化应用功能。将持续、有序地接入更多数据服务 API，进一步丰富可查询的数据维度与覆盖的业务领域。重点增强大模型在特定业务场景下的理解与泛化能力，主要通过引入更丰富的上下文信息、增强对金融领域特定问询模式的识别与处理能力，以提升模型对复杂查询、模糊查询的处理效果，以及在多轮对话中对上下文的精准理解与意图保持能力。

第三阶段应用将向更高阶的智慧赋能与良性生态构建演进。将基于对用户行为和业务数据的深度分析与模式挖掘，探索从传统的“人找数”向更为智能的“数找人、数提策”模式转变，实现主动为用户推送相关的业务洞察、风险预警及决策建议。

业务功能

“大模型驱动的智能问数应用”其业务功能设计深度聚焦于解决一线业务人员在日常工作中面临的数据获取难、理解难、应用难的痛点，致力于通过前沿的人工智能技术，构建一个围绕“便捷交互赋能、智能洞察驱动、安全可控保障”三大核心原则的全新数据服务生态。精确或歧义，模型具备模糊查询处理与智能容错能力，能通过引导式反问或提供候选选项辅助用户明确意图。

1. 自然语言驱动的智能查询与深度交互

该应用的核心革新在于赋予业务人员通过自然语言进行复杂数据查询的强大能力。用户无需掌握专业技术背景，仅需输入日常业务用语，系统内置的“Text2API”引擎便能凭借强大的语义理解和意图识别能力，自动将自然语言精准转换为对后台数据服务 API 的调用。该应用支持流畅的多轮对话，用户可进行追问、筛选或维度深化，模型能智能保持上下文关联，无需用户重复前提，使得数据分析如自然沟通般高效。针对用户表述可能存在的不精确或歧义，模型具备模糊查询处理与智能容错能力，能通过引导式反问或提供候选选项辅助用户明确意图。

2. 结构化结果呈现与交互式智能可视化分析

无论后台 API 返回何种格式数据，该应用均会自动解析并转换为易读的结构化形式，如动态表格、列表或关键指标卡片，用户可进行在线排序、筛选等操作。系统能根据数据特点和用户意图，智能推荐并即时生成多种合适的图表（如折线图、柱状图、饼图等），用户可与图表进行下钻、上卷等交互操作，实现从数据到图形洞察的无缝转换，极大提升数据解读的直观性。

3. 金融级精细化权限管控与全链路安全审计保障

数据安全与合规是平台的生命线。该应用与本行统一身份认证系统深度集成，确保用户身份的严格认证与唯一性。在此基础上，构建了严格且灵活的多维度数据查询权限管理体系，与用户岗位及所属机构层级（总行、分行、支行、网点）紧密绑定，确保用户仅能访问其职责与机构层级授权范围内的数据和服务 API。

四、运营情况

在用户引入、培训赋能及模型交互效果的持续优化方面，运营工作扎实推进。首批试点聚焦于总行部分核心业务部门，如电子银行部等，通过建立响应及时的技术支持与交流群组等方式，有效帮助试点用户快速熟悉各项功能。



通过搭建模型性能实时监控仪表盘，持续追踪包括查询意图理解准确率、API 调用成功率、用户反馈满意度以及无效问答率等在内的关键性能指标（KPIs）。

核心的用数权限管理模块和全流程问答日志查询与审计功能已稳定高效运行，为应用的安全合规运营提供了坚实的技术保障。为拓宽智能问数的服务边界和应用深度，在首批已接入的核心业务 API 基础上，运营团队正根据业务发展优先级和试点用户的实际需求，与数据中台及各业务系统团队紧密协作，动态梳理并有计划地逐步对接更多领域、更深层次的数据服务接口。

目前，在试点初期，该应用日均处理用户各类自然语言查询数千次。试点用户普遍对应用提供的自然语言交互方式的便捷性、数据获取的即时性以及操作的友好性表示高度认可，并积极地为应用的进一步完善建言献策，形成了良好的互动氛围和应用生态雏形。虽然应用尚处于运营初期，但其在赋能业务决策、提升工作效率、降低用数门槛等方面的潜力已初步显现。随着模型交互效果的持续增强、功能的日益丰富、接入数据范围的不断扩大以及用户群体的逐步推广，预计各项关键运营指标将呈现稳步提升的良好态势，该应用的业务价值也将得到更充分的释放。

ID 五、项目成效

尽管尚处早期，该应用已在赋能业务、提升效率、夯实技术基础及培育数据文化等多个维度展现出初步成效，充分预示了其在推动我行数字化转型、提升核心竞争力方面的巨大潜力和深远价值。

首先，在赋能业务人员自主用数、提升数据驱动决策能力方面成效初显。试点业务用户已能突破传统数据获取模式的束缚，通过自然语言直接与该应用进行交互，便捷、高效地获取此前需依赖 IT 支持或复杂 BI 工具操作方能得到的数据。

其次，在关键技术自主可控与应用经验积累方面取得了重要突破。本项目坚定践行国家信创战略，成功依托我行超百卡国产算力集群，完成了大模型在金融核心业务场景下的应用适配、稳定部署与高效运行。这实现了在数据探索与创新应用、提升用数灵活性的同时，有效防范数据泄露和误用风险，充分体现了“安全是发展的前提，发展是安全的保障”的理念。

随着该应用功能的持续迭代完善、模型智能水平的不断进化以及应用范围的逐步推广与深化，预计未来在进一步降低运营成本、提升全行决策质量与效率、创新金融产品与服务模式、增强风险精准防控能力等方面的成效将更加显著和量化。本应用将为我行在日益激烈的市场竞争中注入强大的、可持续的“数智动力”。

ID 六、经验总结

“大模型驱动的智能问数应用”的建设与初步运营，是我行在人工智能时代积极拥抱新技术、探索数据价值释放新路径的宝贵实践。在此过程中，积累的关键经验在于：坚实的数据底座是前提，我行多年数据治理与中台建设为项目成功奠定了基础；场景驱动与业务深度融合是关键，确保技术方案能真正解决业务痛点并创造价值；坚定国产化技术路线不仅保障了项目自主可控，更提升了核心技术能力，对金融安全具有长远战略意义；大模型在金融领域的应用适配是一个持续优化、体系化运作的复杂过程，需结合领域知识与用户反馈不断迭代；同时，复合型人才培养与跨部门高效协同是项目成功的组织保障。这些经验深刻揭示了技术创新与业务实践相结合的内在规律，为我行未来深化数字化转型提供了重要启示。

恒丰银行：恒运 AI 答

一、项目背景及目标

1. 市场 / 需求分析

当前，金融行业的数字化转型浪潮方兴未艾，人工智能技术在优化运营流程、提升服务效率方面的应用日益广泛，智能化已成为金融机构提升核心竞争力的关键驱动力。银行运管条线作为连接前中后台的关键枢纽，其业务具有高度的知识密集性。涉及的政策法规、操作规程、业务指南、风险控制要点等知识点繁多且更新迭代迅速，对信息的准确性和时效性要求极高。传统依赖人工翻阅海量文档、通过邮件或电话层层咨询的知识获取方式，不仅效率低下、响应迟缓，还易导致知识传承断裂、答案口径不一等问题，难以满足快节奏、高标准的业务需求。

具体到用户层面，恒丰银行运管条线的员工，包括柜面人员、后台操作人员、风险管理专员及合规人员等，在日常工作中迫切需要一个能够快速、准确提供业务知识和操作指引的工具，以高效应对各类工作任务和客户咨询。现有的传统知识库系统往往存在检索体验不佳、信息过载严重、难以精准定位答案等弊端，无法有效满足一线员工的实际需求。

从业务挑战来看，主要体现在以下几个方面：

- 运营效率瓶颈：大量重复性的内部咨询和知识查询工作占用了员工宝贵的时间和精力，影响了核心业务的处理效率。
- 知识管理与更新压力：运管相关的制度、金融产品、业务流程频繁调整更新，如何确保每一位一线员工都能及时、准确地掌握最新知识，是一项持续的重大挑战。

服务一致性与合规风险：人工解答可能因个人理解偏差或信息滞后而导致服务口径不一，严重时甚至可能引发操作差错或合规风险。

因此，构建一个智能化的问答系统，以提升运管条线的知识服务能力和运营效能，已成为恒丰银行内部亟待解决的重要课题。

2. 立项背景

为应对上述挑战，恒丰银行以提升运营管理精细化水平、赋能一线员工、有效降低运营成本为内部驱动，积极寻求 AI 技术突破。近年来，大型语言模型 (LLM) 以及检索增强生成 (RAG) 技术的快速发展和逐步成熟，为解决金融领域知识密集型任务的痛点提供了切实可行的技术路径。

尤为重要的是，项目启动之初，恒丰银行已具备了坚实的数据基础。运管条线经过细致梳理和整合，积累了超过 200 份核心业务文档，这些文档涵盖了内部规章制度、标准操作手册、金融产品详细说明、常见问题解答 (FAQ) 等关键信息。同时，还整理形成了超过 10000 条高质量的运营专家知识问答对。这些宝贵的内部知识资产为构建 RAG 系统的知识库提供了丰富且权威的来源，是项目得以顺利启动并取得预期成效的关键前提。

3. 项目目标

“恒运 AI 答”项目的目标是多维度、多层次的，旨在通过技术赋能实现以下目标：



- 业务目标：构建一个高效、精准、易用的运管条线智能问答助手。该助手旨在显著提升运管条线员工获取专业知识的效率和解决实际问题的能力，有效降低人工咨询压力，确保信息传递的准确性和一致性，从而为运管各项业务的高质量、稳健发展提供有力支撑。
- 技术目标：探索并实践一条轻量级、高性价比的 AI 应用落地路径。具体而言，在不依赖复杂知识图谱构建和成本高昂的大语言模型微调的前提下，通过对基础 RAG 技术栈的深度优化，特别是针对金融运管领域文档特性的精细化文档分块策略的应用，实现行业领先的业务问答满意度。

二、创新点

“恒运 AI 答”项目在建设理念、技术实现和应用模式上均体现了显著的创新性，尤其在资源和成本约束下实现了卓越的性能表现。

- 轻量级高效 RAG 实践的典范：本项目最大的创新在于，面对金融行业普遍存在的对 AI 高投入、长周期的顾虑，另辟蹊径地证明了仅通过对基础 RAG 架构的深度理解和精细化应用，特别是针对金融运管领域复杂文档特性进行的多策略组合文档分块优化，即可在不构建复杂知识图谱、不进行大模型微调的情况下，实现高达 97% 的问答满意度。这充分体现了技术选型的务实性和对成本效益的极致追求，为同业提供了宝贵的“以简驭繁”的 AI 落地经验。

- 面向金融文档的深度理解与处理优化：针对运管条线文档普遍存在的多样性（如大量制度条款、操作流程图、数据表格）和高度专业化的特点，项目团队探索并实践了一套行之有效的文档解析、智能清洗、关键信息结构化提取以及多层次语义分块的方法论。这套方法论确保了知识库的质量和检索相关性，是 RAG 系统性能的关键保障。

- 精准聚焦运管核心高频场景的应用创新：系统深度契合运管条线员工在日常工作中面临的政策查询、流程确认、风险咨询等高频场景，提供“即问即答”的精准知识服务。这种以解决一线实际痛点为出发点的应用设计，确保了系统的实用性和高用户粘性。

- 动态交互式知识赋能新模式：“恒运 AI 答”成功将银行内部大量静态的、分散的制度文档和操作手册，转化为动态的、可便捷交互的智能问答服务能力。这不仅极大地便利了知识的获取，更促进了知识在团队内部的快速传播、深入理解和有效应用，从而整体提升了运管团队的专业素养和风险意识。

- 用户驱动的闭环迭代体验创新：项目注重用户体验，提供了简洁易用的交互界面，并设计了清晰的答案呈现方式，如关键信息高亮、精准引用溯源（明确答案来源的运营文档具体段落），显著提升了答案的可信度和用户的核验效率。更重要的是，建立了完善的用户反馈机制，通过对用户提问、点赞点踩、意见建议的持续分析，驱动知识库、分块策略和问答逻辑的快速迭代优化。

- 务实的知识管理与推广模式：在项目推行过程中，坚持“业务主导、技术支撑”的原则，通过试点先行、快速迭代、用户深度参与（如邀请业务骨干参与知识校验和系统测评）等方式，确保了系统功能与业务需求的紧密贴合和持续优化，为系统的顺利推广和广泛应用奠定了坚实基础。

三、项目技术方案

1. 项目规划

“恒运 AI 答”项目的整体规划思路是以恒丰银行运管条线的核心业务需求为根本导向，充分依托现有的 200 余份专业文档和超过 10000 条高质量问答对的知识基础。技术路线上，坚定选择并优化基础的检索增强生成 (RAG) 技术，避免了知识图谱构建和模型微调带来的高昂成本和复杂性。项目成功的关键在于通过对文档处理流程、特别是文档分块策略的持续迭代优化、高效检索算法的选择以及精细化的提示词工程，最终构建一个响应迅速、答案精准、运行稳定且用户体验良好的智能问答系统。

项目的关键实施阶段规划如下：

1. 数据准备与知识库构建阶段：此阶段的核心任务是对运管条线提供的各类文档进行全面的收集、细致的清洗、智能化的解析以及必要的结构化处理。随后，对处理后的文本数据进行初步的文档分块，并利用选定的 Embedding 模型进行向量化转换，构建初步的向量知识库。
2. RAG 核心功能研发阶段：基于业界成熟的 RAG 理论框架，搭建系统基础的 RAG 流程。这包括实现独立的检索模块 (Retriever)，负责从向量知识库中高效查找与提问相关的文档片段；以及生成模块 (Generator)，负责整合检索到的信息并生成自然语言答案。
3. 文档分块策略优化与深度测试阶段：这是项目的技术攻坚核心。针对运管文档的特性 (如长文本、条款密集、含表格等)，设计并实验多种文档分块策略 (如不同大小的固定分块、语义分块、基于章节的分块等)，并通过严格的效果评估 (如召回率、答案相关性分析)，筛选并确定最优的分块方案组合。
4. 系统集成与用户体验优化阶段：开发用户友好的前端交互界面，确保操作的便捷性。优化答案的呈现方式，例如加入引用溯源功能，提升答案的可信度。组织小范围用户进行实际场景测试，收集反馈。
5. 上线试运行与反馈驱动迭代阶段：在特定部门或用户群体中进行小范围推广试运行，密切监控系统表现，广泛收集用户反馈。基于反馈数据和运行日志，对模型参数、分块逻辑、知识库内容进行持续的敏捷迭代和优化。

2. 实施方法

项目在实施过程中，采用了敏捷开发 (Agile Development) 的方法论。我们坚持小步快跑、快速迭代的原则，早期阶段高度关注核心功能的快速实现和用户真实反馈的获取，以便及时调整开发方向，确保最终产品与业务需求的高度契合。

资源投入方面，项目组构建了一支由多领域专业人才构成的复合型团队，其中包括深刻理解运管业务流程和知识体系的业务专家，负责 AI 模型选型、算法优化的 AI 工程师，专攻数据处理与知识库构建的数据工程师，以及负责用户界面和交互体验的前端工程师。特别是在项目初期，银行在运管文档的梳理、人工校验和高质量 Q&A 对的标注与审核方面投入了显著的人力资源，这为后续 RAG 系统的高性能表现奠定了坚实基础。

技术栈选型遵循以下核心原则：

- 成熟稳定性：优先选择在业界得到广泛应用、经过大规模验证的成熟技术和框架。
- 社区支持度：考量技术的社区活跃度、文档完善程度以及问题解决资源的可获得性。
- 金融行业适用性：确保所选技术能够满足金融行业在数据安全、系统合规、可审计性等方面的严格要求。
- 成本效益：在满足功能和性能的前提下，综合评估技术的采购成本、开发成本和长期维护成本。

3. 技术架构

“恒运 AI 答”系统整体采用分层化的模块化架构设计，确保各组件功能内聚、低耦合，易于维护和扩展。其核心数据流和控制流如下所示：

用户输入 → [前端界面 → 应用服务层 → [问题预处理] → [查询向量编码] → AI 核心引擎 (RAG 模块) → [向量数据库检索] → [相关文档块返回] → [上下文构建与 Prompt 生成] → [大语言模型 (LLM) 推理] → [答案后处理与溯源] → 应用服务层 → [前端界面 → 用户获取答案]

核心模块说明

- 数据处理与知识库构建模块



该模块是智能问答系统的基石，负责将原始的运管知识转化为机器可理解和检索的格式。

文档来源：系统处理的主要数据源包括恒丰银行内部权威的 200 余份运管条线专业文档（格式涵盖 PDF、Word、Excel、HTML 等），以及超过 10000 条由业务专家梳理和校准的高质量问答对（Q&A pairs）。这些 Q&A 对对于提升常见问题的解答效率和准确性至关重要。

文档解析与清洗：此步骤致力于从原始文档中提取高质量的纯文本内容。具体操作包括：利用 OCR 技术处理扫描版 PDF，准确提取各类文档中的文本信息；进行细致的格式转换和统一；有效去除页眉、页脚、水印、无关图片等噪音数据；妥善处理特殊字符和编码问题。特别针对金融文档中常见的表格和列表内容，开发了专门的脚本进行结构化提取和转换（例如，将表格转换为易于 LLM 理解的 Markdown 格式）。

知识库构建：经过清洗和结构化处理的文档内容，连同已有的 Q&A 对，将通过精心选择的 Embedding 模型（例如，针对中文优化的 BGE-large-zh 或类似的高质量句向量模型）转换为高维语义向量。这些向量随后被高效地索引并存储在专业的向量数据库（例如，FAISS、Milvus 等，具体选型基于性能、部署便捷性和维护成本考量）中，形成可供快速检索的向量知识库。

• 检索增强生成 (RAG) 模块

这是系统的智能核心，严格遵循基础 RAG 架构，未采用知识图谱进行复杂的图谱问答，也未对所选的大语言模型本身进行任何形式的微调 (Fine-tuning) 操作，力求在基础技术框架内实现性能最优化。

- 用户问题理解与编码：当用户通过前端界面提交问题后，系统首先对原始提问进行必要的预处理，如基本的拼写纠错、去除无关符号等。随后，使用与文档编码阶段相同或高度兼容的 Embedding 模型，将用户的自然语言问题转换为查询向量。

- 核心技术：文档分块策略 (Chunking Strategy) 的精细化实践

文档分块是决定 RAG 系统性能的咽喉环节，其质量直接影响后续检索召回的精准度和最终生成答案的相关性与完整性。不恰当的分块会导致关键信息丢失、上下文割裂或引入大量噪音，从而严重损害问答效果。恒丰银行在“恒运 AI 答”项目中，针对运管文档的特性，探索并实践了如下组合分块策略：

恒丰银行运管文档分块策略实践要点：

1. 混合粒度分块 (Hybrid Granularity Chunking)：认识到运管文档内容多样性（条款、案例、流程、定义等），单一分块策略难以普适。因此，采用多种策略组合：

- 固定大小分块 (Fixed-size Chunking) 结合智能重叠 (Intelligent Overlap)：作为基础策略，设定一个相对适中的块大小（如 300-500 tokens，根据所选 LLM 的上下文窗口和文档特性实验确定），并设置合理的重叠率（如块大小的 10-20%），以确保语义的连贯性，避免重要信息在块的边界被切断。选择此范围是基于对 LLM 有效信息处理范围的考量和初步实验结果。

- 语义分块 (Semantic Chunking)：超越简单的字符或 Token 计数，尝试基于自然语言的语义边界进行切分。例如，优先确保一个完整的政策条款、一个独立的操作步骤说明、或一个 FAQ 的问答对被包含在单个块内。这可能借助 NLP 工具识别句子结束符、段落标记，或利用轻量级模型辅助判断语义完整性。

2. 针对性内容处理与分块：

- 表格内容结构化与独立分块：对于文档中常见的表格数据（如费率表、权限表），首先进行结构化提取（如转为 Markdown 或 JSON），然后可能将整个表格（若大小适中）或表格的关键行 / 列组合视为一个独立的、富含信息的块。

- 列表和枚举项的保留：确保列表项（如操作要点 1.2.3...）尽可能在同一块内，或通过重叠保证连续性。

3. 元数据增强分块 (Metadata-Enriched Chunking)：为每一个生成的文档块附加丰富的元数据。这些元数据包括但不限于：原始文档名称、文档版本、章节标题、发布日期、关键词、主题分类（如“反洗钱规定”、“柜面操

作流程”等)。这些元数据不仅能在检索时用于精确过滤,还能在后续对检索结果进行排序或给 LLM 提供额外上下文时发挥重要作用。(参考:万字长文告别粗暴切分)

4. 父文档与子块关联逻辑 (Parent Document - Child Chunk Association): 虽然系统采用基础 RAG,但借鉴了先进 RAG 策略中的“小块嵌入 - 大块检索”思想的简化版。具体做法是,在将文档分割成较小的、语义集中的“子块”(Child Chunks)进行 Embedding 和初步检索,以提高检索的精确度和相关性。当检索到最相关的若干子块后,系统会智能地将这些子块连同其紧邻的上下文(可能来自同一个“父段落”或扩大到一定范围的“父块”)一并提供给 LLM。这样做的好处在于,检索时利用了小块的定位精准性,而在生成答案时,LLM 能获得更完整、更丰富的上下文信息,有助于生成更全面、逻辑更连贯的答案,有效避免“只见树木不见森林”的问题。

5. 基于 Q&A 对的知识补充: 现有的 10000+Q&A 对直接作为高质量的知识块存入向量库,对于高频或特定表述的常见问题,可以直接匹配并返回标准答案,极大提升了这部分问题的响应速度和准确率。

6. 持续的实验与迭代优化: 强调文档分块策略并非一次设定即一劳永逸,而是项目团队通过大量的实验(对比不同参数设置、不同策略组合的效果)、严格的评估(包括人工评估和自动化指标如召回率、MRR 等)以及听取业务专家的反馈,进行持续调整和优化的过程。

如何通过精细化分块策略提高问答准确率: 上述一系列精细化的分块策略,确保了被检索到的每一个文档块都尽可能地包含与用户问题紧密相关且语义完整的上下文信息。这有效减少了无关信息的干扰,降低了信息噪音,从而为大语言模型准确理解问题、并依据可靠上下文生成高质量、高相关性的答案奠定了坚实基础。这是“恒运 AI 答”在未使用模型微调的情况下仍能达到 97% 高满意率的核心技术保障之一。

向量检索与排序: 当用户查询向量生成后,系统会利用该向量在预先构建的向量数据库中执行高效的相似度搜索(通常采用余弦相似度算法)。系统会召回 Top-K 个与查询语义最相关的文档块。为进一步提升送入 LLM 的上下文质量,系统内部可能集成了一个轻量级的重排序(Re-ranking)步骤(若未使用显式重排模型,则此处强调的是高质量 Embedding 和优秀分块带来的原生高相关性排序)。

上下文构建与提示词工程 (Prompt Engineering): 将经过检索(可能还有重排序)筛选出的相关文档块,与用户的原始问题,按照精心设计的模板整合成一个结构化的输入提示 (Prompt)。Prompt 的设计是 RAG 成功的又一关键,它需要清晰地引导大语言模型 (LLM) 如何利用提供的上下文信息来回答用户的问题,同时可能还需指示 LLM 遵循特定的回答风格(如简洁明了、正式专业)、答案长度,以及强调必要时引用信息来源。

答案生成: 将构建好的提示词 Prompt 输入给选定的大语言模型。模型选型会倾向于选择在中文理解和生成、金融领域知识处理方面表现较好,且部署和使用成本可控的基础大语言模型(如国内主流厂商提供的成熟模型 API,或优秀的开源模型)。大语言模型依据提供的上下文和问题指令,生成初步的自然语言答案。

• 问答处理模块

答案后处理与引用溯源: 大语言模型生成的初步答案会经过必要的后处理环节。这包括: 去除可能存在的冗余表达或模板化语句; 对答案进行格式化,使其更易于阅读; 在某些情况下,可能还需要进行敏感信息过滤。尤为重要的是,系统致力于为答案添加准确的引用来源,例如,明确指出答案依据的是哪份文档的哪个章节或具体条款。这对于运管业务的严谨性要求至关重要,极大增强了答案的可信度和用户进行核实的便利性。

对话管理: 针对可能出现的多轮追问场景,系统具备基础的会话状态管理能力,能够在一定程度上理解并关联上下文中的指代,以提供更连贯的问答体验。

4. 业务 / 商业模式

“恒运 AI 答”项目主要定位为恒丰银行内部的效率提升与知识赋能工具,其核心价值在于内部运营优化而非直



接对外产生商业收入。其业务模式体现在深度融入运管条线的日常工作流程，并通过智能化手段创造间接的业务价值。

四、运营情况

自“恒运 AI 答”正式上线以来，恒丰银行在运管条线内部进行了系统性的推广应用。目前，该智能问答助手已全面覆盖总行运营管理部、风险合规部运管相关处室，以及各一级分行、直属支行的运营管理团队、柜面业务骨干等关键岗位。据初步统计，系统当前活跃用户已超过 1500 人，日均处理各类运管业务相关查询请求达到 2000 余次，高峰期（如新政发布、月末季末）查询量可突破 3000 次。

为确保“恒运 AI 答”上线持续稳定运行，开展如下工作：

- 7×24 小时监控与响应：建立完善的系统监控体系（服务器资源、应用状态、API 响应、知识库更新状态等），设置关键指标阈值告警，确保 7×24 小时快速响应和故障处理。

- 高效用户支持体系：建立清晰、多级（如一线客服 / 内部 IT 支持、二线项目组 / 厂商专家）的用户问题反馈与支持渠道（如在线反馈表单、专门支持邮箱或即时通讯群），确保用户问题得到及时、专业的解答和解决。

为确保系统顺利推广并发挥最大效能，项目组采取了以下策略与方法：

- 试点先行，逐步铺开：在全面推广前，选取了若干业务代表性强、员工接受度高的分支机构作为试点单位。通过试点运行，验证了系统的稳定性和实用性，收集了宝贵的优化建议，并树立了成功样板。

- 高层重视与推动：项目获得了行领导层的高度重视与支持，将其定位为提升运管效能、赋能一线的重要举措，为全行推广营造了积极氛围。

五、项目成效

“恒运 AI 答”的推广应用，已在恒丰银行运管条线的日常工作中产生了显著且多方面的业务价值：

1. 对运管条线工作的直接、高效支持

快速精准解答高频问题：员工在处理日常业务时，可即时通过系统查询各类政策规定（如反洗钱要求、账户管理细则）、具体业务流程（如特殊业务授权、差错处理步骤）、金融产品特性、以及内部 IT 系统的操作指南等。系统能够迅速提供标准答案并附带原文出处，有效取代了以往繁琐耗时的文档翻阅或四处电话咨询。

辅助复杂、偶发业务场景处理：在面对一些非常规或操作较为复杂的业务场景时，系统能提供相关的知识点梳理、历史案例参考（若有）和合规要点提示，帮助员工更全面地理解和应对。

加速新员工培训与业务上手：新入职员工或岗位轮换员工，可以通过“恒运 AI 答”进行自主学习和即时查询，快速掌握所需业务知识和操作技能，显著缩短了培训周期和适应期。

提升合规操作的准确性与一致性：系统提供的答案均源自官方权威文档，确保了信息的标准性和统一性。这有助于减少因个人理解偏差或信息掌握不全导致的操作差错，从源头上降低了潜在的合规风险。

2. 显著的运营效率提升

据抽样调研和用户反馈，员工通过“恒运 AI 答”查找和确认业务知识点的平均耗时，已从原先的平均 10-20 分钟（复杂问题可能更长）大幅缩短至 1 分钟以内，简单问题基本实现秒级响应。

系统上线后，运管条线内部通过邮件、电话等传统方式进行的重复性业务咨询数量显著下降，初步估计降幅超过 60%。这使得资深员工能从繁琐的答疑工作中解放出来，投入到更具价值的核心业务分析和流程优化上。

由于知识获取更为便捷高效，员工处理单笔具体业务（如一笔复杂的对公账户开立审核）的平均处理速度亦有明显提升，初步估算整体业务处理效率提升约 15-20%。

3. 有效的知识沉淀、共享与传承

“恒运 AI 答”构建了一个动态更新的、中央化的运管知识库。它将原本分散存储于各类电子或纸质文档、甚至留存于资深员工脑海中的隐性知识，进行了有效的显性化、结构化和数字化整合，形成了一个全行运管条线统一的、易于检索和持续更新的知识平台。

极大地促进了最新业务知识和操作规范在团队内部的快速、无障碍共享，打破了可能存在的信息孤岛，确保了所有员工都能平等、便捷地获取到最权威、最及时的业务信息。

通过对用户查询热点、反馈问题的分析，系统还能反向揭示出当前业务流程或制度文件中可能存在的模糊地带、理解难点或潜在的改进点，为运管管理部门的业务优化和制度完善提供了有价值的参考，间接支持了业务的持续创新。

六、经验总结

核心经验

1. 基础 RAG 的巨大潜力与卓越成本效益：项目实践雄辩地证明，在知识密集型的特定业务领域（如银行运管条线），即便不投入巨资构建复杂的知识图谱，也不对大语言模型本身进行成本高昂的微调，仅仅通过对基础 RAG 架构的精心设计和深度优化，也完全有能力达到非常高的系统性能（本项目实现了 97% 的用户问答满意率）。这种技术路径不仅大幅降低了 AI 应用的门槛，其开发周期和长期维护成本也相对更为可控。

2. 文档预处理和数据清洗是成功的基石：RAG 系统的核心理念是“检索增强生成”，其效果高度依赖于检索到的上下文质量。因此，“Garbage in, garbage out”（垃圾进，垃圾出）的原则在这里体现得淋漓尽致。在项目早期对原始运管文档进行彻底的解析、格式统一、噪音数据去除（如水印、页眉页脚）、以及关键结构化信息提取等方面投入的巨大精力，被证明是极其值得的，它直接决定了知识库的“纯净度”和可用性。

3. 文档分块策略是 RAG 系统的灵魂与核心技术引擎：这是本项目成功的关键所在，也是投入研发精力最多的环节。不存在“放之四海而皆准”的万能分块策略：必须根据业务文档的具体类型（如制度文件、操作手册、产品说明、FAQ）、内容组织特点（如条款密度、章节结构、图表占比）、大语言模型上下文窗口的实际大小以及最终的应用场景（如精确问答、信息摘要）等多种因素，进行反复的、细致的实验对比和参数调优。

“小块精准定位，大块保障上下文完整”的平衡艺术：在实践中深刻体会到，将文档切分成语义完整且粒度较小的“子块”（Semantic Chunks），更有利于提升向量检索召回结果的精确度；然而，在将这些检索到的信息提供给大语言模型进行答案生成时，又需要确保上下文的相对完整性，以避免模型因信息碎片化而产生误解或生成片面答案。因此，如何动态地提供和组织这些上下文（例如，借鉴父文档检索逻辑，或在召回多个相关小块后进行智能合并与扩展）是一项核心挑战。

元数据的战略价值不容小觑：为每一个文档分块精准地附加丰富的元数据信息（如原始文档名称、具体章节号、发布日期、关键词、所属业务分类等），不仅能够显著提升后续向量检索的过滤能力和结果排序的相关性，还能为用户提供更清晰的答案溯源，增强信任感。

4. 提示词工程（Prompt Engineering）是提升答案质量的“最后一公里”：一个精心设计、持续优化的 Prompt 模板，能够有效地引导大语言模型准确理解上下文信息、聚焦关键点、遵循预设的回答格式和风格，并减少不必



要的幻觉或冗余输出。这项工作需要结合实际运行中出现的 bad cases 不断迭代。

5. 建立多维度、量化的评估体系是持续优化的导航仪：除了最终的用户满意率，还应在开发和迭代过程中建立包括检索召回率 (Recall)、平均排序倒数 (MRR)、答案内容与问题的相关性、答案本身的准确性、信息完整性、语言流畅性等多维度、量化的内部评估指标。这些指标体系能够为系统各模块的优化方向提供科学的指引。

主要教训

1. 初期对金融文档分块策略的复杂度与调优周期的估计不足：项目启动初期，可能认为采用一些通用的固定大小分块方法即可满足需求。但实际操作中发现，金融运管文档的结构复杂性、专业术语的密集性、以及对信息准确性的极端要求，使得分块策略的设计、实验验证和参数优化工作远比预期耗时和复杂。

2. 基础 RAG 在处理深度隐含意图和多步复杂推理查询时的固有挑战：对于那些用户提问非常模糊、需要系统进行多步骤逻辑推理，或者真实意图深埋在字面问题之下的查询，基础 RAG 系统的处理能力仍然存在一定的局限性。这类问题往往需要更高级的查询理解、查询重写或多轮交互澄清机制。

3. 知识更新的及时性、自动化程度与成本之间的平衡难题：运管条线的政策、产品和操作流程更新非常频繁。如何建立一个既能保证知识库内容及时更新，又能将人工干预成本（如文档重新处理、标注、审核）降至最低的高效、半自动化的知识更新机制，是一个持续存在的挑战。完全依赖手动更新不仅成本高昂，也难以保证时效性。

4. 大语言模型上下文窗口限制与输入信息密度的权衡：在向 LLM 提供检索到的上下文时，既要努力保证信息的全面性和完整性，以便模型能做出准确判断；同时也要避免因输入信息过多、过于冗余而超出 LLM 的有效上下文窗口长度，或导致关键信息被稀释，影响最终答案的质量。这需要在文档分块的大小、召回文档块的数量以及上下文组织方式上进行精心的权衡和设计。

财信人寿：“吉小星”AI 助手—— 保险全链路智能增效引擎

一、项目背景及目标

在保险行业数字化转型浪潮中，财信人寿敏锐把握人工智能技术发展机遇，于 2025 年 4 月正式推出面向内部员工的吉小星 AI 助手。这一创新项目标志着公司在智能化运营领域迈出了关键一步，为保险行业内部管理智能化树立了新标杆。

当前保险行业面临多重挑战：代理人队伍专业能力参差不齐、内勤人员工作负荷过重、合规审查流程繁琐耗时、产品开发周期长等问题日益凸显。传统解决方案往往采用分散式系统，导致信息孤岛、效率低下。吉小星 AI 助手应运而生，通过深度融合大语言模型技术与保险专业知识库，构建了统一的智能工作平台，通过新范式激活了公司的内部信息池。

项目自上线以来，已覆盖 1600 名内部用户，支持代理人培训、内勤办公、产品开发、合规审查等核心业务场景。系统采用双模型架构（DeepSeek v3 快速版和 R1 深度推理版），在保证响应速度的同时满足复杂业务场景的高精度需求。值得一提的是，吉小星通过整合内部知识库以及外部实时检索信息，创新实现内外部数据融合技术，以提供更精准的信息检索和决策支持。

二、创新点

“吉小星”AI 助手项目的技术创新体现在多个维度：

智能推理引擎：采用双模型协同机制，DeepSeek v3 模型处理常规查询，响应时间控制在毫秒级；对于复杂逻辑推理需求，模型可以切换至 R1 模型，其思维链能力可深入解析保险条款、评估合规风险。这种弹性模型配置策略使应用场景更加灵活高效。

知识融合系统：构建了动态更新的三层知识架构：基础保险知识库（包含公司制度、产品条款等结构化数据）、实时外部信息索引（监管动态、行业资讯）、员工贡献的实践经验库。系统通过语义理解技术实现三者的有机融合，使回答既权威准确又贴近实务。

安全保护机制：采用更加安全的系统架构，通过本地化、分级权限管控的知识库实现安全问答。通过签订严格的数据处理协议，确保所有交互数据仅用于当前会话，不会用于模型训练或其他用途。

三、项目技术方案

1. 系统架构解析

吉小星的系统架构体现了保险行业特色与技术创新：

应用层 提供多样化接入方式，包括 Web 界面、移动端适配界面以及 API 集成方案，满足不同岗位员工的使用习惯。特别设计了“客户模拟”训练环境，为客服人员提供沉浸式演练平台。

能力层 包含四大核心模块：

- 智能问答系统：支持从简单产品查询到复杂条款解析的多层次问答
- 生产力工具集：涵盖文档润色、Excel 公式生成、可视化图表制作等办公场景



- 专业支持模块：提供话术设计、营销方案策划、合规审查等保险专属功能
- 开发辅助工具：帮助技术人员快速生成代码、设计系统架构、制作项目甘特图

数据层 采用混合存储策略，结构化数据（如产品信息）存储在关系型数据库，非结构化数据（如监管文件）使用向量数据库实现语义检索。通过创新的基于大模型的知识问答对提取技术，将用户问题与分散的保险概念、条款、案例连接成有机网络。

基础设施层 支持容器化技术部署，可根据业务需求弹性扩展计算资源。系统设计了相应监控体系，实时跟踪应用使用量、应用资源、知识更新状态和用户行为模式。

2. 业务场景与应用价值

(1) 代理人赋能体系

吉小星进一步改变了传统代理人培训模式：

- 实时话术支持：当遇到客户异议时（如“重疾理赔条件太苛刻”），系统能生成包含医学通俗解释、真实案例和销售技巧的组合话术。
- 场景化训练：通过角色扮演功能，代理人可与模拟客户进行多轮对话，系统会记录对话过程并提供改进建议。

(2) 内勤工作效率革命

对内勤人员而言，吉小星是全能的工作伙伴：

- 制度查询：将分散在多个系统中的规章制度整合为统一知识库，支持自然语言查询（如“病假薪酬标准”）。
- 文档处理：可自动润色公文，使其符合保险行业公文规范；将复杂表格需求转化为 Excel 公式。
- 跨部门协作：通过共享知识库功能，不同部门可共建专题知识区，减少重复沟通。

使用文档润色功能后，各报告撰写时间平均减少 30%，且格式错误率下降至 10% 以下。

(3) 产品开发与合规风控

在产品创新和风险管理方面，吉小星展现出独特价值：

- 条款智能审查：可识别合同文本中的模糊表述，提出符合监管要求的修改建议，将合规审查时间压缩 60%。
- 产品创意孵化：基于市场数据和公司战略，生成包含目标客户、场景设计、实施路径的完整创新方案。
- 可视化解析：将复杂的责任免除条款转化为阶梯式流程图，帮助非法律专业人员快速理解关键点。

资管部门利用吉小星的数据分析能力，可以将产品募集说明书等内容的投资评估周期从两周缩短至三天。

四、运营情况

吉小星 AI 助手在推广应用及系统运行方面取得了显著成效。

整体运营情况：吉小星 AI 助手自上线以来，在财信人寿内部形成了智能化应用的良性生态。依托大语言模型与知识库技术，能够高效处理多样化自然语言任务，具备跨场景的广泛适用性。同时，系统支持定制化调整，可针对不同部门的业务需求提供个性化服务支持。

应用推广情况：系统采用渐进式推广策略，首批试点覆盖渠道、运营、培训、资管等核心业务部门，通过场景化渗透逐步扩大应用范围。在实际运营中，系统展现出强大的环境适应能力，既能满足日常渠道展业营销素材和方案设计，也可支持资管、精算部门复杂的数据分析任务。吉小星逐步融入员工日常工作流程，通过提升任务处理效率与降低沟通成本，为用户提供了更智能、便捷的服务体验。员工使用习惯正在发生深刻转变，试点用户已逐步形成主动咨询吉小星的工作模式，用户平均单日交互频次达到 9.3 次，充分体现了工具与工作流程的深度融合。

系统运行情况：系统运行表现符合预期目标，在业务高峰期仍保持 99.9% 的服务可用性，首 Token 响应时间低

于 500 毫秒，平均后续单 Token 响应时间控制在 50 毫秒以内。通过创新的双模型机制，既保障了常规查询的响应速度，又确保了复杂业务推理的深度。数据安全方面，全面采用加密传输，实现知识库本地部署并提供多重权限控制，保留完整操作日志。

持续迭代升级：通过资源扩容、模型更新及功能扩展等方式，定期优化系统性能，提升响应速度与处理精度，适应业务发展需求。

综上，内部数字助手在推广与运行层面均表现良好。随着技术迭代与功能完善，其将在保险业务场景中发挥更大价值，为客户提供更高效专业的服务支持。

ID 五、项目成效

吉小星项目产生了显著的经济效益和转型价值。在直接收益方面，系统实现大量人力成本节约，主要来源于重复性咨询工作的自动化和业务流程的智能化改造。更远的价值体现在知识资产转化方面，上线首月，系统已沉淀 30 余个专业知识库，包括 5012 个知识文件，将原本分散在员工个体中的经验转化为可复用的组织智慧，这种知识资本积累对保险企业的长期竞争力构建至关重要。

社会效益维度同样成果丰硕。系统可将代理人平均培训周期进一步缩短，帮助缓解行业人才短缺压力；通过角色模拟实现培训以及辅助式标准化服务话术的输出，可进一步提升客户满意度；在合规风控方面，系统内置的合同条款审查功能已经能为合同预审提供有力协助。

ID 六、经验总结

吉小星项目的成功实施，关键在于把握了三个核心要素：精准的需求锚定、弹性的技术架构和持续的价值运营。在需求分析阶段，项目组没有简单照搬技术方案，而是深入进行行业及业务调研，用一个月时间完成了行业工作场景，跟踪记录各岗位员工的真实工作痛点，最终提炼出“知识即时化、流程智能化、经验显性化”三大核心需求。这种扎根业务的设计理念，确保了系统功能与用户需求的精准匹配。

技术实施过程中形成的“双模驱动、小步快跑”策略具有行业借鉴价值。项目采用“大模型+领域知识”的双引擎架构，既保留了大语言模型的通用能力，又通过保险知识优化打破专业壁垒。项目组形成敏捷组织，在迭代机制上实行更快的更新节奏，始终保持技术先进性与业务适用性的平衡。在推广策略方面，项目组创造的“场景带教法”成效显著，通过选取典型业务场景制作应用示范，让员工在解决实际问题的过程中自然掌握系统使用方法。

面向未来，吉小星将沿着三个方向持续进化：构建 AI 中台，并向业务前端延伸，探索智能核保、智能理赔等深度应用场景；向知识中台升级，构建行业领先的保险知识计算引擎；向生态化发展，通过开放 API 与行业伙伴共建保险智能生态。财信人寿将持续加大 AI 创新投入，计划在未来两年内将吉小星打造为覆盖保险全价值链的智能工作平台，为行业数字化转型提供更多实践样本。



国元证券：燎元智能助手

一、项目背景及目标

1. 市场分析

当前数字化与 AI 成为证券行业构建实现“新质生产力”的重要突破点，证券公司对大模型的应用由早期的对话、智能搜索、“数智人”等初级阶段，逐步向垂类应用、场景融合、生态构建等方向持续拓宽。2024 年以来，大模型在证券行业的应用进入发展新阶段，多家金融机构相继推出金融垂直领域大模型，赋能机构业务各业务场景向智能化、规模化和个性化方向发展。

从大模型应用场景来看，证券经营机构一方面致力于打造覆盖自动化客户服务、研究、量化交易、智能风控、融资融券等机构业务全链条应用场景的垂类大模型应用场景；另一方面逐步从前台业务赋能向算力基建、模型适配、数据闭环、深度应用等方向转型，推动大模型建设从“AI 工具应用”迈向“AI 生态构建”。

2. 项目背景

当前公司的规章制度、金融产品等信息较多，再加上来自互联网的数据（例如行情等），导致公司员工无法第一时间搜索到相关信息，使得员工办公效率低下，且重复劳动的工作让其无法专注于真正需要全神贯注的工作。燎元智能助手是一款基于 AI 大模型的集成式工作效率提升平台，旨在为提升公司员工的工作效率而设计。该系统平台希望通过智能化的公司内外部资源搜索、知识库问答和日程管理等功能，大幅提高员工的工作效率，并简化工作流程。通过减少员工在信息搜索和重复劳动上的时间投入，来提升员工个人和公司团队的整体生产力，助力企业实现更高效、更协同的工作模式，并通过打通内部各系统和外部搜索系统，实现大模型辅助办公，有效提升公司效能。

3. 项目目标

2024 年目标：该年作为燎元智能助手上线的第一年，目标重在该大模型的推广普及，让公司更多的员工知晓并使用该款 AI 产品，对部分活跃用户加强内部培训，让他们更好的通过大模型提高工作效率，节省重复无效劳动时间。例如员工可采用大模型搜索公司的规章制度取代在公司 OA 搜索，也可以利用大模型实现一键订会议室、申请出差等功能，提升工作效率。目前该目标已实现。

2025 年目标：该年度重在燎元智能助手与公司相关系统（产品中心、合规系统、客服系统等）的对接，让大模型帮助员工查询相关系统的数据，让大部分公司员工实现大模型办公并在某些领域实现大模型辅助决策，以提升员工办理业务效率。例如在大模型对接公司的产品中心后，可回答各种产品相关问题，这样业务人员在查询各种产品信息时可不用一个个文件查找，直接问大模型即可，从而节省大量查找时间。

2026 年以后目标：未来将逐步推动燎元辅助员工办公，能将复杂任务自动分解为子任务，并行或顺序执行，直至完成最终目标，可应用于自动筛选简历、财务报表分析、市场调研等企业级任务。也可以从各个用户互动中不断学习，优化过程以提供个性化和高效的响应，可适应用户的特定需求和偏好，如根据用户对数据呈现格式或语气的偏好调整输出，并随着经验增长与员工意图保持一致。

二、创新点

燎元智能助手将在公司各个条线和部门的落地都开创先河：

(1) 公司知识问答模块：燎元可实现公司规章制度、通知公告等信息的大融合，从根本上改变了员工搜索关键字查询公司文件的工作习惯，且大模型可利用自身的归纳总结能力，给予公司员工最准确的回答，节约员工的日常工作时间。

(2) 客服智慧陪练场景：燎元智能助手将根据每个客服人员的表现和需求，定制个性化的训练计划。通过识别客服人员的薄弱环节，系统能够有针对性地提供训练内容，帮助他们快速提升技能。这种个性化的训练方式，使得培训更加精准高效，有助于客服人员在短时间内取得显著的进步。大模型不仅能够提供互动训练，还可以帮助客服人员更新和应用知识库。系统可以记录和分析各个客服人员的训练表现，生成报告并与团队分享。这样，团队成员可以互相学习、分享应对技巧，共同提升整体服务水平。同时，系统还能够根据市场变化和客户需求更新知识库内容，确保客服人员掌握最新的产品信息和服务技能。

(3) 合规风控：燎元智能助手在对接合规系统后，能够辅助公司合规人员进行合规知识问答、合同关键信息抽取等任务，降低合规风险。

(4) 中后台运营：大模型可通过对公司各类运营场景（不限于各类业务流程和服务规范等）的学习，可充分提示各种复杂流程，从而辅助公司一线营业部员工的日常业务办理工作，让他们能够更快捷且精准的帮助广大客户办理业务。

(5) 在金融产品信息问答场景中：燎元计划在对接产品中心数据后，大模型能够融合文本、图像、表格等多种模态的信息，为金融产品信息问答提供更加全面、直观的解答，让公司产品营销人员能在第一时间得到公司金融产品的的相关信息，为他们的客户提供更加准确且便捷的服务。

(6) PDF 助手：该功能支持 PDF 文件的合并、拆分、加水印、扫描等功能，同时支持 PDF 文件与其他类型文件（word, png 等）之间的相互转化，方便公司员工的日常办公。





(7) Dify 智能体平台：支持多模型动态路由，可接入 OpenAI、国产模型等多种大模型，通过 OneAPI 协议实现智能负载均衡。提供 LLMops 工具链，包括模型训练监控、A/B 测试、日志分析等 DevOps 能力，支持 Kubernetes 集群部署。还具备低代码集成优势，可通过可视化编排快速对接 CRM、数据库等外部系统，支持 API 优先开发模式。其核心定位是为开发者提供从模型管理到应用落地的全生命周期服务。

(8) RAGFlow 知识库：是一款基于深度文档理解的开源 RAG 引擎，能够解析 PDF、Word、Excel 等多种格式文档，结合检索与生成技术提供可靠、可溯源的答案，显著减少 AI 幻觉。其采用多路召回机制和融合重排序技术，从海量知识中精准找到与用户问题最相关的部分，同时支持多种主流大模型，还可通过可视化流程从文档预处理到问答测试进行图形化操作，最终让员工建立属于自己的个性化知识库，方便日常工作。

ID 三、项目技术方案

技术架构

1. 燎元助手系统预计采用前端后端分离的技术架构，利用 SSE (Server-Sent Events) 技术实现高效的数据推送机制，确保客户端能实时接收到来自服务端的最新数据，实现流畅的数据流式输出体验。通过集成 OAuth2 认证协议，系统与公司企业微信平台无缝对接，用户仅需通过企业微信单点登录，即可直接进入燎元系统，极大地简化了登录流程，提升了用户体验。

2. 系统内部，我们将深度融合了先进的大模型技术，利用其强大的语义理解与推理能力，精准捕捉用户输入系统内部数据中的关键信息。对这些关键信息进行智能加工与处理，包括但不限于信息抽取、整合、格式化等，以满足多样化的业务需求。借助企业微信提供的丰富 SaaS 服务接口，系统能够自动将处理后的关键信息同步至员工的日程管理与会议系统中，实现个人工作与团队协作的紧密衔接，提升整体工作效率与协同能力。通过这一系列的技术优化与整合，燎元助手系统不仅为用户提供了便捷、高效的数字化工作平台，还进一步推动了企业数字化转型的深入发展。

3. 在本地大模型的部署方面，我们初期采用 Ollama 框架，随着后期燎元助手用户的增加，对大模型的并发量、问答响应时间有了更高要求，我们及时将 Ollama 更新为 vllm。其中 Ollama 是一个开源的大型语言模型服务工具，简洁易用，可以让用户执行一条命令就在本地运行开源大型语言模型，从而实现最前沿开源大模型的本地快速更新；而 vLLM 是一个专注于大语言模型推理加速的高性能开源框架，其核心创新在于 PagedAttention 内存管理技术，显著提升吞吐量和内存效率，从而满足高并发使用需求。

在知识的整合、利用、问答等方面，我们在 LangChain 框架的基础上进行调优定制。LangChain 是一个基于大语言模型 (LLMs) 的应用开发框架，旨在为开发者提供一站式解决方案，以简化和加速 LLM 驱动的应用程序的开发、测试、部署和监控过程，实现了标准化地对模型的集成、接入、组件化等工作。基于 LangChain，我们结合本地非结构化文档知识的特点进行了知识解析、知识切分、知识检索召回、答案生成等模块的定制优化，提升知识的检索问答效果和响应速度。

在大模型的选型方面，我们初期使用通义千问系列大语言模型，DeepSeek-R1 开源后，燎元接入了 DeepSeek-R1-671B 版本，得益于该大模型的算法和架构等创新，燎元的各大模块效果提升明显。例如在合规知识问答等任务中，借助模型的强推理及内容总结能力，回答的准确性、逻辑性等有明显改善，进而显著提升了用户体验和工作效率。此外，在智能识图、语音识别、语音合成等模块，我们也基于使用需要和用户反馈，对前沿多模态大模型保持持续关注，及时进行本地部署更新。

4. 在后期对接内外部系统数据时，我们将会用到当下流行的 DB-GPT 模型。DB-GPT 是一个开源的 AI 原生数据应用开发框架，旨在构建大模型领域的基础设施，通过一系列技术能力让围绕数据库构建大模型应用变得更加简单

和方便。它主要利用 GPT 技术理解 SQL 查询并提供建议，旨在提升数据库操作体验。它通过领域特定训练，提供交互式、易理解且高效的 SQL 建议，适用于学习者和专业开发者。DB-GPT 的目标是构建大模型领域的基础设施，通过多模型管理、Text2SQL 效果优化、RAG 框架以及优化、Multi-Agents 框架协作、AWEL（智能体 workflow 编排）等多种技术能力，让围绕数据库构建大模型应用变得更加简单和高效。

四、运营情况

为了让更多公司员工了解燎元智能助手，在项目的第一阶段我们做了以下推广方案：

(1) 企微宣传：提前一周向全体员工发送企微推广通知，详细介绍燎元的功能、优势、应用场景，以及推广活动安排，吸引员工关注。

(2) 海报展示：在公司的人流密集区域张贴宣传海报，以简洁醒目的视觉设计和核心卖点文案，激发员工好奇心。

(3) 内部网站专题：在公司内部网站开设燎元智能助手的专题页面，展示产品演示视频、操作指南、常见问题解答，方便员工随时了解更多信息。

在进入第二阶段后，我们预计会采用相关使用激励和竞赛活动，去鼓励公司更多的员工使用燎元：

(1) 使用激励：设立燎元使用积分制度，员工每次使用助手完成任务、提交高质量工作成果，根据任务难度与成果价值获得相应积分，积分可兑换公司内部奖励，如食堂卡、培训机会等。

(2) 竞赛活动：公司举办“大模型应用金点子大赛”，以部门或团队为单位，在规定时间内设计一款利用大模型完成一系列复杂工作任务，根据完成速度、质量、创新性等维度评选出优秀团队，给予丰厚奖励与荣誉表彰，在公司内部营造积极使用大模型的氛围。

系统运行情况：

1. 系统性能表现

响应速度：自系统上线以来，燎元在日常办公场景下的平均响应时间保持在 3 秒以内。无论是处理简单的文档生成任务，如回答问题、预定会议，还是面对较为复杂的报告生成请求，均能迅速反馈，有效保障员工工作流程的连贯性，未出现因系统延迟导致的工作停滞现象，满足了员工对高效辅助工具的期望。

2. 系统稳定性

故障率：系统上线至今，共经历了 2 次小规模故障，平均故障间隔时间达到 0.5 小时。这些故障多为短暂的网络波动或个别功能模块的临时卡顿，且每次故障的持续时间均控制在 3 分钟以内，通过技术团队的迅速响应与紧急修复，未对员工的日常工作造成重大影响，系统整体可用性保持在 99% 以上，确保了业务的持续稳定运行。

容错能力：在面对员工输入的复杂、模糊甚至包含错误信息的指令时，燎元智能助手展现出了较强的容错性。它能够通过智能解析与联想，准确理解员工的核心需求，并给出合理的反馈或引导员工进一步澄清指令，有效避免了因指令误解而导致的工作失误，保障了工作任务的准确执行。

3. 系统安全性

数据保护：燎元智能助手系统配备了严格的数据加密与访问控制机制。所有员工使用过程中产生的数据，包括输入的文本内容、上传的文件资料以及系统生成的各类结果，均在传输与存储过程中进行了高强度加密处理。同时，系统依据员工的岗位级别与工作权限，精细划分了数据访问层级，确保只有经过授权的人员能够获取敏感信息，有



效防止了数据泄露风险，为公司的数据资产安全提供了坚实保障。

合规性监控：系统内置了合规性监控模块，能够实时监测员工与燎元之间的交互内容，自动识别并预警可能涉及违反公司内部规定、行业法规或道德准则的对话与操作。一旦发现异常情况，立即通知相关管理人员进行审查与处理，确保公司在利 AI 技术提升工作效率的同时，始终遵循合法合规的运营原则，维护了公司的良好形象与声誉。

ID 五、项目成效

在第一阶段后：燎元智能助手的公司员工使用率能达到 60%，日活平均 100+，月活平均 400+，整体满意度达到 96% 以上。各功能类别估计为：知识库问答占比约 82%，大模型对话占比约 15%，超级助理占比约 3%，各模块综合可减少公司员工 20% 的工作时间。

预计在第二阶段后：燎元智能助手的公司员工使用率能达到 80%，日活平均 200+，月活平均 800+，整体满意度达到 99% 以上，金融产品问答模块可解决 80% 的产品问答问题，合规助手模块可有效帮助合规人员减少不必要的工作时间，客服助手模块可减少客服人员 30% 的工作时间，智能运营场景可减少运营人员 40% 的工作时间。

预计在第三阶段：燎元智能助手的公司员工使用率能达到 95%，日活平均 300+，月活平均 1000+，整体满意度达到 98% 以上。此阶段，燎元的作用不限于通过 AI 节省公司员工的工作时间，在接入足够多的数据后，燎元可以实现局部辅助决策功能，比如智能投顾场景，大模型可帮助投顾人员搜集市场数据并得出相关投资建议，大大缩短投顾人员服务客户的时间。

ID 六、经验总结

(1) 项目成功的关键在于紧密围绕公司的业务需求与战略目标（“产业研究 + 产业投资 + 产业投行 + 综合财富管理”）展开，充分调动各业务部门的积极性与参与度，确保燎元智能助手的功能设计与实际工作场景紧密结合，真正解决公司各部门员工工作中的痛点问题。在项目建设过程中，跨部门协作至关重要，需要建立有效的沟通协调机制，打破部门壁垒，实现技术团队与业务团队之间的无缝对接与协同工作，共同推动项目的顺利实施。

(2) 对于数据的重视与管理是 AI 项目建设的基础性工作，必须提前规划、精心组织，确保数据的质量、完整性和可用性。同时，在数据隐私（公司相关内部文件）保护方面，要时刻保持高度警惕，遵循相关法律法规与行业标准，建立完善的数据安全保障体系，为 AI 项目的稳定运行与可持续发展提供有力保障。

(3) 员工的接受度与使用体验直接影响项目的成败，在项目推进过程中，要高度重视员工的培训与引导工作，采用多样化、针对性的培训方式，帮助员工快速掌握大模型的使用方法，提升员工对大模型技术的认可度与信任度。此外，项目的实施进度与计划安排要合理灵活，充分考虑各类潜在风险与不可控因素，预留足够的时间与资源缓冲，确保项目能够按预期目标顺利完成。

(4) 基于本项目的成功经验，未来将进一步拓展燎元智能助手的功能应用范围与深度，探索其在更多创新业务领域与复杂工作场景中的应用潜力，如智能决策支持、业务流程智能优化、知识图谱驱动的智能知识管理等，持续为公司的业务发展与数字化转型注入新动力。

(5) 未来我们考虑加强与外部科研机构、AI 技术供应商的合作交流，构建产学研用一体化的创新生态体系，共同开展大模型在企业级应用方面的研究与实践探索，推动大模型与企业业务的深度融合与创新发展，为公司在激烈的市场竞争中赢得更大的优势与机遇。

中国银行江苏省分行：知行平台项目

一、项目背景及目标

在我行内外部监管趋严的背景下，业务流程变长、业务系统变多、业务难度变大，导致基层业务人员在业务叙做过程中经常遇到各种问题，条线专业人员疲于应付各类问题。另一方面根据内部管理要求业务人员需要定期轮岗，各岗位上的“新人”难以有效保障业务开展。

为将业务知识与岗位人员有效分离，跳出“积累 - 流失”的恶性循环，满足我行业务拓展需要，打造一个具备自主学习能力、“脑容量”无限、7*24 小时随问随答且服务态度始终如一的 AI 问答机器人，是十分必要且迫切的。一来，可以打破条线甚至产品线竖井，打造“一人多岗”AI 问答机器人，通过制度文件、产品说明、管理办法以及答疑群聊天记录，自主学习业务知识。二来，在提供答疑过程中对不懂的问题可以追问业务专家，获得答案后留存数据库，不断丰富知识库。同时，将业务知识与岗位人员有效分离，跳出“积累 - 流失”的恶性循环。AI 问答机器人以手机端为载体，在大语言模型加持下，实现从“人机交互”到“人人交互”的转变，同时 7*24 小时在线，对每一个问题都能“零秒响应”，做到敏捷赋能。

二、创新点

1. 依托 AI 大模型 (Chat-GLM)，对 LLM (large language model) 大语言模型的深度应用。在 chat-GPT4 问世之后，全社会对 LLM 应用于实际商业场景的期待很高。但是由于国内外网络差异、商用价格以及信息保密等问题，LLM 实际应用特别是在银行的应用场景，至今没有有效推进。而本项目，则是利用开源且可商用的 LLM，进行私有化独立部署，一举解决网络、价格及保密问题，打通 LLM 赋能银行业务的全链路。

2. 依托 Langchain，实现大语言模型与专业领域知识的完美结合。LLM 作为通用模型，对私域专业知识如何快速有效消化吸收，并进行精准输出，这也是 LLM 难以有效落地的核心挑战之一。本项目借助 Langchain 这一技术手段，将私有知识，进行向量化存储，根据问题，检索知识库并形成提示词，交由 LLM 进行 NLP (natural language processing) 自然语言处理，形成最后的问答内容。在此模式下，Langchain 得以与 LLM 完美结合，解决私域知识与通用模型的适配问题。

3. 依托总行提供的 LT-ASR (automatic speech recognition) 语音识别以及 LT-SSC (speech synthesis component) 语音合成技术平台，打造极致的“人人交互”体验。为进一步提升小蜜蜂的拟人化设计，提升基层员工使用的舒适性，在后台服务模型 Langchain、LLM 的技术基础上，项目组拟嵌套 LT-ASR 及 LT-SSC 技术，将前端的文字交互升级为语音交互，为用户打造极致的“人人交互”体验。

三、项目技术方案

1. 技术方案

本项目将融合多项关键技术，涵盖大语言模型、LangChain、语音识别、语音合成、OCR 识别、AIGC (AI generated Content)、NLP、RPA、移动端开发、Web 前端界面开发、用户评价系统、专家反馈机制以及模型微调等领域。

用户端：将提供行信移动端和办公电脑 Web 端作为主要的语音问答应用入口。这些用户端将与后端的语音识



别、语音合成、大语言模型接口以及应用服务接口进行紧密的交互和通信。

管理端：将专门面向 PC 端，作为核心管理用户和专家用户的管理入口，用于管理用户、数据集、数据标注以及模型训练发布管理等关键服务。这一端将为项目的有效运行和管理提供必要的支持。

2. 项目设计

用户界面设计：开发适用于移动端和 PC 端的用户界面，支持文本和语音输入，以及展示文本和语音回答。界面友好易用，适应不同设备和屏幕尺寸。

语音处理流程：整合语音识别技术，将用户语音输入转化为文本，并将回答的文本转化为语音合成以实现语音回答功能。

评价系统：为每个问答结果提供评价功能，让用户对回答的质量进行评分，以持续改进回答质量。

专家修正机制：设计定期的领域专家对问题库的差评问答进行答案修正流程，确保系统回答的准确性和专业性。

模型微调：构建标注后的数据集，用于定期对大语言模型进行微调，以逐步提升问答准确性和适应性。

3. 架构策略

业务架构：根据业务场景需求，搭建业务架构，从知行平台的语料输入、AI 学习、输出、评分机制以及业务专家交互等维度进行设计，整体框架流程图如图 1：

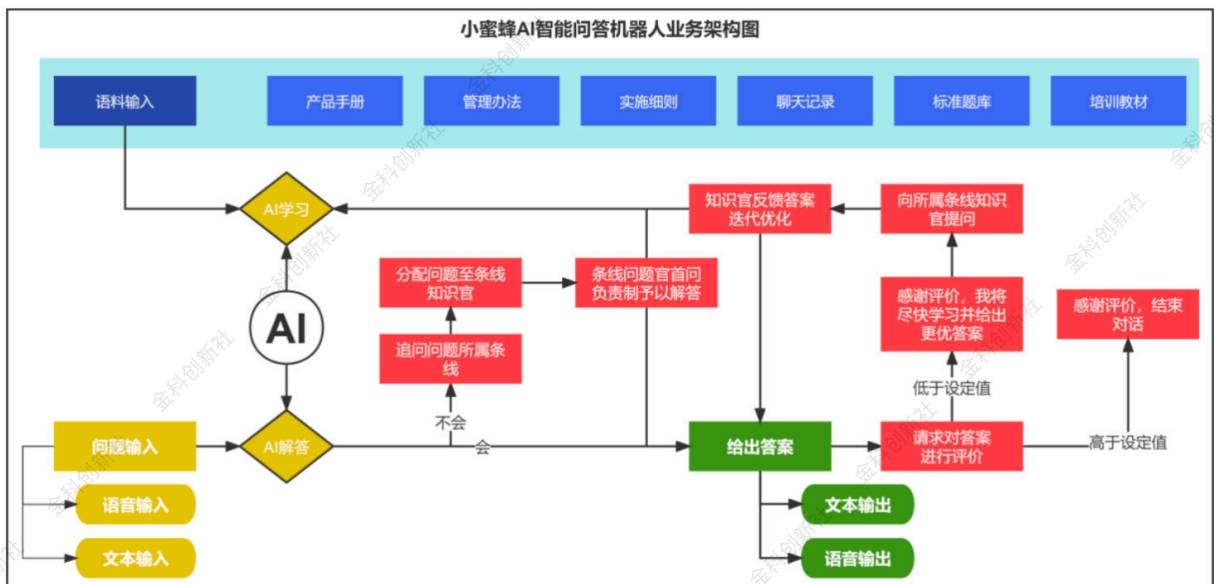


图 1

逻辑架构：小蜜蜂工作逻辑架构包括学习端、问题端以及解答端三方面规划，学习端包括本地文件上传、文本识别、文本切割、向量化存储等环节；问题端包括语音输入、语音识别、OCR 识别、文本向量化、向量检索等环节；解答端包括提示词生成、NLP 处理、语音合成等环节，逻辑架构图如图 2：

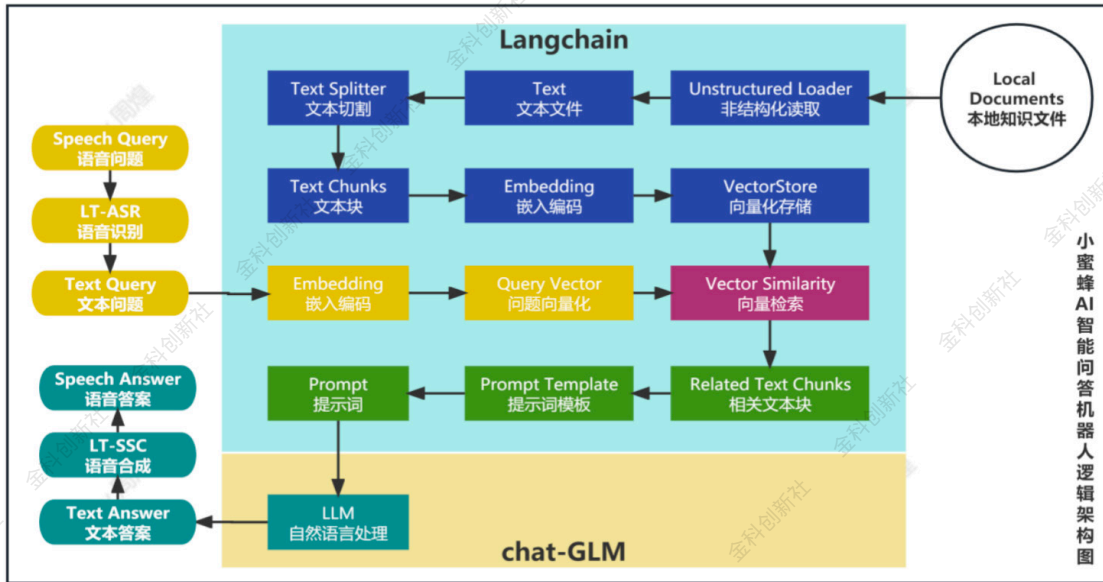
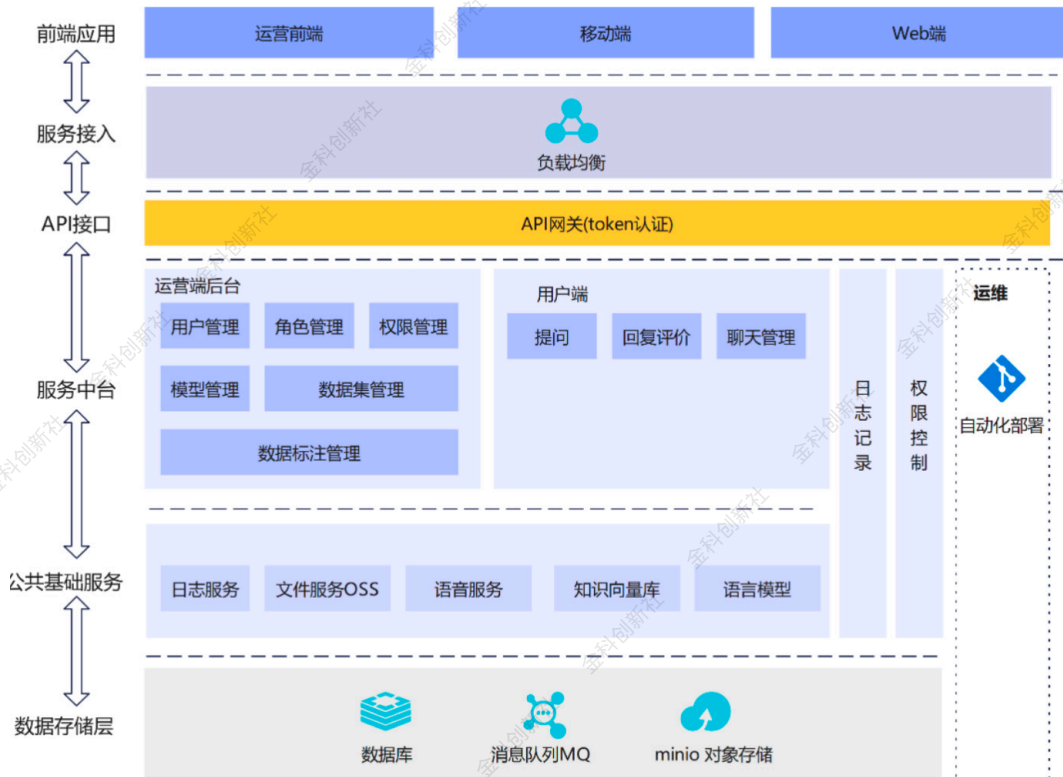


图 2

技术架构：技术架构包括前端架构、后端架构、语音处理架构，其中，前端架构使用现代前端框架，如 React 或 Vue，实现用户界面的开发，确保界面响应迅速且适应不同设备。后端架构主要是搭建后端服务器，用于处理用户输入、调用语音识别和大语言模型 API，生成回答并进行评价记录。语音处理架构旨在整合语音识别引擎和语音合成引擎，确保语音输入和回答的质量和流畅度。





4. 体制机制

专家反馈机制：建立专家团队，定期审查差评问答并提供修正意见，保证回答的准确性。

数据标注流程：设计数据标注流程，将专家修正后的数据用于模型微调，保障模型性能的持续提升。

5. 工作方法

团队协作：项目组组建三个团队，行内技术团队、行外技术团队以及业务专家团队。其中，行内技术团队负责整体技术可行性、安全性审查，负责项目软硬件资源配置；行外技术团队负责项目整体私有化部署及移动端、管理后台等部分功能的个性化开发；业务专家团队负责前期知识库基础资料饱和式搜集以及后期问题的补充回答及部分答案的勘误修正工作。三个团队各取所长、各司其职、同步推进，保证项目进度的平稳快速推进。

敏捷开发：采用敏捷开发方法，将项目划分为多个迭代周期，每个周期内完成一部分功能的开发和测试。

持续集成与部署：建立持续集成与部署流程，确保代码质量和系统稳定性。

6. 迭代方案

迭代周期：将项目分为多个迭代周期，每个周期通常为 2 至 4 周，以逐步实现功能并优化系统。

迭代内容：每个迭代周期内，依次完成一些功能，如移动端界面、语音处理、评价系统、专家修正流程、模型微调等。

反馈机制：在每个迭代周期结束后，收集用户反馈和专家意见，用于指导下一阶段的开发和改进。

五、项目成效

经济效益方面：该项目上线，纵向上有效打通省、分、支及网点四级机构之间的信息壁垒，横向上有效打通各版块、各部门、各团队以及各产品管理人员的业务竖井，以小知行平台一己之力，纵横贯穿全行，以其高度集成的专业知识及友好的交互方式，成为板块业务发展的发动机。在日积月累的问答释疑中，不断积累成长，释放了分行人力，提高了分行的工作效率，赋能中国银行管理运营能力提升，可带来显著的经济效益。

社会效益方面：中行江苏在大模型 + 银行专业知识问答方面做出的探索也为大模型、知识问答行业带来了一些实践经验，为大模型应用于银行业做出了一定的贡献。

六、经验总结

我行在大模型知识问答方面的探索得到一些经验：知识问答主要受两方面的影响：1、模型。采用泛化能力更强、更新、更大的基座模型可以显著提升问答效果；2、知识库。专家知识对于系统的性能提升有着显著的效果，我行在加入行内专家知识后模型的回答准确率得到了大幅提升。

哈尔滨银行：商业银行智能运维体系重构—— AI 技术驱动的数字转型实践

一、项目背景及目标

2025 年作为《金融科技发展规划（2022-2025 年）》收官之年，我国金融业数智化转型进入全面深化、系统推进的关键阶段。2024 年 11 月，中国人民银行等七部门联合印发《推动数字金融高质量发展行动方案》，明确以数字化转型为战略引擎，提出到 2027 年建成“数字金融强国”的路线图，强调数据要素与算法算力对金融服务模式的重构作用。在此背景下，数据要素市场化配置改革深化，以 DeepSeek 为代表的大语言模型等人工智能技术迅猛发展，推动金融服务底层逻辑重构，数字金融向“数智化”质变升级。

本城商行面临传统运维体系在监控告警、作业调度、科技服务等方面效率低下、人工依赖度高、响应不及时等问题，难以适应数智化转型需求，亟需借助 AI 技术重构智能运维体系。

二、创新点

1. AI 大模型深度融合业务场景

将 DeepSeek 等大语言模型创新性应用于运维各环节，如在科技服务台场景，实现知识库信息自动检索与工单自动分配，改变传统人工检索与分配模式，提升服务效率与精准度。

2. 全流程智能化覆盖

实现从事件发现（监控告警、作业调度事件）到处置、服务台支持（知识库检索、工单分配）、数据分析（性能趋势、报表统计）及员工培训的全流程智能化，形成闭环管理，打破传统运维各环节割裂状态。

3. 动态优化与自主学习

建立模型优化与升级机制，基于实时数据反馈，不断优化 AI 模型，使其适应业务变化与数据特征，提升智能化水平的持续性与有效性。

三、项目技术方案

1. 实践场景 - 监控告警事件智能化发现与处置

1.1 实施流程

模型支持解析监控数据，实现告警分级与优先级排序。Zabbix 报警信息通过 DeepSeek 模型与 RAG 向量库中的处置文档结合加持，值班工程师将告警信息发送至 AI 引擎，自动生成根因分析报告及处置建议。例如，针对“Windows 磁盘空间不足”告警，模型不仅定位到具体文件路径，还提供清理脚本和扩容建议，显著缩短 MTTR（平均修复时间）。



1.2 实施结果

检测准确率：从人工检测的误报率明显下降。

处置效率：紧急告警处置率提升较大，平均处置与重要告警工单处理时间显著缩短。

人工成本：监控岗位人力成本投入减少，运维团队可将更多精力投入系统优化。

2. 实践场景 - 作业调度事件智能化发现与处置

实施流程

模型支持解析故障处置手册，显著提升了需作业调度操作的事件流程从发现 - 处置 - 反馈执行结果的时间。各系统基础信息及脚本通过 DeepSeek 模型与 RAG 向量库中的处置文档结合的加持，值班工程师将问题简要发送至 AI 引擎，自动生成根因分析报告及处置建议。例如，针对“系统批量报错”在特殊日期是如何处置等内容回答，模型不仅定位到具体文件路径，还提供处置脚本和相关应用与研发工程师的联系方式建议，显著缩短 MTTR（平均修复时间）。

3. 实践场景 - 科技服务台知识库信息自动检索测试

3.1 实施流程

知识库清洗与结构化：梳理导入历史万条服务记录，清洗整合分散在多个系统中的知识库文档（含服务台基础业务需求操作手册、常见问题解答、故障案例、业务各类需求文档文件模板整理等），使用 NLP 技术提取关键词、实体关系，构建包含标准问答的结构化知识库。

智能检索模型开发：基于 DeepSeek 大模型构建语义检索引擎，支持关键词检索、语义模糊检索（如服务台同事输入“业务数据调取”可匹配“本司人员信息变更”“网络故障维护”等相关问题），并结合用户历史提问记录进行个性化推荐。

接口集成与测试：将检索引擎嵌入工单系统开展多轮灰度测试，累计处理 400+ 模拟工单处置提问，通过此项工作可迭代优化检索算法（如调整 TF-IDF 权重、增加上下文关联度计算）。

用户交互优化：对复杂问题（如涉及多系统联动的故障），大模型添加角色关键字可自动生成“问题拆解员”，引导用户补充关键信息（如报错代码、操作步骤），提升答案匹配精度。

3.2 实施结果

响应速度：平均检索响应时间从人工检索的 2-5 分钟缩短至 4.2 秒，90% 的简单问题（如密码重置、系统登录网址）实现“秒级响应”。

答案准确率：标准问题匹配准确率达 93%，复杂问题引导解决率提升至 85%，服务台人工介入率下降 40%。

用户满意度：服务台工单处理满意度从 72% 提升至 88%，一线客服人均每日处理问题量从 30 单增至 50 单。

4. 实践场景 - 协助科技服务台分配工单测试

4.1 实施流程

1. 工单数据标签化：对历史工单数据进行标注，提取业务领域（网络 / 服务器 / 应用）、问题类型（故障 / 咨询 / 需求）、影响范围（个人 / 部门 / 全行）等 10+ 标签，同时为本部门运维人员建立标签库（含所属组、所关联系统、个人信息等）。

2. 分配模型训练：使用 DeepSeek 大模型构建工单 - 人员匹配模型，输入工单内容、用户等级（普通员工 /

管理岗)、紧急程度等特征,输出最优处理人(支持1对1、1对多回答),通过模拟分配测试优化匹配策略(如优先回答给30天内同类问题解决率>90%的工程师)。

性能趋势分析测试

3. 多源数据整合:接入系统性能指标(CPU/内存/磁盘IO)、业务交易数据(交易量、响应时间)、外部数据(如高峰时段用户访问量),通过数据中台建立统一性能数据集,涵盖多个核心系统的各类性能指标。

4. 特征工程与模型训练:使用时间序列分解算法(STL)提取周期性特征(如日终/月末峰值),结合LSTM神经网络构建性能预测模型,支持72小时趋势预测,重点关注交易响应时间>250ms、吞吐量下降>20%等临界指标。

5. 可视化与协助预警机制:通过BI工具生成动态性能趋势图,当预测到某核心系统(如支付清算系统)未来2小时吞吐量将突破阈值。

4.2 实施结果

预测精度:关键性能指标(如交易响应时间)预测误差率有所控制,在8%以内,提前24小时识别系统瓶颈的准确率达85%。容量规划工作基于趋势分析结果,通过不断优化可在跑批时资源调节到“安全”阈值范围。

5. 实践场景 - 专业报表统计分析测试

实施流程

报表需求梳理:整合运维、科技、业务、合规等部门的报表需求(如《月度系统可用性报告》《故障分类统计报表》),标准化报表指标定义(如可用性=正常运行时间/总时间×100%),建立报表模板知识库。

智能分析与生成:使用DeepSeek大模型解析报表业务逻辑,自动完成数据聚合(如按部门、故障类型分组统计)、趋势分析(同比/环比计算),并生成可视化图表(柱状图/折线图/热力图),支持中文自然语言描述报表结论(如“本月网络故障占比35%,较上月上升10%,主要因为XX交换机固件升级失败”)。

后续计划开展自动化数据pipeline:开发数据采集机器人(RPA)定时抽取核心业务系统(CMDB、监控平台、工单系统)数据,通过AI模型自动清洗异常值(如工单处理时间>24小时的极端数据),数据清洗将大大提高。

6. 实践场景 - 新员工培训测试

实施流程

1. 培训场景建模:梳理运维岗位核心能力矩阵(含监控报警分析、故障发现处置、应急流程操作3大维度与多项技能),基于历史故障案例(如数据库死锁、业务系统紧急处置标准化文档)构建多个虚拟培训场景,每个场景包含事件背景描述、报错处置情形考试、后台记录反馈机制。

2. 智能培训系统开发:使用开源引擎搭建虚拟运维角色+DeepSeek大模型作为“智能导师”,支持文字交互(如学员提问“如何排查服务器无法连接问题”,系统自动引导排查步骤:ping测试→查看防火墙→检查网络配置),并根据学员操作路径生成个性化能力评估报告。

四、运营情况

系统运行稳定性

1. 项目自上线以来,核心业务系统稳定运行率达99.98%,日均处理监控告警事件超2000条,作业调度事件



处置率达 100%，未发生因 AI 模型输出异常导致的生产事故。

2. 通过 AI 引擎动态跟踪模型性能指标（如响应时间、准确率），每周通过各类《智能运维系统数据报告》，累计优化模型参数 12 次，确保系统持续可靠运行。

日常运维机制

1. 采用“AI 初判 + 人工复核”双轨制：简单故障（如磁盘空间不足、密码重置）由 AI 提供标准化处置方法，复杂故障（如多系统联动异常）触发操作提示流程，人工复核平均耗时缩短至 15 分钟。

2. 构建“数据 - 模型 - 场景”闭环优化体系：每月收集一线运维人员反馈的多条优化建议，结合生产数据迭代训练模型，已新增支持多类细分故障场景的分析与标准化处置意见（关键系统异常重启信号灯、网络链路中断执行切换备线操作）。

用户反馈与迭代

1. 后续计划面向全行建立“智能运维体验官”机制，累计收集业务部门反馈问题，优化科技服务台交互界面、工单分配逻辑等功能点，加强用户操作便捷性。

2. 定期开展跨部门联合复盘会，针对极端场景（如月末批量交易高峰）优化 AI 资源调度策略，确保系统在峰值负载下仍保持低延迟响应（平均交易响应时间 < 200ms）。

五、项目成效

效率提升

监控与处置效率：紧急告警平均处置时间缩短 20%，重要告警工单处理时长减少 25%；性能趋势分析可提前 24 小时识别系统瓶颈，准确率达 80%，容量规划效率提升 0%。

服务台效能：科技服务台人工咨询介入率下降 40%，90% 的简单问题实现“秒级响应”回复。

培训效率：新员工培训周期从 4 周缩短至 2 周，通过标准化处置知识库场景完成 80% 的实操考核，降低培训成本。

成本降低

1. 人力成本：值机监控岗位与科技服务岗位同事重复性工作导致的人力成本降低。

2. 运维资源优化：通过 AI 驱动的性能数据趋势分析，调整服务器资源分配，有效提高资源利用率。

质量与体验提升

1. 故障处理精度：监控告警人工误判率下降 50%，根因分析准确率从 60% 提升至 82%，复杂故障引导解决率达 85%，服务台工单处理满意度从 72% 提升至 88%。

行业示范价值

项目成果符合《推动数字金融高质量发展行动方案》导向，验证了 AI 大模型在金融运维场景的可行性，为行业提供可复用的技术路径与实施框架，预计可带动同类机构运维效率平均提升 20% 以上。

六、经验总结

1. 技术适配性：选择国产化开源模型（Deepseek）降低技术壁垒，结合银行特有数据训练提升场景贴合度；
2. 组织协同机制：成立“科技 + 业务”联合专项小组，累计优化流程节点 28 个，需求响应效率提升 50%；
3. 技术能力转型：开展“AI 运维工程师”认证培训的学习工作。

挑战与应对

1. 数据质量瓶颈：建立“数据质量评分机制”，对数据及标准化文档的质量有较高要求，需要实时更新优化；
2. 国产化迁移压力：后续将采用“双轨制”迁移策略，计划业务系统与 AI 模块并行运行较长时间，需确保业务零中断。

未来展望

1. 技术演进：2025 年 Q4 试点“AI 视觉 + 传感器数据”融合的智能巡检，探索多模态大模型在设备运维中的应用；
2. 场景延伸：将连接更多科技赋能的工具，为我行提高工作效率，为业务同事解决繁复工作提供解决方法；
3. 标准共建：为行业的《商业银行智能运维技术规范》出一份力，推动金融业 AI 应用标准化的顺利开展。



国泰海通证券：基于 AI Agent 的金融云平台 全场景运维决策机制研究

一、项目背景及目标

随着金融信创推进与云原生技术普及，国泰海通金融云资源规模与运维复杂度显著提升，微服务化架构下的云网流量、日志数据呈爆发式增长，传统人工运维模式面临多重挑战：海量数据中异常风险排查效率低下，云主机性能波动、网络链路故障等问题依赖人工经验逐层诊断，定界周期长；云服务配置、知识库查询等操作需专业知识，用户自助能力不足，加剧运维团队负荷。与此同时，金融行业对云平台稳定性、安全性的严苛要求，迫切需要智能化技术推动运维从“被动响应”向“主动预防”转型。

在此背景下，项目依托大模型 Agent 技术能力，立项开发“基于 AI Agent 的金融云平台全场景运维决策机制”，旨在构建覆盖云资源智能巡检、云网流量风险分析、故障快速诊断及智能交互的全链路运维体系。目标通过大模型自动化分析能力，实现云资源核心指标智能巡检与异常流量过滤，缩短故障定位时间；集成自然语言交互与工具调用功能，降低操作门槛，形成“监测 - 分析 - 处置”闭环，提升运维响应效率 50% 以上，降低人工成本，同时为金融行业提供可复用的智能化运维范本，推动信创环境下云平台管理能力升级。

二、创新点

1. 多模态监控数据融合的全链路运维分析架构

构建云基础资源视图、实时性能指标、流量日志的多维度数据融合体系，通过大模型算法挖掘云主机 CPU 利用率、Pod 异常数量、跨区域流量等数据间的隐性关联。例如，将集群状态、磁盘吞吐量与历史基线动态匹配，生成资源亚健康风险评分，实现对 Redis 集群主备复制异常、SmartX 资源池节点故障等隐蔽问题的预判。该架构突破传统单点阈值告警模式，形成“全域数据采集 - 关联分析 - 风险建模”的可观测性底座，为智能化运维提供统一数据支撑。

2. 基于大模型 Agent 的工具协同决策机制

引入大模型 Agent 技术，通过 RAG（检索增强生成）整合云平台知识库与实时运维数据，结合 workflow 引擎构建“自然语言意图解析 - 工具链动态调用 - 执行结果反馈”闭环。当用户查询“云主机性能异常”时，Agent 自动触发 CPU / 内存指标分析、网络流量诊断等工具接口，实现多系统协同操作。该机制突破传统运维工具碎片化调用瓶颈，通过模型自主决策优化工具匹配策略，例如优先调用历史成功率高的诊断模块，将问题处理流程自动化覆盖率提升至 30% 以上。

3. 自然语言驱动的运维故障定位范式

建立“自然语言指令 - 多维度数据关联 - 故障快速定界”的智能化路径，大模型通过语义解析将用户模糊需求（如“某 IP 访问异常”）转化为拓扑链路追踪、协议层流量分析、防火墙策略校验等结构化操作。以云机异常

流量异常场景为例，系统可基于自然语言指令自动关联云机资源情况、流量访问列表与数据，快速定位测试环境权限未回收、异常请求持续发送等风险点，可以将故障诊断时间从人工逐层排查压缩至分钟级别，实现从“现象描述”到“根因定位”的直接映射。

ID 三、项目技术方案

技术架构

本项目建设主要围绕三个核心要素展开：数据处理组件、大模型应用组件以及工作流编排组件，三者协同支撑起整个智能运维体系的技术基础与能力闭环。

1. 数据处理组件

该组件以采集器 Agent 为核心，可灵活部署在 Pod、云服务器、虚拟化宿主机、中间件及数据库等多种运行环境中，全面采集各类应用进程的观测性数据。在此基础上，系统还实现了数据标签注入、写入及查询等功能，支撑大规模、多维度的可视化分析服务。经过清洗后的结构化数据，可作为知识推理和工具调用的数据基础，为后续智能分析提供可靠支撑。

2. 大模型应用组件

本组件聚焦于多类型大模型能力的对接，涵盖 LLM、Embedding、Rerank 等模型服务，具备完善的 tool call（工具调用）能力。通过集成多模型融合策略，系统能够灵活调用知识库、执行复杂任务推理，并辅助运维人员完成知识检索、策略生成与故障分析等操作，显著提升人机协同效率。

3. 工作流编排组件

为实现运维知识与经验的标准化表达与自动化落地，该组件支持将大模型节点与工具节点有机编排进统一的工作流中。基于自然语言输入驱动，系统可自动匹配适用的运维规则或流程节点，从而实现面向具体问题的策略下发与任务执行，构建出高度自动化、智能化的运维响应机制。

功能架构

该平台整体架构自底向上分为基础模型层、LLMOps 层、工作流引擎层、用户接入层与工具开发侧能力五大模块，通过插件式接口机制灵活接入主流大模型（如 deepseek-R1、deepseek-V3、qwen3 等）及各类工具服务，在 LLMOps 层封装提示词模板、RAG 检索增强与记忆存储能力，降低大模型使用门槛；工作流引擎作为核心枢纽支持模型调用、工具对接、多媒体能力和开放 API 接入，并通过场景编排能力构建如云机性能分析、智能巡检、NAS 端口分析等典型应用，实现运维逻辑的标准化与自动化；用户接入层则通过智能助手形式分别集成至科技服务云和内部 IM 系统，贴合用户使用场景；右侧工具开发区域进一步提供网盘对接、云网数据与指标分析等功能，结合底层数据采集组件提供的可观测性数据，共同构建了一个可扩展、易用、面向实际问题的智能运维 Agent 闭环。

ID 四、运营情况

目前，在云资源场景中的应用主要集中在故障分析、问题下钻与智能巡检等方面，构建了面向风险预警、故障诊断与运维提效的智能支撑体系。在运维风险分析方面，系统已实现基于云资源视图的智能巡检机制，借助大模型能力对重要指标进行自动巡查与总结分析，定期生成系统健康检查任务，及时识别潜在问题并输出建议。同时，结合云网流量数据，通过每小时超过 75TB 的数据采集与千万级日志处理能力，系统可从复杂流量中自动识别异常访问行为，如服务失联、跨区域异常通信等，显著降低人工排查成本并增强前置风险感知能力。



在故障辅助诊断方面，系统实现了云主机与宿主机多维指标联动分析机制，自动判断云主机在计算、网络、存储等资源维度的运行状态，结合性能特征自动生成排查建议，用于加速问题定界；同时，系统具备云网访问链路分析能力，通过对访问路径中各环节的可用性与性能状态进行全面分析，结合大模型推理能力提供可执行的诊断建议，显著提升了故障识别与处理的效率与准确性。

在日常效率提升方面，已建设面向云用户的科技服务云智能助手，基于文档、知识库整理与结构化处理，结合大模型的语言理解与工具编排能力，辅助用户快速获取操作指引与服务支持；此外，通过对接云平台容量数据接口，系统已实现运维关键数据的自动采集与分析，为资源优化、容量评估及管理策略调整提供数据支撑，进一步提升了日常运维工作的自动化水平与决策效率。

ID 五、项目成效

该项目在经济效益方面体现出多维度的实际价值，显著提升了运维效率、资源利用率和服务保障能力，降低了企业运营成本。通过构建智能巡检、故障辅助诊断与资源分析机制，有效提升了运维支持的响应速度，能够实现分钟级故障识别与初步定界，较传统方式缩短故障排查时间超过 50%，大幅降低因故障长时间未处理造成的业务中断风险。系统对闲置或低效云资源的识别能力，使资源回收更为精准，从而减少不必要的硬件投入与能耗开销，进一步优化云资源池运维结构，提升整体资源使用率。

此外，项目所构建的智能助手和知识驱动的服务支持体系，有效缓解了一线运维人力的服务压力，实现部分问题的自动响应和知识引导，预计可替代约 30% 的重复性人工支持工作，提升服务团队的工作效率与覆盖能力。同时，通过大模型辅助下的流量分析和访问行为异常检测机制，降低了安全事件发生的概率和应急处理的人力投入，提升了系统稳定性与企业数字基础设施的可持续运营能力。综合来看，该项目为公司实现智能化运维转型提供了坚实支撑，具备良好的经济投入产出比。

ID 六、经验总结

本项目自启动以来，通过多阶段、有节奏的系统建设，在平台搭建、数据融合与智能运维场景应用等方面形成了较为完善的工程经验和方法论。围绕智能巡检、故障分析、访问链路诊断等典型场景，项目成功将大模型能力与监控数据深度融合，在实现数据价值挖掘的同时，显著提升了故障识别的准确性与排查响应速度。尤其是在知识沉淀与运维规则构建方面，通过持续收集运维过程中的高频问题、优化用户交互路径，逐步建立起了可迭代、可扩展的智能辅助体系。平台化、模块化的设计理念亦确保了系统具备良好的可迁移性和复用性，为后续在更广泛的云环境和技术服务体系中推广应用提供坚实支撑。

北银金科：北京银行大模型安全测评平台

一、项目背景及目标

1. 项目建设背景

(1) 人工智能引发的安全事件频发

近年来，随着以 DeepSeek、ChatGPT 为代表的大模型技术迅猛发展，人工智能（AI）在提升生产效率、优化交互体验的同时，也带来了前所未有的安全挑战。2025 年初，全球范围内大模型数据泄露、供应链投毒、LLM 劫持攻击等事件频发，暴露出 AI 安全治理的紧迫性。在此背景下，如何构建安全、可控、可信的大模型治理体系，成为 AI 发展的核心议题。

(2) 国家政策高度重视人工智能安全

人工智能是新一轮科技革命和产业变革的重要驱动力量，日益成为全球科技竞争的战略制高点，与此同时，人工智能发展可能会带来前所未有的风险挑战，需要加强人工智能治理。

党的二十届三中全会《决定》提出：“完善生成式人工智能发展和管理机制”、“建立人工智能安全监管制度”。在法治轨道上加强人工智能治理，构建以人为本、智能向善的治理框架，已成为我国抢抓人工智能发展历史性机遇的必然选择。

2023 年 1 月，国家互联网信息办公室等三部门发布《互联网信息服务深度合成管理规定》，针对深度合成技术，明确内容标识义务和用户知情权。

2023 年 7 月，国家互联网信息办公室等七部门发布《生成式人工智能服务管理暂行办法》，要求服务提供者进行安全评估、算法备案及内容合规审查。

2. 现状分析及建设目标

(1) 市场现状分析

① 市场需求快速增长，行业进入高速发展期

随着大模型在金融、医疗、政务等关键领域的广泛应用，其安全风险（如数据泄露、恶意提示注入、伦理合规问题）日益凸显，推动安全测评需求激增。

② 应用场景需求：大模型安全测评平台在金融、政府、能源、医疗等多个行业都有广泛的应用需求。在金融行业，大模型在服务智能化等方面的应用不断深入，需要安全测评平台保障数据安全和内容合规，防止敏感信息泄露等安全问题。

③ 技术创新需求：大模型技术快速迭代，新安全风险不断涌现。当前行业对大模型本质和特征认识不足，缺乏完全科学完善的安全测试方法。后续需大模型安全测评平台持续投入研发，创新测评技术，以应对新风险，提升测评准确性与全面性。

(2) 建设目标

大模型安全测试平台的建设目标聚焦于构建大模型应用的安全防护体系，旨在通过平台化建设构建通用大模型应用解决方案，通过简单对接与参数设置即可一键实现大模型应用的安全测试与报告生成，具体涵盖以下三个核心方向：



① 守护产品合规底线

通过构建多维度筛查机制，对模型的数据采集、算法逻辑及输出内容进行系统性安全审查。例如，利用敏感词过滤、意图识别引擎及合规审计模块实时检测训练数据中的隐私泄露风险（如涉密工艺参数标记），结合对抗样本检测技术拦截恶意指令生成（如违规工艺参数篡改）。同时，依托伦理审查框架评估模型输出的社会影响，确保符合《生成式 AI 服务管理暂行办法》等法规要求，从根源上规避法律与道德风险。

② 提升大模型内生安全能力

建立“测试 - 整改 - 迭代”的闭环机制，通过动态权限控制、对抗性模糊测试等技术识别模型漏洞（如 Prompt 注入攻击、数据投毒），并构建缺陷溯源分析体系。例如，利用安全基线库和意图风险分级模型定位算法缺陷，将测试结果反馈至研发侧优化训练策略（如差分隐私增强、模型水印技术），从而提升模型鲁棒性。此外，通过红蓝对抗演练模拟越狱攻击等场景，持续验证防御机制的有效性，实现安全能力的内生进化。

③ 沉淀大模型安全能力

系统化整合测试案例库（如覆盖数据泄露、对抗攻击等场景的万级样本集）、自动化工具链（如 AST 语法树分析、多模态风险识别引擎），形成可复用的标准化测试流程。实现安全测试能力标准化、体系化，形成可对外赋能的技术服务产品。

ID 二、创新点

项目在建设或推广应用等方面的创新点包括：

① 提出大模型安全测试统一框架：构建大模型安全统一攻击流程，简化构建和评估大语言模型攻击。将大语言模型攻击分解为四个组件：选择器 (Selector)、变异器 (Mutator)、约束器 (Constraint) 和评估器 (Evaluator)，这种模块化设计使得安全研究人员能够轻松地在新旧组件的组合中构建攻击，极大地简化了攻击的构建过程，并且方便对不同攻击方法进行统一的评估。

② 标准定制化可拓展测试：平台支持 10 种以上大模型攻击方法，并在统一框架内进行基准测试、比较和分析，为研究人员提供了一个标准化的评估平台，有助于更全面、准确地评估 LLMs 的安全性，填补了此前缺乏标准实现框架的空白。同时平台提供统一攻击拓展接口，可集成新的大模型攻击方法。

③ AI 生成题库：用户可以根据平台提供的模型和算法对初始题库中的提示词进行语义的识别、分词与变异，并通过内部的评测引擎对生成题目攻击的效果进行评测，不断深度分析与学习，自动生成符合业务需求的衍生题目数据。

④ 动态模型对抗：平台支持在测评过程中选择被测模型、对抗模型，通过模拟大语言模型对话的流程实现模型间的自动化攻防实践，收集模型间的提示词与响应实现动态模型对抗的能力。

⑤ 可视化管理：通过可视化大屏，平台提供了全场景、多视角的大模型应用管理视图，帮助用户实时监控大模型应用安全风险，近期大模型安全风险趋势，有效提升整体大模型应用安全质量。

ID 三、项目技术方案

项目规划

1.1 痛点识别

当前，企业在大模型安全测试方面面临以下痛点：

大模型管理复杂性：随着大模型技术的推广，越来越多的应用开始集成自研或是国内外提供的开源大模型产品，不同的产品根据需求原则的模型、参数、架构、领域千差万别，难以有效管理。

大模型安全测试缺乏统一标准：企业内部大模型安全测试多采用业界提供的少量题库进行抽样测试，测试成功率不高。

缺乏自动化测评手段：企业内部大模型安全测试目前以人工测试为主，测试依赖测试人员的技术与经验，测试过程缺乏自动化。

1.2 产品功能介绍

北银金科大模型安全测评平台项目整体功能设计：实现五大模块，支持从测评题库设置、被测模型管理到测评任务下发的全自动流程，同时北银金科大模型安全测评平台具备丰富的题库维护功能，通过持续更新和优化，确保题库覆盖全面的安全知识和实际应用场景，为各个场景的评测提供支持。

①题库管理

◆ 题库管理

动态筛选：支持题目详情、分类、创建时间等多维度组合查询

可视化展示：ID/详情/三级分类/难度级别（重要/中等/简单）/状态（启用/禁用）等关键字段表格化呈现。

精细操作：单题启用/禁用、详情查看、编辑、删除（二次确认防误操作）

◆ 题目分类分级体系

树形管理：5大项31小类内容维度构建多级分类框架

智能统计：实时展示分类下题目总量及难度分布热力图

权重抽样：任务生成时可基于分类设置题目抽取概率

◆ 风险等级控制

双维度分级：根据题目级别、重要性或危险级别进行分级管理，可分为重要中等简单或者高危中危低危；后续生成报告时可根据不通过问题级别调整评测结果；

◆ 题库生命周期管理

支持题目的全生命周期管理，包括创建、编辑、删除及详情查看；删除时需要检测关联关系；

◆ 题库版本管理

支持通过 excel 模版批量导入题目，前端校验文件格式、必填字段、题型合法性；后端校验数据一致性，自动匹配已有分类分级，且防止重复数据导入；按条件筛选题目并导出为文件，支持 Excel 格式导出；题目每次编辑保存后自动创建新版本，记录修改人、时间及备注；支持一键将题目恢复至指定历史版本；

②评测对象管理

◆ 客户管理：

测评的客户管理：支持多条件动态查询、展示被测客户名称、客户模型数量、模型被测评次数等信息；支持客户增删改查与信息维护：包含新增、编辑修改及删除客户信息；客户与模型关联：支持在该客户下选择模型列表内的关联模型，单个客户可拥有多个模型进行测评；

◆ 模型管理：

支持集中管理所有被评测的大模型实例，提供统一视图进行快速检索、状态跟踪及操作入口；列表支持模型名称、类型等字段查询；列表展示模型名称、添加时间、添加人、模型类型、上次连通时间、状态等字段；支持模型实例的完整生命周期管理，包括新增、编辑信息维护及删除；执行删除前校验关联性（是否关联评测任务）设置模型访问权限的 API key 等配置信息，支持模型需要的鉴权机制；支持通过 WEB 浏览器、API 接口等方式进行模型连接及访问；添加被测模型后，支持通过配置信息测试模型联通状态，通过颜色标签（绿色在线、红色异常）直观展示被测模型连通可用性。

③策略管理



- ◆ 评测方法配置：

支持配置题目集选择时是支持类型选择还是仅支持题目数选择；

- ◆ 评测垂域模型配置：

支持选择评测时使用的 attack 模型和测评模型；

- ◆ 自动衍生算法：

支持选择根据种子题目进行 AI 衍生题目的算法选择及算法组合；

④评测任务管理

- ◆ 自动衍生任务列表：

集中展示所有评测任务的执行状态和基础信息，支持快速检索、筛选及任务操作入口；表格化展示任务名称、关联模型、创建人、任务状态等字段；操作列支持修改和执行评测任务；

- ◆ 任务信息编辑：

创建新评测任务填写基础配置信息，为后续评测流程提供元数据支撑；支持对已创建但未执行的评测任务进行参数修改或删除，确保任务配置灵活性；未执行前可支持修改基础信息（名称、描述）、调整评测维度权重、更新题库选择数据范围；删除时级联删除关联的临时数据；

- ◆ 评测维度及数据范围选择：

在创建评测任务时选择评测题库集类别数量维度等，可选择根据已选范围 AI 生成评测题集；

- ◆ AI 生成评测问题：

通过 AI 模型对题目自定义输入的题目数据进行深度分析与学习，自动生成符合业务需求的衍生题目数据；

- ◆ 评测题库权重配置：

在创建任务时，支持选择题目数和题库类别占比权重；支持根据权重要求使用随机抽样算法抽取对应题目内容进行评测的功能；

- ◆ 随机抽样管理：

从指定数据范围中随机抽取题目组成评测题集，保证随机性，每次评测任务抽样保证平均性，确保评测结果的泛化性与公平性；

任务下发维护：将评测任务分发至目标模型服务，管理任务执行过程中的启停；调用模型连接模块的 API，通过预配置协议发送评测请求，支持异步处理且网络超时或模型异常时自动重试；

- ◆ 任务进程监控：

支持实时追踪评测任务执行状态、资源消耗及异常情况（展示任务进度、已评测题目数）；

- ◆ 任务负载分发：

平衡多模型实例或多节点的评测请求负载，提升任务执行效率与稳定性；

- ◆ 任务定时周期性进行：

按计划自动或者定时执行评测任务；可以配置评测任务的定期执行时间或开始时间。

⑤评测报告管理

- ◆ 评测模型自动评测审核：

支持调用评估模型根据选择评测题目集合自动完成模型问答情况的审核工作，针对回答的准确性、完整性、合规性等方面进行评估给出通过或不通过的结果；

- ◆ 评测结果列表查看：

支持提供评测结果的集中式管理与概览，支持快速定位、筛选及批量操作；列表展示任务名称、评测模型、评测时间、通过状态等字段展示；深度展示单次评测任务的详细数据及展示单个题目的测评过程和详情结果；

◆ 人工审计结果调整：

允许审计人员对自动评测结果进行修正；支持人工针对每个题目的结果进行人工审计修改；审计过程中可查看评测过程中问题的输入输出；

◆ 评测报告生成

评测结果支持按照给出的模版生成报告，平台支持两种固定模版；固定模板内包含数据统计结果、评估分数或评估结果、修改建议等；支持选择对应模版生成报告，报告内含有评估结果且支持评估修改建议输出；生成报告后支持 WORD 版本下载、PDF 版本下载报告。

⑥全平台统计看板

支持展示核心数据的统计图表看板展示，为用户提供全局数据洞察与实时监控能力。顶部核心 KPI 卡片动态呈现题库总量、AI 衍生内容数量、评测任务总数及模型接入数等关键指标，支持点击下钻至明细数据。中部可视化分析区通过折线图、环形图、热力图等多维图表展示任务执行趋势、题目类型分布及模型性能对比等；



大模型测评平台首页页面

项目实施

随着大模型在业务中的广泛应用，公司大模型上线使用后面临模型幻觉、数据泄露、恶意提示注入等安全风险，亟需建立标准化评测体系。本项目通过行业调研及与各方进行学术调研，开源资源研究等手段，制定建设方案后在 3 个月内完成平台开发并落地，实现 模型安全自动化评测，覆盖数据安全、生成内容合规性、抗攻击能力等核心维度。项目实施过程中关键里程碑如下：



阶段	时间	关键成果
需求调研	第 1 周	与行业当前优秀案例进行深入沟通，同时与内部 AI 研发、安全团队访谈，确定评测指标（如提示词注入防御率、数据泄露风险评分）
技术开发	第 2-8 周	搭建平台整体流程及功能，完成自动化评测引擎（支持 LLM 对抗测试、RAG 知识库安全扫描）
内部测试	第 9-10 周	在各个众测项目上进行实施使用，同时丰富种子题库与 AI 衍生算法功能
上线使用	第 11-12 周	全公司内部推广，纳入涵盖 AI 功能及模型上线强制准入流程

业务模式

北银金科大模型安全测评平台计划采用“内部赋能 + 技术沉淀 + 未来商业化”的递进式模式，现阶段以“服务内部 AI 安全治理”为核心目标，同时积累技术能力为潜在的外部输出做准备。当前在公司内部已为各个模型提供上线前及上线后安全评测服务从而保证合规与规避风险，同时提升测评效率实现隐性成本节约，后期计划服务于各个相关单位的模型测评工作，该模式既满足当前内部管控需求，又为未来价值延伸预留空间，符合企业数字化转型过程中“先修内功，再拓外延”的战略逻辑。

四、运营情况

1. 平台推广应用

自大模型安全测试平台上线以来，已在公司环境中得到充分验证。项目团队和开发部门开展了一系列相关预上线大模型的众测活动。并且同步在集团我们在公司 OA 和企微发布了平台上线公告和宣传视频，组织了一系列的培训和分享，确保了平台能得到广泛的应用，快速的提升平台的稳定性和可靠性。随后，我们在集团范围内开展了宣讲营销和市场需求调研，从北京银行及主要投资机构入手，逐步将平台推向市场。

2. 系统运行情况

自平台上线以来，运行状况良好，我们收到了来自用户的积极反馈。用户特别赞赏平台的易用性、自动化功能以及对大模型安全问题检出效率的显著提升。我们的技术团队持续监控系统性能，并定期进行优化更新，确保平台始终保持最佳运行状态。

ID 五、项目成效

1. 经济效益

经过集团内的营销宣讲，我们的产品已经获得了兄弟单位的积极响应，我们计划通过题库、引擎模块、测试模型数量等多类标准进行产品 License 化管理，预计年底可以实现覆盖一期建设成本的经济收益。

2. 间接效益

除了直接的经济效益，我们的平台还带来了一系列的间接效益。我们已着手开展申请软著和专利，以增强知识产权保护。此外，通过不断的市场推广和品牌建设，我们在安全运营领域的经验和影响力得到了显著提升，品牌形象也得到了加强。

我们的平台实现公司大模型安全测试全面覆盖。一是大模型产品统一接入、集中管理，公司内 60% 以上模型均接入平台进行管理与测试；二是自动化测试，减轻安全工程师的工作负担，提升 80% 以上人力；三是在合规管控方面实现从“事后补救”到“事前预防”的转变，大幅度降低后续生成式 AI 违规风险。

ID 六、经验总结

在公司内部大模型安全评测平台的建设过程中，我们通过自主研发攻克了多项技术难题，成功构建了一套全自动化的安全评测体系。我们积累了宝贵的经验，基于对抗训练、动态模糊测试等技术，开发了自动化安全扫描核心算法，支持提示词注入检测、数据泄露风险评估、生成内容合规性检查等核心功能，摆脱对第三方工具的依赖，同时大大缩短模型安全评测的耗时，通过该平台的建设，公司初步实现了 AI 安全治理体系从无到有的突破，更将模型风险管控从人工操作升级为智能化管理，评测效率得到数量级提升。这些成果为公司打造了一套完整的“企业大模型安全基座”，解决了当前面临的实际问题的同时为未来的智能化升级和生态化扩展奠定了一定的基础。希望本案例中平台的成功经验也能为行业同类实践提供有价值的参考，贡献我司在 AI 安全领域的力量。



湖州银行：数据库操作安全管理创新实践

一、项目背景及目标

1. 项目背景

2020年以来，多家大型金融机构出现系统交易故障，发生原因多在IT系统升级或者发生“删库跑路”，对企业造成了灾难性影响。据知名网络安全厂商调查，系统故障事件中约有1/3与人为操作、升级变更有关。同时，信息系统还面临数据泄露风险，调查发现80%的数据泄露来自企业内部。在日常运维、数据查询、数据提取、系统变更等过程中，都涉及到连接数据库进行后台操作，数据库作为价值密度最大的组件，面临着数据泄露、数据篡改、数据销毁等风险。目前以堡垒机为代表的传统运维操作防控手段，主要以事后追溯为主，无法在事前、事中做到风险管控，极易酿成操作事故。近年来，国家为了保护个人及公司数据安全，全国人大颁布《中华人民共和国数据安全法》，国家网信办发布颁布《网络数据安全保护条例》等法律法规。金融监管总局颁布了《银行保险机构数据安全管理办法》，对于银行业数据安全治理提出了明确要求。

在实际IT运营中，湖州银行数据库安全管理面临巨大挑战。随着数字化转型的推进和数智化建设发展，湖州银行数据库种类和数量都在持续增加，包括传统关系型数据库Oracle/MySQL、国产数据库OceanBase/TDSQL，以及大数据平台CDH/StarRocks等。面对越来越复杂的数据库环境，及逐年扩大的研发团队以及不断增加的外包驻场人员，给数据安全管控带来了挑战，并且对于操作行为的监控、账号密码的管理、操作行为的审计等也缺乏抓手进行管理。前期，在数据库管理方面，强化安全管控影响工作效率，放任操作自由又面临安全风险，数据库安全管理难以做到安全与效率的两全，因此鉴于数据库防控需求，发起建设了数据库运维管控平台项目，实现数据库操作全方面安全管理。

湖州银行正全面推进数字化转型，数据库管理作为基础组件管理，也是数字化转型的一个重要领域，需嵌入到湖州银行整体数据防护体系中，实现数据共享与流程通畅，避免出现“数字化烟囱”。这就要求数据库操作平台具有较好的开放性和扩展性，实现与现有数字化平台的融合接入能力。

2. 项目目标

本项目旨在构建湖州银行全生命周期数据库操作安全智能防控体系，实现运维层面和安全审计方面的五大核心目标：

(1) 全面风险管控

通过事前事中事后全域管控 + 动态权限熔断 + 智能行为分析，将无权限、高危等操作及时阻断，显著降低内外部数据操作风险。

(2) 一体化操作运维

通过集成单点登录（SSO）、数据库资源管理、统一账号认证、工单审批及分级分类等系统，实现数据库运维操作自动化，有效消除传统方案中数据库客户端的孤岛问题。

(3) 数据安全治理

创新“安全策略沙箱”模式，在保障数据安全的前提下，支持业务高效迭代，推动数据安全与业务需求协同发展。

(4) 操作追溯机制

实现数据库操作可审计、可关联，满足金融监管合规要求。

(5) SQL 操作智能化

结合私有化 AI 大模型，配置数据库服务智能体，辅助开发、运维人员工作。

二、创新点

新技术方案

1. 实现管控一体的数据库操作运维

平台内置浏览器版 SQL 编辑器（B/S 架构），无需依赖传统数据库客户端，打开浏览器即可使用，实现了数据库管理和控制在同一个平台。客户端可兼容行内所有数据库，可平替传统客户端 95% 以上的常用功能，为开发、测试、运维团队提供了统一数据库管理工具。

平台通过事前、事中、事后的全周期风险管理机制，保障操作的安全性。一是事前强化认证。实行数据库操作的用户实名制，通过和单点登录等系统的对接，从技术上对平台操作人员的身份进行实名制认证。人员不接触数据库账号密码，从根本上杜绝多人共用权限、混用权限、盗用权限、交叉使用权限和越级访问权限等多发问题。二是明确权限管理。建立“三员分离”权限管理制度，按照《银行保险机构数据安全管理办法》要求设立系统管理员、安全管理员和业务操作员。三是定岗定员定权。严格按照“以责定岗、权随岗定、人岗相配、最小权限、权责一致”等原则对人员操作权限进行授予，确保岗位职责和操作权限相一致。四是事中风险拦截。通过内置风险识别引擎识别人员风险操作，在 SQL 操作提交给数据库之前进行拦截，确保提交给数据库语句的合规性。让风险看得见、拦得住、不扩散。五是事后追踪审计。建立端到端的双向快速追踪审计能力，即支持以人员为出发点检索其操作记录，又支持以操作结果为出发点反向检索人员及其相关操作。实现行为可审计，风险可测量。

2. 构建闭环、可脱敏、数据不落地的取数流程

通过对接拉通工单系统和数据库操作防控系统，形成线上闭环取数用数流程。从业务发起取数用数工单，到技术撰写 SQL 获取数据，到系统执行 SQL 获取数据，再到将数据链接推送回工单系统，形成完整的取数用数闭环。需要执行的 SQL 语句可被系统自动审核，规避高危语句带来的操作风险。通过规则 + 算法的方式，实现了所取数据的分级分类动态脱敏、文件水印、文件加密等效果，防止个人信息等敏感数据对外泄露。

3. 采取沙箱式操作验证的生产变更操作

拉通测试环境和生产环境协同变更，视测试环境为安全沙箱，允许提前在测试环境进行高风险 SQL 验证，避免生产环境直接发布带来的风险。从测试环境发起变更流程，系统自动同步生产安全策略及数据进行测试环境的验证，并提供验证结果给流程审批人参考。审批通过后，生产环境可自动同步验证结果，并按照安全策略在生产环境中进行变更。

4. 借助 AI 大模型能力辅助 SQL 编辑及分析

结合内部私有化大模型，打造 AI 数据库智能体，提升数据库智能化运维能力。赋能所有技术和非技术人员，在日常 SQL 查询、故障分析等多个运维场景下产生作用，支持 SQL 生成、SQL 调优、SQL 解释等多种能力，可有效提升数据库运维效率。

新技术能力

通过数据库云客户端工具，兼容更多的数据库类型，兼容不同数据库的 JDBC 协议，支持不同 SQL 语法。在本项目使用的解析器，具备国产自主知识产权，支持国内外 40 多种数据库，具备传统数据库工具 95% 的以上的 SQL



执行功能，支持超万行或者超大单条 SQL 的正确解析与执行能力。

实现预防为主、事前事中事后相结合的安全管控策略。平台以预防为主，防止无授权数据库操作，避免发生数据库安全事故。通过行为监控和事后审计，将事中事后结合起来，形成全面的数据操作管控措施。通过动态智能权限熔断机制，引入实时行为分析。对高风险操作（如批量导出、敏感表访问）进行动态拦截，且支持权限自动回收，权限粒度从“角色级”细化到“会话级”，异常操作阻断响应时间从分钟级缩短至毫秒级。

采取“安全策略沙箱”模式。允许开发人员在测试环境中测试高风险 SQL，自动生成合规改造建议，替代传统“一刀切”封堵策略。实现 SQL 语句从操作人到操作结果的端到端式全链路信息采集，支持双向秒级检索，解决传统审计关系链绑定难题。结合私有化 AI 大模型、RAG 检索增强生成等技术，实现数据库智能体。在开发、运维、业务查询等多个场景提升数据库智能化操作运维能力。

三、项目技术方案

1. 应用架构

为全面提升湖州银行数据库操作安全防范能力，遵循“高内聚低耦合”的模块化建设理念，以及信创安全要求，构建了统一的数据库运维操作管控平台。

系统功能架构图如下：

(1) 接入纳管能力层：实现与不同类型数据的接入，主要通过 JDBC 实现连接；(2) 业务功能层：主要实现平台的各项功能处理，主要包括工作台、SQL 窗口、数据资产管理、数据库维护、安全中心、系统管理等；(3) 统一管理门户：提供系统框架，包括用户中心、权限中心、流程中心、消息中心以及应用门户。



2. 部署架构

数据库运维操作管控平台的部署架构方案特点如下：一是系统架构具备良好的伸缩性，可根据负载横向扩展；二是具备良好的信创兼容性和全栈信创部署能力，平台部署采用鲲鹏 CPU 服务器与麒麟 V10 操作系统，运行在国产信创虚拟化云平台上，后台数据库则使用国产 TDSQL 分布式数据库。

在湖州银行生产环境运维管理区部署平台服务器，并将所有需要管控的数据库实例作为资源加载到数据库运维管控平台并开启 Web 服务。基于此，运维人员即可通过数据库运维管控平台页面进行数据库访问操作和运维管理。

四、运营情况

在湖州银行生产环境和开发测试环境分别部署了数据库运维管控平台，自系统上线以来，已实现规模化落地与持续价值输出，主要运营成果如下：

1. 规模化应用

已覆盖包含核心系统在内的所有生产系统数据库，纳管数据库实例 200 多套，日均监测数据库操作超 1000 次，系统运行总体稳定。

2. 运维效率提升

项目自上线以来，通过对接科管平台等系统，使原有维护及提取流程更加安全合规高效，截止目前通过数据库运维管控平台完成的数据维护流程为 769 条，数据资产提取流程为 841 条。

3. 风险防控实效

通过数据库运维管控平台的事前审核及风险拦截功能，多次帮助研发及运维人员识别出数据库语句的问题和风险，有效避免数据维护操作风险。累计拦截无权限操作 800 次，高危操作 500 次，数据操作事故和数据泄露事件实现零发生。

五、项目成效

1. 提升数据库运维效率

建立规范化、标准化、可信任的运维操作流程，缩短业务取数用数、业务上线周期，加速业务迭代。通过结合大模型等人工智能技术解析 SQL 语句，支持 10000 行以上单条 SQL 语句的极速解析。通过智能扫描识别敏感数据，比人工设置效率提升 10 倍以上。通过低代码方式实现常用 SQL 的封装和执行，大幅提升对业务的响应效率；通过反向索引技术实现快速审计定位，比传统方案效率提升 3-5 倍。

2. 实现数据库运维降本增效

通过建设数据库运维管控平台，每年为行内节省外购正版化数据库客户端工具费用近 5 万，同时通过标准化流程提升运维及开发人员效能，合计每年节省 10 万元费用支出，降本增效效果明显。



3. 技术自主可控

平台全面采用全栈信创基础软硬件，同时替换掉了各类数据库客户端工具，减少对国外技术依赖，符合金融信创要求，减少外围风险。

六、经验总结

本建设项目以“安全、高效、智能”为目标，基于国产化信创基础架构，以实际业务需求为导向，通过技术创新解决运维具体痛点，通过系统间的融合解决业务闭环流程问题，最终实现安全、可控、高效的数据库操作及审批为一体的数据库运维管控平台。这不仅是对数据库生态工具国产化改造的一次探索，也是在当前日益严峻的数据安全环境下面对挑战做出的一次重要实践。该数据库运维管控平台的应用，使我行在生产运维过程中对于数据安全有了更强的把控力，同时对于生产数据的维护和提取效率也得到了显著的提升。

接下来，我们会基于项目实践过程中的经验，结合我行现有实际，充分分析利用平台上沉淀的审计数据，进一步细化管控平台人员权限管理，完善平台数据安全风险策略，规范平台用数流程，持续提升我行生产数据用数安全及用数效率。

青岛银行：基于安全大模型 + 零信任的新一代内网访问控制体系建设

一、项目背景及目标

1.1. 需求分析

我行在实现数据变现、确保业务系统稳定安全运行的同时，为保障我行的业务及数据安全，需达成以下目标：

- (1) 满足合规要求，如数据安全法、国密改造、信创及银行业政策标准等；
- (2) 通过覆盖数据全生命周期的安全保护与管理，支撑我行远程访问开发平台接入安全、内网业务数据访问保护，保障全行现有业务的稳定运行及新业务上线；
- (3) 通过代理技术降低行内业务系统的对外暴露面，实现对行内业务系统的数据保护；
- (4) 通过沙箱等技术手段保障数据安全，实现数据泄密可追溯；
- (5) 通过 GPT 技术的推理能力实现将访问关系、网络暴露面和识别的单点风险有机串联，构筑我行员工行为建模及行为可视化能力，帮助我行管理人员提前预测潜在的多条攻击路径，形成面向业务资产视角的攻击路径全景图，从而完善事前加固措施，有效规避安全事件发生。

1.2 项目立项背景

(1) 访问流量缺乏身份：在标准流量数据中只有基础的五元组信息，缺乏完整的访问主体、访问客体的身份信息，无法快速对应到人、终端和业务。如源 IP 对应的我行员工身份、终端资产信息，目的 IP/ 端口对应的业务资产身份信息（服务类型、业务名称）。

(2) 多源身份割裂：虽然建设了身份基础设施，也普遍是多源割裂的状态，既有 IAM 账号体系，也有独立认证的业务系统。同一个人不同应用系统、数据库的账号并不一样，平台则无法基于统一身份对全网访问日志进行身份化的行为分析。

(3) 访问链路难以串联：在 SNAT、Nginx、跨局域网等环境中，基于网络的访问行为分散在不同的终端安全、网络安全设备上，只能依赖我行人工专家进行访问链路的溯源、分析，日常访问中难以看清完整访问流。

(4) 忽视潜在攻击路径：尽管边界防护建设较好，但是在办公跨网访问、应用系统权限管理上存在管理盲区，导致暴露面未及时收敛。单点的暴露面和资源脆弱性（漏洞、弱口令、基线）容易被安全管理人员忽视，未及时修复，但可以被攻击者利用完成最终的攻击。过往的安全运营中，更多的是发生安全事件后，再通过人工溯源回溯整个的入侵的路径，整体偏置后，事前无感知，无法提前加固进行规避。

(5) Web 应用缺失管控：我行存在大量 B/S 应用，承载大量核心数据，管控粒度粗，审计不完整、风险难追溯，具体表现如下两点，首先是业务难审计：大多数应用存在风险操作审计短板，操作日志审计缺失，比如某员工查询了客户的个人资料行为，同时存在大量老旧应用，业务改造工作量大，内部推动困难；其次是风险难追溯：出现应用风险操作或数据泄露风险事件后，溯源取证困难，缺少直观的举证，我行希望做到像 ECC 运维室摄像头，可把人员访问应用的所有操作都记录下来。

(6) 数据落地缺失高效轻量保护机制：过往基于文件加密或桌面云等办公方式过重，严重依赖于应用层，存在



诸多绕过风险，且对应用程序具备侵入性，对信创 PC 等兼容性较差，我行缺失接近操作系统层面具备一定抗破解能力及绕过能力的数据保护机制。

1.3 项目目标简介

(1) 构建我行基于零信任的人 - 应用的访问控制体系

通过本项目构建我行人 - 应用的 SDP 访问控制体系，以实现我行员工在访问被允许之前，都必须经过身份认证和授权，避免过度的信任；我行员工对资源的访问权限是动态的；分配访问权限时应遵循最小权限原则；尽可能减少我行资源非必要的网络暴露，以减少攻击面；尽可能确保所有的我行员工、资源、通信链路处于最安全状态；尽可能多地和及时地获取可能影响授权的所有信息，并根据这些信息进行持续的信任评估和安全响应。

(2) 构建我行行为建模能力

从我行员工访问地点、访问时间、访问频率、访问路径、操作习惯等维度对我行员工访问行为进行建模分析，通过应用安全大模型以这些访问属性为基础进行持续的学习和行为画像构建，为每个员工构建独特的访问行为基线画像。基于访问行为基线画像为入口点，结合以降维、聚类、决策树为主的计算处理模型发现异常行为，对我行员工 / 资产进行综合评分，识别内鬼行为、已入侵的潜伏威胁、外部入侵行为提前预警。在我行员工访问过程中，对于偏离正常基线的行为，具备根据事先配置的动态访问控制 ACL 规则联动实现对异常访问行为的阻断、告警或二次认证。

(3) 构建我行行为可视能力

通过可视化技术将我行员工访问路径、访问流量、我行员工异常访问行为直观展示，同时也将在线设备状态、统计情况、策略执行情况、执行路径、资源访问等统计数据可视化呈现，帮助我行安全运维人员更为直观、更为全面地了解访问主体的安全状态和行为，从而更快速、更精准地找到风险点，触发安全响应，支撑安全决策。

(4) 构建基于驱动层的数据保护机制

引入沙箱技术，沙箱在 Windows、Mac、统信 UOS、银河麒麟 Kylin 等终端设备上创建一个或多个与本地 PC 环境逻辑隔离的安全工作空间，在工作空间中运行的软件（应用）具备 SSL 通信商密加密、落地文件加密、文件外发管控、内外网络访问隔离、程序管控、剪切板拷贝控制、外设管控、屏幕水印等数据保护功能。与其他数据防泄密方案相比，具备部署成本低、用户体验好、安全性高等优点。与我行现有的桌面虚拟化也不同，沙箱使用终端的硬件资源，业务数据隔离加密保存在本地，同时部署时考虑和我行现有桌面虚拟化进行客户端融合，在终端安装一个客户端软件，不需要安装额外的操作系统。

(5) 构建 B/S 应用保护机制

针对于行内现有的 B/S 应用构建保护机制，可实现免客户端访问业务，最好实现无感知代理访问，同时具备“透明代理为主，按需改写”的代理思路，尽量降低对业务系统的影响，兼容绝大多数 B/S 应用。可以为 B/S 应用提供禁止打印、禁止调试、录屏审计等能力。

(6) 满足合规性、兼容性及未来延展能力

本项目需满足合规要求，如数据安全法、国密改造、信创及银行业政策标准等；同时需要考虑对我行现有域控、全终端（含信创）、桌面虚拟化、现有数据中心应用等的充分兼容性，无需改造即可上线，在未来延展性上，具备敏感数据 API 接口用数活动识别、数据资产分级分类等能力。

二、创新点

1. 采用分步建设方案

我行基于过往同业实践经验，结合行内实际需求，采用先构建统一身份体系（域控）、统一访问平台（桌面虚拟化）、再从远程办公场景到内网逐步延伸，总结实现出了一套良好的落地方案，通过分场景、分阶段建设，助力我行实现零信任进展顺利。在方案架构的选择上，以 SDP 控制中心、SDP 代理网关、客户端（含沙箱）、SDP 分析中心为核心组件的极简架构，帮助我行更容易分场景、分阶段实践落地。

2. 实现纯 WEB 接入

可实现在免客户端情况下，B/S 资源采用“透明代理为主，按需改写”的代理思路，尽量降低对我行现有业务系统的影响，兼容我行绝大多数 B/S 应用。对于已经实现域名化的业务系统，提供了智能改写黑名单、手动改写等能力来解决此类兼容性问题，保证业务系统的正常访问。同时零信任将为其提供暴露面收敛、细粒度管控和数据访问保护三个层面的安全性保护。能够将业务系统收敛至零信任的保护之下，有效的防扫描、防 0day/1day 攻击；同时支持 URL/API 级别的应用权限管控，针对不同的 URL 路径分别授权，实现应用权限的精细化管控；零信任还支持为 WEB 应用添加页面水印、禁止复制、禁止下载、禁止打印、禁止鼠标右键、禁止浏览器调试等数据安全保护能力，能够有效的防止业务数据的泄露。

3. 全功能跨平台兼容

建设时就需适配 windows、mac、UOS、麒麟、中科方德、Ubuntu、iOS、安卓、鸿蒙等主流终端及国产终端；Edge、Chrome、Firefox、Safari、IE、Opera、搜狗、360、密信、红莲花、赢达信等主流浏览器及国密浏览器。提供多个用户自服务能力简化管理运维难度，如提供权限申请自服务工具、终端环境诊断工具等。我行员工可通过自服务工具自助申请权限，自助诊断终端环境，减少管理员配置及管理难度，提升管理效率。

4. 统一工作台

零信任客户端可根据实际需求灵活扩展能力，包括沙箱、我行现有桌面虚拟化客户端、未来的终端安全杀毒等；我行员工只需要安装 1 个客户端即可满足安全便捷的入网和访问业务的需要，规避员工重复安装的动作，优化员工使用体验，简化管理员客户端推广工作，一定程度上减少我行员工投诉。

5. 业务发布性能

系统可以更加高效灵活地调度 CPU、内存等硬件资源，针对资源发布、用户认证、日志审计等消耗性能较高的功能模块重点进行性能优化，从而支持 10 万级别的应用、10 万级别的用户数、上万级别的用户组，相比传统的访问控制设备性能获得极大提升。满足金融级高性能、高稳定性要求。

6. 大并发高可用

系统在架构上采用了控制平面与转发平面分离的架构，不依赖于外置负载均衡即可实现分布式部署和水平扩容，极大地提升了并发接入能力，满足我行大并发、高性能的需求。对于“早高峰”特殊场景，大量用户同时接入时，将根据优先级不同，动态调整不同类型请求的处理速率，使得接近高优先级业务请求先被处理，有效防止“踩踏现象”，实现 2-10 倍压力状态下，用户上线速率可保持超压前的 95% 以上。当控制面故障时，不会影响代理转发面转发已



认证用户的流量，保证业务的连续性和可靠性。

7. 系统主动防御能力

在安全检测和防护能力上，基于主动防御的设计思想，围绕设备、账号、终端三个核心保护对象构建针对性的“三道防线”，并融合多维度诱饵和原生轻量蜜罐等欺骗诱捕技术实现主动防御。

8. 事件和数据运营

在安全事件和数据的运营上，系统通过对登录、访问大数据的二次运营、关联分析，实现全网数据的攻防态势可视化，发现关键异常，自识别误报，缩减告警数量，进一步增强实战攻防场景下的信任评估和访问控制运营能力，并通过可视化大屏提供整个访问控制体系的运行状态直观监测。

9. 引用大模型技术的分析中心

围绕关键业务资产，通过大模型的强大推理能力，将访问关系、网络暴露面和识别的单个风险有机串联，帮助管理人员提前预测潜在的多条攻路径，形成面向业务资产视角的攻击路径全景图，从而完善事前加固措施，有效规避安全事件发生。利用应用内容日志和行为分析模型，识别应用内与应用组合的行为模式差异，同时具备关联推理能力：通过 AI 大模型深入理解攻击者技战法和 ATT&CK TTP 战术，借助大模型独特的行为链研判和情景关联分析能力，从海量的正常访问行为中精准发现无攻击特征的高隐秘行为。在安全制度理解与运营：借助 AI 大模型对于语义的深度理解能力，既能让 AI 学会通用安全规范知识，也能不依赖安全人员制定规则，自动学习我行内部安全规范，自动理解规范并识别违规，发现点对点的风险。

三、项目技术方案

1. SDP 控制中心

SDP 控制中心，是整个零信任方案中控制平面。控制中心负责认证、授权、策略管理与下发，是整体的调度与管理中心。负责控制建立连接和切断主体（用户）与客体（应用）之间的通信连接（通过给网关发送控制指令）。它负责生成客户端用于访问应用的身份验证凭证。

控制中心支持自适应身份认证、动态权限控制，对接入的身份、终端、环境、行为进行信任评估，基于策略引擎配置的策略结果，决定最终允许或拒绝会话。如果会话被授权且请求已被认证，则控制中心通知网关允许代理访问。如果会话被拒绝（或之前的批准被拒绝），则控制器向网关发出指令以切断连接。同时，控制中心受 SPA 单包授权技术对设备本身的服务进行隐身保护——只有已授权的客户端接入，才能打开认证页面，未授权的客户端将无法接入控制中心对外提供的任何服务，认证页面都无法打开。

此外，在主动防御能力体系中，控制中心负责威胁诱捕策略配置和下发，并提供丰富分析和可视化能力，如行为洞察分析、防线可视分析、实体调查分析；同时，控制中心还需对接我行现有身份管理体系，实现整个访问控制体系的身份唯一化。

2.SDP 代理网关

零信任代理网关，是整个 aTrust 方案中的数据平面。安全代理网关负责建立、监视及切断访问主体（用户）和客体（应用）之间的连接。它与控制中心通信，从控制中心接收策略和指令。代理网关支持多种访问协议代理：7 层 WEB 代理、4 层 TCP 代理、3 层 IP 代理，可以满足多种类型应用的发布，所有经代理网关代理转发的流量均基于国密协议加密。同时，代理网关受 SPA 单包授权技术对设备本身的服务进行隐身保护，只有已授权的客户端才能通过代理网关代理访问业务，否则将无法连接代理网关对外放通的端口服务。由于所有流量都经过安全代理网关，代理网关会记录所有的访问请求，包括源 IP、目标 IP 以及访问的 URL 的路径，可进行日志审计，同时支持通过 syslog 将平台的所有日志对接给第三方日志平台。

3.SDP 分析中心

SDP 分析中心旨在打造全场景可视可控的零信任安全框架，以身份和资产为中心，以持续审计分析和持续验证为手段的细粒度访问分析和控制。零信任分析中心通过开放日志接入能力，打造全面可视、信任评估等核心能力，并且能够拓展数据安全场景，最终帮助我行在持续开放的环境中有信心（能见性和可预测性）和有力量不断收敛权限（而不影响业务），并不断确保每次访问真实可靠。同时对接安全大模型，安全大模型外挂向量数据库，保障具备我行管理制度等内部数据理解能力的同时不会将我行内部管理制度等数据训练使用。同时提供大屏展示等能力，确保整个访问控制体系可视化能力形成。

4.SDP 客户端

系统的 PC 面相市面主流的操作系统均有对应的客户端，PC 客户端支持国密隧道访问。同时客户端提供终端安全检测的能力，对接入的终端当前的环境（如操作系统、防火墙、杀毒软件、应用进程）进行收集和上报，上报给控制中心进行信任评估的策略管理。开启 SPA 服务隐身，只有授权过的客户端才能连接控制中心和代理网关，才能进行认证、授权和代理访问。此外，客户端具备沙箱能力，保障敏感业务限制在仅能在工作空间中访问，沙箱可提供数据防泄密能力。其在终端上创建与个人环境完全逻辑隔离的安全工作空间，在工作空间中运行的软件具备链路加密、文件加密、文件隔离、网络隔离、剪切板隔离、进程保护、屏幕水印、防截屏录屏、外设管控等数据保护功能，与其他数据防泄密方案相比，具备部署成本低、用户体验好、安全性高等优点。

四、运营情况

1. 应用情况

目前已经面向全行正式发布 12 套系统，可以满足全行员工日常办公需求，满足信息技术部员工开发测试需求。详见图 1：

2. 运行情况

目前，系统已正式上线运行近 2 个月，日活用户超过 1000 人，系统运行平稳，负载正常，未发生系统故障、安全策略异常和功能异常，整体情况符合预期。



图 1

五、项目成效

1. 经济效益

数据安全建设间接辅助效益主要是保障我行内部业务持续、安全、稳定运转,由业务产生效益,安全作为辅助保障。数据已成为驱动业务、经济发展的核心要素,数据作为驱动业务进步的“发动机”,保障其安全是我行应重点关注的。数据安全以数据为核心,通过对业务或场景梳理,建设贴身的、动态的、可持续性的数据安全体系(技术、管理、运维),保障业务相对安全稳定运行,保障我行发展战略更好落地。通过对业务及战略的有效支撑,从而实现我行业务线可持续、稳定的产生价值。同时有了数据安全保驾护航,减少因数据破坏造成的经济损失,提升我行数字竞争力和拓展数字经济空间潜力。

2. 技术效益

我行通过建设数据安全体系,加强了业务保护能力、加大了数据防护范围、加深了数据保障能力,对我行原有以网络为中心的安全体系进行扩充完善,弥补传统防护技术短板,提升我行整体(可视、可控、可管)应对能力。新技术的应用是为了更好支持业务不断发展,在应用过程中,如何解决新技术产生的新风险则尤为重要,通过新技术的探索、建设和使用,提高了我行员工的技术水平,以数据为核心的安全体系建设,在业务技术与数据之间建立“护城河”,有效解决新技术风险对数据的破坏行为,也提升我行对业务的持续保障和管控能力。

3. 社会效益

我行通过建设数据安全体系，保障了我行客户的数据安全，大大减少了我行客户数据的泄露风险，为社会稳定与公共网络空间安全贡献了我行的一部分力量。同时我行对数据安全建设探索的经验在行业内可复制性较强，对于行业内的安全体系建设探索提出了一个可行方向，也完成了一次较为成功的实践。

六、经验总结

通过内网零信任新一代访问控制体系的建设，确保我行的业务系统安全和数据安全，在网络访问安全、业务数据安全及访问行为安全等方面，通过对不同业务数据、不同人员身份、不同访问环境的敏感级别的定级和管控，有效收缩业务系统对外发布的网络暴露面，减少被攻击的突破口，降低网络安全和业务数据泄露风险。

通过此项目的建设和经验积累，给予中小规模金融银行单位，在网络安全边界模糊化后，内网新一代访问控制体系如何保障业务安全的经验和方向。建立一套行之有效的底层基础建设与相配套的管理体系是极其必要的，通过软硬配套，保障数据安全。保证机密不离行，数据不落地，以提前应对于可能会带来的业务风险，保障数据安全性的情况下拓展业务能力。



青岛银行：基于 Dify 的钓鱼邮件分析助手

一、项目背景及目标

传统规则引擎难以识别新型语义钓鱼邮件，邮件网关多依赖特征匹配，容易被伪装巧妙、特征未知的钓鱼邮件绕过，造成安全隐患。员工在面对可疑邮件时，缺乏便捷的分析工具，需依赖安全团队逐一研判，响应效率低、资源消耗大。

为弥补传统网关的检测不足，降低人为误判风险，故设计并构建了一个基于 Dify 低代码平台的钓鱼邮件分析助手。该助手可自动解析邮件内容并结合大模型能力辅助判断，提高识别精度与响应效率，有效防止钓鱼邮件造成信息泄露。

二、创新点

1. 整合钓鱼邮件研判的信息流

开发 Python 解析工具，自动化提取 .eml 文件关键信息（收发人、主题、超链接、附件），并联动 WHOIS、发件人域名黑名单查询及域名备案分析，识别异常发件来源与高风险链接。

2. 构建训练数据的多样性

建立钓鱼邮件主题关键字黑名单库，结合历史攻击数据沉淀风险特征，强化规则引擎预筛能力。

3. 实现自动化、本地隔离沙箱检测机制

引入附件本地沙箱检测机制，实现对可疑邮件附件的自动化行为分析。

4. 设计结构化提示词框架

设计结构化提示词框架，引入 DeepSeek 大模型，聚焦发件人可信度、链接合规性、语义诱导性等维度，输出标准化分析结论。

5. 建立钓鱼邮件的风险评估模

采用积分制量化风险，通过多阈值判定邮件风险等级，能够准确识别出邮件网关漏拦的钓鱼邮件，提升钓鱼邮件的识别率，降低误报率。

6. 性能优化

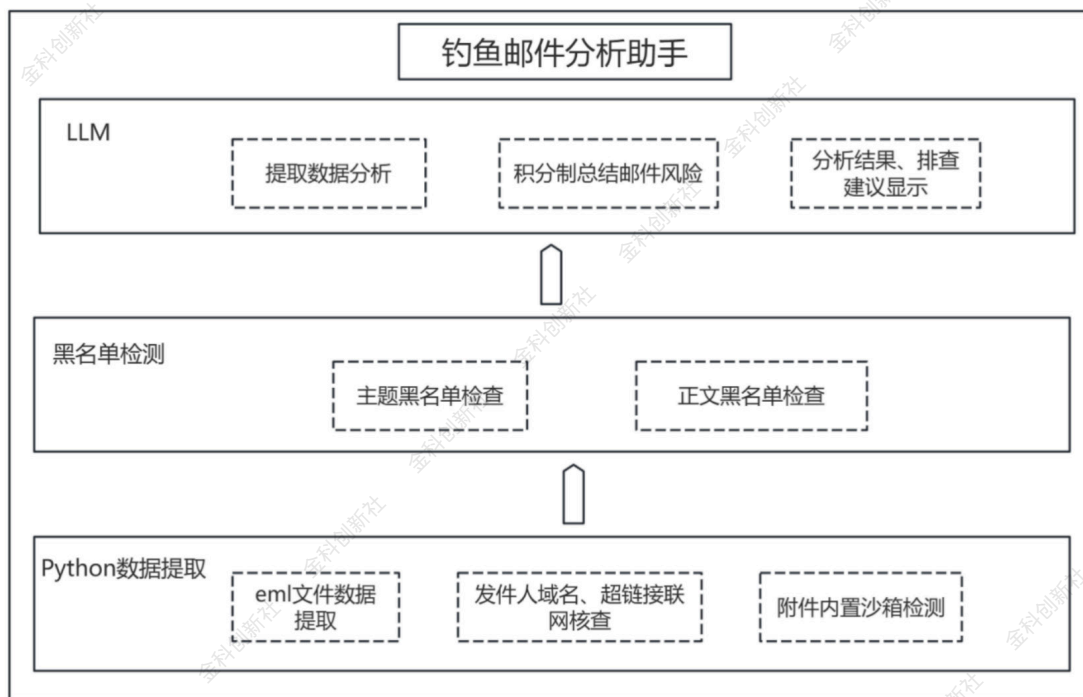
本地脚本对邮件进行结构化解析和安全特征提取，在进入大模型前已完成大部分决策性分析，确保输入语义模

型的数据完整、可信。这种“前重后轻”的设计降低了对 LLM 性能的依赖，使得低成本模型即可胜任风险识别任务。

三、项目技术方案

1. 项目规划

本项目围绕钓鱼邮件检测流程进行模块化设计，采用“提取 - 分析 - 判定”三阶段架构。首先通过脚本提取邮件中的发件人、域名、URL、附件等关键信息，并进行结构化处理。随后结合备案信息、黑名单库及外部情报源对提取结果进行初步分析识别风险。最后将分析数据输入至大模型进行语义理解与综合判定，输出风险等级与建议。该架构灵活可扩展，便于后续接入更多检测能力。



2. 项目实施

首先通过 Python 脚本自动解析 EML 文件，提取邮件头、真实发件人、发件人域名信息，并对域名备案状态及邮件内超链接进行联网安全性核查。其次，将结构化数据传入 Dify 工作流，由黑名单检测器模块对邮件主题与正文进行历史风险库比对，识别已知恶意信息。最终，调用 LLM 大模型对全部分析结果进行语义融合判断，生成详细风险评估报告。

3. 技术架构

系统采用分层结构，前端接口支持人工提交或自动收件；中间层以 Python 为核心处理语言，结合 Dify 工作流与 LLM API 调用完成智能分析；后端依托本地沙箱、域名信息查询接口、黑名单数据库和模型服务。



ID 四、运营情况

系统供青岛银行员工内部使用。运行期间系统稳定，钓鱼邮件网关漏拦的钓鱼邮件能够精准识别。

ID 五、项目成效

1. 精准防御：融合规则库与 AI 推理，能精准识别出钓鱼邮件网关漏拦的钓鱼邮件。
2. 隐私保护：本地化部署确保数据零外泄。
3. 优势互补：Python 提供结构化风险特征，LLM 输出上下文推理结论，两者结合实现数据和语义双维度研判，形成技术闭环，全面提升钓鱼邮件检测率与场景适应性。
4. 能力提升：提供可视化风险评分（如“高危”“可疑”“安全”等级），降低非技术人员使用门槛，提高行员邮件分析能力。

ID 六、经验总结

在该钓鱼邮件分析助手项目中，围绕邮件威胁检测流程进行了系统化设计与实践，重点体现了数据驱动与模块化架构的优势。通过 Python 脚本完成邮件结构信息的精准提取与预处理，有效减轻了大模型的负担，实现了“轻模型 + 强数据”的组合优化。利用 Dify 工作流串联黑名单检测、语义分析等能力，实现了可复用、低成本的自动化判断流程。

镇江农商银行：基于大模型的银行网点的物理攻击行为预警探索与应用

一、项目背景及目标

随着农村金融机构数字化转型进程加快，营业网点智能化建设已进入深水区。在中小型农村金融机构网点，LED 信息屏、智能广告机、自助体验终端等公共显示设备的使用率显著提高，在提升服务效能的同时也暴露出物理安全防护短板。攻击者可能通过伪装成正常客户，利用物理端口接入进行网络安全攻击或信息窃取。根据银行业监管最新要求，各金融机构需履行好网络安全主体责任，做好网络安全防护工作，其公共显示设备更是重点关注对象，需各机构熟悉掌握每块屏幕的基本情况，确保在联网期间有人值守并对联网显示设备实时监控。

二、创新点

技术应用：现有视频分析仅识别基础行为，课题针对性的对社会工程学行为进行精准识别。

业务模式：现有模式通常为安全事件发生后人工溯源录像，项目研究实时性对视频流进行预警。

三、项目技术方案

本项目核心在于构建一个基于开源大模型技术栈的智能视频分析系统，实现对银行网点内针对公共显示设备（如 LED 屏、广告机、自助终端）的物理攻击行为（如非授权接入、恶意设备插入、屏幕遮挡破坏、异常靠近滞留等）进行实时监测与预警。方案充分利用开源大模型在多模态理解（视频 + 文本）和复杂推理方面的优势，建立 "AI 识别 - 实时告警 - 日志存储 - 智能追溯" 的闭环主动防御体系。

1. 视频流接入与预处理

输入源通过标准 RTSP (Real Time Streaming Protocol) 协议接入网点部署的高清摄像头视频流。支持主流安防摄像头品牌。为降低后续 AI 分析的计算负载，保证实时性，同时又能够捕获关键动作瞬间，项目采用了自适应帧采样策略，即默认每秒随机抽取 1-2 帧，检测到运动突变时，短暂提高采样率。对抽取的帧进行必要的尺寸缩放后按时间顺序拼接成短时帧序列片段。这些片段是后续多模态分析的基础单元。

2. 基于视觉类大模型的视频内容理解

采用视觉类大模型作为视频理解的引擎，本项目采用的是 Qwen 2.5 Max，该模型擅长将视频帧序列转化为连贯、详细、富含语义的自然语言描述，预训练超 20 万亿 tokens，知识覆盖面极广，符合项目对于公共显示屏视频分析的针对性。使用提示词明确告知模型任务目标。例如：“你是一个银行安全监控系统，专注于识别针对公共显示设备（如广告屏、信息屏、自助终端）的异常物理行为。详细描述画面中人物的动作、位置、与设备的交互，特别是任何疑似攻击、破坏或非授权接入的行为（如插拔不明设备、遮挡屏幕、异常滞留、破坏设备等），忽略正常客户操作。”

模型对每个输入的视频片段会输出一段结构化的自然语言描述文本。例如：“画面中，一名身穿灰色外套的男性（约 30 岁）长时间（>1 分钟）徘徊在自助服务终端左侧。他身体微微前倾，左手似乎遮挡了终端侧面的某



个区域（疑似接口位置），右手在口袋中摸索。周围无其他人员。行为可疑，疑似尝试隐蔽接入设备。”

3. 基于文本类大模型的异常行为检测与风险评估

利用强大的文本大语言模型对先前生成的文本描述进行深度语义分析和风险评估，本项目采用的是 DeepSeek R1，该模型具有强大的通用文本理解、推理和生成能力，能够高效的进行快速推理。设计提示词，引导模型执行风险评估任务，例如：“你是一个银行物理安全风险评估专家。根据以下监控画面描述，判断是否存在针对公共显示设备的物理攻击或高风险行为。如有，请明确指出行为类型（如：非授权设备接入、屏幕遮挡破坏、异常滞留窥探、设备破坏尝试等），评估风险等级（高 / 中 / 低），并简要说明判断依据。如果行为正常或低风险，请回答‘无异常’。监控描述：[输入先前生成的文本]”。

4. 基于评估结果的处置

评估结果若为中高风险行为时，系统将触发实时告警。项目对接企业微信推送接口，将通过企业微信发送告警信息至指定负责人。

5. 智能追溯 (RAG)

将每一次告警事件及其相关的视频片段、文本描述、分析结果作为案例知识存入向量数据库。安全人员可通过基于 OpenUI 开发的友好前端界面，按时间、地点、风险类型、行为标签等条件查询历史告警记录和原始分析数据（文本描述、关联视频片段）。当发生安全事件需要回溯时，安全人员可以输入自然语言查询（如“查找上周所有在 XX 终端附近发生的高风险接入尝试”）。

四、项目成效

1. 产业效益

技术溢出效应：形成可复用的视频分析中台，多金融场景，技术迁移成本降低。

数字经济赋能：通过安全底座建设，助推网点开展远程视频银行等创新业务，衍生数字服务市场。

2. 社会效益

公众权益保障：构建客户信息“防火墙”，将敏感数据接触人员缩减至必要岗位，符合《个人信息保护法》第五十五条要求。

就业结构升级：培养“AI 训练师 + 安全分析师”复合型人才。

区域示范效应：形成农村金融机构数字化转型样板。

3. 经济效益

合规成本优化：单网点年度等保合规审计成本降低，满足关于自动化风险管控的要求

服务连续性保障：减少因安全事件导致的营业中断时长，提升网点服务可用性。

中国邮政储蓄银行：AIGC 重构质量基石——智能测试自主进化实践

一、项目背景及目标

近年来，金融行业数字化转型加速，银行业务系统复杂度与多样性攀升，同时监管部门对银行业的风险管理、合规运营等方面提出了更为严格的要求，并相继出台了一系列金融监管政策，如《关于规范金融机构资产管理业务的指导意见》《商业银行互联网贷款管理暂行办法》等。在此背景下，软件系统的稳定性和安全性显得尤为重要，对软件质量的要求也越来越高。软件测试作为保证软件产品质量的重要环节，直接关系到银行产品服务的可靠性和用户体验，其重要性不言而喻。然而，在业务需求增长与产品频繁迭代的双重压力下，传统测试方法面临诸多挑战：

一是测试设计难全面。传统测试用例设计依赖人工经验，受认知与精力限制，难以全面覆盖日益复杂的软件系统测试点，影响测试效果与质量评估。

二是测试经验难继承。软件测试经验积累性强，但新员工适应慢，人员流动与项目变更易导致测试知识资产流失。

三是测试交付压力大。银行业务发展快，项目敏捷度提升，传统测试流程耗时长，难以平衡测试质量与交付效率。

四是现有工具局限。现有测试工具依赖模板与规则，使用成本高、拓展性差，智能化和自动化程度不足。

随着全球科技和产业变革的加速推进，人工智能作为关键驱动力，受到了各国政府的高度重视与广泛支持，我国更是将人工智能视为国家战略新兴产业的重要组成部分。2023年7月，国家互联网信息办公室等多个部门联合发布了《生成式人工智能服务管理暂行办法》，积极鼓励生成式人工智能在各行业、各领域的创新应用。邮储银行刘建军行长在2024年全行工作会议上强调了深化大模型技术和生成式人工智能技术应用的重要性，明确提出了开展大数据技术攻坚、构建数据探查和机器学习建模等工具生态的战略要求。

大语言模型（LLM）作为人工智能重要突破，在软件测试领域潜力巨大。通过训练海量数据，LLM具备强大语言理解与生成能力，可用于自动生成测试用例、模拟测试场景，解决传统测试痛点难点，重塑测试设计模式。通过打造智能测试引擎，将AI能力无缝接入现有测试管理平台“星辰平台”，在功能测试用例生成场景实现测试设计自动化与智能化管理，在提升测试效能与质量的同时降低运营成本，为银行业数智化转型提供关键技术支撑。

二、创新点

1. 架构创新，协同增效

为实现测试用例生成方案的快速进化迭代，本案例创新采用“工程化平台+智能体”的设计架构，该架构将工程化单元和智能体单元彻底解耦。工程化单元专注需规条目化解析等流程化任务，降低模型不稳定带来的风险；智能体单元则聚焦测试要点分析等核心生成任务，提升智能生成效果。这种解耦架构使得双方独立演进，灵活适应业务需求变化，高效推动测试用例生成方案快速迭代升级。

2. 多元融合，技术突破

依托邮储银行研发测试大模型平台提供的智能体框架，融合模型调用、知识库管理等多种能力，实现智能体工作流的快速编排与调优。智能体采用原子化任务链解耦机制，通过渐进式调优策略提升各模块性能，在确保系统可



扩展性与可维护性的架构优势下，实现与工程化平台的低耦合对接。

区别于同业普遍采用的“Prompt 工程+ workflows编排”的智能体调优方案，我们创新性地引入检索增强生成（RAG）技术，将历史测试资产盘活利用并引入生成流程，让团队经验得到传承，摆脱对个人能力的过度依赖，通过此技术，部分项目生成用例的幻觉率降低超 30%。

3. 平台赋能，数智转型

邮储银行自研的“星辰平台”作为同业领先的企业级测试管理平台，在国有大行中首批通过 Devops 持续测试能力认证，日常服务邮储银行的上万名科技人员，是邮储银行建设三大核心系统的重要基础设施。在“星辰平台”上自主研发的基于大模型的智能测试引擎提供 AI 能力无缝接入和科学的指标看板，支持智能测试场景运用效果的自动采集、实时反馈，不仅支撑了功能测试用例生成场景快速落地，也加速了整个“星辰平台”的功能测试全场景数智化转型进程。

4. 流程革新，体验升级

本案例将智能体深度融入测试工作流程，精心打造了一步式生成和分步式生成两种特色工作模式，为用户带来沉浸式体验。其中分步式生成模式以直观的脑图形式，让智能生成测试用例的过程思维链可视化，测试人员可以清晰看到用例生成的逻辑脉络，并通过便捷友好的人机交互，更好地掌控智能测试设计过程，提升用例质量。此外，全线上化设计用例模式打破传统线下设计的局限，生成的用例支持一键导入下一工作环节，测试人员可随时随地在线操作，极大提升了工作效率与灵活性，为测试工作带来全新变革。

三、项目技术方案

本案例的整体架构如图 1 所示，智能测试引擎通过统一接入模型库为智能测试场景提供强大的 AI 能力，同时还提供数据看板功能，实时直观地呈现智能测试场景中的各项关键指标。

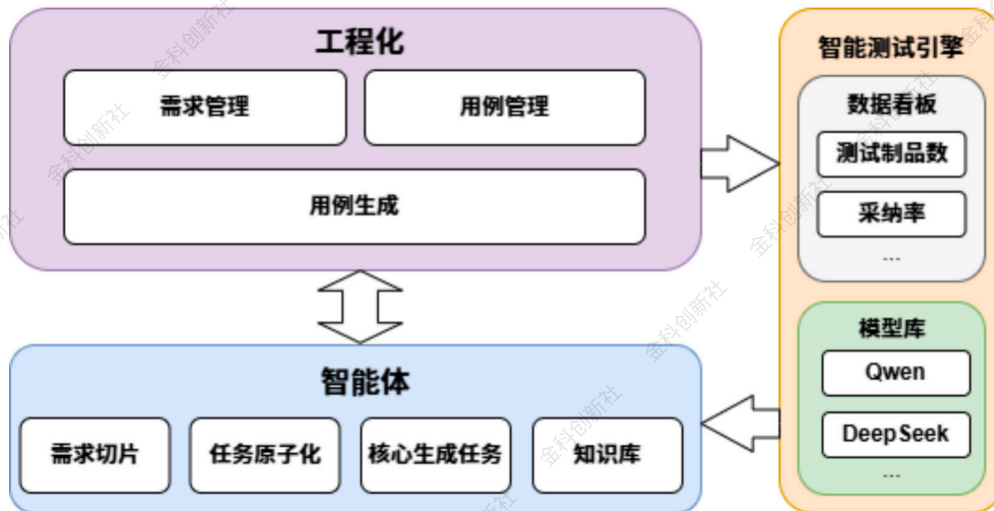


图 1 整体架构图



例平均采纳率从推广初期的 10% 在几个月内快速提升至 28%，在多媒体智能分析等项目采纳率已超过 60%，100% 需求实现全线上化和常态化应用，技术应用深度和应用效果已迈入国有大行第一梯队。部分项目的测试人员已将用例设计工作主要交由模型来完成，自身仅需针对生成的用例进行部分调整以及评审工作，这极大地减轻了测试人员的设计负担。

2. 传承历史经验，减少个人依赖

借助分析需求文档、历史测试用例等多种数据源，大模型能够生成更加全面和深入的测试用例，覆盖更多的测试场景和边界条件，在历史测试资产建设完善的项目中测试覆盖度平均能提升 12%，有效减少漏测的风险。

3. 重塑设计模式，推进数智转型

依托大模型强大的自然语言理解能力，功能用例设计模式发生了根本性变革。一是重塑测试设计工作模式。传统测试用例设计依赖人工逐行分析需求文档，逐条编写测试用例，耗时费力且对人员技能有较高要求。而重塑后的智能测试设计工作，只需输入需求文档即可自动完成需求分析并生成测试用例，显著降低了工具使用门槛与人力成本。在当前技术成熟度下，测试用例设计效率可提升 10%。二是重构测试资产使用模式。通过学习历史测试资产，大模型能够生成更具高效性与针对性的测试用例，在提升测试效率的同时，进一步优化测试质量。经过人工确认的高质量用例，也可反哺测试资产库。

ID 五、项目成效

在智能功能测试用例生成领域的深入探索与应用实践中，本案例成功提炼出一套系统化、科学化的方法论，全面覆盖了从场景挖掘到应用推广的全过程。

首先，通过细致的场景挖掘，深入剖析当前软件测试面临的挑战与需求，精准定位智能功能测试用例生成的突破方向与价值锚点。随后，在实验评估阶段，设计并实施了一系列详尽的评估方案，对主流大模型在测试用例生成能力上进行综合评估，筛选出最优的基座模型，为后续工作奠定了坚实基础。在系统建设环节，依托先进的 AI 能力构建智能测试引擎，并通过“工程化平台 + 智能体”架构方案在功能测试用例生成场景实现了测试用例的自动化生成、优化与管理。在项目试点阶段，通过收集反馈与持续优化，验证了功能用例生成场景在提升测试效率与覆盖率方面的显著效果。最后，在应用推广阶段，积极推广功能用例生成场景在更广泛范围内应用，助力实现软件测试的智能化升级。

通过本案例总结的方法论不仅为智能功能测试用例生成领域提供了系统化的指导框架，也为 AI 技术在软件测试领域的全面应用与发展提供了优秀案例和落地思路。

下一步，我们还将持续关注行业动态和技术发展趋势，不断优化功能用例生成场景的性能和功能，探索更多创新应用场景，如测试用例评审、回归用例推荐等，推动软件测试向更高水平的智能化演进。

中信建投证券: AI 大模型驱动的自动化测试平台建设与应用

一、项目背景及目标

市场分析

证券行业系统迭代频繁,规则日益复杂(如科创板规则更新、北交所业务上线等),传统人工测试难以应对高并发、强合规的业务场景,存在用例覆盖不全、回归效率低以及人工遗漏风险高等痛点。

立项背景

随着人工智能技术的快速发展,大语言模型引领多个领域技术变革。在软件测试领域,为用例的生成、自动化脚本的编撰、测试执行与监测都提出全新的解决方案。同时也为响应证监会《证券期货业科技发展“十四五”规划》的要求,提升系统稳定性与交付效率。

项目目标

1. 构建 AI 赋能的自动化测试体系,覆盖 95% 以上证券核心业务;
2. 实现测试设计效率提升 80%,测试周期压缩 50%;
3. 减少因人工遗漏导致的线上故障;
4. 实现需求规则实时监测,自主识别并优化受影响测试脚本;
5. 具备测试结果智能分析,动态调整测试执行策略;
6. 实现智能回归测试,动态调整测试资源。

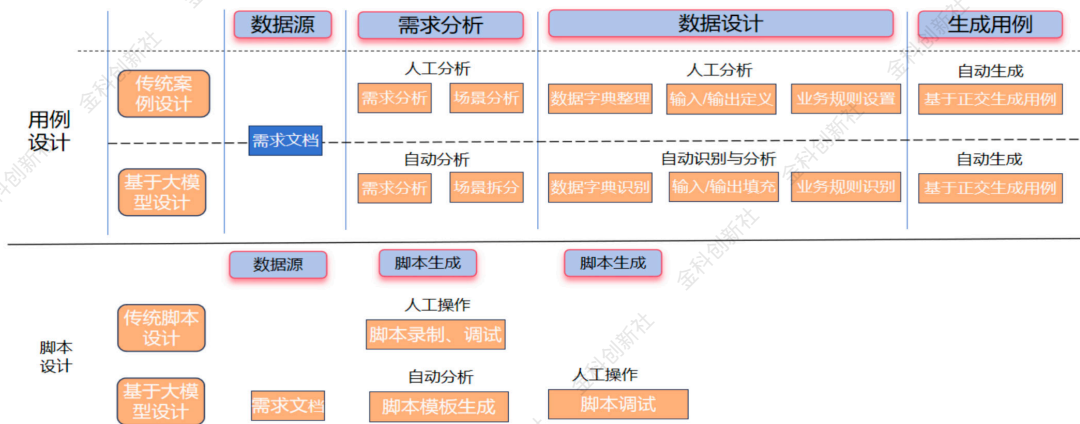


图 1 传统测试与大模型智能测试对比



二、创新点

本项目在建设与应用方面具有以下核心创新点：

1. 提出“AI-测试智能体”概念：实现需求监听→用例生成→脚本修复→回归执行→报告分析的全链路闭环；
2. 证券业务规则深度适配：针对两融交易、新三板、北交所等场景定制数据规则引擎，支持实时合规校验（如投资者适当性规则）；
3. 动态脚本修复技术：基于 DeepSeek 等模型自动识别需求变更并修复受影响脚本，适应证券规则高频迭代；
4. 智能回归调度与数据预埋：根据系统发版计划自主编排回归任务，并自动预埋所需测试数据（如模拟大宗交易数据）。

三、项目技术方案

1. 技术架构

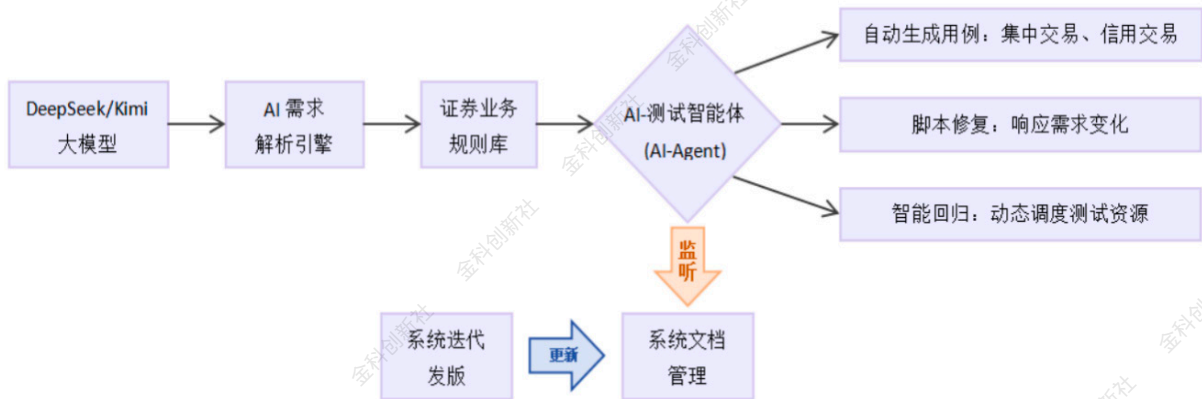


图 2 大模型集成技术架构

2. 业务模式

大模型智能构建证券专属数据字典（如股东账号编码规则、委托价格校验逻辑）与业务知识库；支持上交所 / 深交所差异化业务场景（如深 A、沪 A 风险警示股的市价交易规则），实现完整业务匹配；平台适配与扩展性强，通过调整数据规则引擎和测试用例模板，可快速响应并适配新业务需求。

四、运营情况

推广范围：覆盖中信建投 PC 卓越版、VIP 快速交易版等 6 大核心客户端版本；
 运行效能：日均自动生成用例 1200+ 条，脚本自主修复成功率 85%；
 系统稳定性：平台可用率 99.50%，持续监测调整回归计划。

五、项目成效

1. 经济效益

测试人力成本降低 60%，年节约投入约 800 万元；
系统上线周期从 10 天缩短至 5 天，支持业务快速创新。

2. 社会效益

通过 100% 合规用例覆盖，保障投资者交易安全；
通过 AI 深度赋能测试全流程，助力企业实现测试智能化转型，提升测试效率和质量；
推动 DeepSeek 等国产开源大模型在金融关键领域的产业化应用与验证，助力金融科技自主创新。

六、经验总结

1. 关键经验

以证券业务规则为锚点训练模型，显著提升场景适配性；
采用“人工审核 + AI 执行”的双轨制模式，有效确保了项目初期的落地可靠性与风险可控。

2. 推广建议

优先在高频迭代、规则复杂的业务领域（如量化交易系统）部署 AI 测试智能体；
探索建立跨金融行业的测试知识库与智能体经验共享机制，加速全行业测试智能化进程。



晋商银行：基于 AI 与业务组件扫描的 智能化质量防护体系构建

一、项目背景及目标

针对研发测试工作过程中面临的研发设计评审粗放（依赖经验判断，缺陷早发现能力弱）、代码分析低效（人工梳理易漏易错，版本变更响应滞后）、业务场景覆盖不全（核心交易链路复杂导致漏测）、测试案例评审低效（依赖人工经验判断，覆盖度验证耗时）、自动化闭环断裂（变更分析与测试执行脱节）等痛点，本项目创新打造“AI 驱动中小银行核心系统测试范围精准定位与自动化闭环构建”，实现从代码级差异识别、组件级影响评估到业务级测试覆盖验证的全流程智能化，确保“该测的不遗漏，测过的精准有效”，有效解决测试中“测不全、评不准”的核心痛点。

二、创新点

技术创新：全维度测试范围精准定位

1. 业务组件化建模，轻量化代码差异分析，锁定核心测试对象

多维度组件建模引擎：开发“核心业务组件化定义工具”，通过业务组件化建模与轻量化代码差异分析，精准锁定核心测试对象。工具支持自定义业务场景层组件（如“账户开户/维护/销户”“行内转账/跨行支付/通用记账/批量记账”“客户控制/账户控制”等高频交易场景组件）、入口层组件（接口对应 API 服务端点）、类方法编排层组件（Gravity 编排、Stria 流、Flow 流编排等）及更细的业务原子层（如风险敏感类“资金计算”“账户余额更新”“黑名单校验”等）或技术原子层组件（如 RPC、Repository、异步消息、TAE 记账引擎等），通过灵活的组件层数定义，更能贴合核心业务需求及测试分析。

代码静态智能差异分析：采用 GITDIFF+ASM+ 字节码比对三重技术构建，实现 AVA 工程调用链路静态追踪（支持接口/抽象类深度解析）；通过抽象语法树级代码变更捕获，从而精准识别版本迭代中涉及的组件级差异（如：修改“账户开户组件”中的“客户信息校验方法”）；借助变更图谱自动生成引擎自动关联该组件上下游依赖关系（如“开户组件→账户初始化组件→风控校验组件”），输出细粒度变更影响域报告，解决传统“孤立测变更、漏测关联模块”的问题。

智能过滤引擎：特别针对 SpringBoot+Maven 的项目架构，通过自定义规则智能过滤公共组件与无关组件（如模型层、接入接入层、工具层以及三方依赖代码等），实现真正意义上影响范围的灵活控制和精准定位，仅扫描核心涉及业务代码（工程下的账户、交易模块），避免无效扫描，将代码分析效率提升 50% 以上。

自动化分析扫描：通过 DevOps 流水线嵌入并集成组件化扫描，在持续交付过程中做到全流程的风险把控，摆脱开发过程黑盒的困境，真正做到交付过程实现静态分析前置卡点、自动化扫描分析、风险组件变更预警、变更图谱可视化展示、测试用例智能推荐的自动化精细交付。

AI 智能分析：将静态扫描分析的基础数据落入关系型数据库中，依赖资产数据通过引入推理模型在代码设计、组件设计、影响范围分析等方面进行智能分析对话。技术上服务端引入 Spring WebFlux 流式响应对接 DeepSeek-R1 推理模型，前端通过 FetchAPI 进行流式请求，同时前端采用 Markdown 解析、d3 图谱、prism 高亮代码块、防抖动及缓冲延时响应实现前端的交互式输出，通过自研原生 AI 交互式前端为未来智能化提供扩展保障。

通过技术创新实现了金融核心系统测试从经验驱动到数据驱动的转变，为关键业务系统构筑起智能化质量防护体系。

2.AI 驱动测试覆盖度智能评审

构建行内专用 AI 评估模型，基于测试功能点、测试案例、历史缺陷数据、业务风险等级（如监管合规类、资金交易类）及组件调用频次，自动生成《测试覆盖度分析报告》：

横向覆盖：检测变更是否涉及未覆盖的业务分支（如“开户失败回滚逻辑”是否缺失测试案例）；

纵向深度：评估核心组件的方法级覆盖完整性（如“转账接口”是否覆盖异常参数、超时处理等边界场景）；

风险加权：对高风险组件（如涉及客户资金、账户安全的模块）标记“强制覆盖”，确保其测试用例覆盖率。

通过可视化图谱实时查看“未覆盖节点”“低覆盖风险区”，评审效率提升 80%，覆盖度验证耗时从 2 天压缩至 2 小时。

3. 业务链路深度追踪，保障核心场景覆盖

“工程结构 + 业务调用链路”双维度扫描模式：

工程结构扫描：基于行内核心系统分层架构（表现层、业务逻辑层、数据层），自动识别入口层方法（如柜面交易入口、API 接口入口），标记业务抽象层组件（如“活期开户组件”“转账交易组件”），确保测试范围聚焦核心业务模块。

调用链路追踪：依托 ASM 技术解析复杂交易链路（如“开户→账户初始化→额度校验→流水记账”全流程），精准定位代码变更对上下游业务的影响（例如修改“账户余额计算方法”是否影响“利息结算”“风险预警”等关联模块），生成《链路影响热力图》，解决传统测试“只测表面、漏测关联”的问题，使业务场景覆盖完整性提升 40%。

自动化闭环：从分析到验证的全流程贯通

1.AI 智能评估，降低人工依赖

测试范围圈定：引入场景定制化 AI 模型，基于历史测试数据（如缺陷分布、业务风险等级）训练，自动生成组件调用关系图谱，提取所有受影响的组件、方法及关联链路，AI 智能评估变更影响优先级，同时，对高风险业务组件（如涉及资金动账交易、客户敏感信息的模块）标记“强制覆盖”，生成《必测清单》，确保核心场景 100% 测试；

脚本智能匹配：对接银行现有自动化脚本库，自动关联匹配度 $\geq 80\%$ 的已有脚本，对缺失场景触发“脚本生成预警”；

执行策略优化：对高频交易组件优先执行，低影响变更合并执行，减少 30% 的无效测试任务。

2. 测管平台无缝闭环，实时验证覆盖效果

构建“差异扫描→链路分析→AI 风险分级→自动化任务生成→自动化测试执行→智能报告反哺”的自动化链路闭环：

数据流转：将 AI 分析的“未覆盖节点”“高风险组件”推送至测管平台，自动创建覆盖验证任务；

智能验证：执行自动化测试后，系统对比实际覆盖结果与 AI 预期覆盖范围，生成《覆盖度偏差报告》（如“某转账接口遗漏‘余额不足’异常测试”）；

持续优化：将验证结果反哺 AI 模型，动态更新组件风险权重及覆盖规则，形成“分析 - 执行 - 优化”的螺旋提升闭环。

ID 三、项目技术方案



技术方案

1. 技术架构

分层架构设计：

层级	技术组件	功能说明
基础设施层	Git、Maven、SpringBoot	代码管理、依赖构建、微服务框架
核心引擎层	GITDIFF+ASM+字节码比对、组件建模引擎、变更图谱引擎、AI 推理模型 (DeepSeek-R1)	代码差异分析、组件依赖追踪、影响域可视化、智能决策
数据层	关系型数据库	存储静态扫描数据、测试资产、缺陷历史等
服务层	Spring WebFlux (流式响应)	支持 AI 模型实时交互,处理高并发推理请求
应用层	自研 AI 交互前端 (Markdown 解析 /D3 图谱/Prism 代码高亮)	变更图谱可视化、覆盖度报告生成、智能测试建议推送

2. 关键技术突破

- 智能范围定位：
 - 组件化建模：支持 4 层业务组件定义 (场景层 / 入口层 / 编排层 / 原子层)。
 - 三重代码分析：GITDIFF (版本差异) + ASM (字节码解析) + 语法树比对。
 - 变更影响图谱：自动生成组件级上下游依赖链路 (如 "开户→风控校验→记账")。
- AI 测试评审：
 - 多维度覆盖模型：业务风险等级 (监管 / 资金类) x 缺陷历史 x 调用频次。

- 动态热力图：标记未覆盖节点 / 高风险组件（如资金计算模块强制覆盖）。
 - 链路追踪：
 - 双维扫描：工程分层结构（表现层 / 业务层 / 数据层）+ 业务调用链路（ASM 解析交易流）。
3. 自动化闭环设计



业务与商业模式

1. 业务痛点解决

痛点	解决方案	效果
业务场景覆盖不全	组件化建模+链路追踪	场景覆盖完整性 ↑ 40%
案例评审低效	AI 覆盖度报告+热力图	评审效率 ↑ 80% (2 天→2 小时)
自动化断裂	变更分析→用例推荐→执行闭环	无效扫描 ↓ 50%

2. 商业模式

- 价值主张：测试成本节约（案例：传统分析 12.5 万 → AI 方案 5 万，节省 60%）
- 收费模式：项目制（当前方案 10 万元），可扩展 SaaS 订阅（按扫描次数 / 系统规模计费）
- 客户价值：
 - 规避资金风险（100% 覆盖高风险组件）
 - 加速交付周期（评审耗时从日级压缩至小时级）

创新点总结

1. 首创组件化影响分析

支持自定义业务原子层级与技术原子层，精准控制扫描范围

2. AI 评审双维验证

横向（业务分支覆盖）+ 纵向（方法级边界场景）+ 风险加权三维度报告

3. 流式 AI 交互架构

Spring WebFlux+FetchAPI 实现低延迟流式响应，支持秒级生成变更图谱



四、运营情况

目前项目已在我行核心系统开发迭代中成功试点，覆盖多个研发团队。未来，计划推广全行信贷、手机银行、集中作业等多个系统的项目开发测试中，预计服务超过 200 余人，覆盖代码规模超 5000 万行。通过在不同项目中的广泛应用，帮助更多项目团队解决软件设计评审、代码评审、测试范围分析和测试案例评审难题，构建“小而精”的测试能力矩阵，提升项目的开发交付质量和测试效率。

五、项目成效

基于 AI 技术通过多维度创新优化，在人力、测试、研发等领域实现显著降本增效，显著降低成本、提升效益，每年创造经济效益 214.7 万元。未来，随着应用规模拓展，经济效益将持续攀升。具体分解如下：

1. 人力成本降低（核心公式：节约工时 = (传统耗时 - 项目耗时) × 团队人数 × 年迭代次数）：节约 18.8 万元

传统模式工时计算：5 人测试团队，单版本传统测试范围分析和案例评审需 2.5 天，每天按 10 小时计算，则单版本耗时为 $5 \times 2.5 \times 10 = 125$ 小时。年迭代 12 次，那么传统模式下一年的总工时为 $125 \times 12 = 1500$ 小时。

项目模式工时计算：项目实施后，单版本仅需 5 小时，年迭代 12 次，项目模式一年的总工时为 $5 \times 12 = 60$ 小时。

工时节约计算：年节约工时 = 传统模式年工时 - 项目模式年工时，即 $1500 - 60 = 1440$ 小时。

人力成本节约计算：按中小行测试人员时薪 100 元计算，人力成本节约 = 年节约工时 × 时薪，即 $1440 \times 100 = 14.4$ 万元。

培训成本节约计算：新人上手周期从 14 天缩短至 3 天，年均入职新人 5 人。节约的培训工时为 $(14 - 3) \times 8 \times 5 = 440$ 小时，培训成本节约 = 节约的培训工时 × 时薪，即 $440 \times 100 = 4.4$ 万元。

人力成本合计：人力成本节约与培训成本节约相加， $14.4 + 4.4 = 18.8$ 万元。

2. 测试效率提升（核心公式：缺陷修复成本节约 = 传统修复成本 - 项目修复成本）：节约 9 万元

传统模式年修复成本计算：单版本平均有 15 个缺陷（根据历史缺陷数据），其中 30% 在生产环境修复，每个修复成本 8000 元；70% 在测试阶段修复，每个修复成本 3000 元。则传统模式下一个版本的修复成本为 $15 \times 30\% \times 8000 + 15 \times 70\% \times 3000 = 36000 + 31500 = 67500$ 元。年迭代 12 次，传统模式年修复成本为 $67500 \times 12 = 81$ 万元。

项目模式年修复成本计算：项目实施后，20% 的缺陷在生产环境修复，80% 在测试阶段修复。项目模式下一个版本的修复成本为 $15 \times 20\% \times 8000 + 15 \times 80\% \times 3000 = 24000 + 36000 = 60000$ 元。年迭代 12 次，项目模式年修复成本为 $60000 \times 12 = 72$ 万元。

年节约修复成本计算：年节约修复成本 = 传统模式年修复成本 - 项目模式年修复成本，即 $81 - 72 = 9$ 万元。

3. 研发成本控制：节约 102.9 万元

通过技术复用、流程优化、效率提升三大路径，实现同类项目研发成本的显著节约。核心在于构建可复用的智能化测试分析体系（扫描规则、AI 智能体调优、组件化工具），避免传统模式下每个项目重复投入的“烟囱式开发”，

从代码分析、测试范围定位到自动化闭环形成标准化能力输出，直接降低新增项目的前期设计、开发及验证成本。

按每年新增 3 个项目，共节约（技术资产复用 27.5+ 效率提升 3.1+ 标准化交付 3.7）×3=102.9 万元。

3.1 技术资产复用规避重复开发：单项目成本节约 27.5 万元

扫描规则库扫描规则库：沉淀 300+ 金融核心业务场景扫描规则（如“账户开户组件变更影响扫描”“资金转账链路风险校验”），覆盖 90% 以上高频交易场景，可直接应用于同类项目；

AI 分析模型：训练成熟的“缺陷早发现模型”“覆盖度评估模型”，基于历史 200+ 版本缺陷数据持续优化，避免每个项目重新建模调参；

组件化工具：自主研发的“核心业务组件化定义工具”“代码差异分析引擎”“链路影响热力图生成器”，形成标准化工具集，支持一键导入新项目。

传统模式 vs 项目模式对比：

成本项	传统模式 (单项目)	项目模式 (复用后)	单项目 节约成本
需求分析与规则设计	8 万元	0 (直接复用)	8 万元
AI 模型搭建与智能体调优	12.5 万元	5 万元	7.5 万元
工具开发与集成	12 万元	0 (工具集复用)	12 万元
合计单项目成本	32.5 万元	5 万元	27.5 万元

说明：

需求分析与规则设计（8 万元）：在传统模式下，需求分析与规则设计通常需要 1-2 名资深业务分析师及测试专家协作完成，结合中小城商行市场薪资水平，资深业务分析师月薪约 2-3 万元，专家月薪约 2.5-3.5 万元，完成单项目该阶段工作，平均耗时约 20-30 天，综合人力成本及其他资源消耗，估算单项目费用约 8 万元。

AI 模型搭建与智能体调优（12.5 万元）：依赖组件化扫描基础源数据，可根据用户要求结合 AI 推理能力和智能体开发的工作流实现组件化深度优化建议、调用链路设计合理性建议、版本影响范围场景测试建议等结构化产物，拜托了繁琐结构化代码设计 - 开发 - 测试的环节，由 AI 智能编排分析，提升效率并保障专业性的同时又节省了开发的成本。传统分析代码开发工作量在 5 人月，每人月成本 2.5 万元，需要 12.5 万，模型搭建 + 智能体开发工作量在 2 人月，每人月 2.5 万元，大概需要 5 万元，节省成本近 7.5 万元。

工具开发与集成（12 万元）：依据开发团队规模、开发周期及技术难度确定。开发代码分析、测试覆盖度评审等工具，通常需组建包含 2-3 名软件开发工程师、1 名测试工程师及 1 名项目经理的团队。软件开发工程师月薪约 2-3 万元，测试工程师月薪约 1.5-2.5 万元，项目经理月薪约 3-4 万元。项目开发周期约 45-60 天，结合工具与现有系统集成所需的调试、适配工作，综合估算单项目成本约 12 万元。



3.2 效率提升减少人力投入：单项目成本节约 3.1 万元

关键技术创新提效：代码分析效率提升（2 万元）+ 测试覆盖度评审效率提升（0.14 万元）+ 业务链路追踪精度提升（0.96 万元）= 3.1 万元。

代码分析效率提升 50%：通过“GITDIFF+ASM + 字节码比对”三重技术，自动识别代码变更影响域，替代人工逐行梳理代码的低效工作。传统模式下，单项目代码分析需 5 人 × 10 天 = 50 人天，现在仅需 25 人天，单项目节约 25 人天，按人均日成本 800 元计算，单项目节约 2 万元（25 × 800）。

测试覆盖度评审效率提升 80%：AI 驱动的覆盖度智能评审将“人工经验判断 + 逐案例校验”的 2 天耗时压缩至 2 小时，单项目节约 1.75 天（约 14 小时），节约 0.14 万元（14 × 100 元 / 时，时薪按 100 元计算）。

业务链路追踪精度提升 40%：减少因“漏测关联模块”导致的设计返工，传统模式平均每个项目因漏测引发的返工需 3 人 × 5 天 = 15 人天，项目模式下返工量降至 3 人天，单项目节约 12 人天，年节约 0.96 万元（12 × 800 元）。

3.3 标准化交付降低返工及沟通成本：单项目成本节约 3.7 万元

质量前置减少后期返工：通过代码级差异识别和组件级影响评估，在设计阶段提前拦截 70% 以上的潜在缺陷，避免传统模式中“开发→测试→返工”的循环损耗。据历史数据，传统单项目后期返工成本平均 5 万元，项目模式下下降至 1.5 万元，单项目节约 3.5 万元。

自动化闭环减少沟通损耗：“差异扫描→AI 评估→自动化执行”全流程贯通，消除传统模式中开发、测试、需求三方的反复沟通成本（单项目平均节约 20 小时沟通时间，时薪按 100 元计算，折合 0.2 万元）。

合计节约费用 3.5+0.2=3.7 万元。

3.4. 隐性成本节约：节约 84 万元

版本缺陷遗漏率降低 70% 以上，原本生产环境中因缺陷遗漏导致的隐性成本（如客户流失、业务中断损失、监管处罚等）平均每个版本为 10 万元（基于城商行历史事件损失进行估算）。

年迭代 12 次，那么年节约隐性成本 = 每个版本隐性成本 × 年迭代次数 × 缺陷遗漏率降低比例，即 10 × 12 × 70% = 84 万元。

六、经验总结

产业辐射效应：本项目的标准化扫描规则和 AI 分析接口，为测试工具链、代码分析组件等相关产业发展提供了新机遇。推动了“AI + 测试”生态的构建，其组件化分析方法和调用链路建模技术，为低代码开发、DevOps 集成等新兴技术提供关键支撑，促进软件产业向自动化、智能化方向转型。

示范效应：

1. 技术普惠：让“精准测试”触手可及

人才赋能：通过“可视化 AI 分析界面 + 标准化组件资产库”，降低技术门槛，使初级测试人员也能完成复杂业务覆盖度评估（传统需 5 年以上资深工程师），缓解高层级人才缺口的困境；

2. 行业范式：从“跟跑”到“并跑”的转型启示

可复制模式：形成“技术工具（扫描引擎）+ 场景化 AI（银行专用模型）+ 本地化适配（流程定制）”的三位一体方案，提供“轻量化创新模板”；

生态共建：推动测试工作范畴从“依赖人工经验”向“数据驱动测试”转型，建立“精准分析 - 智能决策 - 高效验证”的新型研发范式。通过输出组件化分析方法、调用链路建模技术，为低代码开发、DevOps 集成提供关键支撑，在金融科技竞争中实现“技术能力并跑、响应速度领跑”，推动整个测试工作从“粗放测试”向“智能测试”生态升级。

中信建投证券：基于 AI 大模型的多智能体技术在投顾领域的应用案例

一、项目背景及目标

1. 项目背景

中信建投证券打造了“找好投顾——到中信建投”的行业品牌。形成了人工投顾 + 智能投顾 + 基金投顾三位一体的、涵盖股债基的完整的投顾业务体系。特别是基于大数据与人工智能技术的智能投顾，为全体客户提供投前、投中、投后全投资周期内的个性化的智能化投顾服务。但传统智能投顾服务，因为受限于 AI 技术的发展水平，侧重于“智能投”而非“智能顾”，只能通过分析客户画像、产品画像，基于客户自选、持仓、交易行为、使用行为，结合投研端策略信号提供预测型“智能投”服务，而提供陪伴、心理按摩的“智能顾”服务深度不足。

核心原因是传统 NLP 技术在意图识别、多轮对话、上下文语义理解等方面的能力不足，难以支撑“智能顾”服务的预期体验。随着以 OpenAI 的 ChatGPT 为代表的大语言模型技术的突破，其在多轮对话、意图识别、上下文语义理解等自然语言方面的能力远超传统 NLP 技术，可以大幅度提升“智能顾”方面的服务体验。同时，agent 在规划、反思等方面的能力可以进一步提升大语言模型处理复杂任务的能力，弥补大语言模型的技术不足。比如 AutoGPT 能够自主进行问题识别和拆解，分步骤解决复杂问题，FinGPT 可以作为投资顾问自主输出投资建议，并能够根据用户的反馈自主改进建议。但大模型同时带来了幻觉、数学计算能力差、可控性差、道德偏见等问题。比如 AutoGPT 自主处理问题经常陷入自我反思的无限循环导致问题处理失败，FinGPT 输出的投资建议缺少依据，不能说明其投资原理。

2. 项目目标

本平台旨在搭建一个集大模型训练、微调、开发、测试、验证为一体的完整的开发框架，通过研究投顾领域大模型的开发方法，以及基于 RAG、Agent 等技术的智能体改进方法，全面提升“智能顾”的能力。通过模型训练实现道德对齐、提升模型的业务专业性。

(1) 使用 RAG 实现答案引用的知识可溯源，通过固定场景人工辅助 Agent 进行业务编排、发散场景注入专家知识引导 Agent 进行自主规划来指导模型进行正确的问题处理。通过取长补短，利用大模型优势的同时实现模型输出可信、可控，以技术创新实现 AI+ 业务创新。

(2) 在 B 端提高投顾人员在客户问答、技术分析、话术生成、日报生成、热点分析、报表生成、策略推荐等方面的一致性、专业性、可信性，提高专业下限。

(3) 在 C 端提高自主对客户业务的智能化水平，为自主客户分析、交互、产品解读、推荐、服务等方面实现突破。

二、创新点

1. 解决现有投顾业务体系难点

实现多技术融合的智能投顾体系。大语言模型与 Agent 协同，基于 KGPT 大语言模型突破传统 NLP 限制，结合



Agent 技术实现复杂任务规划与自主决策。RAG 技术实现知识可溯源，通过检索增强生成（RAG）技术整合金融知识库（如研报、财报、产品库），确保回答内容引用来源透明，解决大模型“幻觉”问题，提升可信度。

2. 全流程开发框架与场景覆盖

模型工厂一体化开发，搭建集大模型训练、微调、测试、部署于一体的开发平台，支持自定义 Agent 工作流编排（如人工预设流程或基于意图理解的智能规划），降低技术门槛，加速业务落地。多场景智能化服务，覆盖投顾业务全链条。

3. 业务赋能与用户价值

通过产品助手实现投顾产品问答自动化，解决客户经理长尾问题，释放运营专家精力；策略助手整合选股、诊股工具，通过 API 调用与大模型汇总，提升投顾响应速度。多模态与多渠道支持，支持文本、图片、视频等多格式知识库，适配企微群、APP 等多终端交互，例如企微机器人实时答疑、APP 端智能诊股报告生成。

4. 行业价值与创新示范

树立金融科技应用标杆，突破“重投轻顾”瓶颈，通过“智能投 + 智能顾”双轮驱动，构建覆盖投前、投中、投后的全周期服务，为行业提供可复用的 AI+ 财富管理实践范式。探索大模型在金融领域的合规应用，通过 RAG、Agent 等技术平衡创新与风险，为证券行业智能化转型提供技术参考。

三、项目技术方案

1. 项目规划主要分为以下几个阶段

(1) 基础构建与智能体设计，包括需求分析与场景定义，调研投顾业务痛点，技术选型与模型训练，单智能体功能验证。

(2) 多智能体协同与系统集成涉及，包括多智能体通信机制开发，实时数据融合与决策链路，合规与可解释性增强。

(3) 投顾场景落地与优化，主要包括首个客户场景正式上线，性能优化与规模化，以及打造一个安全、高效、可解释的 AI 投顾平台。

2. 技术架构图如下所示

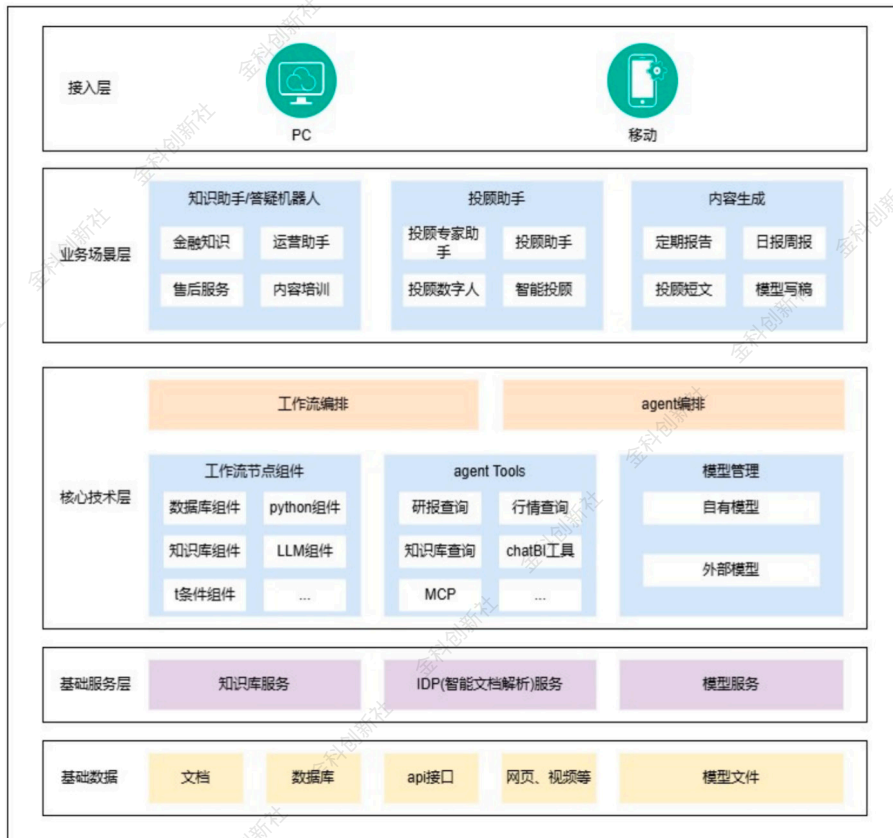
(1) 基础设施层

计算资源：GPU 服务器（如 NVIDIA A100）支持大模型训练与推理，CPU 服务器部署 Agent 调度、数据服务等轻量级任务。容器化部署（Docker）实现资源弹性扩展，通过 Kubernetes 管理容器集群，支持动态扩容与负载均衡。

存储系统：数据库：MySQL/MongoDB 存储用户会话、权限数据；Redis 缓存高频访问数据（如客户画像、历史分析结果），缓存有效时长 5 分钟。

非结构化数据：研报、财报、公告（PDF/Word/Excel 格式），通过 IDP（智能文档处理）技术解析为文本，存入向量数据库（如 Milvus），支持 Embedding 检索。

网络与安全：内外网隔离，API 接口通过 HTTPS 加密传输，采用 CAS 统一身份认证，防止未授权访问。部署 Web 应用防火墙（WAF），定期进行渗透测试与代码安全扫描。



(2) 核心技术层

基础模型基于 Transformer 架构的自研模型（如 K-GPT），混合金融领域数据（研报、政策、交易数据）预训练，支持多轮对话与意图理解。技术增强中 RAG（检索增强生成）针对用户问题，先通过向量数据库检索相关知识库内容（如产品 QA、研报片段），再由大模型生成回答，确保内容可溯源。

Agent 技术中主 Agent 作为交互入口，负责意图识别（如判断“选股”“诊股”需求）、子 Agent 调度、结果汇总与风控审核。子 Agent 中产品助手基于 RAG 解析投顾产品知识库，解决标准化问答（如“Level2 黄金版是否含白银版服务”）。策略助手调用选股 / 诊股 API（如同花顺因子选股接口），生成策略建议（如“成交量 <3000 手且市盈率 >50 的股票推荐”）。个股分析助手整合多源数据（研报 RAG+ChatBI 数据库查询），输出财务分析、行业竞争格局等报告。默认助手兜底处理非核心问题（如金融知识问答），直接调用大模型生成回答。

(3) 应用服务层

投顾 Copilot 平台前端交互中提供智能诊股、策略推荐、投后分析报告。企微群机器人支持多群并发答疑，界面类似 ChatGPT 的 LUI 交互，显示知识源引用。

后端服务模型工厂支持可视化 Agent 创建、 workflow 编排、知识库管理（上传、解析、版本控制）。评估服务三级评测体系（通用能力 + 金融能力 + 场景测试），结合 GPT-4 初评与人工专家终评，优化模型准确率。

(4) 业务场景层

投顾服务中产品问答覆盖 20+ 投顾产品，通过 RAG 快速定位知识库 QA 对，秒级响应。



投资分析中个股诊股调用技术指标（MACD、KDJ）、主力资金流向数据，生成多维度分析报告。行业分析基于研报 RAG 提取行业政策、竞争格局等信息。策略推荐根据用户风险偏好（如“保守型”），调用多因子模型生成选股组合，支持因子参数自定义（如“市盈率 < 20 且股息率 > 4%”）。客户运营中个性化匹配基于五维投资性格标签，提供投资总结。

3. 商业模式

收入来源方面，C 端增值服务定制化报告，个股深度分析、客户画像投资总结。B 端解决方案向中小券商输出投顾大模型能力，收取技术服务费或分成，与金融机构合作，提供定制化产品解读服务。

成本结构方面，技术投入包括大模型训练算力、数据标注、API 接口采购。运营成本有知识库维护、合规审核、客户服务。

4. 业务模式

服务场景方面，投顾服务产品问答，覆盖 Level2、点金投等 20+ 投顾产品，支持 7×24 小时自动答疑；投资分析中提供个股诊股（技术面 / 资金面）、行业分析、券商研报解读；策略推荐中多因子选股、主力动向跟踪、资产配置建议（需人工审核后输出）。

客户分层方面，B 端客户包括客户经理、投顾专家，通过平台提升服务效率与专业度。C 端客户包括普通投资者（免费基础服务）、高净值客户（付费深度分析），差异化定价。

四、运营情况

目前推广到投顾与客户，B 端投顾赋能中用户规模：已覆盖数客户经理与投顾人员，通过企微群机器人、内部管理平台提供实时支持，日均处理产品咨询、技术分析等问题超万次，累计调用次数 7500 频次，系统运行稳定。

场景落地方面，产品培训中产品助手解决 80% 以上的高频问题（如“Level2 港股版签约限制”），释放总部运营专家 30% 的精力。投研助手可以提供金融知识问答、投顾产品问答、个股基本面分析、技术面分析、资金面分析、选股、诊股等投资顾问服务，投顾观点生成效率提升 80%。C 端用户中企微生态中接入企微客户群，机器人自动答疑覆盖率达 30%，客户咨询响应时效缩短至秒级。

五、项目成效

1. 经济效益

提升投资决策效率：多智能体系统可并行处理海量市场数据、宏观经济指标和投资者偏好，快速生成个性化投资建议，降低人工投顾成本，提高金融机构的运营效率。

优化资产配置与风险控制：通过多智能体协同分析（如交易策略智能体、风控智能体、客户画像智能体等），动态调整投资组合，减少人为情绪干扰，提升长期收益稳定性。

创造新型金融服务模式：支持 7×24 小时智能投顾服务，覆盖长尾客户群体，扩大金融机构的客户基数与收入来源（如订阅制、绩效分成等）。

2. 社会效益

普惠金融推广：降低专业投顾服务门槛，让中小投资者以更低成本获得高质量的财富管理建议，缩小投资信息鸿沟。

增强市场稳定性：多智能体系统可通过实时监测市场异常（如波动预警、舆情分析），减少非理性交易行为，促进资本市场健康发展。

推动金融科技伦理实践：通过可解释 AI 技术（如智能体决策溯源），提高投顾透明度，帮助监管机构防范算法操纵风险，保障投资者权益。

六、经验总结

(1) 技术融合驱动创新，以大语言模型为底座，结合 RAG、Agent 技术构建分层架构，实现知识可溯源与任务自动化，解决传统投顾“重投轻顾”痛点。

(2) 分阶段精准落地，从单一场景（产品助手）逐步扩展至多场景协同（个股分析、策略分析），通过小步快跑降低风险，确保技术与业务深度适配。

(3) 生态化推广策略，内部聚焦于智能投顾业务线，外部探索 SaaS 化输出与第三方合作，构建开放生态，推动技术价值最大化。



中信建投证券：基于财富管理业务的 员工协能全链平台

一、项目背景及目标

随着大模型技术在证券行业的深化应用，财富管理智能化转型成为行业核心趋势。国内资本市场扩容与居民财富管理需求升级，推动海量投资者通过线上渠道寻求服务，传统模式在客户个性化服务、营销效率及风险管控上的短板日益显著——人工服务难以满足差异化需求，合规框架下的降本增效需求迫切。员工也因大量重复繁琐的基础工作，难以投入精力提升专业能力与服务水平。

从市场现状看，国外头部券商如摩根士丹利、高盛已通过 AI 技术优化投资建议与量化策略，但复杂产品适配和跨模态数据融合存在瓶颈；国内华泰证券、国泰君安等依托大模型提升了智能客服与投研效率，却面临大规模数据实时处理和极端场景风险响应的挑战，行业普遍存在智能与业务协同效率问题，这些问题也限制着员工工作效能的发挥。

中信建投证券针对传统模式中客户服务同质化、营销精准度不足、风险评估滞后等痛点，基于大模型技术构建基于财富管理业务的员工协能全链平台。该体系旨在通过大模型及智能化技术整合客户行为、市场舆情与产品数据，突破传统数据处理边界，不仅为数字金融转型提供技术支持，更通过智能化手段为员工赋能。

本项目目标聚焦核心场景智能化与员工能力提升的深度融合。在客户服务上，员工借助大模型分析结果，实现需求预判与精准触达，提升服务专业性；市场分析环节，员工利用动态情绪图谱获取深度市场信息，增强投资决策分析能力；产品推荐过程中，依据客户画像生成的个性化方案，助力员工提升资产配置专业度；风险评估方面，实时智能预警让员工及时掌握风险动态，降低风险识别压力；业务协同构建的“人机协作”生态，优化跨部门资源调度，使员工能够将更多精力投入到高价值服务与个人能力提升中。

本项目致力于实现以下三大价值：通过技术赋能打造智能服务闭环，提升客户满意度与服务效率，同时让员工从基础工作中解放；降本增效释放人力，使员工能够投入更多精力开展高价值服务，实现个人能力与职业价值的提升；形成可复制的解决方案，推动行业向“数据驱动、智能决策”转型，助力证券财富管理进入技术引领新阶段，为行业员工发展提供新范式。

二、创新点

当前，证券行业在金融科技领域取得显著进展，以大模型、云计算及人工智能等为代表的数字技术创新发展不断冲击着证券公司的传统商业模式。在金融科技发展浪潮下，数字化转型成为证券公司适应数字经济、谋求生存发展的必由之路。随着资本市场的不断发展，证券公司的管理规模也愈发庞大，根据中国证券业协会发布的最新数据显示，证券资产管理规模节节攀升。各证券公司在面临庞大的客户群体时，如何提供高效、专业且优质的服务已然成为证券行业目前的业务痛点之一。

为解决上述问题，中信建投证券坚持以大模型驱动，实现证券财富管理数智化转型，利用大模型技术和人工智能技术，建设了基于财富管理业务的员工协能全链平台。平台通过用户的全景分析画像，在获得客户允许的情况下进行客户行为的实时预测与分析，及时识别业务机会，同时融合知识图谱、机器学习和推荐算法构建了一套证券领域的专业数据管理中台，利用文本向量计算技术为员工和客户提供专业的知识推荐和解析，为员工提供机器人协同服务、智能机器人电话服务、智能机器决策等诸多能力，有效提升了员工服务效率，平台所具备的业务创新点和技术创新点如下：

平台业务创新点

中信建投证券财富管理业务的员工协能全链平台，以多模态异构融合技术为核心，整合语音、文本、交易行为等全场景数据，解决了传统业务中数据孤岛与实时性不足的技术难题，形成“数据-服务-知识-智能体”一体化体系，推动客户服务精准度提升与业务连续性强化，成为证券行业数智化转型的标杆实践，其业务创新点如下：

1. 多元业务覆盖，提升员工综合服务效率

平台覆盖多种证券业务类型与客户交互方式，涵盖投资顾问、业务办理、客户答疑、证券产品营销、音视频服务等，能够为客户提供丰富且优质的服务。平台协同能力支持多业务线合作服务，丰富的员工端功能有效提升员工的综合服务效率，实现从传统人工服务向“人机协同”模式的进阶。

2. 动态金融智能知识库提升员工服务专业化程度

基于人工智能技术的数据管理中台实现了组织内各部门及分支机构数据库、表格、文本、图像、音视频文件的聚合接入，为员工提供高效的知识推荐和检索，支持服务过程中的高效人机协同。通过实时同步证券行业政策、市场动态及产品信息，保障了员工服务内容的专业化与时效性。

3. 稳定合规控制与全方位质检能力

智能合规质检实现针对服务和营销内容的实时文本、语音和视频 100% 覆盖质检，涵盖事前、事中和事后的全流程合规控制。平台语义解析技术能够根据上下文内容实现模糊内容质检，在关键词质检基础上大大提升质检内容深度、广度、准确率，同时支持自定义质检目标内容和业务范围，适应企业业务变化需求，有效提升质检员工作效率，筑牢证券业务合规防线。

4. 新一代平台大幅提升智能比重

新一代平台重点关注智能科技，利用大数据存储海量用户行为数据，利用并行计算将过去难以实现的客户分析行为进行落地，针对证券开户、产品销售、浏览兴趣等众多场景进行实时的监听和分析，及时反馈员工进行跟进服务。云原生能力保障系统的稳定运行和动态扩容能力，能够应对行情波动导致的客户流量骤增。大模型智能化底座为全部服务过程进行效率提升，为每个员工打造个性化服务驾驶舱，让科技的力量落地一线，极大提升证券客户服务质量。

平台技术创新点

中信建投证券基于财富管理业务的员工协能全链平台以多模态数据融合与智能体协同为核心，构建四层递进式技术链路，并以反馈大模型实时优化策略闭环，最终实现“数据治理-需求洞察-决策生成-行动追踪”全流程智能化的技术突破，其技术创新点如下：

1. 智能员工赋能辅助

平台创新性地引入基于多模数据的人机协同服务机制，在技术上引入知识推荐、情感识别等智能化手段赋能人工投顾，结合语音转写、图像理解、意图识别进行自动检索数据和最佳匹配答案，实现客户自动对话服务和高效的服务辅助。针对证券客户的投资咨询、交易指导等高频需求，系统可毫秒级响应并生成专业建议，极大提高金融客户服务与营销效率，提升客户体验。

2. 面向证券垂直领域的智能数据中台

平台引入面向证券垂直领域的智能数据管理中台服务，实现各部门及分支机构间的多模数据聚合接入；支持图、FAQ 问答对、向量等多样化知识的定义、生产、管理及应用的全流程无代码图形化工具化操作；结合强化学习驱动的服务分流模型，实现关键节点客户服务并实现潜在需求挖掘，为证券投研、客服、营销等业务线提供全场景知识支撑。



3. 多模态异构融合的智能金融服务能力

平台搭建多模态融合的智能金融客户服务体系，通过引入多模态融合技术构建基于多媒介互通的客户服务方式，将语音、视频、文本等多模态输入实时转换为目标语言文本并解析，增强不同特征数据融合能力，并支持数据隔离与混合检索机制，实现跨模态数据的安全存储与毫秒级特征提取，从而为客户提供更精准、多元的证券服务，提升客户满意度。

4. 融合大数据与人工智能的智能体平台基座

平台不断拓展和完善人机对话解析的能力，创新性地引入智能体协同机制，构建语义搜索、情感分析、会话意图识别等智能应用组件，建设运营高效的统一非结构化 AI 分析平台，实现“数据治理 - 需求洞察 - 决策生成 - 行动追踪”全流程智能化的技术突破。该基座支持证券业务全链条自动化，在节省人力资源成本的同时提升客户服务体验，为证券财富管理业务构建技术护城河。

三、项目技术方案

中信建投证券基于财富管理业务的员工协能全链平台，采用“四层智能闭环”架构。底层依托混合云多模态数据管理仓库，整合公司内跨模态数据，借 Data Agent 实现超融合存储与隐私计算的安全治理。中台部署 AI 模型算法库，为各业务系统及场景提供通用智能服务，避免算法重复开发，降低模型运维成本。应用层通过多智能体协同（如投顾 / 风控 / 交易 Agent）接入关键业务系统，结合 Agent 自动化与数据管理中台形成策略闭环，提升传统业务系统智能化水平。顶层以大模型赋能体系贯通“数据 - 服务 - 决策”全链条，实现智能投研、精准推荐与风险穿透管理。平台全栈集成零信任安全架构与 AI 主动防护机制，支撑亿级实时交互，保障业务连续性与服务精准度，成为证券行业数字化转型核心引擎，推动客户体验与业务效能跨越式提升。

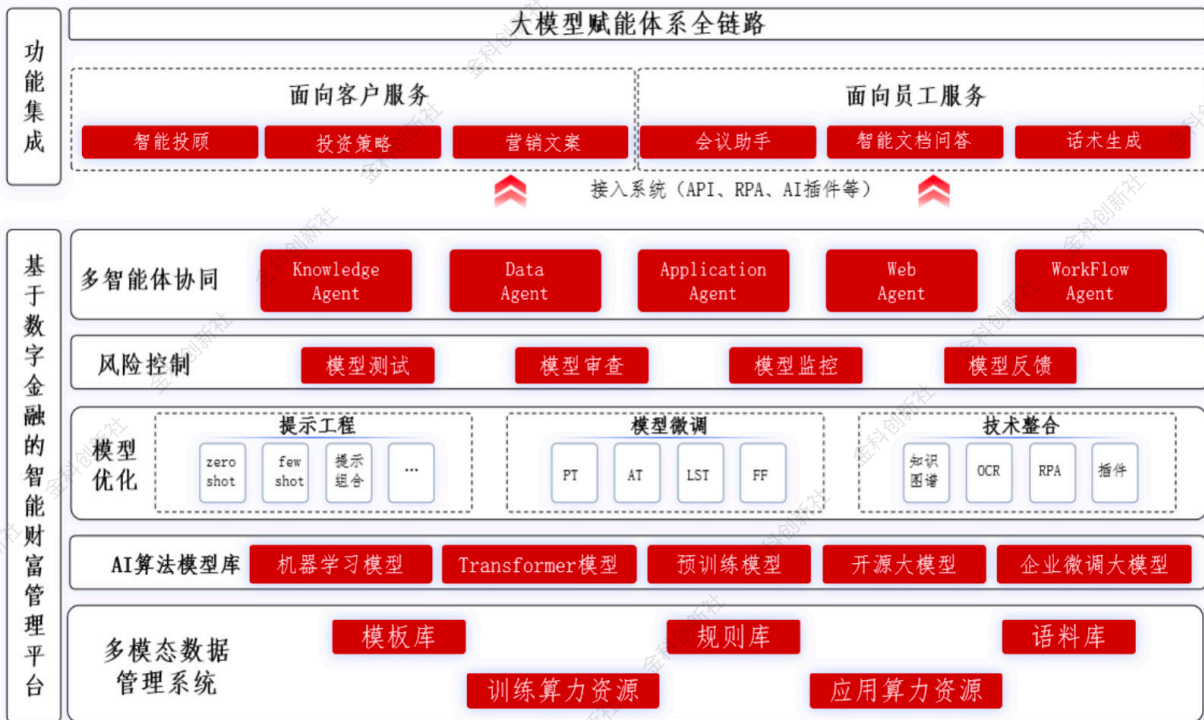


图 1 总体架构图

技术方案

(1) 系统业务架构

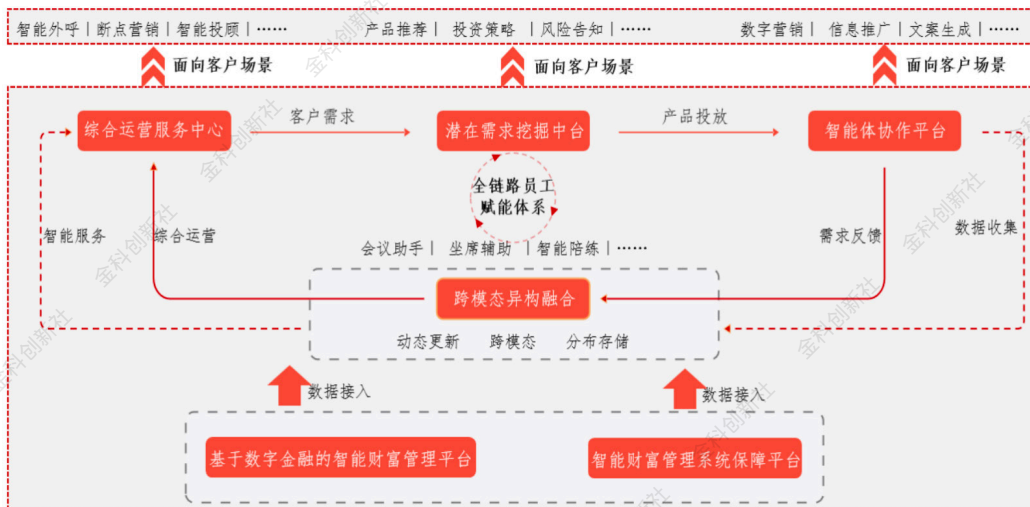


图 2 业务架构图

中信建投证券基于财富管理业务的员工协能全链平台在数据存储端，以跨模态异构特征融合技术整合全场景数据，利用动态更新算法与跨模态分析引擎，实现数据的深度关联。平台基于综合运营服务中心动态调度客户交互系统（如智能外呼、断点营销）并动态捕获客户需求，潜在需求通过智能数据管理中台系统对客户提供智能化后续服务（产品推荐、投资策略、风险告知等）。在智能体应用端，智能体协作平台的智能体应用通过嵌入式方式与核心业务系统交互，构建营销、风控等场景的 Agent 协同网络，实现数据互通与智能决策闭环。对内业务场景中，财富管理平台依托全链路员工赋能体系（集成 NLP、研报生成）进行研报生成、坐席辅助等全流程提效，形成“数据 - 服务 - 知识 - 智能体”一体化体系，推动客户、员工服务精准度提升与业务连续性强化，成为证券行业数智化转型的标杆实践。

(2) 应用架构



图 3 应用架构图



平台在应用层的建设通过“分层协同、业务驱动”模式构建端到端智能化体系。前端聚焦客户交互，整合客户画像、行为分析及实时响应模块，实现精准推荐与定制化服务；中台为业务中枢，集成智能投研、组合优化、合规风控三大引擎，通过算法模型预测市场趋势、优化持仓结构，自动化生成财富管理策略并实时监控交易异常；后台强化运营支撑，基于 Agent 技术自动化处理交易清算、信息披露及数据归档，确保流程高效透明，最终形成“客户需求洞察 - 服务策略匹配 - 风险穿透管理”的闭环生态，推动证券业务在投资决策精准性、业务执行效率及合规管理深度上的全面提升，助力数字化转型目标落地。

(3) 系统技术架构

本系统技术架构为递进式技术链路：多模态异构数据融合仓库（提取文本 / 语音 / 视频 / 时序等多样化数据），支持数据隔离（如差分隐私与同态加密）与混合检索机制（RAG 增强检索 + 向量化索引）实现跨模态数据的安全存储与毫秒级特征提取；客户运营服务阶段依托多模态解析引擎（集成 ASR 语音转写、VIT 图像理解、NLP 意图识别）实时构建客户交互知识图谱，实现关键节点客户服务并实现潜在需求挖掘；需求挖掘层创新采用联邦学习框架，联动深度行为序列分析模型（LSTM+GraphSAGE）与市场情绪感知 Agent，从持仓变动、宏观经济指标等维度挖掘客户隐性需求（如风险偏好迁移预警），生成预见式资产配置建议；智能体协同层通过集成任务编排引擎与规则引擎，驱动智能投顾、合规审核、交易执行等 Agent 的联邦式协同，并以反馈大模型实时优化策略闭环。

(4) 系统数据架构 - 多模态异构融合数据架构

多模态异构融合数据仓库通过深度融合多种模态处理引擎，结合决策层隐空间融合技术（Latent Fusion 3.0），构建了覆盖证券全业务场景的智能数据处理体系。系统采用 NLP 4.0 框架实现金融语义解析高准确率，ASR&TTS 双通道实现低延迟，并支持 16 类证券文档的智能图像结构化解析，形成多模态数据特征矩阵。在存储架构层面，通过关系型数据库保障交易数据强一致性，依托分布式文件系统实现 PB 级非结构化数据存储，结合图形数据库 Nebula 构建实体关联网络，并集成搜索数据库 ES 实现毫秒级检索响应。

(5) 系统安全架构

本系统安全架构采用动态化、智能化的多层级防护设计，通过数据加密、分布式存储与实时脱敏技术保障数据安全，结合多层次身份验证、细粒度权限管控及全操作日志追溯构建零信任访问体系，并依托 AI 大模型实现合规检测、风险预判与自动化应急响应。通过深度融合云原生灾备能力，在满足证券行业强合规要求的同时，实现安全防护从被动防御向主动感知的范式升级，为高并发交易场景下的资产安全与隐私保护提供全栈式技术支撑。

ID 四、运营情况

本项目在发展数字金融指导下，坚持以客户为中心，致力于基于以大模型为代表的人工智能技术构建智能化金融服务生态，实现了金融科技与业务创新的协同发展，大幅度提高公司金融服务效率和质量。

1. 全渠道多模态客户触达

系统在传统 IM 交互方式的基础上，搭建了基于实时音视频通信的全媒体交互平台，为员工与客户提供基于音频、视频、桌面共享、同屏交互等多模态交互方式，突破客户触达的时间与空间限制，实现了投顾顾问与客户之间随时随地、面对面的沟通，解决行业客户触达及客群覆盖的重点问题，提高服务便捷性，是金融普惠服务的一种典型实践。平台上线以来，现实现年均智能触达数量 400 万通以上，员工多媒体渠道通话数量 300 万以上。特别在本次市场行情波动期间，累计高效完成 200 余万通电话服务，有效缓解开户回访峰值压力。

2. 全流程智能化辅助

平台依托私有化部署大模型基础平台，集成了智能语音、意图识别、会议助手、智能知识检索、智能质检等多种智能化工具，构建了一整套完备的业务流程辅助体系，通过实时通信服务过程中，洞察客户需求和行为方式，完成高效信息检索和智能化信息推荐，实现全业务流程对员工及业务场景的持续赋能，为客户提供私人订制的个性化服务。并通过事中风险预警和事后智能质检，充分保障了服务内容安全合规，助力公司财富管理业务的可持续增长。

3. 全场景金融服务赋能支持

平台上线以来，基于综合智能服务生态能力的建设，实现呼叫中心、客户回访、企业微信、优问、断点营销等多维度客户服务场景的接入，并完成对多家分支机构、子公司的业务支持，向千余名投资顾问提供业务服务，支持了公司多业务联动，助力传统金融服务的数字化转型升级。智能语音模块年均调用次数达 2000 万次以上，助力数百万客户完成开户业务。平台提供的全流程智能辅助工具，实现近百人年工作量提效，为公司节省直接通信和人工成本达千万元以上，为公司带来显著的降本增效成效，并间接赋能了非现场开户、营销等核心业务场景，创造巨大的业务价值。

ID 五、项目成效

该平台自上线以来，逐步实现个性化服务、精准营销、智能回访等核心场景的全面覆盖。月均辅助员工回访业务 15 万通，触达业务 4 万通，企微机器人问答服务 30 万次以上；基于大模型实现日均生成 2 万通客服录音对话摘要、800 余工单，相关智能化赋能场景累计节约人力成本 300 余万元。智能音视频通信平台每年稳定支持 260 万通以上通话，为公司节约直接通信成本 500 余万元。智能语音服务实现全渠道年均 1800 万次以上的调用，协助完成开户 8 余万户，引入客户资产 120 亿元以上，会议助手年使用人次超 10 万+，在智能议程生成、实时语音转写、会议纪要自动生成等场景中实现高效赋能，为公司带来显著的降本增效成效，并间接赋能了非现场开户、营销等核心业务场景，进一步提升企业协同效率，创造巨大的业务价值。

ID 六、经验总结

1. 落地实践端：从技术整合到业务闭环的全链条经验

需求驱动的分层架构设计：以“员工服务痛点-业务场景需求-技术能力匹配”为逻辑链，采用“数据中台+智能引擎+应用场景”三层架构，避免技术与业务脱节（如智能质检模块同步对接合规政策与服务流程）。

敏捷试点与梯度推广策略：分四阶段迭代（智能语音→数据中台→员工助手→大模型升级），每阶段聚焦 1-2 个核心场景（如试点期优先落地智能外呼缓解回访压力），通过小步快跑降低实施风险。

人机协同的组织适配改造：同步推动业务流程重构与员工能力升级，例如通过“智能体辅助+人工复核”模式，将投顾从基础数据处理中解放，聚焦高净值客户策略定制，实现人力结构优化。

2. 行业价值端：从单点创新到生态赋能的范式突破

技术融合的证券服务新基建：验证了“大模型+多模态数据”在财富管理中的可行性，尤其在跨模态特征提取（如语音情绪+交易行为联动分析）与实时策略生成方面形成差异化优势。



中信建投证券：长尾客户普惠数智金融服务平台

一、项目背景及目标

在金融开放趋势和全球化竞争压力下，金融机构综合财富管理业务将成为未来个人金融服务的主战场，众多机构纷纷将财富管理转型提上日程，未雨绸缪地在线上线下布局财富管理相关的产品和服务。中证协数据显示，截至2024年6月30日，全国券商营业部总量为11,647家，较2023年5月的10,233家有所增长，近年来券商不断加速财富管理转型，营业部的智能化程度持续增加，伴随服务模式转变，不少券商裁撤A型营业部，增设C型营业部，顺应行业发展趋势。同时，营业部职能逐步向为客户提供全产业链综合金融服务转变。

中信建投证券以近6000名前台员工服务千万级客户群体，存在明显服务缺口，长尾服务客群作为财富管理业务的挑战之一，其具有的特点包括三无客户占比高、风险承受能力低、年龄大、学历低、开户时间久、0资产客户占比高，较难触达服务客群显著占用员工精力，随着公司客户规模加速增长，服务缺口还将进一步扩大，未来每年的客户规模快速增长，都将进一步放大服务缺口矛盾，如何服务好长尾服务客群是我司持续探索的方向。为此，我司不断应用技术创新，建设了长尾客户普惠数智金融服务平台，重点关注针对“交易少、资产少、活跃少”的长尾客户，通过行为分析寻找交易出问题、系统不会用、操作遇障碍、产品不理解等众多服务痛点，探寻客户紧急需求，优化过去服务触达效率低、精准度差、专业度低等问题，快速帮助长尾客户解决关键问题。

二、创新点

中信建投证券长尾客户普惠数智金融服务平台以“大模型技术基座+场景化智能中枢”为核心架构，通过技术创新与业务模式重构，打造了覆盖客户全生命周期的智能服务生态体系。平台深度融合大语言模型（LLM）、自然语言处理（NLP）、向量知识增强（RAG）等前沿技术，突破传统金融服务中客户洞察浅层化、服务响应滞后、合规管控低效等瓶颈，实现多维度创新突破：

1. 智能交互辅助

依托自然语言处理与RAG向量检索技术，平台构建了覆盖公司预存知识库、法规制度库、内外数据资源的智能交互辅助中枢，能够在毫秒级响应中完成多维度知识检索，通过动态关联TOP3最优知识条目与实时对话语境，经大模型深度语义加工生成精准应答方案，具备知识覆盖全面、快速检索、智能整合、精准推荐等特点，不仅实现了全类型知识资源的智能化调度，更通过语境感知与内容边界控制技术，确保服务话术的专业严谨性，有效降低合规风险。

2. 画像标签挖掘

平台的画像标签挖掘引擎突破传统结构化数据处理范式，运用大模型对每日数万条非结构化对话进行实时语义解析。通过建立标准化标签体系，平台能够自动提取客户投资偏好、风险特征等多维度画像，同步生成员工服务能力画像与商机标签，并通过跨系统标签统一管理，实现客群分类模型与触达策略的动态优化，具备“快速感知，辅助运营”的特点，使营销转化率提升显著。

3. 智能教练分析

平台智能教练模块构建了对话质量评估的数字化标准，通过实时语义分析技术对服务过程进行多维评价，不仅能够识别客户性格特征与潜在风险，更能结合经典营销方法论生成个性化改进建议，实现话术合规指引与服务质量闭环优化。同时，平台的反馈迭代机制将员工实践经验反哺算法模型，形成能力提升的良性循环，辅助缩短新员工培训周期，持续提升员工专业度。

4. 智能合规质检

智能合规质检体系创新性地采用大模型与小模型协同架构，通过多级检测机制突破传统质检规则复杂难实现、需提前训练、运营成本高瓶颈，支持正则匹配、语义理解、策略组合等混合检测模式，业务人员可自主配置质检规则逻辑关系，既满足通用合规要求，又适配各业务线特性，从而大幅提升质检准确率及规则维护效率。

5. 大模型基座支持

平台的大模型智能底座构建了统一的能力输出平台，通过标准化接口封装会话小结生成、智能标签提取等核心功能模块，支持各业务系统按需调用 AI 能力，既保障了技术服务的集约化管理，又实现了智能营销组件的灵活部署，为公司智能营销提供了可扩展的大模型技术支撑底座。

三、项目技术方案

1. 项目规划与实施路径

① 总体规划目标

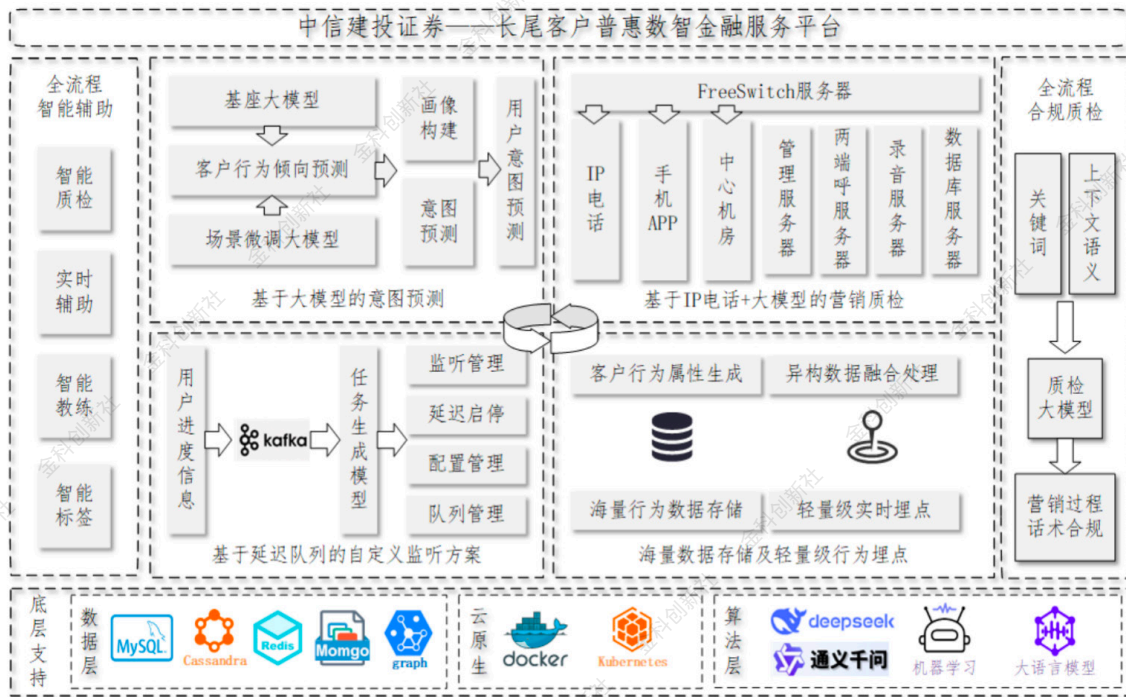
构建以“数据驱动、智能决策、精准服务”为核心的长尾客户普惠数智金融服务平台，通过全链路智能化改造，解决长尾客户服务中存在的触达效率低、服务精准性不足、合规管理复杂等痛点，实现客户服务从“被动响应”向“主动关怀”的范式升级，推动证券行业普惠金融服务能力跃迁。

② 具体实施方案

平台以“数据贯通 - 智能赋能 - 生态融合”为核心路径，整合了优问、互联网营销、智能外呼、断点营销、企业微信、两端呼等全域数据源，构建多渠道全域历史服务记录数据，提高数据深度及覆盖面。平台同步部署了 DeepSeek-R1 671B 等大模型基座，结合 RAG 增强检索技术搭建智能决策中枢，完成智能外呼、坐席辅助、合规质检等核心模块构建，赋能长尾客户服务全流程。在业务侧建立“智能触达 - 服务沉淀 - 策略迭代”闭环，动态优化营销策略库，形成可弹性扩展的智能化服务生态体系。

2. 技术架构设计

中信建投证券长尾客户普惠数智金融服务平台通过平台利用大模型、大数据等技术精准分析客户画像，建设基于智能外呼的多层级营销决策与涵盖事前事中事后的全流程智能坐席辅助模块，提升洞察分析与触达能力，实现了基于短信、电话、在线客服等多种形式的一键触达，解决过去服务触达效率低、精准度差和专业度低等问题，平台技术架构如图 4 所示。



中信建投证券长尾客户普惠数智金融服务平台技术架构图

长尾客户普惠数智金融服务平台基于微服务架构设计，采用 DevOps 技术部署，利用 MySQL 实现数据存储，引入 Redis 实现数据缓存，同时支持 Kafka 消息队列实现用户状态监听，设计架构大大简化系统基础设施构建，通过流水线实施自动代码拉取、自动测试、自动安全监测和自动部署等环节，实现无人工接入的自动化部署，平台各模块以大语言模型技术为核心，围绕智能化进行优化，在实现助力营销决策的同时保障了“事前 - 事中 - 事后”的全流程合规，大幅节省资金和人力投入，技术方案可划分为智能服务、任务管理、行为监控及基础设施四个部分，全流程智能辅助、全流程合规质检两个流程。

①智能服务

智能服务部分以应用智能化为核心，提供安全、高效、便捷的智能服务，利用客户大规模、多模态数据，同时基于基座和场景微调大语言模型、计算机视觉、多模融合及深度学习技术，以预置 Prompt 的方式对客户行为数据对客户进行预测，获取该对话所表达的意图，最终能够预测用户行为，筛选高价值用户，实现营销决策过程精确画像构建和意图识别；同时引入智能决策、智能外呼，主要用于客户运营、智能服务等场景，深入外呼中心痛点，基于语音合成、语音识别、语音理解等 AI 技术建设智能外呼机器人，能够与金融业务完美整合，分担人工坐席压力，在降低成本的同时大幅提升金融营销能力。

②任务管理

任务管理部分包含任务管理、数据抓取、智能分配及一键拨打等功能。任务管理能够从行为监控单元中抓取数据实时生成任务，并将该任务智能分配给营销服务人员，同时内置任务生成引擎，采用可配置模型实现任务的自动生成和匹配。任务管理则是为管理员用户提供的核心功能，可实现任务的重定向、删除等操作。一键拨打支持客户数据一键导入在系统中，由坐席直接点击外呼呼出，从而提高外呼效率，同时能够记录客户需求、备注下次跟进时间、推荐话术及区分出无效客户与意向客户等，从而有效管理外呼数据，提高成单率，同时也避免重复外呼发生的封号、投诉风险。

③行为监控

行为监控部分包含异构数据处理、客户行为属性生成、海量行为数据存储及轻量级用户行为埋点等功能。异构数据处理将不同产生方式的金融数据按照一定的程序和要求进行处理，从而给出针对性的客户行为属性。海量数据存储监控为埋点奠定基础，轻量级用户行为埋点通过状态埋设，收集各界面上客户的状态信息，并发送至统一后台消息队列，消息队列提供全流程客户信息供任何监听消息队列的营销人员使用，为其决策提供数据支撑。

④基础设施

基础设施是传统 IT 基础设施的延伸，它以数据为中心，让数据存得下、流得动、用得好，使数据要素价值实现最大化。平台的基础设施单元主要提供数据基础设施的底层支撑，包含 Spring Cloud、MySQL、Redis 等数据服务及 Docker 等云服务设施和英伟达 A800 系列显卡等硬件算力设施，能够巧妙简化分布式系统基础设施及构建必要工具集，如服务注册与发现、熔断机制、网关路由、配置中心、消息总线、负载均衡、链路追踪等子系统开发，同时基于 DevOps 的部署方式以低成本的方式实现员工、渠道、总览的报表数据汇总服务，为运营决策提供底层支持。

⑤全流程智能辅助、全流程合规质检

全流程智能辅助贯穿客户服务的“事前规划 - 事中执行 - 事后复盘”全生命周期，基于大模型技术构建动态知识引擎，在服务启动前通过客户历史行为与实时状态预生成个性化沟通策略，服务过程中依托实时语义分析实现智能话术推荐与风险预警，对话结束后自动生成服务小结并提取关键标签反哺画像体系。全流程合规质检采用“规则引擎 + 小模型初筛 + 大模型语义核验”的三级架构，支持多维度检测项灵活配置：事前通过正则表达式匹配敏感词库，事中结合意图识别模型实时拦截违规话术，事后运用大模型深度解析对话上下文，识别隐性合规风险（如诱导性表述、信息遗漏等），最终实现服务合规率提升与人工复核工作量下降的双重突破。

3. 业务模式创新

①数据闭环驱动

通过“客户行为采集 - 智能分析 - 策略生成 - 触达执行 - 效果反馈”五步循环，持续优化服务模型，形成“越用越智能”的业务增强回路。

②场景差异化服务

针对高频痛点场景（如交易异常指引、产品认知盲区），封装标准化服务解决方案，支持按客户分群自动触发差异化流程导航。

④人机协同模式

智能外呼 + 人工触达的模式，遇到复杂问题无缝转接人工坐席，并通过实时辅助系统提升服务专业度，实现服务成本下降与质量提升的双重目标。

⑤客户价值深挖

通过商机标签预测模型，辅助客户经理精准识别潜在高净值客户，促进长尾客户资产转化率提升。

四、运营情况

本项目通过“总部策略中枢 - 互联网运营中心执行枢纽 - 分支机构落地终端”的三级协同机制，构建了覆盖全业务链的数字化运营体系，实现服务精准度与运营效率的双重提升：

1. 推广应用成效

总部基于大数据分析构建“客户价值 - 服务需求”双维度分类模型，发布服务任务、内容与标准，覆盖交易异



常处理、产品认知提升等高频场景。互联网运营中心通过智能任务分发系统，日均处理派单量达 4000 通，借助智能外呼机器人实现标准化服务任务的自动化执行，客户触达效率较传统模式明显提升。分支机构依托系统智能筛选的高价值客户清单，开展定制化服务触达，服务营销转化率较人工模式显著提高，形成“总部策略输出 - 中心批量处理 - 机构精准攻坚”的差异化运营闭环。

2. 系统运行效能

平台日均处理客户交互请求超 4000 次，智能辅助模块单日峰值承载量连续上升，语音识别准确率达 92%，对话意图理解精度提升，智能质检系统累计完成十余万次服务质检，违规话术检测响应时间压缩至 500 毫秒内。系统稳定性方面，核心模块可用性达 99.99%，服务响应时效从传统模式的分钟级优化至秒级，客户问题解决率显著提升，投诉率大幅下降。

3. 生态协同成果

通过标准化 API 接口，平台的大模型基座服务已对接公司多个业务系统，实现客户行为数据跨域流动与策略联动。互联网运营中心累计沉淀多项客户行为标签，反哺总部策略迭代效率提升，分支机构通过系统自动化生成的客户跟进建议，人均服务客户量和高净值客户识别准确率上升，形成“数据驱动服务 - 服务反哺数据”的价值闭环。

累计目前，平台累计服务客户 37.36 万人，服务客户数量同比增长 300%，新增 APP 活跃客户接近 20 万人，长尾客户问题解决数量大幅增加，复杂问题处理时效提升，客户满意度显著上升。

ID 五、项目成效

中信建投证券针对长尾客户“交易频率低、资产规模小、活跃度不足”的特征，通过用户行为深度分析，聚焦交易异常、系统操作障碍、产品认知不足等多类高频服务场景，建设了实现“智能分析 - 精准触达 - 服务闭环”的智能服务平台。平台人工外呼累计触达客户 40.4 万人，服务客户数量同比增长 300%，新增 APP 活跃客户接近 20 万人，长尾客户问题解决数量大幅增加，复杂问题处理时效提升，客户满意度显著上升。

中信建投证券大力推动财富管理数字化转型，通过建设长尾客户普惠数智金融服务平台，为综合财富管理业务提供有力的支持，在长尾客户服务场景下取得显著成效。项目不断突破创新，项目累计获得国家发明专利授权三项，先后获得中关村数字经济产业联盟、中国信通院、服贸会等机构颁发的多项荣誉，在《金融科技时代》、《财经》等杂志刊发多篇文章。

ID 六、经验总结

本项目以“技术架构先行、数据驱动决策、生态协同落地”为核心建设经验，通过微服务架构与 DevOps 自动化部署实现系统高弹性扩展，验证了大模型与业务场景深度耦合的技术路径；推广层面创新“总部策略标准化输出 - 运营中心智能批处理 - 分支机构精准攻坚”的三级协同机制，以智能质检与实时辅助系统解决规模化服务中的合规与效率矛盾，最终形成可复制的“AI 能力中台化、服务策略场景化、运营管理闭环化”方法论，为金融行业长尾客户服务数字化转型提供实践范式，其“技术赋能 + 机制创新”双轮驱动模式使服务成本下降，客户留存率显著提升。

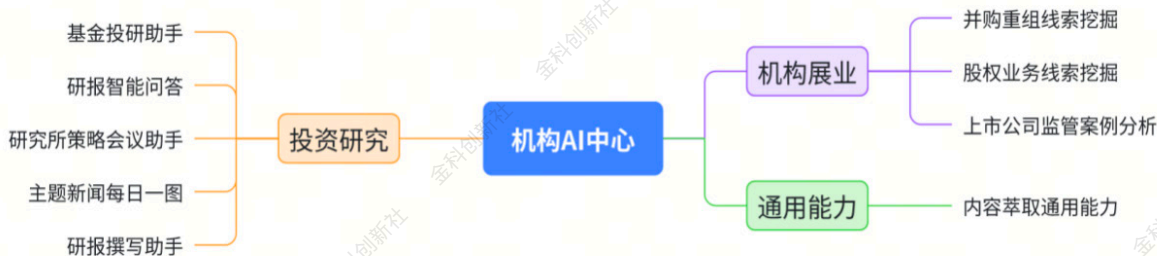
中泰证券：基于大模型的证券机构业务 AI 赋能平台

一、项目背景及目标

近年来，我国金融大模型市场规模快速增长，2024 年已达 26.46 亿元，年复合增长率超 60%，技术应用从基础客服向投研分析、展业辅助、合规管理等核心场景加速渗透。头部券商如中信证券、国泰海通证券已通过自研大模型在智能投顾、舆情分析等领域建立优势，行业竞争进入“技术+场景”深度融合阶段。

中泰证券在内部运营中发现，投研业务高度依赖人工经验积累，从市场数据采集、研究物料生产到研究报告总结分析均需投入大量时间进行手工处理，特别是在处理非结构化数据，如电话策略会议录音时，人工转录与信息提炼过程存在效率瓶颈与信息失真风险；机构展业系统功能分散，客户经理需跨多个平台检索信息，商机响应周期长，缺乏数据驱动的精准营销支持工具，制约了机构客户综合服务能力的持续升级。为破解这些痛点，公司启动机构 AI 中心建设项目，项目覆盖基金投研助手、研报智能问答、主题新闻每日一图、研究所策略会议助手、研报撰写助手、股权业务线索挖掘、并购重组机会挖掘、上市公司监管案例分析、内容萃取通用能力九大场景应用。

目标通过大模型技术实现投研全流程自动化（数据采集→总结分析→报告生成）和展业场景智能化（线索挖掘→精准匹配→跟踪执行），缩短业务响应周期，提升员工效率与客户服务精准度，打造行业领先的智能化服务体系。



机构 AI 中心业务场景赋能目标图

二、创新点

1. 技术创新

融合私有化大模型、RAG 增强检索、AI Agent、多模态内容生成技术，通过动态更新知识库与大模型多模态解析（文本、图表、语音）技术，降低模型分析幻觉，解决金融领域知识时效性与合规性难题，提升展业效率与服务专业性。

2. 展业模式创新

利用 AI 大模型精准分析收购方与求购方企业财务数据、行业要求、企业资质等关键信息，智能匹配并购重组潜在标的；实时监控上市公司股份回购、股东变动等公告，通过大模型提取回购目的、实施进度等关键信息，生成机构业务展业服务线索，辅助一线展业人员提前介入客户需求，掌握先发优势，提升工作效率及展业成功率，为业务侧提供新的增长点。



3. 业务交互创新

将AI大模型能力嵌入机构CRM、投研中台等业务系统，通过自然语言交互（多轮问答）简化操作，降低学习门槛，方便一线员工使用，提升系统利用率。

4. 知识管理创新

构建结构化知识库与会议摘要提炼能力，支持跨场景、跨部门的知识复用，提升研究人员工作效率和非结构化数据复用效率。

5. 安全创新

采用私有化部署大模型，数据交互严格限制于公司办公网域内，结合全程数据留痕与敏感信息脱敏技术，实现零数据泄露风险，确保金融数据安全可控。

ID 三、项目技术方案

1. 技术架构

基于微服务理解设计，分为四层架构——基础能力层（公司大模型MaaS平台）、组件能力层（领域小模型、向量数据库）、应用服务层（意图识别、多轮对话、内容抽取、函数调用）、下游接入层（API对接业务系统）。

2. 业务模式

通过API为投研中台、机构CRM等系统提供智能化服务，支持基金投研助手、研报智能问答、主题新闻每日一图、研究所策略会议助手、研报撰写助手、股权业务线索挖掘、并购重组机会挖掘、上市公司监管案例分析、内容萃取通用能力九大场景。

3. 商业模式

聚焦内部赋能，降本增效，间接提升客户服务能力与商机转化率，形成差异化竞争力。

ID 四、运营情况

1. 基金投研助手

该功能模块面向基金产品分析与客户服务场景，一线投研业务人员可以通过自然语言对话交互实现基金信息快速查询与多维度对比，同时投研助手还具备自动生成结构化分析结论的能力。基于大模型语义理解能力解析模糊需求，结合净值、评级、持仓等多源数据生成可视化报告，实现一线员工从查询到输出的全流程自动化，降低操作复杂度并提升服务专业性，提升投研服务工作效率。

2. 研报智能问答

该模块基于私有化部署的大模型与 RAG 增强检索技术，实现研报语义解析、产业关联推理及图表混合内容理解。针对研究人员高频查询场景（如“某行业政策影响分析”），通过知识库精准匹配历史研报核心观点与数据结论，支持跨文档关键信息提取与逻辑串联，输出可追溯的答案来源，解决传统检索模式信息碎片化问题，强化研究结论的连贯性与可信度，提升相关研究工作的效率。

3. 主题新闻每日一图

该模块服务于一线研究人员，整合了监管动态、行业资讯与市场舆情等多源非结构化数据，通过大模型自动解析政策核心要点、产业链影响范围及关联标的，结合预设可视化模板生成包含数据图表、趋势曲线与摘要文本的日报图文。每日一图功能支持按行业标签体系分类呈现，帮助研究人员快速定位关键信息，减少人工筛选与图表制作等重复性工作，可将主题新闻每日一图制作耗时由小时级缩短至分钟级，提升研究人员工作效率。

4. 并购重组线索挖掘

该助手通过整合买方需求标签（如“行业偏好”“估值区间”等）与卖方企业画像（如“主营业务”“盈利能力”等），利用大模型能力构建多维度匹配模型，智能匹配潜在并购标的，同时生成业务线索及买方卖方匹配度。展业人员可根据匹配度分级，自上而下开展工作，提升工作效率、撮合成功率。

5. 上市公司监管案例分析

该功能模块自动抓取官方监管案例，通过大语言模型解析违规类型、法规依据、处罚措施等要素并生成结构化标签，构建多维查询系统和监管知识库。支持用户按证券代码、处罚日期等条件筛选，并提供数据导出与批量下载功能。大幅节省人工整理监管案例的时间成本，通过服务内部员工，提供专业的培训材料和案例库，进而可服务上市公司充分了解行业标准和竞争对手的合规状况，为上市公司提供警示预警并辅助其提升自身的合规管理水平。

6. 项目运营推广

在机构 AI 中心建设完成之后，研发部室与相关业务部室协同合作，积极推进用户培训工作，确保系统正式投入运营前，用户能够全面掌握系统功能及其显著优势。培训工作参考运营推广服务小组制定的工作方案，具体措施包括：按分支机构组织培训研讨会、在用户服务群中分享应用案例、通过 E-learning 平台发布在线培训视频、编撰用户使用手册等。在培训实施过程中，密切监控用户反馈，以便及时优化培训内容和方式。此外，在推广培训阶段，设立专门的服务群组或沟通渠道，为用户提供辅导和问题解答服务，并通过定期问卷调查和电话回访等方式，积极收集用户的宝贵意见和建议。

五、项目成效

1. 业务价值效益

机构 AI 中心上线后，显著提升赋能中小企业综合金融服务能力。通过大模型生成股权业务线索及并购重组推介，



按照当前业务开展情况测算，预计年度可实现近百条展业机会推送，有效提升一线员工展业成功率，为业务带来新的增长点。

2. 时间价值效益

机构 AI 中心上线前，研报阅读与分析、电话策略会议复盘回顾、“主题新闻每日一图”素材物料生产、上市公司监管案例搜集整理等工作大约需要耗费数小时工作量。机构 AI 中心上线后，可在分钟级别完成信息检索、总结分析、内容生产等工作，降低 60% 的基础性、重复性工作耗时，大幅提升工作效率。

3. 战略效益

通过建设机构 AI 中心，实现大模型在机构业务各场景的应用实践，提升了机构业务相关系统的智能化水平。助力公司沉淀大模型应用落地的实践路径并形成方法论指导，有助于催生更多大模型场景应用与产品创新。

六、经验总结

1. 建设经验

以“解决一线痛点”为出发点，技术团队与业务部门紧密合作共建需求清单，确保功能实用性与易用性。

2. 推广经验

通过小范围试点快速验证价值，降低全面推广风险。上线后通过分阶段培训与深度调研相结合的方式，确保用户能够全面掌握系统功能及其显著优势，并持续提供反馈意见优化迭代相关产品。

3. 改进方向

未来将探索与外部数据源的合规对接，扩充数据来源与知识规模，进一步提升大模型结果的时效性与专业性。同时加强跨机构技术协作，共建行业级大模型应用实践标准。

中原银行：大模型信贷助手

一、项目背景及目标

当前，大模型正掀起新一轮智能化发展热潮，赋能千行百业。大模型具备优秀的理解、学习、生成和推理能力，其工程化应用包含数据构建、模型算法、模型训练、模型压缩与加速、模型评测、模型运营和安全可信多个环节。大语言模型凭借其强大的语言理解、文本生成、知识储备以及多轮对话等能力，在知识问答、文案生成、代码生成等方面表现出跨越式提高，大幅提升了生产效率。

国内外金融机构和科技企业正积极探索大模型在金融业的合理应用，已试点应用于智能客服、智能办公、智能研发、智能投研等多个业务场景，进一步推动金融服务的智慧再造，加速 AI 技术赋能金融业务提质增效。

为提高银行风控智能化水平，提升信贷业务处理质效，完善信贷报告分析维度，中原银行基于行内实际业务需求，启动大模型信贷助手项目建设。本项目建设范围主要包括系统管理、基础设置、智能写作、智能问答、数据整合、场景开发等 6 大部分。项目投产后，大幅提升了信贷报告撰写质效，实现了业务数据智能分析、智能核验，促进业务知识有效沉淀，丰富了银行业大模型应用实践，提高了中原银行信贷业务智能化水平。

二、项目方案

基于信贷业务痛点现状，本项目以业务需求为导向，构建具备报告内容智能生成、辅助审批、智能决策等能力的信贷助手系统。本项目一期主要完成公司信贷调查报告生成能力的构建，二期渐次推进辅助审核、智能决策等能力的构建。

1. 报告智能生成

报告智能生成模块是大模型信贷助手系统的核心组成部分，利用大模型的自然语言理解和生成能力，对原始文本进行内容润色和智能改写，以提升报告的专业性和可读性。此外，段落生成功能可以在用户需要的情况下自动生成相关段落，辅助完成报告；素材参考和模板填充功能则为用户提供丰富的资料和基于预设模板的智能填充，加速报告的撰写过程；内容抽取和数据溯源功能可以自动从大量数据中提取关键信息，并确保报告中使用的数据来源可靠；在线编辑功能则让用户能够直接在系统中对报告进行编辑和修改，方便快捷。

2. 辅助审批

辅助审批通过财报校验、数据核验、引用检索和合规审查等功能，强化报告的合规性和数据的准确性。财报校验和数据核验功能可以自动核对财务报告中的数据，确保其准确无误；引用检索功能检查报告中引用的信息是否来自可靠来源，是否需要进一步的授权或许可；合规审查功能则确保报告符合相关的法律法规和行业标准，减少合规风险。

3. 智能决策

智能决策利用财报分析、风险评估、关联分析和授信评估等工具，对信贷申请进行全面的智能分析。财报分



析功能对财务报告进行深入分析，识别公司的财务状况和趋势；风险评估功能评估信贷申请者的信用风险，为授信决策提供支持；关联分析功能分析信贷申请者与其他实体的关联性，识别潜在的关联风险；授信评估功能则综合考虑各种因素，对信贷申请进行评估，给出授信建议。

随着本项目首期工程的完成，项目的整体愿景、架构设计和执行策略已经确立，为多项工作的并行推进和协同作业提供了坚实的基础。项目依照“分段执行，逐步显现成效”的原则，分阶段展开大模型信贷助手项目建设，每个阶段都依照“需求明确、功能开发、试点运行、推广运营、功能优化”这五步建设流程闭环迭代进行。

ID 三、创新点

大模型信贷助手系统采用先进的大语言模型技术，基于大模型强大的自然语言理解、生成和推理能力，结合 RAG 技术，构建了大模型信贷助手系统。针对行内报告智能生成、辅助审批、智能决策等实际应用场景构建智能应用，一期主要完成了报告内容智能生成能力的构建，实现客观数据自动引入和主观分析智能生成，赋能银行信贷业务提质增效，辅助业务决策，助力知识沉淀。

ID 四、运营情况

目前系统已完成所有技术开发工作，并完成内部性能、功能等测试，各模块上线有序进行。

ID 五、项目成效

在经济效益方面，本项目是行内降本增效的重要举措，大幅提升了行内尽职调查效能和业务办理质效，提高了客户的满意度和一线员工的幸福感。运用多种智能化技术实现多源数据的自动解析、接入及智能分析生成，增强了公司信贷业务线上化和智能化水平，极大减轻了客户经理资料搜集、信息整理、材料撰写工作量，实现调查报告初稿的智能化撰写比例超过 70%。

在社会效益方面，本项目为中原银行积累了大模型的金融场景应用经验，也为银行同业在大模型技术探索与应用方面提供了有益参考。后续将通过积极参与同业交流、标准制定等方式输出中原经验，与同业及相关科技企业一起推动大模型赋能千行百业，促进金融数智化转型走深向实。

在金融科技方面，本项目是丰富金融科技赋能业务运营优化场景的一次有益探索，实现了“行内信贷数据资产 + 外部数据资产 + 客户离线数据 - 尽调报告”的有效沉淀，形成了基于数据驱动的智能尽调闭环，构建了中原银行信贷尽调新范式。同时，为包括信贷助手在内的大模型应用的行内规模化推广奠定了基础。

ID 六、经验总结

通过大模型信贷助手系统的建设，一期率先实现客户尽职调查报告的智能撰写生成，完成了对内外部多源客户数据的自动填充、智能分析，并支持生成内容的定点更新、二次编辑、下载等。该创新型项目在建设方面需重点考虑高频场景和功能实用性，适宜采用敏捷开发模式快速迭代；在推广方面强调影响范围适度性、需求适应性、数据合规性。这些经验对于类似金融科技项目的成功实施具有重要的参考价值。

中国大地保险：数字审计员项目

一、项目背景及目标

随着数字化转型的深入推进，审计行业面临数据体量激增、风险复杂性加剧的挑战，对审计效率和质量提出更高要求。传统非现场审计高度依赖人工抽检，存在效率低、覆盖不足的问题。为应对这些挑战，中国大地保险于2024年启动“数字审计员项目”，旨在借助技术手段实现审计流程的自动化与智能化转型。该项目融合生成式大模型、机器学习方法和RPA等技术，重构审计流程，实现高效、精准的审计问题识别与管理，推动审计职能向前延伸，提升整体审计监督能力。

二、创新点

“数字审计员项目”在多个维度上实现了创新突破。项目引入生成式大模型并通过人类反馈强化学习（RLHF）进行微调，有效实现了审计问题的自动分析与整改意见的辅助生成。同时，将机器学习技术引入审计经营分析，辅助审计人员对数据进行健康何判断。又将RPA技术全面引入非现场审计流程，首次实现了审计风险点的全流程自动化排查。自主构建的审计数据集市及标准化审计问题知识库，使得机器学习模型在数据挖掘与风险识别中发挥更大效能。同时，项目还推动了“人机协作”模式建设，释放审计人力用于更高价值工作。值得强调的是，项目整体架构及关键技术路线，包括大模型的微调工作均由公司自主设计开发，确保合规、安全、可控。

三、项目技术方案

本项目构建了集数据整合、风险识别、报告生成于一体的智能审计平台。在数据层面，项目搭建了覆盖多业务系统的统一审计数据集市，整合承保、理赔、反洗钱等核心数据。在模型层面，运用统计学习算法与AI模型，提升风险判断的准确性与效率，辅助给出一些分析何建议。在工具层面，RPA负责样本提取、数据对比与底稿整理等操作，全面覆盖高频、规范流程。在文稿生成方面，基于标准化知识库与生成式大模型，自动完成审计方案、征求意见稿及整改建议的撰写工作。平台整体采用模块化架构设计，形成数据采集、分析、报告输出的闭环流程，支撑审计工作高效开展。

四、运营情况

截至目前，“数字审计员”已在三级机构的高管任中 / 离任审计中全面投入使用，常规项目的自动化审计流程已成为标准工作模式。系统实现了智能外呼、自动生成审计方案与意见稿等功能，运行稳定，反馈良好。项目落地后，审计工作效率显著提升，审计覆盖率大幅提高，有效支撑了高质量审计工作的开展。同时，审计人力资源投入减少30%-40%，进一步释放了审计队伍的能力潜能。

五、项目成效

“数字审计员”显著提升了审计工作的经济效益与社会效益。在成本方面，通过自动化处理常规事务性工作，显著减少了人工重复劳动，实现人力资源的有效节约。到目前为止，已协助完成百余审计项目，合计处理方案、底稿四百余项，平均相当于增加5到6名初、中级员工。在问题防控方面，上线当年即实现追回违规问题资金近20万元。基于大模型微调技术的问题分析及整改意见功能，削减了90%的硬件开销，同时生成准确率达86%，已达到初级



审计员水平，既降低了使用成本，又提升了问题处置效率和质量。最后，通过机器学习技术开发的监控模型，构建了更加稳健的防控体系。项目的成功运行不仅推动了公司内部管理提质增效，也为保险行业的数智化转型提供了案例。

ID 六、经验总结

“数字审计员”项目的成功建设离不开技术与业务的深度融合。审计人员需具备数据思维与智能工具的操作能力，才能实现审计模式的升级。此外，审计规则与问题知识的标准化建设为系统自动化提供了基础支撑，需持续迭代与完善。项目建设过程中同步强化了数据安全与合规管理，确保了数据的安全使用与传输。采用模块化开发、分阶段落地的推进方式，有效降低了替换风险，提升了项目稳定性与可维护性。“人机协作”的审计模式有效缓解了人力资源压力，为未来审计工作的智能化演进积累了宝贵经验，具备广泛的行业推广价值。

山能财务有限公司: 基于 DeepSeek 构建 财务公司智慧金融大模型

一、项目背景及目标

1. 需求分析

当前，金融行业正面临深刻的数字化转型挑战：一方面，客户需求日益多样化，传统人工服务模式难以满足日均 5000+ 次的业务咨询量；另一方面，监管趋严要求金融机构提升风险识别与合规审查能力。以煤炭行业为例，能源集团内部资金流动复杂度高，关联交易频繁，传统的贷前审核依赖人工经验判断，风险滞后性问题突出。此外，业务文档编制占员工总工时的 35%，成为效率提升的瓶颈。在此背景下，财务公司亟需通过智能化升级实现降本增效与风险管控的双重目标。

2. 项目目标

基础目标：构建覆盖金融领域 90% 专业知识的专用大模型，集成 15 万条政策法规、内部制度及业务案例，形成可动态更新的知识中枢；

效能目标：

智能问答系统响应时效从小时级压缩至 30 秒内，年处理能力达 200 万次咨询；

贷前审核效率提升 60%，单笔贷款审批时间从 3 天缩短至 4 小时；

智能写作工具实现 12 类标准文档自动生成，文档一次通过率达 92%；

质量目标：专业问题回答准确率 >95%，文档合规率 >98%，AI 决策过程可追溯。

二、创新点

1. “三位一体”智能服务模式

知识获取：首创 AI 自动推送关联政策条款功能，替代传统人工查阅模式。例如，当用户询问“2025 年新版资金管理法”时，系统自动关联集团制度、监管要求及 8935 例历史案例，生成结构化解读；

风险决策：采用“模型量化评分 + 人工复核”双轨机制，通过图计算引擎识别担保圈、实际控制人网络等隐蔽风险，风险识别准确率提升 25%；

文档生产：基于 200+ 金融模板库，实现信贷审批报告、财务分析报告等文档的智能生成与迭代优化，人工修订耗时减少 60%。

2. 动态政策适配引擎

对接银保监会、能源局等监管平台，新政策发布后 24 小时内完成解读生成，并自动筛查受影响业务清单。例如，2024 年某项新能源补贴政策出台后，系统 48 小时内完成对 237 笔存量贷款的风险重估。



3. 混合推理架构

结合规则引擎与 DeepSeek 大模型，实现复杂问题的分层处理。简单问题（如“贷款利率计算”）由规则引擎即时响应，复杂问题（如“跨境并购融资风险”）则触发多模型融合分析，兼顾效率与专业性。

4. 私有化知识增强

构建包含 1200+ 集团制度文件、8500+ 行业政策法规的私有化知识库，通过向量数据库实现毫秒级检索，确保数据安全合规。

三、项目技术方案

1. 技术架构

采用“1+1+N”分层架构：

基座模型层：基于 DeepSeek-32B 金融领域微调，通过 LoRA 技术仅训练 0.1% 参数即达到专业领域最优效果，硬件成本降低 80%；

知识中枢层：

构建金融知识图谱，包含实体节点 15 万个、关系边 32 万条；

集成 RAG 检索增强生成架构，支持多模态数据（文本、表格、图表）的深度理解；

应用场景层：

智能问答：采用多轮对话与意图识别技术，支持复杂问题拆解（如“根据 2025 年集团新版资金管理法，子公司资金应如何归集？”）；

贷前审核：融合 DeepSeek（非结构化文本分析）与 XGBoost（结构化数据分析），结合图计算引擎识别关联风险；

智能写作：预设 200+ 金融文档模板，支持动态段落组合与智能校对（如勾稽关系检查）。

2. 关键技术突破

动态政策适配引擎：通过 BERT 模型提取政策关键词，结合知识图谱实现影响范围自动推理；

混合精度训练：在保证模型精度的同时，将训练时间缩短 40%；

安全沙箱机制：建立模型决策的可信执行环境，确保数据不出域。

四、运营情况

系统覆盖：智能问答日均处理量达 2000 次，峰值并发量 100+；

知识库规模：累计收录 200 余份内部制度文件、7452 条行业政策、8935 例业务案例，月更新频率 15%；

应用效果：

贷前审核模型累计处理 1200 笔贷款申请，拦截高风险申请 47 笔，挽回潜在损失 3.6 亿元；

智能写作工具生成文档 200 余份，其中资金计划报告、信贷审批书等核心文档占比 85%。

五、项目成效

1. 经济效益

直接效益：年节省人力成本 120 万元（按 5 人团队计算），贷款审批周期缩短 86%，资金周转效率提升 40%；
间接效益：通过精准风险识别，不良贷款率下降 1.2 个百分点，年增利息收入超 2000 万元。

2. 社会效益

推动财务管理数字化转型，形成可复用的金融知识资产；
培养复合型数字化人才超过 10 名，其中 5 人获行业资格认证；
接待各类参观调研 5 批次。

六、经验总结

业务与技术深度融合

从高频场景（如贷前审核）切入，确保技术落地实效；
建立“业务需求池”机制，动态调整优先级；

创新技术架构

混合推理架构平衡效率与专业性，简单问题响应速度提升 90%；
私有化部署保障数据安全，通过等保三级认证。

持续优化机制

建立“AI 初筛 - 人工复核 - 模型迭代”闭环，每月更新模型参数；
开发 AI 效能看板，实时监控问答准确率、文档生成效率等 20+ 指标。

七、应用前景与推广价值

1. 技术延展性

可快速复制至煤炭、电力等能源细分领域，适配供应链金融、碳资产管理等新兴场景；
支持多语言版本开发（如英文版跨境业务支持）。

2. 推广价值

已与省内多家财务公司达成合作意向，预计带动年间接经济效益超亿元；
形成《能源行业 AI 治理白皮书》，提出金融大模型合规评估标准。



平安信托: DeepTrust 大模型平台及应用项目

一、项目背景及目标

近年来,以大型语言模型(LLM)为代表的AI技术发展迅猛,其强大的泛化能力和创造力正深刻变革各行各业的运行模式。在金融领域,智能化应用潜力巨大,能够有效提升服务效率、优化客户体验、提升运营效率并强化风险管理。与此同时,技术栈日益成熟、开源生态活跃与部署成本优化,显著降低了AI模型的开发与应用门槛,为金融科技的创新与规模化落地创造了有利条件。

平安信托秉持创新驱动发展的理念,积极把握这一数字化机遇,致力于通过深度应用AI等前沿技术,构建更智能、高效、安全的金融服务体系,巩固并提升核心竞争力。本项目旨在整合并高效输出先进AI技术能力,构建一站式大模型应用平台。支持各AI应用场景快速落地,助力公司实现数字化升级与核心竞争力提升。项目目标:

(1) 加速场景落地与创新:为公司各类业务需求提供快速构建、部署和迭代AI应用的标准化能力,缩短从想法到应用的周期。

(2) 深化核心业务赋能:覆盖营销、客服、运营、管理、经营等全业务环节,通过科技赋能,显著提升效率、优化体验、强化风控,构筑差异化竞争优势。

(3) 驱动数字化转型纵深发展:以平台为载体,规模化推进全业务线的智能化升级与模式创新,打造可持续核心竞争壁垒。

二、创新点

该项目创新点主要体现在“一体化集成”、“一站式实施”、“全方位应用”三个方面:

(1) 一体化集成:构建一体化AI智能底座,实现组合式AI架构设计,深度融合大小模型及Agent技术,并搭载轻量级全流程工具链,有效降低智能体构建门槛。在插件生态层面,平台集成语音识别、OCR、NLP等功能插件,支持多模态交互场景及低代码功能组装能力,满足多样化业务需求。同时提供开箱即用的RAG框架,支持多模态数据解析、索引构建及溯源召回,实现知识库内容的动态更新与高效利用。

(2) 一站式实施:依托全链路一体化架构实现一站式实施,通过可视化流程编排与标准化工具链,系统性解决AI落地复杂度问题,提升开发效率,加速业务场景落地。针对业务集成场景,支持API/H5等标准模式与企业现有系统无缝对接,通过低代码配置能力与轻量化分发机制,快速完成应用部署与功能适配。支持灰度发布验证,覆盖从智能体开发到跨终端部署的全流程。

(3) 全方位应用:基于DeepTrust大模型平台提供的能力,平安信托实现了多场景应用落地,推动业务智能化升级。

数字化服务提升服务便捷性与安全性:通过证照信息智能识别、人脸识别核身、AI面签等技术手段,为客户提供更加便捷、安全的服务体验,同时确保信息的真实性与合规性。

数字化运营提升运营效率:通过RPA和NLP技术打造的运营助手,实现了网银流水回单归集、外部产品单位净值解析、券商对账单解析等,大幅降低运营成本,提升运营效率。

数字化管理强化风险管理能力:打造公司合规知识库,实现智能合规问答,为业务提供实时、准确的合规指导;使用AI技术实现合同差异的自动识别与处理,进行文本纠错和敏感词审核,显著提升了合同审查的效率和准确性;搭建了风险智能引擎,通过智能预警模型实时识别潜在风险并发出预警,由专业团队跟进处理,形成闭环管理。

三、项目技术方案

项目技术方案主要从下面几方面进行设计和实施：

1. 打造一体化 AI 平台：平台实现开发协作、模型训推、运营运维的一体化，通过可视化监控看板实时追踪模型性能与资源利用情况。平台采用模块化 AI 架构设计，融合大小模型与 Agent 技术，搭载轻量化工具链降低智能体开发门槛；插件生态集成语音识别、OCR、NLP 等功能模块，支持多模态交互与低代码功能组装；内置 RAG 框架实现多模态数据解析、智能索引构建及动态知识库更新，满足多样化业务场景需求。

2. 构建可视化 Agent 编排能力：通过可视化拖拽界面实现智能体触发条件、对话流程、工具调用策略的图形化设计，集成细粒度权限管理体系，支持基于角色的访问控制与功能模块分级授权。配套低代码应用配置工具，可快速完成表单设计、状态机管理、多轮对话逻辑定制，无缝对接业务系统。

3. 支持多维度发布渠道管理：提供多渠道发布管理控制台，支持 API 接口调用、Web 端（PC/H5 自适应页面）、移动端 SDK 等多形态输出，集成一键打包、灰度发布、版本回滚等自动化部署工具，配套渠道流量监控，实现智能体应用的快速分发。

4. 构筑高水平信息安全体系：平台对客户敏感信息实施加密存储和传输。通过严格的数据隔离机制、访问控制和审计日志等多重安全管控措施，为平台的安全稳定运行提供有力保障。

5. 通过 AI 平台能力，结合业务流程实现在多个场景落地。

(1) 数字化服务：运用证照信息智能识别、人脸识别核身、AI 面签等技术手段实现便捷服务，不仅提高了服务效率，还确保了信息的真实性、合规性和安全性。以家族信托为例，我们为客户提供了 AI 面签，双录，保单录入助手，资料智能识别等工具大幅提升客户设立的便捷性。我们还搭建了一套智能化的流动性测算管理管控工具，能够精准测算未来现金流入流出情况，提升现金头寸安排的合理性、资产兑现安排有效性、以及资产收益预测准确性。通过这些智能化手段的综合运用，平安信托在提升服务效率的同时，也确保了服务的高质量和客户满意度。

(2) 数字化运营：借助 RPA、NLP 等技术，打造了智能化的运营助手，实现了外部信息的自动下载、解析和使用。例如，通过 RPA 自动归集各类文档，进行智能识别，包括：网银流水回单、单位净值行情、券商对账单、投资建议等，大幅提升了运营效率和数据处理的精准度。通过 OCR 技术，实现合同与影印件扫描件的比对，自动完成有效性验证，大幅提升人工核查效率。同时，基于已有文档知识、问题处理过程建设了运营专有作业知识库，打造运营作业问答助手，辅助日常作业中同类问题的快速响应。

(3) 数字化管理：智能文档助手运用 OCR 技术完成电子合同高效比对，单份审核效率提升 65%，可视化呈现版本差异；通过模型进行智能合同审核，纠错与敏感词筛查效率提高 20%，并分析差异识别法律风险。合规助手整合政策法规、监管处罚及内部制度，通过大模型 RAG 技术，构建合规知识库，提供全天候智能问答与指导，赋能业务合规开展。智能预警借助大模型与知识图谱技术，强化舆情监控及交易对手风险预测，提升预警精准度与闭环管理能力。

四、运营情况

项目于 2023 年底开始分批持续交付功能，并开始逐步推广应用，截止目前，大模型平台已上线，在数字化服务，数字化运营，数字化管理，等方面有较好的应用效果。在数字化营销和数字化经营方面，也有一些积极的探索。系统自首批功能上线以来，已经持续运行约 2 年多，支持 7*24 小时无间断服务，运行稳定未出现因系统异常的中断情况，系统稳定率达 99.99%。



ID 五、项目成效

DeepTrust 大模型平台为公司业务的全面数字化提供了强有力的技术支持，成效显著。

(1) 业务发展：在 AI 技术的加持下，通过数字化赋能，保险金信托业务快速增长，总规模接近 1800 亿，市场总份额超过 50%，遥遥领先。

(2) 降本增效：通过 AI 技术落地运营场景，实现替代人力手工工作约 300+ 人月 / 年，节约成本约 1000+ 万 / 年。

(3) 风险管控：通过 AI 技术强化风控防线，实现全业务品种 19 类交易对手的风险智能预警，早介入早防范。通过智能审核，辅助合同纠纷与敏感词筛查，降低合规风险。

(4) 创新示范：构建一体化的大模型平台并基于业务场景应用实践，形成可复制的数字化解决方案，为行业智能化转型提供实践样本。

ID 六、经验总结

通过对项目的跟踪、复盘和总结，有如下 4 点经验值得借鉴：

(1) 以实效彰显价值，聚焦场景落地：AI 项目需将技术转化为直观可见的成果，通过精准匹配业务场景实现落地。优先选取痛点突出、见效快的场景，快速形成可感知的应用成果，让用户切实体验技术带来的便利与效益，以此建立信任基础，为后续深化应用奠定良好开端。

(2) 坚持整体规划，夯实平台支撑：做好顶层设计，明确 AI 项目的整体目标与实施路径。搭建灵活高效的 AI 平台，发挥其资源整合与能力复用优势，缩短应用开发周期，提升落地效率。确保平台具备良好的扩展性与适应性，以应对业务需求的动态变化。

(3) 融入业务流程，推动应用推广：将 AI 应用深度嵌入现有业务流程，实现无缝衔接，降低用户使用门槛与学习成本。通过流程优化与智能化改造，提升业务处理效率与质量，使 AI 应用自然成为业务运转的有机组成部分，从而提高用户接受度，促进技术的广泛应用与持续迭代。

(4) 遵循渐进发展，稳步替代人工：AI 发展是长期演进过程，初期应定位为人工辅助工具，与人工操作协同运行，在实践中验证技术稳定性与可靠性。待应用效果稳定、性能达标后，再逐步减少人工干预，实现从辅助到主导的平稳过渡，保障业务连续性与优化效果。

平安信托将持续以创新驱动发展，积极拥抱数字化浪潮，通过前沿技术与业务深度融合，为客户提供更优质、更可靠的金融服务体验，推动行业高质量发展。

北银金科: 北京银行水晶球 ChatBI

一、项目背景及目标

在竞争激烈的市场环境中，银行对于能够快速、直观地获取数据洞察的需求日益增长。传统的 BI 工具操作复杂且仅提供数据可视化输出，难以满足企业对于一站式数据分析的需求。具体表现为：

1. 业务响应不及时：传统数据分析工具往往需要经历“分析师提出需求 - IT 人员开发投产 - 表哥表姐制作报表 - 决策者决策”的冗长流程，导致企业无法迅速响应市场变化，错失决策良机。

2. 数据洞察不全面：数据探索受限于预设报表，深度分析依赖个人经验，难以实现全面、深入的业务洞察。

3. 战略传导不通畅：总部决策层战略意图到分支机构、一线人员的传导不畅，难以上下对齐，指挥效能层层衰减。

随着大数据技术的不断发展，数据的收集、存储和处理能力得到了显著提升，而人工智能技术的飞速发展，特别是大模型技术的突破，为水晶球 ChatBI 的产生提供了技术基础。使得计算机能够理解和生成自然语言，从而实现与用户的自然交互。

而北京银行水晶球 ChatBI 融合大数据技术和人工智能技术，通过对话式的交互方式，能够迅速生成分析结论、进行根因分析并给出决策建议，并将结果通过图表、文字、语音等形式自适应的直观展现，从而满足了企业对数据洞察的迫切需求。

二、创新点

在技术架构创新方面，水晶球 ChatBI 采用大小模型协同技术，让不同规模和功能的模型相互配合，能够发挥各自的优势，在复杂任务处理中，大模型可以负责全局性的分析和规划，而小模型则专注于专业性、局部性的执行和监控，从而实现更高效、更精准的任务处理。

在计算性能方面，水晶球 ChatBI 不光支持传统 MPP 的数据架构，还在数据层引入多维大数据计算引擎，通过行列混合计算与内存计算优化的方式满足维度、指标分析的数据实时处理高性能，其亿级数据量可实现秒级响应。同时通过自研动态解析的语义映射，保障指标分析的灵活性，其支持动态计算指标维值过滤、时序对比、指标间加减乘除、复杂函数，并实现一次定义全局使用。

在智能分析深度方面，水晶球 ChatBI 搭载了智能归因分析能力，通过集成蒙特卡洛树搜索（Monte Carlo Tree Search，简称 MCTS）。MCTS 通过模拟随机策略来探索可能的决策路径，并利用统计方法评估各个节点的潜力，从而逐步构建和优化搜索树，在解决状态空间巨大的多维度、多因素的归因分析领域中实现了效率的代际式提升。这种算法的应用，使得智慧水晶球能够更准确地识别影响业务的关键因素，为决策者提供有力的支持。

交互体验创新方面，水晶球 ChatBI 引入自然语音对话式交互，融合领域知识（包含 5 亿 + 结构化知识，3 万篇非结构化知识）实现 93% 的意图识别准确率，并支持图表 + 文本 + 语音的多模态交互，覆盖各类应用场景，为用户提供了更加便捷、高效的交互体验。

水晶球 ChatBI 使用主流先进的应用框架，其核心设计理念在于通过分层解耦实现系统的模块化和灵活性，同时依托数据驱动机制确保从原始数据到指标分析全链路贯通，支持实时与批量混合计算。该架构强调智能融合技术，AI 能力与业务系统深度融合，以洞察业务深层规律。此外，其弹性扩展能力能够灵活应对业务增长和变化，确保系统始终保持在最优运行状态。这一系列技术设计共同作用于业务架构，高效支撑并优化各类复杂场景需求，推动业务创新与持续发展。水晶球 ChatBI 核心技术创新点：



1. 大模型 + 小模型技术

旨在通过不同规模和功能的模型相互配合，发挥各自的优势，从而实现更高效、更精准的任务处理。系统集成基础大模型，结合金融领域知识图谱与强化学习能力，构建垂直金融大模型基座，通过自然语言处理（NLP）实现对话交互式检索数据。基于大模型的推理能力，自动识别业务现象与核心变量的潜在关联，生成可解释性强的动态指标模型。同时系统搭载动态指标解析、蒙特卡洛树搜索算法、时间序列算法、实体识别模型等多种小模型，实现指标解析、快速最大路径归因分析、预测等特殊任务。

2. 智能指标引擎

平台采用了自主研发的指标计算引擎和指标展示组件，实现了指标的实时计算和快速展示。实现指标的智能建模、自动加工计算，大幅降低指标数据开发的工作量，加快指标开发和上线的效率；用户可基于原子指标实现衍生/派生指标的从创建、试算、发布、检索、维护、下线等全生命周期一站式管理。

3. 统一语义层

基于集市模型实现上下口径统一的数据语义层，基于维度实现指标的穿透式分析。通过这一层实现指标的一次定义、多处使用。统一语义层为大模型的意图识别、拖拉拽式数据可视化，业务 workflow 和高级分析操作等应用提供坚实保障。

4. 多模态技术

系统可根据用户需求通过多模态的方式输出可视化图表、文字、语音满足用户需求，实现数据与信息融合，并通过丰富的人机交互形式，使得用户可以获得更加自然和人性化的体验。

5. 项目技术方案

水晶球 ChatBI 项目是金融业“数智化”转型的里程碑实践，通过技术创新与业务赋能的深度耦合，推动银行从“经营驱动决策”向“数据驱动决策”跃迁，从“被动识别问题”向“主动归因问题”跃迁，从“手动督办管理”向“精细化全流程敏捷管理”跃升，从“全员表哥表姐”向“全行数据分析师”跃迁，为行业树立了“数据、智能、敏捷、普惠”的数字化标杆。

水晶球 ChatBI 使用主流先进的应用框架，其核心设计理念在于通过分层解耦实现系统的模块化和灵活性，同时依托数据驱动机制确保从原始数据到指标分析全链路贯通，支持实时与批量混合计算。该架构强调智能融合技术，AI 能力与业务系统深度融合，以洞察业务深层规律。此外，其弹性扩展能力能够灵活应对业务增长和变化，确保系统始终保持在最优运行状态。这一系列技术设计共同作用于业务架构，高效支撑并优化各类复杂场景需求，推动业务创新与持续发展。

系统采用分布式微服务架构与前后端分离设计，对服务进行隔离分层，通过容器化编排与 API 网关实现业务逻辑层与基础设施层的解耦。技术架构深度集成 AI 技术、实时指标计算引擎及动态可扩展服务框架，构建起支持智能归因分析、决策推演的数字化体系，为用户提供全场景智慧经营分析，支撑从数据洞察到精准决策的业务闭环。

数据层，实现数据一体化管理。结构化、非结构化数据的统一贯标、统一处理、统一调度、统一存储，并实现数据质量的持续监控。

指标能力层，实现指标管理与计算，支撑业务分析。多数据源插拔式接入，通过多维预计算技术实现大数据量

的数据高性能实时计算，通过语义解析和血缘管理，确保数据自动识别性和可信度，并为上层 AI 和可视化提供高质量数据基础，搭载的 20 余中 BI 控件，为用户提供最直观、最合适的数据呈现。

AI 能力层，为业务提供智能化支持。整合多种 AI 技术，支撑各场景的智能决策、指标预测和个性化服务。

应用层，基于容器化技术实现金融应用的快速部署、弹性扩缩容及资源隔离，支持高可用和微服务架构。

可视化层，实现用户界面渲染、数据可视化及业务逻辑处理，为用户提供直观的操作体验。

水晶球 ChatBI 为用户搭载以下核心能力：

指标引擎语义层：

亮点概述：以指标引擎为核心，将用户输入的自然查询语言进行意图识别，转换成指标引擎可调用的方法指令，再由指标引擎与基础数据库进行交互并返回结果，通过指标引擎的语义解析提升数据查询的准确性与适用场景的广泛性。

具体优势：这一功能极大地提升了查询准确性；可覆盖各类数据分析场景，并可简单的实现多维归因分析；基于指标引擎加速，可实现秒级响应；通过指标引擎语义层识别指标与底层数据关系，使得训练成本较低；指标引擎可内置和隐含权限管理，方便进行权限管控。

自然语言数据查询：

亮点概述：水晶球 ChatBI 支持自然语言数据查询功能，用户可以通过自然语言与系统进行对话，获取所需的数据信息，实现人机交互无缝对接。

具体优势：这种交互方式更加直观和便捷，用户无需学习特定的查询语言或操作界面，只需用自然语言描述自己的需求即可。此外，自然语言查询还支持模糊匹配和智能联想，进一步提高了查询的效率和准确性。同时本方案可实现文字、数据、可视化控件、语音等不同样式的结果以最直观、最恰当的方式自适应输出给用户。

总结与预测：

亮点概述：水晶球 ChatBI 具备强大的数据总结和预测能力，能够根据历史数据自动生成分析报告，并预测未来的数据趋势。

具体优势：这一功能有助于企业及时发现数据中的规律和异常，把握市场趋势和机遇。同时，预测结果还可以为企业制定战略规划和决策提供有力支持。确保用户 10 秒以内洞察数据趋势，把握未来机遇。

AI 归因分析：

亮点概述：AI 归因分析是 ChatBI 的另一大亮点，它能够通过智能模型算法自动分析数据中的因果关系，精准定位问题的根源。

具体优势：这一功能有助于企业深入了解业务运营中的关键因素和影响机制，从而制定更加有效的优化措施和改进方案。

AI 决策建议：

亮点概述：基于大模型与知识库技术结合的决策建议功能，ChatBI 能够根据数据分析结果自动生成决策建议，为企业的运营和管理提供智能辅助。

具体优势：能够为银行提供科学、客观的决策依据，以及切实可行的改进策略，帮助企业提升运营效率和市场竞争力。

四、运营情况

系统自启动建设以来，通过“迭代与推广同步、总分协同共进”的模式，实现了 4 次大版本升级，成功推广十余次，成为企业经营分析的核心平台。其显著成果体现在数据分析及管理效能的大幅提升，将原本线下的零散数据及分析工作系统化，实现了经营分析“零手工”操作，数据获取达到亚秒级，数据分析效率提升 80%。同时，系统深度



固化经营管理意图，在经营策略制定和分析中发挥关键作用。

目前，系统覆盖各层级管理者及分析师，全方位赋能经营管理，推动决策数智化与精细化。此外，系统在行业内积极宣传交流和活动中的专题分享，进一步提升了其影响力。

主要性能指标：

- 同时在线用户数峰值在 200；
- 简单查询、系统功能模块响应时间小于 1s，复杂查询响应时间小于 2.5s；
- 系统登录、打开、刷新页面的响应时间在 1 秒内；
- CPU 占用率小于 60%；
- 内存占用率小于 70%；

AI 模型召回率和精准度大于 80%。

ID 五、项目成效

1. 经济效益

①直接经济效益：降低投入与运营成本

研发费用节约。系统仅需加工少量基础指标及标签，就可“零投入裂变”生成大量指标，满足差异化应用需求，节约研发费用。预计未来研发费用节约潜力巨大。

人力成本节约。主要体现在：

- 系统提供全景数据视图，替代过去人工查询整合数据再制作报表过程，每日节省至少总分支各层级经营分析人员人工劳动。

- 自动化流程：系统覆盖了各类功能点，将大量重复性、标准化的工作自动化，减少了人工干预。
- 错误率降低：系统通过标准化操作和实时校验，显著降低了人工操作中的错误率，减少了纠错成本。

随着系统使用率的提升和功能优化，单位成本将进一步降低。同时系统将不断迭代升级，提升自动化水平和智能化能力，进一步减少人工依赖。

②间接经济效益：效率提升与业务增长

- 重塑传导机制，业绩提速增效。系统优化经营分析体系，重塑管理传导机制，成为全行经营管理分析主阵地，有力推动经营业绩稳步增长，推动业务高质量发展。

- 流程优化升级，工作效能提升。系统提供原子、衍生、派生指标组合，为业务人员提供了全面的数据支持。系统上线以来已成为业务人员日常工作的重要工具，显著提升了工作效率。

- 资源动态调配，成本效益双赢。系统通过实时数据分析和可视化功能，帮助管理者更科学地配置资源，优化业务决策。例如，系统可以精准识别高潜力客户和高收益产品，从而引导资源向高价值领域倾斜，提升整体业务效益。

- 移动渠道接入，实现自由用数。多端渠道接入能够为用户提供更加灵活、多样的访问方式。无论是在 PC 端、移动端还是大屏端上，用户都能够轻松获取所需数据，并进行相应的操作。

2. 社会效益

①培育复合人才与科学管理范式

项目建立“总部大脑-分支协同”的数字化作战机制，通过灵活分析工具与标准化管理模块，培养兼具业务洞

察与数据思维的复合人才。其自动化看板、可视化工具将管理决策效率大幅提升，形成的“数据驱动+人才赋能”管理模式为行业提供了数字化转型样本，带动金融业管理科学化转型。

② 优化资源配置与低碳运营模式

通过对话式交互与拖拉拽看板结合的模式，大幅降低了科技开发成本。通过低碳运营模式的引入，系统实现指标开发成本大幅降低。通过数据资源整合与流程再造，减少重复性系统建设投入，推动金融机构向绿色集约化运营转型。

③ 夯实金融安全与主动风险防控

实时监测关键风险指标，通过指标穿透分析提前识别业务异常，辅助管理层建立动态风险预警机制，防范信用风险、防范客户流失。

该项目通过技术创新与模式重构，在行业标准建设、普惠服务深化、人才生态培育等方面产生溢出效应，为金融业数字化转型贡献了“优秀样本”。

六、经验总结

1. 核心经验

在项目启动与需求分析阶段，本项目成立了专项小组，通过市场调研和结合行内实际需要做出了合理的需求方案，并对项目范围、时间表、预算及关键里程碑做出了科学合理的评估。在设计与开发阶段，专项小组针对项目需求，进行了充分的技术攻关，进行了详细论证，完成了系统的整体技术方案设计和原型开发。期间进行了多次内部评审和迭代，确保设计满足功能、性能要求且成本可控和易于扩展。在测试和优化阶段，对开发出的产品进行严格的功能测试和性能测试，收集用户反馈，针对发现的问题进行了优化调整。在试点上线与推广应用阶段，项目组完成前期开发与测试阶段成果的可行性与稳定性的验证，收集真实环境下的用户反馈。为此，项目组精心挑选了具有代表性的一批用户作为试点对象，制定了详细的试点计划，包括试点范围、时间节点、预期成果以及风险评估与应对措施。基于试点用户的反馈意见，项目组完成了迭代优化，并制定了详细的推广策略。通过计划的执行，达到了既定的推广成效，并实际为业务创造了价值。

2. 展望

通过水晶球 ChatBI，我们建立数据分析的新范式，这种范式具备社会传播作用。传统数据分析的范式决策层或者数据分析师需要用数需要让“表哥表姐”去制作报表，表哥表姐给开发人员提需求开发报表或者通过拖拉拽的方式自主制作报表，这样会导致决策层决策数据少，分析师分析效率慢，表哥表姐意图理解难，开发人员开发任务多；而我行的数据分析新范式，是让“表格表姐”这个工种消失，每个人都是数据分析师，决策人员和分析师直接对话 ChatBI，分析师能够及时认知，决策者可以及时决策。



国元证券：面向 AI 大模型时代的新一代移动端架构 ——组件化与端智能深度融合实践

一、项目背景及目标

在移动端开发领域，随着 AI 技术不断及大模型场景应用的蓬勃发展，我们正迎来一场架构智能化革新。国元点金 APP 是国元证券股份有限公司重点打造的智能化综合理财投资服务平台，项目工程中存在大量代码冗余且代码复用率低，业务模块间高耦合，在开发效率、质量与安全管控、团队协作、新技术支持等方面逐渐遇到瓶颈，并且缺乏对于大模型技术应用落地实践的支持，影响项目长远健康发展。

为解决以上问题，同时贯彻执行证券业协会《证券公司网络和信息安全三年提升计划》中“持续提升代码开发效率及安全”的相关要求，以及践行公司数字化转型实践的相关内容，项目组经过详细调研并结合项目自身情况，基于组件化架构与大模型场景需求，对项目工程进行架构改造整合打造新一代移动端架构，设计并开发 RMD 2.0 智能化核心引擎组件，引入跨平台开发技术框架，以更好地支持业务侧需求，达到降本增效与安全可控的目的，不仅解决了传统开发中的诸多痛点，更为应用的智能化、个性化发展注入强大动力。

二、建设内容

总体架构设计

组件化是目前业界主流的项目架构方式，适用于多人合作开发的场景，核心思想是组件独立、解耦、复用，具有开发与协同效率高、有利于管控代码权限等诸多优点，可以很好地解决国元点金 APP 客户端项目中目前存在的各种问题。本项目基于组件化架构与大模型场景需要，进一步将不同功能组件根据职能分为应用层、业务支持层、基础层、组件化工程底座四层架构。

组件化工程底座作为项目的容器，用来集成项目组件，管理项目依赖，提供组件开发模式相关配置与切换功能，负责项目初始化、服务启动、路由注册等基础功能。

基础层主要包括开发常用的基础框架，与业务功能无关，主要包括网络请求、图片加载、通用的工具类、UI 组件、异步线程池、数据存储、Flutter 引擎等。

业务支持层的组件均为具备支持应用层特定功能的能力的特殊组件，部分需要中后台服务进行动态配置，主要包括移动分析、移动发布、运营能力以及业务公共组件等四大类型的能力组件，其中统一调度引擎（Router Engine）、模板引擎（ModuleEngine）、数据引擎（DataEngine）等的核心引擎组件结合 AI 场景需要，对现有 App 服务及数据进行原子化封装，打造成 RMD 2.0 智能化核心引擎组件簇，提供强大的端智能支持。

应用层的组件包含 APP 中具体的业务功能模块，负责 APP 主体业务实现，如行情、交易、理财、个人中心等。

技术方案

1. 组件化核心技术实践

组件化架构的核心包括组件化开发模式、组件化通信技术、组件化路由技术等关键技术。

组件化开发模式包括集成开发模式与组件开发模式。在组件开发模式下，组件可作为独立的 APP 进行开发和

调试、发布，提升开发效率；集成模式下，组件作为依赖库，被组件化工程底座管理，组成一个具有完整业务功能的 APP。组件经提前编译发布成二进制库，可以节省大量的编译耗时。

组件的独立性决定了应用层组件间不允许相互依赖，通过组件化通信组件完成相互通信与交互。本项目基于开源通信组件 ARouter 的基本原理，自研开发 CrossKit 通信组件，用于支持应用层组件间的通信与交互。

路由组件负责完成应用内界面跳转功能。本项目中路由组件为团队自研，运用反射技术做到组件间无需依赖即可完成跨组件界面跳转，支持可配置、白名单权限控制、路由降级等进阶功能。

2. 组件全生命周期管理

组件生命周期管理包括五个方面：组件开发规范、组件代码仓库、组件制品仓库、组件版本管理、安全管控等。

组件开发规范包括代码编写规范、代码优化规范、资源管理规范、命名规范等代码级别的规范。组件提供详细的使用指南、开发指南、历史记录等文档，方便使用者集成与使用。

组件代码仓库包括本地化部署开源的 Gitlab 私有环境作为代码仓库与版本控制工具，支持代码管理、版本控制、代码审核、CI/CD 集成等。

组件制品仓库：本地化部署 Nexus 作为组件制品仓库，作为企业内部组件制品仓库的和三方厂商 SDK 仓库，方便易用。

组件版本管理制定严格的三方厂商 SDK 更新制度与完善的项目管理流程，要求三方厂商提供 SDK 更新检查表，内容包括申请的权限列表及说明、权限申请频次说明、更新内容、更新版本、涉及开源组件说明、SDK 测试报告、环境配置信息等，项目管理人员检查通过后，通过项目管理工具提交集成 / 更新流程，由配置管理人员进一步核查 SDK 相关信息后，提交至组件制品仓库。

安全管控机制包括组件权限管控、权限申请流程制度、组件代码混淆技术等方面。组件处于独立的代码仓库，独立设置访问与修改权限，实施权限精细化管控。组件权限的申请需要通过办公系统提交正式的权限申请流程，经过严格的审批流程并书面留痕。权限的授予做严格的过期时间设置，做到权限最小化。运用代码混淆技术保证组件制品无法被反编译，运用加固与签名验证技术，避免 APP 被二次打包与恶意篡改，同时经过权威安全检测机构全面检测，修复潜在的组件安全问题，进一步保证 APP 的安全。

架构核心模块：端智能的 RMD2.0 多擎驱动

架构核心模块主要包括调度引擎（Router Engine）、模板引擎（ModuleEngine）和数据引擎（DataEngine），打造端智能的 RMD2.0 多擎驱动体系，为 AI 大模型场景应用赋能。

1. 调度引擎

在大模型应用场景下，将采用大量人机对话模式，AI 在识别用户意图后，应用端需能够提供快速跳转或直接在对话框中完成相应业务工作，传统路由方式，只覆盖简单的模块跳转，而涉及复杂页面交互、数据传递等场景，未能统一封装。

调度引擎实现应终端数据、交互、跳转进行统一标准封装，整合三方、H5、原生服务，为对话及服务场景提供统一调用方法，同时基于任务跑批能力，在终端完成特定场景数据加工处理，从而充分利用终端算力，提升对话响应速度。

2. 数据引擎

数据引擎提供统一的应用内部网络请求、数据存取、数据缓存、数据加解密等全局数据能力，屏蔽了数据使用底层逻辑，提供了协议化的数据能力 API。

基于统一接口调用 API，协议，降低三方 SDK 接入的学习成本，无需关心内部实现，面向接口编程，所有的接



口遵循一个协议方式调用，返回相同的数据格式。配合线程调度引擎，请求自动切换到规定的线程池中执行，成功之后自动序列化，解析，然后再自动切换到 UI 线程中刷新 UI。配合线程调度引擎的多线程并发操作符，可以实现多种场景的并发请求。

基于数据引擎能力，可为大模型对话提供数据服务，并同时保障数据隐私安全，如用户希望查询持仓个股当前盈亏情况，可基于数据引擎，客户端作为桥梁，查询持仓结果后渲染显示，可以避免用户个人数据泄露。

3. 模板引擎

基于 JCE 数据私有化传输协议，提供客户端动态装配页面的能力。客户端无需开发即可通过模板组件复用，根据中台下发的配置动态装配展示业务页面，做到轻量级业务快速上线，结合运营中台用户分群功能，可以做到千人千面的业务配置。

服务原子化与策略化：将各类服务拆解为细粒度的原子服务单元，如用户认证服务、推荐服务、服务等。基于大模型对用户行为、业务场景的深度分析，制定灵活的服务调用策略。例如在国元点金 APP 上以 ETF 全策略和网络百宝箱两个特色业务功能为例，通关配置不同内容，实现模板复用，同时单个页面中卡片可以做到根据不同客群展示不同运营素材。

在大模型对话模式下，基于调度引擎拉起数据引擎提供的数据，让模板引擎按照特定样式渲染，实现对话及服务模式快速落地，同时基于 RMD2.0 多引擎驱动体系，能够解决特定场景数据安全问题，

4. 其他

主题换肤组件：基于面向切面的 Hook 技术开发，支持多套皮肤动态切换，无需开发代码，通过中台配置下载，即可一键完成主题皮肤切换，换肤切换过程迅速，不影响用户体验。

Flutter 跨平台开发框架：将 Flutter 引入项目作为原生跨平台开发框架。Flutter 相比其他跨平台开发框架，在渲染性能方面更接近纯原生开发，有效提升用户体验的统一性和开发效率。

1D 三、运营情况与项目成效

本次实践中的组件化改造是一个庞大且复杂的工程，改造完成的组件上线的同时要兼顾业务正常迭代，应避免影响业务迭代的正常推进。按照项目整体分层架构，经过分析评估，自顶向下的四个层级对于业务迭代的影响逐渐减弱，因此组件采用分层分批上线方案，从基础组件、业务支持组件、应用层组件逐步依次上线。组件上线前需要经过严格的评审，根据其改造难度、改造影响范围、改造代码量等做评估，确定是否需要做灰度发布上线，联合信息技术部与业务部门商议确定其上线的版本。最终全部经改造的组件均上线后，集成组件化工程底座形成完整的组件化架构版本国元点金 APP 客户端，并在用户群体中进行小范围使用，稳定运行半年后逐步切换到组件化架构版本。

经过组件化改造与大模型技术的融合应用，开发过程可以灵活组合模块、快速进行开发调试，日常业务迭代开发过程中编译调试速度有了 5 倍的提升，组件化整工程与原始整工程编译速度相比提升 42% 左右，极大地提高了开发效率，降低了开发成本。打造了完善的代码安全管控体系，保证了公司核心资产安全与 APP 运行安全。

业务支持方面，模板引擎组件配合运营中台做到前端页面可配置化动态装配，并可根用户分群做到千人千面，提升轻量级业务上线速度与业务推广的灵活性。引入 Flutter 跨平台开发技术框架，做到一套代码可在安卓与苹果两端运行，降低因双端开发导致的用户体验不统一、业务逻辑偏差等问题，提升业务迭代效率，降低人力成本。大模型技术的深度融入，使应用在个性化推荐、智能客服、市场预测等关键业务环节实现突破性进展，显著提升用户活跃度、留存率和业务转化率。

四、项目总结与展望

基于组件化架构及未来大模型场景需要打造的完善的代码安全管控体系与组件全生命周期管理体系，一方面为公司核心技术的安全保驾护航，并在一定程度上保证研发质量，另一方面是贯彻执行证券业协会《证券公司网络和信息安全三年提升计划》中“持续提升代码开发效率及安全”的相关要求，具备可复制性，有较强的示范性和较高的推广价值。面向大模型场景的新一代移动端架构，通过端智能的多擎驱动（模板引擎、调度引擎、数据引擎），深度融合组件化开发模式，在应对千人千面需求、功能大爆炸挑战以及优化对话响应时间等方面展现出强大的优势和创新能力。

未来，随着大模型技术的持续进步和移动端开发实践的不断深入，我们可以预见，这一架构将在以下几个方面进一步发展和完善：一是大模型与移动端硬件的深度融合，利用设备端的 AI 算力加速模型推理，实现更实时、更高效的智能化应用体验；二是跨应用、跨平台的组件共享与协同机制将更加成熟，构建起一个开放、互联的移动应用生态系统；三是大模型将不仅仅是技术工具，更是应用的核心价值创造者，基于模型的创新业务模式和用户体验将不断涌现，重塑移动互联网的行业格局。让我们共同期待并积极拥抱这一移动端架构变革带来的无限可能，为用户创造更加卓越、智能的移动应用世界。



西南证券：证券公司大语言模型中台建设实践

一、项目背景及目标

自 ChatGPT 发布以来，大语言模型在全球范围内迅速掀起了人工智能的浪潮，引发空前关注。证券公司的主营业务涉及大量文本内容处理与生成的工作，如提供投资分析报告书、发布投资研究报告等；同时在公司内部管理过程中也积累了大量的专业知识文档如合规管理法规制度等。可以说天然为大语言模型的落地实践提供了丰富的应用场景。然而，基于证券行业的强监管特性，证券公司无法使用公有云 API 的调用方式使用大语言模型，需本地部署。根据调研结果显示，目前市场主流 AI 厂商、互联网厂商的大语言模型私有化部署方案成本较高，动辄数百万起，对于中小型券商来说不得不谨慎考虑相关投入产出比。另一方面，据统计，在 2023 年共有一百余个中文大语言模型完成了技术开源，众多的开源大语言模型为私有化部署与应用带来了机遇。

综上，我们从公司实际情况出发，采用自主研发的方式基于开源大语言模型规划并建设了公司级的大语言模型中台。目前搭建的大语言模型中台将实现公司大语言模型技术的首次落地应用，为多个业务场景实现降本增效；且大语言模型中台支持新应用的敏捷开发、快速迭代，未来可高效赋能更多业务场景。

二、创新点

在大语言模型应用建设过程中，我们遇到了若干挑战：①如何选择合适的大语言模型？②如何快速完成大语言模型的落地？③如何降低因模型变更或闭源给自身造成的影响？

基于这些挑战，我们设计并建设了大语言模型中台。从大语言模型应用的落地流程来看，主要的步骤包括：大模型测试、大模型微调、大模型开发、大模型量化部署这四个大的模块。在本课题建设的大语言模型中台技术组件中，完成了对大语言模型测试时对应用效果、复杂指令理解、回答安全性等方面的测评工具落地，从而实现根据不同应用场景挑选出综合能力较强的大语言模型。大语言模型的微调工具可使用下游应用场景的训练数据，支持使用全参数量、LoRA（低秩适应）或 RLHF（基于人类反馈的强化学习）的方式调整模型参数，使大语言模型适应下游应用场景的特定表述方式和表达习惯。大语言模型的开发工具落地了不同的开发框架，如 Langchain、Llama-index、AutoGen 等，实现了针对不同的应用场景选择更匹配、更完善的定制化开发框架的效果。大语言模型的部署落地了不同的部署框架，如 vLLM、Tensorrt-LLM 等，可提供不同的量化方式与加速方式，节约推理时的显存资源耗费并提高推理并发量。

本项目形成了公司大语言模型应用落地的参考标准，可为后续大语言模型项目缩短建设周期，使项目建设聚焦于应用场景本身，挖掘出更多的业务价值。

三、项目技术方案

我们按照大语言模型应用落地的步骤对大语言模型中台的架构设计，总共划分为四层，自底向上分别是大语言模型训练、大语言模型开发、大语言模型部署、大语言模型应用。

大语言模型训练层包含了基础设施、数据源管理与训练框架，这一层主要用于管理 GPU 与高性能网络，并在硬件之上搭建分布式训练环境，为后续大语言模型的测试或者 PEFT（Parameter-Efficient Fine-Tuning）微调提供基础环境。大语言模型开发层包含了不同的开发框架与技术组件，这一层用于在大语言模型参数确定后，根据业务场景定向封装大语言模型的能力；同时设计出可切换不同大语言模型的适配层，降低大语言模型变更为项目建设

带来的风险。大语言模型部署层包含了不同的分布式推理加速框架，这一层主要用于对大语言模型进行统一的量化加速，提高大语言模型推理时的最大并发量，有效提高硬件使用效率。最上层通过统一的 API 设计为大语言模型应用提供服务，包括单次调用 API、批处理 API、流式处理 API，满足不同应用场景与大语言模型的交互需求。

四、运营情况

大模型中台已在公司金融科技部 AI 开发团队广泛应用，基于该中台建设的 3 个不同类型的大语言模型应用已应用于投顾展业、合规管理、财务管理相关部门。大模型中台在不同应用场景建设过程中运行稳定，各技术组件根据开源框架更新速度稳步升级，可满足典型自研大模型应用落地的全流程技术组件需求。

五、项目成效

基于大语言模型中台，我们落地了 3 个不同类型的大语言模型应用，包括了基于思维链、提示工程、Agent 调用的报告生成类应用——证券投顾分析报告书智能生成，基于 RAG（检索增强的文本生成）的应用——智能合规助手、智能财务问答。

利用思维链、提示工程与 Agent 能力，可实现对投资研究信息的整理与总结，结合市场信息与新闻舆情信息生成相关投顾报告，大幅减少人工浏览信息、整理信息的时间。

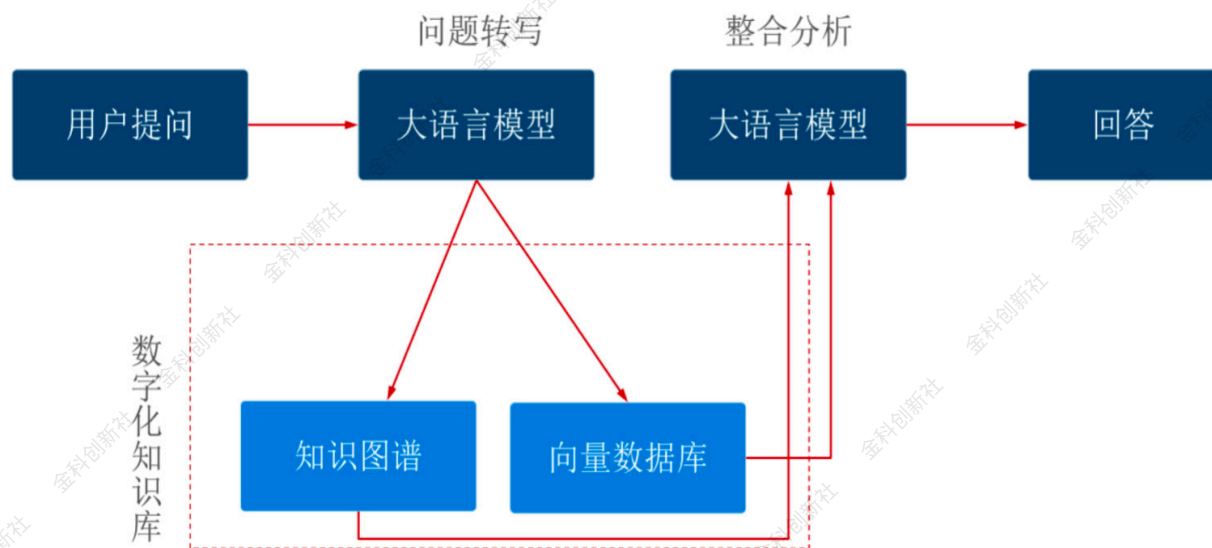


图 RAG 流程示例

图为基于检索增强的知识库问答流程示例图，使用不同的知识库组件完成多重召回，如向量数据库、知识图谱、传统检索等技术，结合大语言模型完成了对公司数千份文档按业务条线分类的整理，形成了数字化的合规知识库与财务知识库，实现了用户通过自然语言与知识库助手、问答机器人进行流畅的交互。

人保寿险：“AI 保宝”大模型平台赋能公司数字化转型

一、项目背景及目标

为提升业务服务质量与办公效率，人保寿险创新性的提出“AI 保宝”大模型平台创新应用项目，协同集团公司引进大语言模型，切实解决外勤、内勤两大群体在公司日常经营、管理中的痛点问题，在产品咨询、坐席助手、办公助手三大类场景下实现了降低人力成本、提高运营效率、提升服务水平。助力公司业务模式变革和管理模式创新。目前已在公司内网门户、移动互联、人保 E 通、费控系统、企业微信等多系统上线。

面向内勤建设办公助手提升办公效率、助力风险防控。一方面实现从日常办公到专业业务操作的全方位支持，打造“人人有助手，人人是专家”的人保办公工具，为每一位员工配备专属的数字化助手。具体上线企业制度问答、营销海报/文案自动生成、个人/群助理机器人、智能报表生成。另一方面面向企业风险合规条线，推动打造审计分类、内部合规助手、合同审核助手、金融政策动态、声誉风险问答等产品，全方位提升企业风控能力，确保企业运营安全合规，助力高效决策。

面向外勤建设坐席助手助力降本增效，推动客户服务、内部管理升级，为外勤提供生产力工具，实现人人有助手、人人是专家，为保险行业的数字化转型探索新模式。具体上线产品条款解读、坐席话术推荐、电话工单生成、智能陪练智能建课、坐席服务支持。建设产品宝典支持公司所有产品问答。支持保险基本信息咨询、投保条件咨询、投保流程咨询、保险责任咨询等场景。

二、创新点

创建行业领先的“2+1+3”大模型平台架构

“AI 保宝”大模型平台采用“2+1”大模型智能平台以“两大用户群体 + 一个大模型平台 + 三大创新场景”为核心架构，旨在满足不同用户群体的需求，提供全方位的智能化服务。

1. 一个大模型平台

“AI 保宝”大模型平台是整个架构的核心，它集成了自然语言处理、机器学习、知识图谱等先进技术，为用户提供智能化的服务。平台通过不断学习和优化，能够更准确地理解用户需求，提供更精准的建议和解决方案。

2. 两大用户群体 + 三大创新场景

产品咨询场景：针对保司外勤快速了解公司产品的需求，平台提供产品咨询功能。外勤可以通过自然语言输入查询需求，平台智能识别并返回相关的产品条款、费率等信息，帮助用户快速了解产品特点和优势。

坐席助手场景：针对保司外勤的工作特点，平台提供坐席助手功能。通过自然语言处理和机器学习技术，平台能够自动生成电话工单、推荐坐席话术等，帮助外勤人员更好地与客户沟通，提高销售转化率。

办公助手场景：针对保司内勤的工作需求，平台提供办公助手功能。平台能够智能回答公司知识问答、进行合同审核对比等，同时提供贺报、营销文案的自动生成能力，帮助内勤人员提高办公效率。

自主创新研发多项大模型应用技术



人保寿险推出的 AI 保宝产品体现了公司在自主研发方面的独特优势，具备高度的先进性和行业领先性。

1. 文档版面知识采编解析：对于不同版式的文档进行版面解析和建模。通过大模型进行文档数据的规整，减少人工介入，实现了数据采编的自动化与高效管控，为业务问答数据的快速更新提供了强大的技术支持。

2. 问答检索模型：训练 1024 字长检索模型以配合长文本大模型与长文本采编，突破了传统检索模型对于字长的限制，减少了 embedding 阶段的语义信息损失，在长文本场景下拥有更好的检索效果

3. 问答回复模块：创新的采用了思维链的训练策略，以此指导大模型学习思维链的思考方式，来激发大模型对问题和答案的思考判断，提升对专业领域的语义分析能力和总结能力，保障回复效果的正确性。

4. 数据隔离：采用多租户方式强化身份认证和访问控制，对 API、用户界面以及后端服务的访问权限进行精细控制，提升了问答效率及效果，减少不同业务间数据误吸风险

5. 灵活的对话流程：提供多种系统意图，例如打招呼、再见语、感谢语、超时、意图拒识、系统异常等常用意图，更好的捕捉用户意图，流程执行更加灵活，流程设计更便捷。

融合 DeepSeek 进一步提升平台赋能

上线 DeepSeek 产品条款解读功能，营销员通过人保 e 通或人保友客，在对应产品界面点击“AI 条款解读”按钮即可畅享智能服务。

发起“A 计划”，开发 DeepSeek 系列培训课程，全方位赋能营销队伍发展。计划书系统上线基于大模型的名词解释，辅助营销员制作建议书过程。

上线基于 DeepSeek 的反洗钱审计可疑交易案例自动审核，实现对可疑交易分析中客户信息的完整性、客户身份 / 收入水平与保费规模的适配性、收入水平的合理性、资金去向的合理性等维度的批量自动化审核。涉及保费金额数十亿元，发现的问题线索如机构妥善整改将大大减少给公司带来的处罚风险。预防监管处罚风险，助力机构有针对性的加强洗钱风险管控。

三、项目技术方案

（一）项目规划

战略目标确立：人保寿险立足行业数字化变革趋势，将大模型建设作为推动公司战略转型的关键举措。首要目标是借助大模型强大的数据分析与智能处理能力，重塑核心业务流程，实现降本增效。如在客服场景，期望通过大模型驱动的智能客服，提升问题解决效率，降低人工客服成本；在营销领域，利用大模型精准洞察客户需求，提高营销转化率，挖掘潜在客户资源。长远来看，项目致力于打造行业领先的智能化寿险服务体系，增强人保寿险的市场竞争力与品牌影响力，巩固行业领军地位。

分阶段规划：

基础构建阶段（第 1-6 个月）：此阶段重点搭建项目基础架构。一方面，组建专业的跨学科项目团队，涵盖人工智能专家、数据科学家、保险业务专家等，明确各成员职责，确保项目具备技术与业务双重专业视角。

模型研发阶段（第 7-12 个月）：聚焦大模型核心研发。利用前期收集的数据，运用数据清洗、去噪、标注等技术手段，构建高质量训练数据集。结合保险业务特性，选取如 Transformer 等先进架构，进行模型搭建与初步训练。在训练过程中，不断调整模型参数，优化模型性能，同时引入迁移学习等技术，加速模型收敛。期间与外部机构合作，共同攻克技术难题，如针对寿险业务的复杂风险评估模型优化等。

试点应用与优化阶段（第 13-15 个月）：挑选智能客服、代理人展业辅助等核心业务场景开展试点。在智能客服场景，部署大模型驱动的智能客服系统，实时收集客户咨询数据与客服人员反馈，评估模型在问题理解、解答准确性等方面的表现；在代理人展业辅助场景，通过为代理人配备大模型支持的移动应用，收集其在产品推荐、客户沟通环节的使用数据。根据试点反馈，针对性地优化模型，调整业务流程，如优化智能客服话术生成策略、完善代理人展业助手知识检索算法。

全面推广阶段（第 16-24 个月）：将优化后的大模型及配套应用全面推广至全国各分支机构与更多业务场景，如核保风险评估、理赔智能定损等。建立推广实施计划，明确各阶段推广目标、责任主体与时间节点。同时，搭建全国统一的大模型应用监控平台，实时监测模型运行状态，确保推广过程稳定、有序。

（二）技术架构搭建

数据层：构建湖仓一体化架构，整合人保寿险内外部数据。内部涵盖结构化数据以及非结构化数据；外部引入行业市场数据、宏观经济数据等。同时，建立流批一体化的数据架构，大幅提升数据采集与计算效率，其中数据采集达到秒级响应，数据计算效率提高 83%，为模型训练提供及时、高质量的数据支持。

模型层：以自研大模型、DeepSeek 通用大模型、讯飞星火、阿里千问为底座，通过对海量保险行业语料、数据及专业知识的学习，具备基础的语言理解、知识储备与逻辑推理能力。在此之上，面向寿险营销、客服、核保、理赔等垂直业务领域，自研多个场景大模型。

应用层：孵化多款大模型应用产品。“AI 保宝”面向客户与代理人，提供保险类知识问答服务，提升知识获取效率，文生图、图生图等跨模态生成能力，助力营销设计物料制作；代码生成、代码补全等能力，加速软件开发。这些应用产品通过 API 接口等方式，与公司现有业务系统深度集成，为员工与客户提供便捷、智能的服务。

四、运营情况

目前人保寿险“AI 保宝”大模型已在移动互联、人保 E 通、费控系统、企业微信、办公门户、e 掌柜、核心系统运维机器人、智能陪练、双录、OA、核心系统等 10 多个系统上线，23 年启动建设至今共调用几十万次。累计激活用户数量数百万。知识检索准确率大于 95%，问答准确性大于 94%。

五、项目成效

“AI 保宝”大模型平台快速响应、高效处理、稳定服务、实时更新的优势，免去了公司人员工作中繁琐的检索过程、耗时的海报制作过程，让工作更便捷，决策更迅速，为公司提供了更高效、更精准、更个性化的营销手段，优化了营销流程，提升了用户体验。同时“AI 保宝”问答准确率也在连续的迭代和升级中不断提升，知识检索准确率大于 95%，问答准确性大于 94%。近期审计部依托大模型上线反洗钱案例筛查，发现的问题件涉及保费金额共数十亿元，线索如机构妥善整改将大大减少给公司带来的处罚风险。

六、经验总结

1. 聚焦业务需求，精准锚定方向

多场景调研：项目初期，组建跨部门联合调研团队，深入寿险营销、客服、核保、理赔等核心业务场景。与一线员工、管理人员、客户进行广泛交流。例如在营销场景中，发现代理人在产品知识讲解、客户需求挖掘方面存在挑战；客服场景里，对复杂保单问题的快速准确解答需求迫切。基于此，明确大模型需具备精准的保险知识问答、客户意图理解、智能营销辅助等关键能力。



动态需求更新：建立需求动态跟踪机制，每月收集业务部门反馈，根据市场变化、监管要求及新技术发展，及时调整优化大模型功能。如随着健康险市场的快速增长以及新的健康险产品条款更新，迅速将相关知识纳入大模型训练体系，确保其能为业务提供实时、有效的支持。

2. 打造技术基石，筑牢模型根基

强大算力支撑：构建专属算力集群，引入高性能 GPU 服务器，保障模型训练与推理的高效运行。根据业务负载动态调配算力资源，降低成本的同时提高响应速度。例如在营销活动高峰期，将更多算力分配至大模型的客户需求分析与产品推荐模块。

数据治理与标注：整合人保寿险海量数据，涵盖保单信息、客户资料、理赔记录等。运用数据清洗、去噪、脱敏等技术，确保数据质量。

3. 试点先行，稳步推进

试点区域与场景选择：选取业务规模适中、需求多样的产品解读作为首批试点地区，在智能客服、代理人展业辅助两个核心场景开展试点应用。通过试点，深入了解大模型在实际业务环境中的表现。

优化与复制推广：根据试点经验，对大模型进行针对性优化，如改进智能客服的话术生成策略、提升代理人展业助手的知识检索准确性。优化后，智能客服问题解决率提升 42%。随后，将成功经验与优化后的大模型逐步推广至全国其他地区及更多业务场景。

大家保险集团：大家灵图影像平台

一、项目背景及目标

大家保险集团业务体量庞大，产生了海量的非结构化影像文档数据（包括各类证件、保单、理赔资料、视频等），存量规模达 10 亿级，且日均增长数十万份。传统模式下，这些数据的价值难以有效释放：

- 操作体验差：文档上传、查看效率低。
- 智能化不足：依赖人工识别、检索、分类，耗时耗力。
- 安全挑战：海量敏感文档的安全存储与访问控制需求迫切。
- 价值挖掘难：数据蕴含的业务洞察因无法高效检索分析而被埋没。

为应对上述挑战，充分挖掘非结构化数据的巨大潜力，提升业务效率、节省人力成本、优化客户服务体验并强化安全合规能力，大家保险集团决定建设大家灵图影像平台。项目核心目标包括：实现影像文档的高效智能管理、构建安全可靠的数据访问体系、深度赋能核心业务流程、释放非结构化数据价值。平台通过引入前沿 AI 技术，提升了图片文件检索能力，挖掘影像价值。同时，智能检索与分析功能大幅缩短业务响应时间，助力决策精准高效，全面推动集团数字化转型进程。平台上线后，显著提升了数据处理效率，智能检索与分析功能大幅缩短业务处理时间，以图搜、图片相似度检索等功能提升了财务、理赔环节中的风控能力

二、创新点

1. 架构创新

- 文件流与数据流分离：解耦文件物理存储与元数据管理，提升系统灵活性、可扩展性与安全性，减少机房带宽压力。
- 多存储源动态插拔：支持公有云（互联网业务）与私有云（内部业务）存储的无缝接入与动态切换，并且可以动态增加新的存储源。
- 分布式智能存储与检索：采用成熟稳定的分布式检索和存储技术，保障海量数据存储与高性能检索。

2. 功能创新

- 智能处理引擎：深度融合多模态大模型、OCR、图像特征提取、向量化技术，实现图搜图、文搜图、智能分类、特征提取等高级能力。
- 精细化访问控制：支持基于时间有效性、访问次数、长期有效等多维度的文件访问权限管理。
- 安全加固：集成安全签名、文件加密技术，确保数据安全。
- 高效访问体验：应用短链接重定向技术优化访问效率。
- 在线处理能力：支持文档在线编辑预览、图像标注、压缩包在线预览等便捷功能。

3. 推广实施创新

超高速海量小文件迁移方案：针对亿级海量小文件迁移瓶颈，创新采用多线程与文件直接定位法，迁移速度



较传统方式提升数百倍（日均可达 3T），极大缩短系统切换周期，降低迁移风险。

ID 三、项目技术方案

1. 平台整体架构

- 分层设计，接入层、应用层、AI 能力层、数据层、存储层；
- 微服务化；
- API 网关。

2. 平台存储架构

- 分布式对象存储 / 文件存储；
- 元数据库；缓存层；
- 多存储源适配器；
- 智能分层存储策略。

3. 网络机制

- 文件流和数据流分离。

ID 四、运营情况

平台自上线以来运行稳定高效，表现优异：

架构健壮性：应用层、数据层均采用集群部署，支持横向扩展；采用分布式存储，支持在线扩容；双集群架构保障高可用，满足 7 乘 24 小时运行要求。

安全合规：通过安全漏洞扫描，符合金融行业严格规范要求；项目实施遵循 CMMI3 过程标准。

性能与规模（截至申报时）：

对接业务系统数量：65 个

管理文档总量：12 亿 +

总存储量：130T+

系统压测指标：830 TPS

接口平均响应时间：~60ms

日均访问量：~2000 万次

日均上传文档量：~20 万个

ID 五、项目成效

平台产生了显著的经济效益和社会效益：

保障业务稳定，提升用户满意度：每天承受 2000 多万次的调用，网络带卡为 4g/ 秒，保障影像文档的上传和查看速度，提供高清、水印在线预览等各种贴心服务，客户出单迅速，业务员审核快捷。

强力护航营销，提升公司业绩：在 2 首季营销关键期，平台为子公司提供了坚实的技术支撑。累计调用服务超 2 亿多次，助力完成保单成交 22511 笔。

全面提升效率与体验：实现了影像文档的智能化处理与秒级检索，大幅减少人工操作，显著提升内外部用户（员工、客户、审计等）操作体验和满意度。

释放数据价值：使海量非结构化数据变得可检索、可分析，为风险控制、产品优化、客户服务等提供数据洞察基础。

降本增效：自动化处理替代大量人工操作，有效节省了人力成本。

提升风控管理能力：为集团业务部门提供了多次业务影像文件的快速分析，并发现多笔问题案件，有效提升审计时效。

六、经验总结

大家灵图影像平台的成功建设和推广，为大型金融集团构建智能化、安全高效的影像文档管理基础设施提供了宝贵经验：

1. 前瞻性技术架构是基石：采用“文件流与数据流分离”、“多存储源动态插拔”等创新架构设计，为平台应对海量数据、复杂业务场景和高并发访问提供了坚实的基础，确保了系统的灵活性、扩展性与长期生命力。拥抱 AI 技术（多模态大模型、向量化等）是释放非结构化数据价值的关键。

2. 业务深度协同驱动价值落地：平台建设始终紧密围绕核心业务痛点（如出单效率、理赔时效、审计合规）展开。通过深入了解业务需求，将 AI 能力精准嵌入业务流程（如 OCR 用于自动录入、智能检索用于快速调阅），确保了平台建设的实用性和价值转化效率。与 40+ 业务系统的快速对接验证了协同的有效性。

3. 科学严谨的项目管理是关键：采用分阶段、小步快跑的敏捷实施策略（先试点后推广），有效控制了风险。制定清晰的里程碑（如 1 个月完成核心架构与 SDK），并严格执行 CMMI3 过程标准，保障了项目在紧张周期内高质量交付。特别是在海量历史数据迁移中，创新的高速迁移方案解决了关键瓶颈。

4. 安全与性能并重是保障：在金融行业，安全是生命线。平台从设计之初即融入多重安全机制（加密、签名、精细权限控制、漏洞管理），并通过分布式架构、集群部署、性能压测等手段保障了在高并发、大数据量下的稳定高效运行，日均 2000 万次访问的稳定表现即是证明。

5. 可推广性设计提升投入产出比：平台定位为集团级统一基础设施，通过标准化 API、易用 SDK、可配置策略等设计，极大降低了各业务系统的接入门槛和改造成本，实现了建设成果在集团范围内的快速复用和规模化效益。创新的海量小文件迁移方案也具有行业推广价值。

6. 持续优化提升用户体验：项目并非上线即结束。根据上线后的用户反馈和性能监控（如 6 月的性能与体验升级），持续进行功能优化和性能调优，是平台保持活力、不断提升用户满意度和粘性的重要因素。

总之，大家灵图影像平台的成功实践，不仅解决了大家保险集团自身的核心痛点，其先进的技术架构、高效的实施方案和显著的业务价值，也为金融行业乃至更广泛领域的非结构化数据智能化管理提供了可借鉴的标杆案例。



大家保险集团：大家保险人事办公智能化项目

一、项目背景及目标

传统保险行业在人事管理、行政办公等后援服务领域普遍存在效率低下、流程繁琐、服务质量参差不齐等问题。伴随数字化转型的推进，企业亟需通过技术创新提升内部管理效能，降低运营成本。同时，随着大模型技术的快速迭代，企业对 AI 技术在业务场景中的应用需求日益迫切。

大家保险集团在推进服务降本增效与管理模式进化的过程中，意识到 AI 技术在提升办公效率和服务质量方面具有巨大潜力。通过对内部办公流程的深入调研，集团发现人事、行政、法务等后援服务领域存在大量的标准化、重复性工作，非常适合通过智能化手段进行优化，以实现如下目标：

- (1) 通过 AI 大模型技术赋能人事办公场景，打造智能化办公工具，提升工作效率和服务质量。
- (2) 建设中台化、组件化的 AI 应用服务体系，实现技术与业务的深度融合。
- (3) 推动企业内部管理的数字化、智能化转型，为行业提供可复制的实践经验。

二、创新点

1. 技术创新

(1) 多模态大模型融合：集成 GPT-4、千问 72b 等多模态大模型，实现文本、语音、图像等多种数据形式的智能处理。

(2) 中台化架构设计：构建智能交互中台，支持模型、知识库、提示词等组件的灵活配置，确保系统的可扩展性和可维护性。

(3) 知识库优化：通过知识加工、向量化、RAG（检索增强生成）等技术，提升知识检索的准确性和效率。

2. 应用模式创新

(1) 一站式智能办公平台：覆盖知识问答、文档校对、会议管理、合同要素提取、智能出题、阅卷辅助等场景，打造全场景智能办公新生态。

(2) 场景定制化服务：针对不同业务系统的特点，提供定制化 AI 服务，如接入 OA 系统实现公文校对润色。

3. 管理机制创新

(1) 权限控制体系：通过细粒度的权限管理，实现知识库与 AI 应用的多维度安全管控。

(2) 智能体管理中台：支持模型的灵活配置与优化，确保 AI 交互的合规性与稳定性。

三、项目技术方案

1. 架构体系

基础层：混合部署 GPT-4o/ 千问 72B/Whisper 等多模态模型。

数据层：构建超 10 万条制度文档向量知识库。
服务层：RAG 增强引擎支持语义检索与结构召回。
应用层：提供 API/SDK/ 低代码三种接入方式。

2. 关键技术

知识加工：采用语义切片 + 上下文扩展技术。
权限体系：三级组织架构权限映射模型。
流程引擎：支持跨系统服务组装编排。

四、运营情况

目前，人事办公智能化平台已在集团内部全面上线，覆盖人事、行政、法务等多个部门。通过企业微信、PC 门户、统一用户中心等多个入口触达用户，日活跃用户数达到 500+ 人。系统运行稳定，用户反馈良好，已成为集团员工日常办公的重要辅助工具。典型应用运营情况如下：

- (1) 智能会议助手（办会助理）：语音指令识别准确率 95%，预定耗时从 5 分钟降至 30 秒，联动企业微信日程，会议提醒准时率达 100%。
- (2) 公文核校系统（大家润笔）：集成 WPS 文档中台，实现格式 / 文字 / 合规三重校验，关键错误检出率提升至 98%，核校时效缩短至 3 分钟 / 份。
- (3) 制度知识中枢（大家千问）：构建包含 15 万条制度的动态知识库，问答准确率 92.3%，替代 83% 人工咨询。
- (4) 智能课程生成（大家 E 学）：日均处理音视频课程 20+ 课时，生成课程脑图准确率达 89%，内容纠错反馈采纳率 76%。

五、项目成效

1. 降低人工作业投入：通过自动化处理会议预定、公文校对、课程制作等任务，减少人工操作时间 50% 以上。
2. 提升业务处理效率：合同要素提取、会议纪要生成、课程制作、智能出题、阅卷辅助等场景的自动化处理，使相关业务处理效率提升 70%。
3. 节省人力成本：通过智能化改造，减少部分岗位的人力投入，节约成本约 100 万元 / 年。

六、经验总结

1. 坚持以业务需求为导向，场景驱动，小步快跑：项目从需求调研出发，结合实际场景痛点，设计解决方案，确保技术与业务的深度融合。优先选择高频、低风险场景（如会议室预定）快速验证，积累信心后再拓展复杂场景。
2. 注重中台化建设：通过中台化设计，实现技术能力的复用与扩展，为后续应用开发奠定基础。
3. 重视安全与合规：通过权限控制、敏感词管理等机制，确保 AI 应用的安全性与合规性。未来，将继续优化大模型技术在业务场景中的应用，探索更多智能化应用场景，为企业的数字化转型注入更多活力。



中国大地保险：AI 中台建设项目

一、项目背景及目标

随着人工智能技术的飞速发展，特别是大模型的崛起，为各行各业带来了前所未有的创新机遇。保险业作为传统与现代技术交汇的领域，正积极寻求通过大模型技术重构业务流程，优化运营效率，提升客户体验，以适应行业变革。

中国大地保险从 2023 年初启动大模型应用探索，联合国内大模型厂商、解决方案厂商开展大模型能力验证及业务场景落地探索。为进一步加速大模型等 AI 与业务的深度融合，2024 年公司启动“灵山界”AI 中台项目建设，以大模型为核心，整合各类资源，在企业内部构建开放、高效的 AI 应用体系，同时在组织层面建立创新机制，推动人才培养及场景落地，形成“平台、场景、人才”三位一体的智能基座，赋能企业数智化转型向价值深水区迈进。

二、创新点

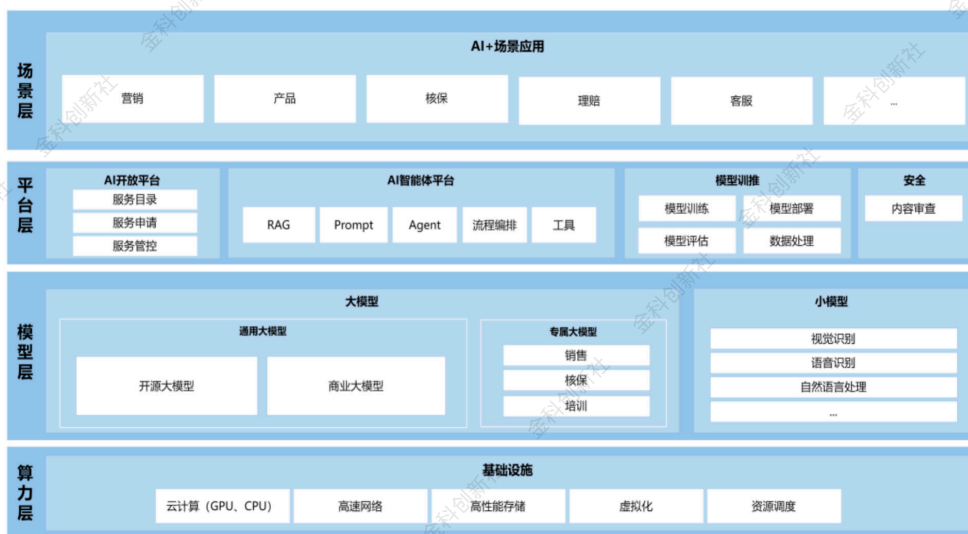
通过 AI 中台项目建设，形成“平台、场景、人才”三位一体的智能基座，并通过建立长效的创新机制，赋能企业数智化转型。主要创新点：

(一) 平台筑基：构建 AI 大模型驱动下的“云数智”一体化，将模型、数据、算力和框架有机结合，打造资源可共享、模型可复用、服务可编排、应用更灵活的 AI 中台，为业务创新、数智化转型提供坚实基础。

(二) 场景赋能：聚焦保险核心价值场景，实现多样化的全链路 AI 赋能。基于 AI 智能体平台，员工根据自身业务领域及场景，快速构建企业级 Agent 应用。基于 AI 开放平台，实现 AI 能力与业务深度融合，驱动业务模式革新。在办公领域，打造 AI 超级助理，赋能办公提效，优化内部服务流程。

(三) 人才创新：建立全员 AI 共创机制，激活组织创新势能。公司成立 AI 创新专项工作组，整合业务与科技资源，面向全员提供 AI 智能体平台，员工能够根据自身业务领域和场景，自助式构建 AI 大模型应用，激发全司创新活力。

三、项目技术方案



面向员工创新，提供 AI 智能体平台，该平台集成大模型、知识库、工具等各类能力，员工可以根据自身业务领域、业务场景，快速构建企业级 Agent 应用，激发员工的创新精神，使企业能够更敏捷地响应市场变化和客户需求。

面向业务应用，通过 AI 开放平台，提供快速共享 AI 能力。AI 能力与业务深度融合，驱动业务模式革新，持续激发公司业务与服务模式的创新潜能。

面向办公提效，依托大模型等 AI 能力，打造 AI 超级助理，覆盖业务作业、知识服务、员工服务、办公服务等全方位服务，赋能办公提效，优化内部服务流程。



同时，通过 AI 中台建设，建立长效运营机制。从运营组织建设、价值模型评估、工作流程设计等三个方面，建立组织级的 AI 创新行动计划。在公司内部，成立跨部门、跨领域的专项工作组，负责各业务板块的创新应用孵化及运营，建立从创新构思、预研探索、落地实施以及推广普及等阶段的高效协作流程，实现 AI 创新的闭环管理，推动持续的人才培养及场景落地。

四、运营情况

业务场景落地：基于 AI 中台，支持场景的快速落地。在销售辅助场景，利用大模型的识别及生成能力，为销售人员提供全方位的支持，包括智能报价、营销文案，企业画像等；在智能理赔场景，基于大模型等 AI 能力实现了单证智能分析、疾病智能判别、条款解读等功能，辅助理赔审核自动化，显著提升了理赔效率；同时，基于大模型 RAG 技术，在企业内部构建财务、人事、客服、产品以及系统支持等知识问答服务，互动交互等客户体验显著提升。此外，面向员工，构建多个办公助理，包括 AI 文档校对、会议纪要、招聘 JD、智能文档、差旅报销等全方位支持服务，赋能办公提效。

员工创新实践：通过培训、竞赛等方式，在全司积极推广 AI 智能体平台，目前总、分公司员工广泛参与，在销售、承保、理赔、财务、客服以及办公等领域积极实践，用户数超 10000 人，创建智能体 700 多个。

五、项目成效

业务场景快速落地：AI 中台的建设，实现了资源的高效共享和能力的快速复用，使得业务场景能够迅速从概念验证阶段过渡到实际应用。在销售辅助、智能核赔、知识助手、智能办公等业务场景中积极实践，取得显著成效。



组织级业务创新：建立长效的 AI 创新行动机制，激发全员的创新活力，各部门积极踊跃，参与公司数智化升级，促进业务流程的优化和业务模式的创新，为企业的长期发展注入了新的活力。

ID 六、经验总结

目前，AI 大模型的发展处于初级阶段，存在广阔的想象空间，大模型的分析理解，内容生成等能力在保险营销、承保、理赔等业务场景存在极大的变革可能，在客服场景有望重塑交互流程，提升客户体验。在 AI 技术革命及公司数字化转型之际，顺势而为，抢占先机，具有重大的战略意义及实际应用价值。

通过 AI 中台的建设，实现资源的高效共享和能力的快速复用，使得业务场景能够迅速从概念验证阶段过渡到实际应用。这种灵活性和敏捷性，使得企业能够迅速响应市场变化，快速部署新服务，从而在竞争激烈的市场中保持领先地位。

各行业仍在积极探索价值场景，在企业内部，通过建立组织级的 AI 创新机制，推动人才培养及场景落地。引入 AI 智能体平台，员工根据自身业务领域、业务场景，亲身参与构建企业级大模型应用，激发全员的创新精神，为企业的长期发展注入新的活力。

基于大模型构建银行统一知识库管理体系

北京知化时代科技有限公司

一、项目背景及目标

1. 背景

- (1) 知识孤岛严重：银行内部知识（产品、制度、手册、FAQ 等）分散在各业务部门、各系统中，缺乏统一管理和有效整合。
- (2) 信息检索低效：员工查找所需知识耗时耗力，依赖个人经验或多次询问，影响业务处理效率与客户服务质量。
- (3) 知识质量不一：缺乏统一标准和持续维护机制，知识内容存在过期、错误、表述不一致等问题。
- (4) 服务渠道单一：知识获取方式传统（如文档库、邮件），缺乏智能化、便捷化的服务手段。
- (5) 管理规范缺失：缺乏全行级的知识管理规范、流程和制度，知识资产价值未得到充分挖掘和利用。

2. 目标

- (1) 构建统一平台：建立全行级、集中化的统一知识库系统，打破信息壁垒。
- (2) 建立规范体系：制定并推行全行知识管理规范、制度和流程，实现知识全生命周期管理。
- (3) 整合与标准化：全面梳理、整合、清洗、标准化全行各业务领域的知识资源，确保知识全面、准确、一致、可用。
- (4) 多端智能服务：基于大模型技术，打造强大的智能搜索和智能问答能力，通过多终端（PC、移动端、内嵌业务系统）为员工提供高效知识服务，显著提升工作效率和决策支持水平。
- (5) 奠定知识基础：形成可持续的知识资产积累、更新和应用机制，提升组织整体知识能力。

二、项目 / 策略方案

本方案采用“管理先行、技术赋能、分步实施”的策略。

1. 需求调研与蓝图设计（前期核心）

- (1) 深度访谈与问卷：覆盖公司金融、零售金融、运营管理、风险管理、科技、客服中心等关键部门，深入了解各部门知识痛点、使用场景、内容类型、更新频率、权限需求等。
- (2) 现状分析与差距评估：全面盘点现有知识资源（文档系统、邮件附件、部门共享盘等），评估其管理状态、质量、可用性及与目标的差距。
- (3) 制定顶层设计：基于调研和分析，明确知识管理组织架构、知识分类体系（多级分类、标签体系）、知识管理流程（采集、审核、发布、更新、归档、权限）、知识质量标准、安全策略等核心规范草案。
- (4) 大模型选型与场景定义：评估主流大模型（deepseek-R1、Qwen3 等），结合银行数据安全要求，选择



部署模式。明确大模型应用的核心场景：智能搜索（语义理解、精准召回）、智能问答（多轮对话、上下文理解）、知识摘要、知识推荐等。

2. 知识资源整合、梳理与标准化 (核心基础工作)

(1) 建立采集机制：明确知识来源（业务部门、制度发文系统、产品管理系统、培训系统、历史案例库等）、责任人、采集频率（如新制度发布即触发采集）、采集格式要求。

(2) 大规模内容迁移与清洗：制定迁移规则，将历史有价值知识批量导入临时库。进行数据清洗、去重、格式转换（统一为 Markdown/ 结构化数据）、基础元数据补充（标题、作者、部门、生效日期等）。

(3) 构建知识图谱框架：设计核心实体（如产品、客户、流程、风险点、制度条款）及其关系，为后续智能应用打基础。

(4) 深度内容加工：对关键知识（如复杂产品说明、操作流程、制度要点）进行结构化拆解（FAQ 化、步骤化、要点化），提炼核心知识点。

(5) 统一元数据标准：定义强制性和可选性元数据字段（如知识类型、适用部门、适用岗位、密级、有效期、版本号），确保知识可管理、可检索。

(6) 质量评估与持续优化：明确知识的责任部门及责任人，负责内容的准确性维护和定期审核更新，结合用户检索问答后评分等机制的反馈，触发知识 Owner 的修订流程。

3. 统一知识库系统与大模型平台建设 (技术实现)

(1) 核心平台架构



(2) 核心功能模块

- 智能搜索：结合关键词、向量语义搜索、知识图谱关联，返回精准、相关、结构化的结果。
- 智能问答：理解自然语言问句，基于知识库和精调后的模型生成准确、简洁、有引用的答案；支持多轮对话澄清意图。

- 知识关联与推荐：根据用户查询或当前浏览内容，推荐相关知识卡片。
- 知识管理后台：知识录入、审核、发布、版本管理、权限配置、统计分析（知识热度、质量、用户行为）。
- 员工门户：统一搜索框、智能问答机器人、知识分类浏览、个人收藏、反馈入口。
- 系统对接模块：实现与行内统一身份认证、消息通知等系统的集成。
- 大模型监控与调优台：监控模型性能（响应时间、准确率）、用户反馈、安全风险（幻觉、敏感信息），支持持续迭代精调。
- 多渠道服务接入：独立 Web 门户、嵌入核心业务系统（如柜面系统、信贷审批系统）、对接客服工作台、企业微信 / 钉钉机器人等，同时提供标准 API 供行内各业务系统调用知识服务。

ID 三、创新点

1. 大模型驱动的知识服务范式升级

(1) 从“关键词匹配”到“语义理解与多模态生成”：突破传统搜索局限，精准理解员工复杂、口语化的业务问题（如“小微企业信用贷需要哪些材料且利率最低？”），直接生成结构清晰、要点明确的答案，同是结合音视频内容答复，而非仅提供文档链接。

(2) 主动式知识推荐：基于用户画像（岗位、业务场景、历史查询）和上下文，在业务流程关键节点智能推送相关知识和操作指引（如客户经理在录入贷款申请时，自动提示该产品最新制度要点和风险提示）。

(3) 多轮对话交互：支持追问和澄清（如“那这个流程和之前版本的主要区别是什么？”），提供拟人化的交互体验。

(4) 动态知识图谱与传统知识库的深度融合：将结构化、半结构化知识构建为动态更新的知识图谱，显式表达实体关系（如“产品 A” - 适用 -> “制度 B” - 关联 -> “风险点 C”）。大模型利用图谱进行深度推理和关联检索，提升答案的准确性和可解释性。

2. 全栈式、标准化的知识治理体系及运营机制

(1) AI 辅助知识加工：利用大模型能力辅助知识 Owner 进行知识摘要生成、FAQ 自动生成、内容润色、疑似错误提示等，大幅提升知识生产和维护效率。

(2) 用户反馈驱动闭环：用户对问答结果的点赞、点踩、修正建议直接反馈至知识质量评估体系及大模型优化流程，形成持续改进的飞轮。

(3) 从源头的采集规范、到内容的结构化 / 标准化加工、再到明确的责任人制度和质量评估体系，构建覆盖知识全生命周期的、可落地的治理框架，确保知识资产的长期价值。

ID 四、项目过程管理

1. 阶段划分与里程碑

(1) 启动 (1 个月)：立项、组建跨部门项目组（业务、科技、合规、风险）、完成详细需求调研与确认。

(2) 蓝图设计与规范制定 (2 个月)：完成知识管理体系规范（分类、流程、标准、安全）、系统架构设计、大模型 POC 验证。

(3) 知识梳理与系统开发 (5 个月)：并行开展，大规模知识采集清洗、结构化加工、图谱构建；知识库平台



与大模型引擎开发、集成测试。

(4) UAT 与上线准备 (1.5 个月): 用户验收测试 (UAT)、操作手册编写、全员培训准备、上线部署方案制定与演练。

(5) 试点上线与推广 (1 个月): 选择 1-2 个典型部门 (如客服中心、零售信贷) 试点运行, 收集反馈并优化, 随后全面推行。

(6) 持续运营与优化 (长期): 正式移交运营团队, 建立持续优化机制。

2. 组织保障

(1) 高层指导委员会: 由行领导牵头, 负责战略决策和资源协调。

(2) 核心项目组: 项目经理 (科技部)、业务负责人 (各主要部门代表)、知识管理专家、大模型算法工程师、系统架构师、数据治理专家。

(3) 知识 Owner 网络: 各部门指定专 / 兼职知识管理接口人。

3. 风险管理

(1) 数据安全与合规风险: 严格遵循数据安全规范, 大模型采用私有化部署或可信行业云, 问答结果不包含敏感信息, 建立内容安全审核机制。

(2) 大模型“幻觉”风险: 设计“引用溯源”机制 (答案需标注来源文档), 结合规则引擎和置信度阈值过滤不确定回答, 引导用户查看原文。

(3) 知识质量风险: 强化知识 Owner 责任, 建立严格的多级审核流程 (业务初审、合规风控复审), 持续质量监控。

(4) 用户接受度风险: 加强培训和宣传, 展示实用价值, 设立便捷的反馈渠道, 快速响应问题。

五、运营情况

1. 运营体系建设

(1) 组织: 成立专门的“知识运营小组” (归属科技部或办公室), 负责平台日常维护、用户支持、知识质量监控、培训推广。

(2) 流程: 知识更新流程: 知识 Owner 发起 -> 业务审核 -> 合规 / 风险审核 (如涉及) -> 知识运营小组发布。

(3) 用户反馈处理流程: 用户反馈 (错误 / 建议) -> 知识运营小组受理 -> 转知识 Owner 核查处理 -> 反馈结果给用户 -> 更新知识库 / 优化模型。

(4) 大模型监控与优化流程: 定期分析日志 (问答准确率、热点问题、失败案例)、收集用户反馈 -> 算法团队调整模型 / 规则 -> 测试验证 -> 发布更新。

(5) 工具: 利用后台的统计分析功能监控知识库活跃度 (访问量、搜索量、问答量)、知识质量评分、用户满意度。

六、项目成效

1. 效率提升显著

(1) 员工知识获取效率：平均搜索 / 查找知识时间下降 80%+（从平均 15-30 分钟降至 <3 分钟）。智能问答首问解决率 >85%。

(2) 知识生产与维护效率：AI 辅助工具使 FAQ 生成、摘要提炼效率提升 50%。

2. 成本节约与价值创造

(1) 人力成本：降低了员工因信息查找而产生的低效工作时间。

(2) 培训成本：新员工利用知识门户自学效率提升，上岗培训周期缩短 30%。

(3) 风险成本：因信息滞后导致的操作失误差错率下降。

3. 业务赋能与体验提升

(1) 决策支持：为前台营销、中台风控、后台管理提供实时、精准的知识支持，提升决策质量和响应速度。客户经理产品推介准确性和专业性获客户认可度提升。

(2) 员工体验：员工满意度调查显示，知识获取便捷性提升。成为员工日常工作的“智能助手”。

七、经验总结

1. 关键成功因素 (KSF)：高层坚定支持与跨部门协同，行领导亲自挂帅，打破部门墙，业务部门深度参与（需求、知识梳理、担任 Owner）是项目成功的基石。

2. “管理 + 技术”双轮驱动：完善的、可执行的知识管理体系规范是平台发挥价值的前提，先进的大模型技术是提升体验的关键引擎。二者缺一不可。

3. 数据治理与知识质量是生命线：投入充足资源进行知识梳理、清洗、标准化和持续的质量管控，确保“输入”的高质量，才能保证“输出”（智能服务）的可靠性。

4. “小步快跑，价值驱动”：采用敏捷迭代，优先聚焦高价值业务场景上线核心功能，快速见效，增强信心，再逐步扩展。

5. 挑战与应对：

挑战：初期知识梳理工作量大、复杂度高。

应对：采用“分阶段、抓重点”策略，优先梳理高频、核心业务知识；利用工具辅助（如 OCR、文本分析）提高效率；充分动员业务部门力量。

挑战：大模型在专业领域的精准度和幻觉问题。

应对：严格限定知识来源范围（基于权威文档）；深度精调模型；设计强力的“引用 + 置信度过滤 + 人工审核”机制，持续优化模型和规则库。

6. 结语

本项目的成功实施，不仅构建了一个智能化、便捷化的知识服务平台，更重塑了银行的知识管理文化和能力。基于大模型的统一知识库门户已成为支撑全行业务高效运转、提升员工能力、保障合规经营、优化客户体验的重要基础设施。其经验表明，将先进 AI 技术与扎实的管理基础、深入的业务理解相结合，是金融行业推进数字化转型、打造智慧银行的有效路径。未来将持续深化大模型应用（如知识推理、预测性知识推送），拓展知识服务的广度和深度，释放知识资产的更大价值。



扫码联系了解更多详情



领雁科技“智鉴”反洗钱智能体

北京领雁科技股份有限公司

一、智能体方案背景

在全球金融监管日趋收紧的背景下，反洗钱 (AML) 已成为金融机构维护金融安全与市场秩序的核心任务。然而，传统反洗钱机制高度依赖人工分析，不仅效率低下、成本高昂，且主观性强，严重制约了反洗钱成效。

随着大语言模型 (LLM) 在自然语言理解与推理生成方面的突破，反洗钱领域迎来了全新智能化转型契机。领雁科技打造的“智鉴”反洗钱智能体，基于金融领域指令微调模型，结合 RAG (检索增强生成) 架构与行业知识库，不仅提升了数据处理与分析效率，更显著增强了对可疑交易的上下文理解、逻辑链推演与报告生成能力，重构了传统“人审+规则”的作业范式。

二、“智鉴”：反洗钱智能体创新解决方案

当前银行反洗钱面临的核心挑战

挑战一：异常交易案例激增，人工不堪重负

金融机构普遍采用“案例特征化、特征指标化、指标模型化”的思路构建可疑交易监测模型。尽管监管要求模型持续优化，但受限于多种因素，模型准确率提升有限，导致系统每日产生海量异常交易案例。这些案例需经人工多环节审阅，即便采用集中作业，面对手法不断翻新的洗钱行为，平均每人每日约 30 笔的处理能力已远不能满足需求。

挑战二：数据整合匮乏，辅助严重不足

异常交易案例分析高度依赖对客户历史交易、风险评级等信息的综合研判。现实中，往往缺乏有效的计算机辅助手段来快速整合分散数据并进行深度识别。分析人员需不断拓展数据探查的广度和深度，对现有技术和工具提出了更高要求。

挑战三：人员经验不齐，识别效率低下

许多金融机构反洗钱案例甄别工作并非总能专人专岗，可能导致分析人员思路固化、手段单一。若缺乏丰富经验和多样化分析工具，便难以在海量数据中迅速锁定可疑特征，也无法高效地将当前行为与历史场景综合对比，影响风险识别的及时性与准确性。

“智鉴”核心业务框架

面对上述挑战，领雁科技推出的“智鉴”反洗钱智能甄别助手，依托大模型强大的语言理解与生成能力，“智鉴”可在结构化数据预处理基础上，自动提取核心交易要素与行为特征，并根据上下文关系生成逻辑严谨、层次清晰的案件初稿。通过“专家经验+提示工程+案例微调”，系统可不断吸收历史标注案例，实现持续优化与类比学习，显著减轻分析人员重复劳动强度。

“智鉴”的业务框架围绕智能化、精准化和高效化三个核心维度构建，旨在全面提升反洗钱工作的能力和水平。



图注：“智鉴”反洗钱智能甄别助手方案示意

1. 案例处理智能化

- 自动化流程：实现从初步分析、信息汇总到报告初稿生成的自动化，减少人工重复劳动。
- 人机交互：构建深度分析场景，分析师可利用助手对客户尽职调查及复杂行为信息进行深度挖掘。
- 长效训练：通过持续学习已处理报告内容，不断优化逻辑思维与分析方法，生成更高质量的分析报告。

2. 逻辑判断精准化

- 异常识别点分析精准化：快速定位关键信息，精准识别并汇总异常点对应的证据链路。
- 历史信息分析精准化：辅助分析人员在海量历史数据中快速发现潜在关联和例外情形，为风险判定提供更广泛深入的数据支撑。

3. 处理效率高效化

- 异常交易案例处理高效化：高效辅助人工获取初步结论，显著提升整体甄别效率。
- 可疑交易报告编写高效化：辅助生成结构完整、内容详实的可疑交易报告。

“智鉴”分析思路与价值提升



图注：“智鉴”分析思路流程



“智鉴”借鉴并优化人工甄别逻辑，通过结构化数据处理和智能分析实现高效精准判断。核心流程包括：通过构建“语义标签识别—知识图谱匹配—推理链生成”三阶段模型逻辑，“智鉴”能够识别案例中隐含的可疑行为模式，并利用大模型构建的多跳与证据合成机制，还原洗钱交易链条的潜在动因，极大提升可疑点分析的可解释性与可溯源性。

业务价值提升

序号	“智鉴”甄别助手优势	说明
1	提升人工甄别效率	传统模式日均处理约 30 笔，通过“智鉴”辅助，预估可提升至每日 50-60 笔。
2	扩充数据范围，提供强有力数据支撑	自动扩充至客户历史交易、行为等多维度信息，更有效判断客户行为异常性。
3	完备的人机交互模式	人机交互模式（基于 LLM）：分析人员通过自然语言与“智鉴”交互，利用其记忆和问答功能，完成复杂关联分析，形成符合专家逻辑的流程。
4	完整的可疑交易报告结论	综合多渠道信息，辅助生成逻辑严密、要素完整的可疑交易报告结论。

“智鉴”成果展现：张小伟案例剖析

案例背景：客户张小伟，身份信息完整。反洗钱系统生成其异常交易案例，主要特征为：对私客户短期内频繁收付款，对私账户当日连续多次一进多出、交易对手众多、日终账户余额基本为零，且全天无现金交易。

（一）客户身份信息及账户信息分析：

客户张小伟，证件类型：居民身份证，证件号码:622****233，证件有效期至2042年8月10日，职业为技术人员，联系电话为166****1023，联系地址为甘肃省兰州市城关区一只船街，基本信息完整。在我行开立了3个账户，分别为：

1. 账号 ACCT-01，开户日期为2023年7月5日，账户类型为个人活期结算账户，已开通手机银行，限额为 999999，目前状态为销户；
2. 账号 ACCT-02-01，开户日期为2023年12月10日，账户类型为个人活期结算账户，已开通手机银行，限额为999999，目前状态为正常；
3. 账号12/10/2023(疑似日期误填)，开户日期为2023年7月5日，账户类型为个人定期账户，未开通手机银行，目前状态为正常。

根据尽职调查情况：2024年2月28日下午3点30分拨打预留电话，电话接通，客户配合尽调。客户职业为三林科技有限公司员工，交易用途为开火锅店需开立对公账户，账户有一定流水信息，故客户让其朋友(雅丹、赵弘、诸葛原、尹大目等)帮忙转账做流水。

(二)资金交易整体情况分析:

张小伟在2024-02-22至2024-02-25期间,共发生交易52笔,总计590836元。从交易对手分析:付方交易对手23名,交易笔数31笔,共计291088元;收方交易对手6名,交易笔数18笔,共计299283元。

从交易类型分析:现金交易0笔,金额0元;转账交易52笔,金额590836元;POS消费交易0笔,金额0元;数字人民币交易0,金额0元。另外,日均交易笔数13笔,非柜面交易渠道占比100%,夜间交易占比1.92%。

(三)疑点分析:

根据提供的材料信息,结合反洗钱特征、历史案例以及人工分析内容,可以总结出以下疑点:

1.客户基础信息疑点

- 账户开立频率及状态:客户张小伟在我行开立了3个账户,其中1个账户已销户。短期内开立多个账户,尤其是销户后再次开立新账户,可能表明客户试图规避监管或掩盖真实意图。

- 账户功能不符:客户称其交易用途为开火锅店器开立对公账户,但其本人开立的。

- 客户让朋友帮忙转账做流水,这种行为可能进一步掩盖其真实资金来源和用途,增加资金链条的复杂性。

客户张小伟的账户交易行为存在明显的洗钱嫌疑,主要表现为短期内频繁交易、对手众多、快进快出、非柜面交易为主等特点。结合其挂失划卡和小额测试交易的历史行为,可以认为该客户可能涉及资金过渡或洗钱活动。建议进一步核实其资金来源和用途,并考虑提交可疑交易报告。

“智鉴”的未来展望与行业赋能

随着通用大模型向金融垂直模型不断演进,未来反洗钱场景将融合多模态识别(图像/语音)、跨语义链推理、多Agent协同处理等技术,构建具备自学习、可解释、可审计能力的智能审查体系。

在这一演进过程中,垂直领域的业务规则、专家经验与判断逻辑将成为智能体精准识别和决策能力的核心基座。只有将金融一线专家的知识与反洗钱监管要求系统化嵌入模型训练与推理机制中,智能体才能真正“懂业务、可落地”,避免“懂技术但不懂金融”的偏差风险。“智鉴”也将进一步探索专家知识与大模型融合的训练机制、跨系统对接的业务流适配路径,确保在安全合规前提下,智能反洗钱能力能够真正嵌入实际业务流程并持续演化。

领雁科技将持续深化人工智能大模型在金融领域的应用探索,致力于以AI之力助推金融科技创新。期望通过“智鉴”等智能化产品,为金融行业的稳健运营与持续成长保驾护航,共同迎接更智能、更安全的金融新时代。



AI 大模型在金融业务场景的应用实践

深圳市金证优智科技有限公司

随着人工智能（AI）技术的快速发展，AI 大模型正深刻改变金融行业的业务流程与服务模式。金融行业作为数据密集型领域，对技术的进步尤其敏感，天然适合大模型技术的落地应用。2024 年以来 AI 大模型已在金融行业多业务场景应用，成效显著，大幅度提升了工作效率和产能。

成立于 2018 年的深圳市金证优智科技有限公司（以下简称“金证优智”或“公司”）是一家专注于人工智能在金融行业应用的高新技术企业，是国内领先的金融 AI 服务提供商，产品在垂直领域应用优势明显。公司产品包括可私有化部署的金融大模型、金融问答系统、智能文档处理（IDP）、智能撰写，以及针对各种金融业务场景的智能应用产品和解决方案，涵盖信用业务、投行业务、风险管理等。

目前，金证优智的产品已经在金融监管机构、交易所、券商、银行、保险、资管、期货等多个领域场景应用，覆盖 100 余家客户。包括公安部、证监会、香港交易所、伦敦金属交易所、中信证券、国泰君安、申万宏源、华泰证券、平安银行等众多头部金融机构。以下是金证优智产品三个典型的应用场景：

1. 投资研究：智能化的投研

在投资研究领域，大模型能够快速收集和分析海量的财务及市场信息，为投资研究人员提供全面的数据支持。在研究准备阶段，大模型可以协助收集各类信息，包括非结构化的文档数据，将其转化为公司级的知识沉淀。在调研执行阶段，大模型可以根据机构的风格和对企业的初步了解，生成针对性的访谈提纲，并在调研后根据相关资料生成符合机构风格的会议纪要。在研究报告的生产和发布阶段，大模型能够根据搜集的材料完成撰写各种统计报告和深度投研报告，也可以生成适合不同受众群体的摘要或简讯，从而提高投研效率和质量。

2. 财富管理：全面、个性化的客户服务

在财富管理领域，AI 技术的推广和产品应用为投资者提供了更加个性化和高效的服务。智能体作为 AI 技术的重要应用形式，能够为投顾人员提供专家级的助手，整合资讯、行情、舆情等信息，为投资者提供综合服务。多智能体架构可以为不同类型的投资者提供定制化的服务，提高服务的覆盖面和质量，提高效率，平均可节约人工成本 70% 以上。此外，AI 技术还可以通过分析投资者的行为数据和偏好，为其提供个性化的投资建议和产品推荐，使服务更具有针对性，投资者的体验更好。

3. 风险管理：精准识别与风险控制

风险管理是金融行业的核心，大模型能够分析海量历史数据和实时市场数据，识别潜在的风险因素并提供预警和决策支持。例如，金证优智的信用风险评估产品可以通过对企业财务数据、行业趋势、业务内在逻辑等的综合分析，准确评估标的证券或企业的信用风险，支持金融机构提前进行风险预警和风险控制。同时，AI 大模型产品也能通过对市场行情的实时监测和分析，预测市场波动趋势，帮助金融机构制定更加科学的风险管理策略，有效应对市场风险。

金证优智具有平台化产品，可降低 AI 落地成本，赋能不同业务场景，产品的竞争优势体现在以下几个方面：

1. 文档解析更精准

公司深耕垂直领域场景，多源多格式文档处理能力突出，不仅能更精准识别、解析行业文档，还能输出具备行业认知的结果，真正满足业务需求。

(1) 具有完备的 IDP 技术体系。累计储备大量权威与规范数据，深耕行业多年；支持 PDF、图片、网页等多种格式，兼容多种复杂文档类型。

(2) 定制化行业算法，融合多种技术。针对行业语义，OCR / NLP 准确率远超通用算法；融合 LLM、多模态等最新模型算法，更好支持复杂版面分析、文字识别、表格识别、抽取等功能。

(3) 行业领先性。对垂直领域语义理解更深，实际应用准确率更高；处理复杂文档时表现稳定、可扩展。

2. 混合检索机制更强大

针对金融行业文档体量巨大、格式多样，如何有效建库及如何在多知识库间高效检索，保证准确率和召回率、传统搜索技术难以支撑大模型时代的 RAG 需求等行业痛点，公司产品具备完整的解决方案。

(1) 针对行业文档，设计最优向量化、切分与索引策略，异构数据接入+多数据源协同管理。进行知识库建设。

(2) 以深厚的搜索技术积累打造高标准检索体系，在 RAG 场景下更能“搜得准，答得好”，支持 ChatBI 数据库检索与数据分析；保证召回准确与响应效率，检索更准、更快、更易用。

3. 场景应用运营经验更丰富

金证优智不仅有清晰的知识库边界、完善的 bad case 管理机制，还具备多年行业项目实战经验，产品能落地、能运营、能长效升级。

(1) 建立知识库运营策略，明确文档更新与维护职责；不断扩充行业权威数据，实现动态更新。

(2) 结合业务场景，梳理 AI 需求；通过小步快跑、可视化评估的方式，管理用户期望。

(3) 提供标准化评测集与行业评估集；bad case 收集与持续迭代机制，保证模型精度不断提升。

4. 平台化研发 + 一站式交付

公司提供“全流程、低门槛”的平台化支持，通过自动化流程+行业专属语料库，让客户快速上线，并持续优化。

(1) 一体化平台端到端支撑，将数据清洗、标注、训练、评估、部署一次性打通；自带运行环境+高端装备制造行业知识库+通用/行业评估集。

(2) 配置化 RAG 机制、模型精调、自带通用语料和行业语料，实现模型训练、知识库构建、评估等流程可自动化，缩短迭代周期，并可以快速产出实验结果，自动化评测和对比。

(3) 无需处理各种开源组件对接，专注业务实现与迭代，实现“从玩具到工具”的进化。

人工智能技术的发展和應用必将推动金融行业深刻的变革，大模型作为 AI 领域的重要突破，为行业的智能化转型提供了强大的技术支持，将为金融行业带来更加智能化、高效化和个性化的服务体验。

金证优智始终坚持以技术创新为核心，以客户需求为导向，致力于推动 AI 技术在金融行业的应用和发展，将与对行业的理解深度融合推出专业化、标准化的产品，与金融机构携手推动行业的技术发展，共同迎接智能化金融时代的到来，同时金证优智也会紧跟技术发展趋势，保持行业领导者的地位。



跬智信息 Data Agent: 首个企业级 自主深度推理和洞察产品在金融业的落地实践

上海跬智信息技术有限公司

在人工智能深入发展的当下，金融企业的数据价值正加速从“存量资产”向“实时能力”演进，智能决策正成为组织提升经营质量的关键能力。无论是营业网点、手机 App，还是业务条线与后台运营，组织中的每一项决策——无论大小——都不再依赖经验拍板，而是需要基于数据的支持与智能的引导。

尤其是在与销售业绩、客户运营、风险控制等高度相关的场景中，管理者往往面临以下挑战：一方面，难以及时获取完整、可信的数据支撑，判断易延迟；另一方面，一线员工常常错失关键客户意图信号，科技与业务之间难以形成有效联动。随着金融业务体量增长，数据量激增、系统割裂、协作链条拉长，进一步加剧了决策迟缓与响应失真。

而今，AI 的快速发展正在重构这一格局。相比以往依赖复杂流程、专业分析师和专属工具才能实现的“数据赋能”，基于 Data + AI 的融合体系，正使数据价值实现从“可见”向“可用、可行”跃迁。它不仅极大降低了数据使用门槛，更显著提升了组织的管理效率与决策速度，真正拓宽了企业的“管理半径”，释放人效与系统协同的潜力，推动金融机构实现从“信息获取”向“智能行动”的升级跃迁。

ID 从数据到决策：金融业务智能化的三大断层

1. 信息散、节奏慢：经营管理难以统一指挥

面对不断变化的市场环境，金融企业对高效响应的依赖越来越强。但管理者在做关键决策时，往往难以一眼掌握全局。数据分散在多个系统、多个角色之间，信息上收流程冗长，反应速度明显滞后。想快速判断一个分支机构的业绩走势，仍需层层提数、逐级确认，管理节奏被动，极易错失调整窗口。

2. 分析难、归因弱：运营效率与人效提升无从下手

一线业务执行节奏快，但数据反馈节奏慢，管理者无法真正“看见问题”。当指标下滑时，无法准确判断是团队执行力不足，还是市场结构出现变化，还是客户画像偏离。即便企业配备了数据工具，也难以实现多角色、多业务环节之间的高效协同与自动分析。归因困难，分析脱节，直接影响人效提升与资源配置优化。

3. 判断慢、执行断：数据到决策存在结构性脱节

很多企业在月度或季度经营决策上投入巨大，会议前紧急调数，会议中反复讨论，但会后方案却难以迅速落地。原因在于：从数据分析到业务执行之间，缺乏系统性的闭环机制。数据只是“参考”，洞察无法驱动行动，系统不能自动响应——结果就是“做了决策，没跟上行动”，经营效率在不经意间被拉低。

总体而言，这三大断层使得金融企业“有数据却难用、能决策却难落地”。要提升管理效率，关键在于打通从数据到洞察、从判断到执行的智能闭环。

ID 从“人传人”到智能决策，跬智信息以“三步走”打造金融业务实战型 AI 能力

为破解“数据看得到却用不动、分析做得出却落地不地”的管理难题，跬智信息（Kyligence）率先构建了 AI 数据智能能力分级框架，通过“三步走”逐步实现从基础问答支持，到主动洞察推理，再到自动生成与闭环执行的全面升级，推动企业从“人找数据”迈向“数据找人”的智能化跃迁。

目前，Kyligence 的 AI 数据智能体，具备数据检索、可视化分析、波动归因、逻辑推理与报告生成等综合能力，全面支撑企业实现实时、精准、自动化的业务决策。

集成企业数字资产，打造数据分析 AI 智能体，引领&释放潜在价值

为用户提供数据检索、可视化、归因洞察、总结推理，生成报告等的数据分析综合能力



对话式分析，让一线业务“开口即得数”

通过自然语言交互，AI 数据智能体能理解用户意图、保持上下文记忆，并自动联动多系统数据，生成可视化图表与结论摘要。例如用户仅需一句话：“查询团队和销售代表业绩趋势”，系统即刻给出分析结果。实践中，某区域的工单量减少了 70%，一线人员自主调取数据的比例提升至 72%——组织的“反应力”正被重新定义。

数据波动自动归因，管理不再拍脑袋

借助指标血缘识别与多维度深度扫描，AI 数据智能体能对业绩异常快速定位根因，从客户结构、团队执行到外部变化进行自动归因。原本需要 2~3 天完成的归因分析，如今仅用 20 分钟即可交付，真正实现从“经验拍板”转向“数据驱动决策”。

自动生成战报，助力策略评估与落地执行

区域经理只需输入：“生成 Q3 各区域销售分析报告”，系统便自动完成包含动态图表、对策建议与战略评估的完整材料，并同步挂载至督办系统，实现从分析、输出到执行的一体化联动。原需 5 天的流程，如今压缩至 2 小时——效率提升的背后，是组织结构与工作方式的重塑。



总体来看，跬智信息通过搭建从数据到行动的数字化管理机制，帮助金融企业实现更敏捷、面向未来的管理方式。系统每天会自动推送关键信息预警，比如“昨日外呼中静默时长超标的销售人员”或“续签率异常需重点关注的员工”，让管理从过去的事后复盘，转变为实时预警、即时干预，把管理重点真正落在业务一线，实现高效、精准的运营。

ID 从被动到主动，跬智信息推出最新智能决策利器

4月，跬智信息推出业内首款具备自主推理与洞察能力的产品——Kyligence DeepInsight（AI深度洞察）。该产品融合 DeepSeek 推理框架与 Kyligence 的专利多维分析技术，高效检索、精准分析，并结合企业知识体系和管理规范，提供专业、全面的决策支持。产品不仅优于传统专家分析，且部署快速、成本可控，被业内称为“真正符合企业级需求的大模型应用”。

Kyligence DeepInsight 独特优势在于多层次下钻的分析能力，不仅能对关键业务指标进行深入归因，还可结合大语言模型整合外部数据（如天气、竞品、行业趋势），为企业提供前瞻性的决策支持。依托统一语义定义与大模型计算等专利技术创新，DeepInsight 还能进一步生成涵盖 6 个以上维度的行动建议，助力企业实现更高效、可靠的决策。

经过验证的AI落地经验 + 实施方法论

快速交付能力

“我AI用”MVP项目历程

2周内快速迭代交付



近期，Kyligence 助力某全国性综合保险集团，在春节后基于 DeepSeek 快速落地了一项业绩管理与提升 AI 项目。仅一名工程师在 2 周内完成 4 个数据源的接入，处理上千万条数据，并搭建 100 多个 AI 智能体，成功赋能近千名一线管理人员，大幅加速该集团的 AI 战略落地。使用过程中，Kyligence 产品自主识别了团队及具体成员在关键指标上的提升空间并推送可执行方案。当某组 / 成员新单下滑时，系统不仅预警“周环比降 15%”，更精准定位“因高端客户触达频次不足”，并自动生成《高净值客户激活清单》和 OA 跟进任务，让管理者第一时间获得“懂业务”的决策支持。

凭借领先的数据与人工智能技术及扎实的业务能力，跬智信息（Kyligence）先后荣获国家高新技术企业和上海专精特新企业等称号。在“卡脖子”的关键技术领域，跬智信息积极推动国产替代，针对 Oracle Essbase、IBM Cognos、微软 SQL Server Analysis Service 等国外软件，提供业内最完整且先进的国产替代方案，已成功应用于多家大型国内金融机构。

目前，跬智信息 Data Agent 已广泛覆盖金融、零售、制造、医药等多个行业。凭借多年丰富的企业级服务经验，团队始终聚焦 AI 产品的“可靠、安全、可控”三大核心能力，作为技术优化与功能完善的基准。通过一年内三次重大升级及数十次小版本迭代，这三大能力不断趋于成熟，推动产品持续进化。



云杉：DeepFlow 金融核心业务系统可观测性及运维智能体建设方案

北京云杉世纪网络科技有限公司

一、方案概述

随着金融数字化转型的深入推进，银行、证券、保险等金融机构的核心业务系统正面临前所未有的复杂性挑战。云原生改造、分布式架构迁移、信创适配等多种技术路线并存，导致传统运维手段已无法满足金融级稳定性要求。

DeepFlow 基于“可观—可控—全自动”一体化运维新范式，通过零侵扰的全栈可观测性技术和 AI 智能体，为金融核心业务系统提供从被动响应到主动保障的全方位解决方案。该方案已在人保集团、国泰君安等头部金融机构成功落地，实现了故障平均恢复时间（MTTR）降低 70%、系统可用性提升 0.42 个百分点的显著成效。

二、金融行业的“完全可观测性”困境

2.1 为什么金融核心系统难以实现完全可观测？

根据《人工智能：现代方法（第四版）》，“完全可观测的环境很容易处理，因为智能体不需要维护任何内部状态来追踪世界。”然而，金融核心系统恰恰面临着可观测性的重大挑战：

灰盒化运维：应用拓扑复杂，组件多样，传统 APM 工具插桩困难，监控盲点与误报频发；

监控工具割裂：监控、日志、告警、追踪各自为阵，存在严重的数据孤岛问题；

运维经验传承难：依赖个人经验的“人海战术”模式难以标准化和规模化；

未知故障响应不足：常规 SOP 仅覆盖已知故障场景，对突发事件缺乏针对性策略。

2.2 金融监管与业务要求

监管合规：满足金融监管对系统可用性、数据安全、审计追溯的严格要求。

业务连续性：核心交易系统要求 99.99% 以上可用性，故障恢复时间窗口极短。

成本优化：在保障稳定性的前提下，优化 IT 资源利用率，降低运维成本。

三、DeepFlow 破局之道：三大原创技术

3.1 零侵扰采集技术——让可观测性“零负担”

通过融合 eBPF、cBPF、Wasm 等前沿技术，实现对大规模分布式系统的零侵扰数据采集。这项发表于 SIGCOMM 2023 会议的技术，解决了传统监控需要修改应用代码的难题，让企业能够在不影响业务的情况下获得全栈可观测性。

技术优势：

- 无需修改应用代码，避免影响业务性能

- 支持热加载，无需重启服务即可部署
- 覆盖应用、系统、网络、存储全栈数据

应用场景：

- 云上云下全景覆盖，支持信创环境
- 自动发现服务依赖关系和 API 调用链路
- 实时采集性能指标、日志、事件等多维数据

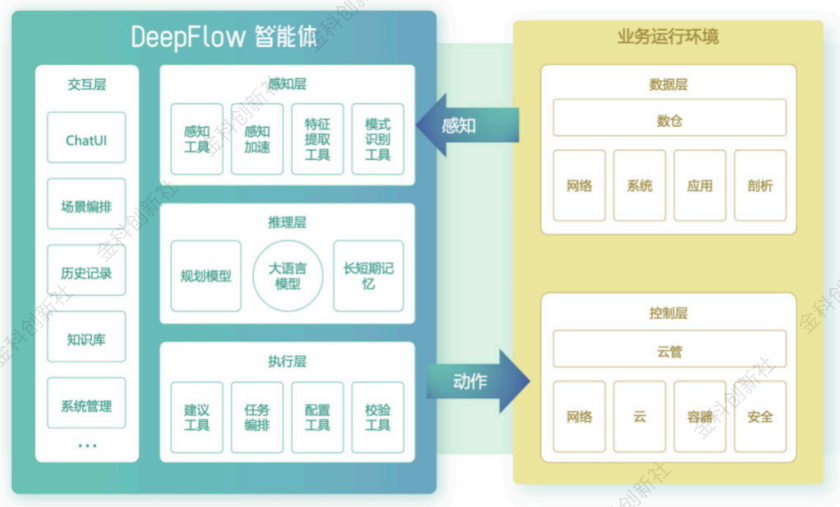
3.2 思维链状态机技术——让 AI 推理“不出错”

通过思维链（Chain of Thought）指引，可以解决由大模型推理带来的幻觉问题。然而，随着业务和场景的不断变化，思维链的复杂性呈指数级上升。通过使用基于 DFA + NFA 的混合状态机技术，可以有效解决思维链复杂性带来的状态空间膨胀问题。

3.3 自适应感知技术——让资源消耗“最优化”

自适应感知技术实现了推理前感知和推理中感知的混合感知技术。推理前感知技术包括对数据的实时特征提取和分类，推理中感知技术则结合业务场景对数据进行按需的特定特征提取和分类。自适应感知技术使得用户可以在成本和性能之间进行不断优化，避免不可控的算力资源投入。

四、DeepFlow 智能体：实现金融运维的“自主化”



图：DeepFlow 智能体架构图

4.1 智能体的三大关键要素

DeepFlow 智能体通过三大关键要素实现金融运维自主化：

完全可观测的环境：基于零扰动技术实现全栈数据采集，为智能决策提供完整的环境感知。



可靠的推理机制：采用思维链状态机技术，确保 AI 推理的准确性和可控性，避免幻觉问题。

量身定制的场景方案：针对金融行业特点深度优化，覆盖分钟级根因分析（1-5-10 快速响应）、7×24 不间断巡检（主动发现隐患）、一句话问数（自然语言交互）三大核心场景。

通过“感知—推理—学习”的闭环机制，DeepFlow 将传统依赖人工经验的被动运维模式，转变为 AI 驱动的主动保障模式，真正实现了让机器像资深运维专家一样思考和决策，为金融核心业务提供智能化、自主化的运维能力。

4.2 核心应用场景：创造量化的业务价值

1. 分钟级根因分析

基于多维数据实时关联和故障模式匹配，通过知识图谱建立症状、原因、解决方案的关联关系。当故障发生时，智能体自动执行：1 分钟定位异常组件，5 分钟溯源根因，10 分钟提供修复方案。某银行核心支付系统故障案例中，快速识别 SSL 握手异常并追溯到安全组误删避免了数千万元损失。

2. 不间断巡检

采用时序数据建模和预测性分析，对业务健康度进行全天候监控，主动发现性能退化趋势和潜在隐患。通过故障传播分析和资源依赖分析，实现风险早期预警。某跨境支付系统通过持续巡检，及时发现并跟踪了 API 超时从 5 次递增至 271 次的趋势，预防了 900 万美元交易的潜在损失。

3. 一句话问数

基于多源数据实时整合和自然语言理解技术，通过智能化特征提取和复杂查询自动编写，实现秒级响应。消除技术门槛，支持“当前 tps 余量”、“预测资源瓶颈”等自然语言查询。某券商在开盘前 12 分钟通过询问“系统容量预测”，8 分钟内完成 Redis 扩容，确保了万亿成交量下零故障运行。

ID 五、建设实施路径

5.1 第一阶段：全栈可观测性建设

目标：解决“看不见”和“看不清”的问题

- 部署 OneAgent 采集器，覆盖核心业务系统
- 构建统一的可观测性数据平台
- 实现应用性能指标、分布式追踪、日志聚合
- 建立业务黄金指标体系和告警规则

5.2 第二阶段：混沌工程与稳态建模

目标：构建“已知问题快速处理”能力

- 基于混沌工程验证系统稳定性
- 建立核心场景故障模型库
- 生成标准化应急处置预案
- 构建自动化故障演练体系

5.3 第三阶段：智能体赋能

目标：实现“面向未知情况”的智能处理

- 部署 DeepFlow 智能体
- 定制金融场景推理模型
- 实现故障自动诊断和修复建议
- 建立持续学习和优化机制

六、成功案例：金融行业的最佳实践

6.1 国泰君安证券：基于大模型驱动的云网可观测智能体创新

案例概述

国泰君安证券在数字化转型过程中大规模部署 AI 大模型训练和推理业务，但面临 GPU 利用率不足 50%、性能瓶颈定位困难、智能算力监控盲区等挑战。传统工具如 NVIDIA Nsight 无法提供 CPU 函数调用栈，PyTorch Profiler 需要侵入式插桩且性能影响大。为解决这一难题，国泰君安与云杉联合开展金融科技创新研究，构建基于大模型驱动的云网可观测智能体。

应用成果

项目通过四大创新实现突破：一是利用 eBPF 技术实现零侵扰的 GPU 持续剖析，无需修改代码即可热加载部署；二是首创 CPU+GPU 全栈追踪能力，通过 Frame Pointer、DWARF 等机制获取完整调用栈；三是部署智能分析 Agent，使用公司自有通义大模型进行故障智能分析；四是实现毫秒级性能瓶颈定位。

- 部署云网可观测智能体 Agent，实现 GPU 资源池 100% 监控覆盖
- 利用 eBPF 技术实现零侵扰的 GPU 持续剖析能力
- 开发 CPU+GPU 全栈持续剖析能力，解决大模型训练中的性能瓶颈定位难题
- 基于自有通义大模型构建智能分析 Agent 总结

该项目开创了金融行业智能算力可观测性的先河，不仅解决了大模型应用的性能优化难题，更为金融科技创新提供了坚实的技术保障。通过原创的零侵扰剖析技术和 AI 驱动的智能分析，实现了从“看不见 GPU 瓶颈”到“精准优化每个函数”的跨越，为行业树立了 GPU 资源高效利用的标杆。

6.2 人保信息科技：构建集团级“可视—维稳—智能”一体化平台

案例概述

人保信息科技作为人保集团全资科技公司，承担着为集团及财险、寿险、健康险等多个子公司提供 IT 服务的重任。面对万量级容器云主机的复杂环境，各子公司因业务特点差异导致云原生改造、分布式架构、传统迁移等多种技术路线并存，原有监控工具各自为阵形成数据孤岛，跨团队协作效率低下。为破解这一困局，人保科技基于 DeepFlow 构建“可视—维稳—智能”三阶段运维体系。

应用成果

项目分三阶段推进：第一阶段通过 eBPF 零侵扰技术实现万量级节点全链路透视，将追踪覆盖度提升 5 倍，彻底解决“看不见”问题；第二阶段结合混沌工程构建近 50 类场景稳态模型，生成标准化应急预案，实现已知问题



快速处理；第三阶段部署运维智能体，基于评测与反思的强化学习，实现 7×24 持续巡检和故障自动恢复。

- 故障恢复时间 (MTTR) 降低 70%，实现 1-5-10 分钟快速恢复；
- 系统可用性提升 0.42 个百分点达 99.95%；
- 告警噪声率减少 60%；
- 链路追踪覆盖率从 20% 提升至 100%；
- 年节省千万级基础设施投入，运维人效提升 30%。

总结

该项目成功验证了“数据 + 算法 + 自动化”三位一体的新型 IT 生产力范式，为保险行业构筑了竞争新优势。通过 AI 驱动的全栈可观测性体系，人保科技不仅大幅提升了集团 IT 运维效率，更重要的是形成了可复制、可推广的智能运维标准，加速了保险业务的数字化、智能化进程。

ID 七、总结

DeepFlow 金融核心业务系统可观测性及运维智能体方案，通过将原创的零侵扰采集、思维链状态机、自适应感知等技术与可观测性场景融合，为金融机构提供了从“看得见”到“管得住”再到“自动化”的完整智能运维能力。该方案不仅解决了当前金融 IT 运维的痛点，更为未来智能化运维转型奠定了坚实基础。

选择 DeepFlow，让金融核心业务在数字化转型的征途上行稳致远。

云杉简介

云杉成立于 2011 年 12 月，为企业提供自主研发的可观测性产品，消除 IT 系统运营的技术代差，保障客户核心业务的稳定与自主运行。

特别鸣谢以下单位对本案例集的支持

 知化时代

 领雁科技
Leadingtek

优智 Youzhi

 KYLIGENCE®

 YunShan

 金科创新社
Fintech Innovation in China



办公地址: 北京市朝阳区北苑路170号凯旋中心C座1501

电子邮箱: gaoyangchun@fintechinchina.com

企业官网: www.fintechinchina.com