AI对抗与基准评测平台

为银行内部提供统一的大模型测试与对比环境

项目背景

当前痛点

银行在智能问数、客户咨询、贷款审批、营销沟通等场景中,选型多依赖厂商演示和小规模人工抽测,缺乏统一标准,难以评估幻觉风险和真实业务效果。

解决方案

建设统一的**AI**评测平台,通过"回合制"自动对话和任务执行,统一评测不同模型及智能体的关键指标,支持**AB**测试。

服务对象与核心价值

业务条线

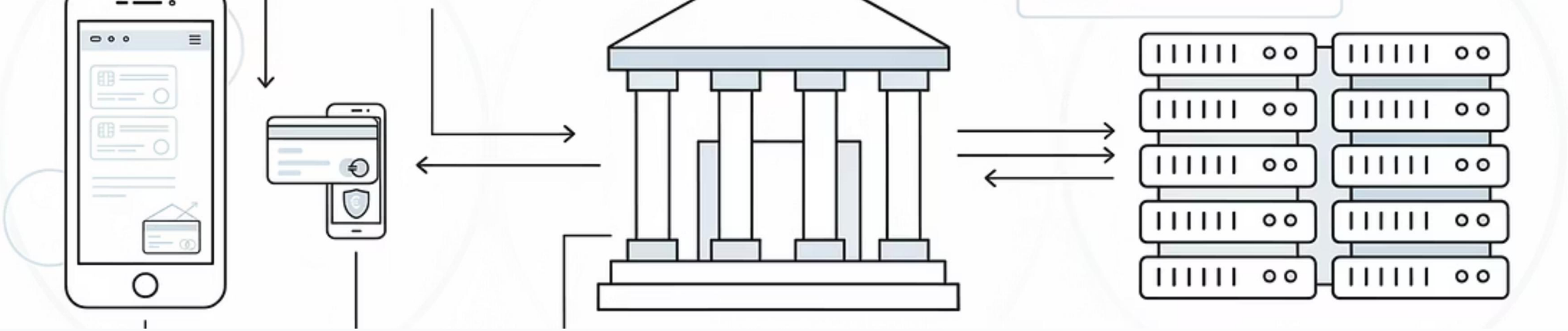
零售、对公、信用卡、理财、客服中心等,获得统一的选型、验收和持续评估工具。

科技团队

标准化接入和评测平台,避免每个项目单独搭环境、写脚本。

风控与合规

量化评估幻觉、违规话术和决策可追溯性,提供可解释的评估证据。



市场现状与发展趋势

当前银行在智能客服、智能问数、智能审批、智能营销等方面投入快速增加,但评测方式仍以厂商**Demo**、小规模人工抽测和通用大模型榜单为主,缺乏统一口径和场景化评估。

1

现状

各项目自建评测小工具,通用能力评分

2

趋势

全行统一评测底座,业务适配度评估

核心竞争优势

1

业务抽象能力

将智能问数、客服、审批、营销等流程转成可重复执行的标准测试场景。

2

评测方法体系

量化幻觉率、任务完成度、多轮对话稳定性、合规命中率等关键指标。

3

技术与安全

支撑多模型接入、稳定算力,满足金融级数据安全、隔离与审计要求。

4

中立性公信力

独立于单一厂商,结果在行内形成共同认可的标准。

商业模式与运营服务

平台定位

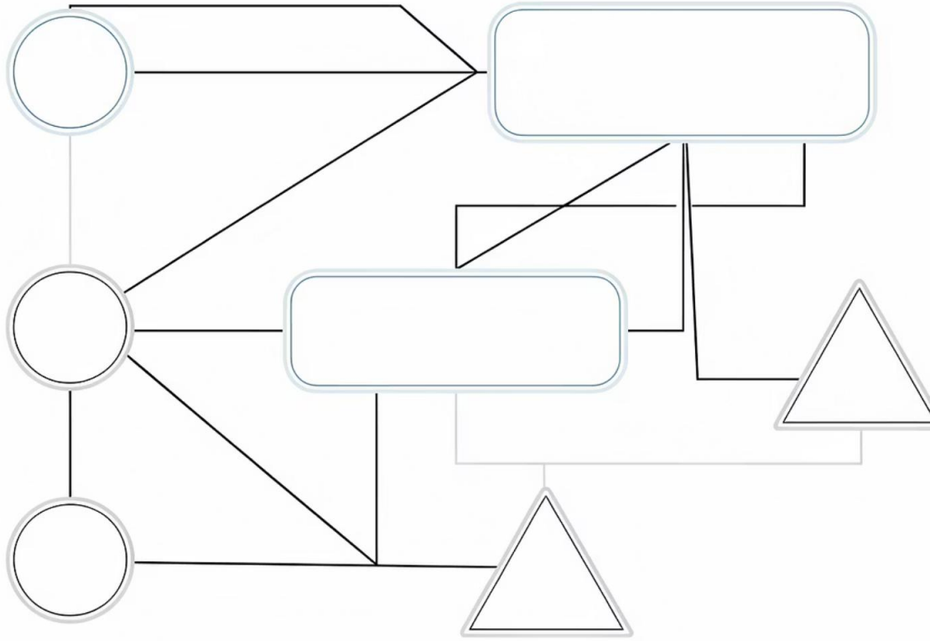
打造"全行级AI基准评测平台",作为统一的模型选型与效果验收基础设施。

服务方式

- 项目评测服务:提供评测方案设计、执行与报告
- 自助评测平台:支持自助配置场景、上传测试集

价值实现

- 用统一评测标准支撑模型集采与对外合作谈判
- 通过AB测试提升上线效果,减少返工
- 周期性评测发现模型效果衰减和潜在风险



商业模式图

客户细分

业务条线，科技团队，风控合规

价值主张

统一评测标准，量化风险，持续优化

渠道

项目评测服务，自助评测平台

客户关系

场景设计，定期报告，技术支持

评测闭环流程

01

场景建模与测试设计

与业务、风控、合规联合梳理典型流程,形成测试脚本和标准答案。

02

模型与智能体接入

通过统一网关接入行内外大模型、**RAG**应用和智能体。

03

回合制对抗评测执行

引擎自动驱动脚本,与各模型多轮对话与任务交互。

04

指标计算与对比分析

计算正确率、幻觉率、合规命中率、响应延时等,生成对比报表。

05

优化与回流

根据结果优化模型选型、提示词与策略,回流样本库。

核心技术架构

多层评测架构

基础能力评测、场景化流程评测、提示词与策略**AB**评测三层体系。

正反向样本库

沉淀高质量回答与问题案例,支持一键加入新问题,持续迭代升级。

回合制对抗引擎

支持多角色、多轮对话和不同客户画像,评估关键节点稳定性。

指标与报告引擎

可配置指标体系,自动生成项目级、模型级、场景级报告。

安全与合规机制

数据脱敏、最小必要数据集管理、全量审计日志与访问控制。

核心团队

胡刚

CEO

原**IBM**软件服务部华东区负责人,近**20**年企业级市场服务经验,曾服务工商银行、人民银行等标杆客户。

梅菊花

CTO

华中科技大学硕士,主导研发核心企业级开发框架,已在外汇交易中心、东方证券等项目落地。

张逊

项目总监

原**IBM**软件服务华南区项目负责人,曾服务中国银行、兴业银行等,擅长解决高复杂度架构难题。

王鑫阳

AI负责人

负责平台**AI**技术底座构建,致力于将前沿**AI**技术与客户业务场景深度融合。

产品核心优势

业务导向

从"测模型能力"转为"测业务好用程度",直接支撑立项、选型和验收决策。

技术优势

多层评测+回合制对抗+样本库迭代,评测内容随业务演进而升级。

安全部署

支持行内专有环境部署,与现有安全、审计、脱敏体系衔接。

流程管理

统一指标体系、评分规则和报告模板,评测过程全程留痕可追溯。

项目进展与落地

平台已完成第二轮迭代,具备场景管理、回合制对抗引擎、指标与报告模块、正反向样本库等核心能力,正在部分金融机构试运行。

1

头部证券公司

智能问数与投顾问答场景评测,完成多模型对比和策略**AB**测试

2

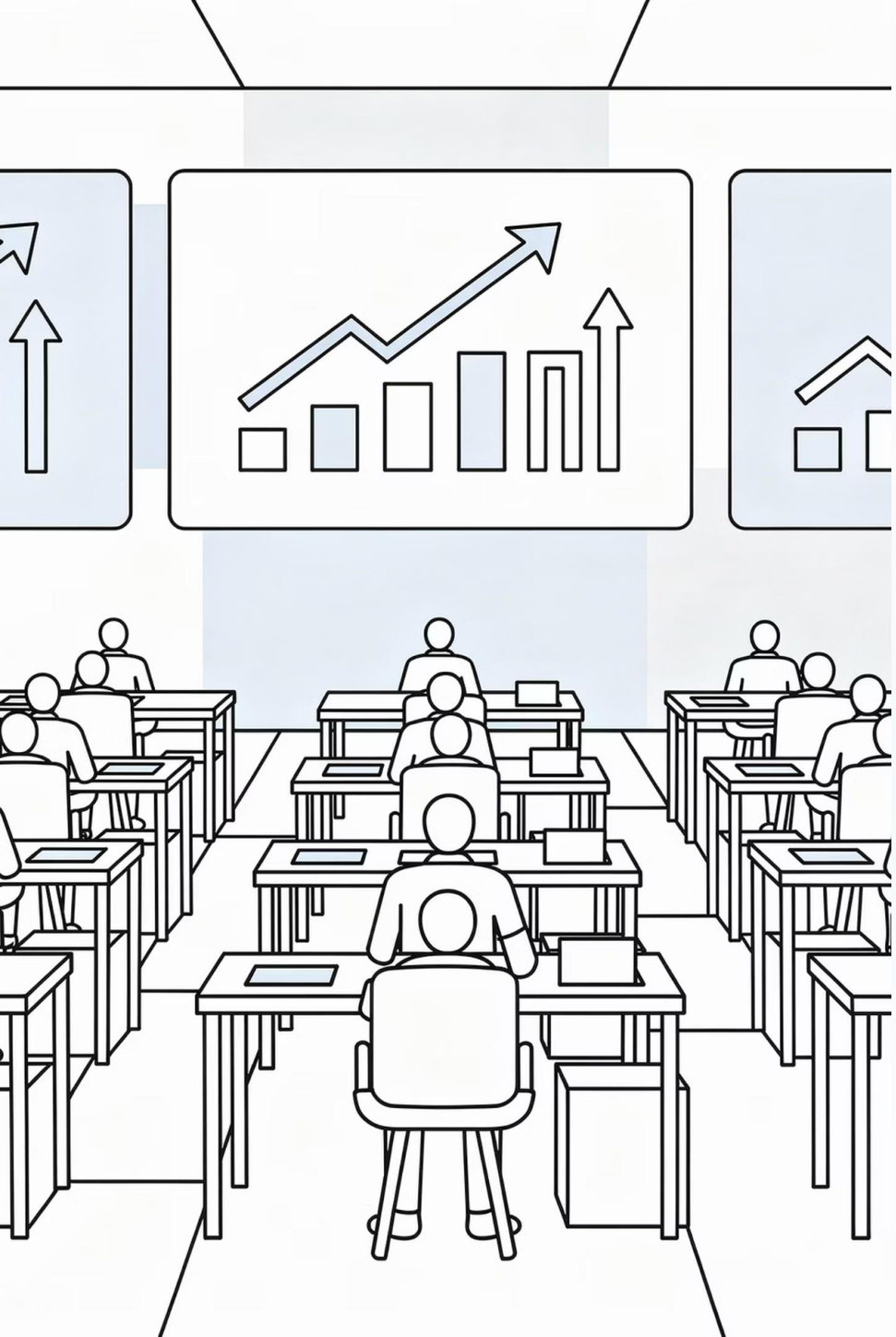
股份制商业银行

智能客服与信贷审批机器人评测**POC**,用于多厂商模型选型

3

城商行/互联网银行

零售营销话术、智能外呼场景样本共建和方案评审



典型案例:某头部证券公司

项目背景

该券商计划上线智能问数与投顾助手,需要在多家大模型和不同投顾话术策略中选出既专业合规又稳定的方案。

实施内容

- 建立覆盖开户、交易规则、产品咨询、投顾说明等场景的标准测试集
- 对多家模型进行多轮对话评测,考察正确率、幻觉率、合规命中率等
- 针对不同提示词和话术策略进行**AB**测试

20%

正确率提升

智能问数场景正确率提升约**20%**

显著

合规改善

违规和不当表述明显下降

发展规划

近期:做好行内"AI质检中心"

建立清晰的评测流程、模板和报告格式,让业务和科技同事"拉得动、用得上"。

中期:把能力做成标准产品

将核心能力沉淀成标准化平台功能,降低使用门槛,支持自助评测。

远期:形成行业影响力

参与行业标准、评测规范制定,升级为"金融AI评测标杆方案"提供方。

年度发展目标

1 服务好重点项目

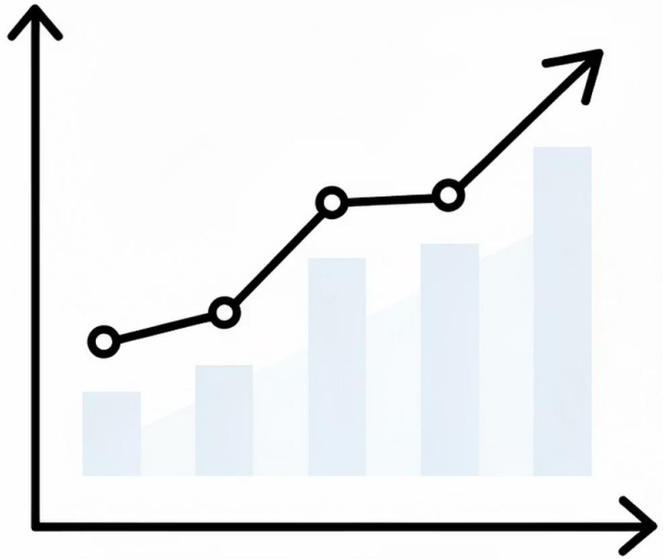
年内重点服务不少于**3**条业务线,为其**5**个以上**AI**项目提供正式评测报告,用于立项、选型或验收。

2 沉淀可复用场景模板

年内完成**10**个以上金融典型评测场景模板,形成"拿来即用"的场景库。

3 提升评测效率和体验

推出自助评测功能,评测配置和执行时间较当前人工方式缩短**50%**以上。



合作诉求

我们希望借助兴业银行及大赛联办单位的资源和专业能力,共同把"金融AI评测"这件事做深、做实。

1

联合试点

选择若干真实业务场景作为联合试点,使用本平台完成模型评测与选型。

2

标准制定

共同制定适用于兴业银行的AI评测指标和报告模板,为后续项目推广提供统一标准。

3

合规对接

在评测数据脱敏、安全审计等方面对接兴业现有体系,形成可在全行推广的合规方案。