面向生产服务的大模型评估体系探讨

邓超

中国移动研究院 2024年1月

人工智能产业链联盟

星主： AI产业链盟主

知识星球

微信扫描预览星球详情

内容提纲

中国移动通专大模型体系

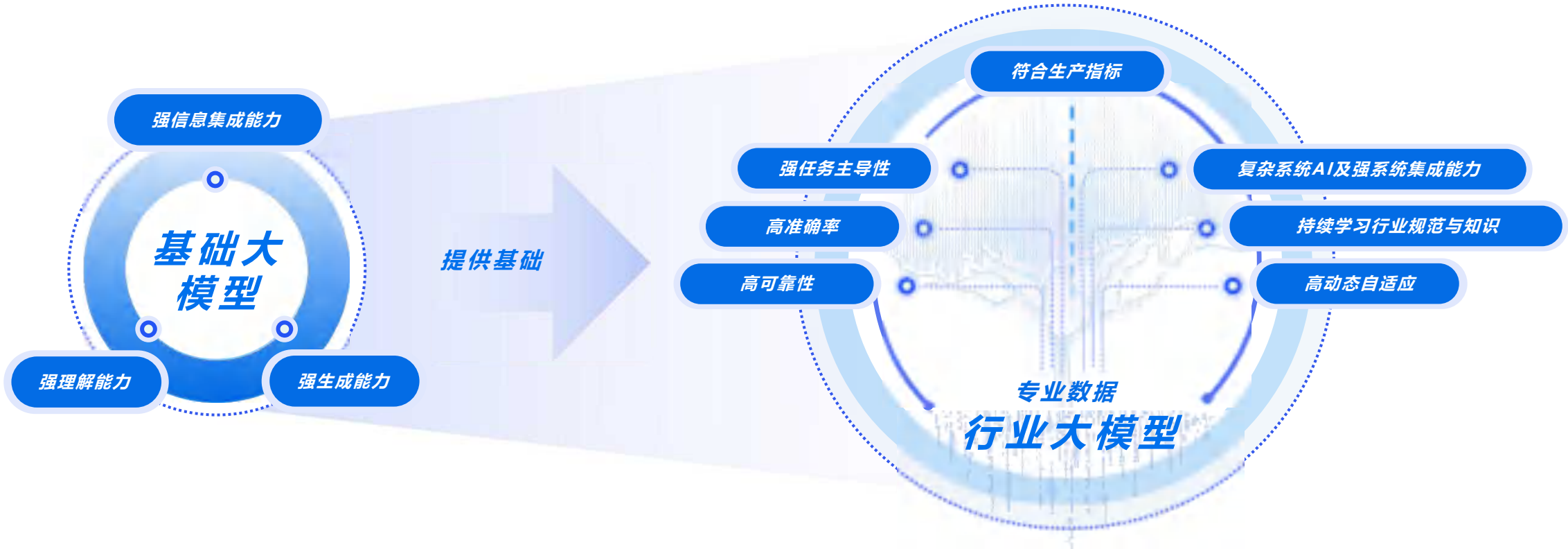
中国移动“九天”基础大模型：既能“作诗”，更能“做事”

中国移动自主构建语言、视觉、语音等多种类型大模型，具备跨行业供给侧增强、高可控性、异构软硬件灵活部署几大显著的技术特色，整体性能指标实现国内主流水平，能更好满足企业全场景全部署的大模型落地需求

联合骨干企业，共建共享九天·众擎基座大模型

以九天基础模型为基础，联合通信、能源、航空等行业的骨干企业，共建共享九天·众擎基座大模型，加速国民经济主体行业的智能化转型升级，促进我国战略性新兴产业发展，带动我国整体生产力提升

面向生产服务的九天行业大模型建设理念：“+AI”向“AI+”转型



客服	政务	医疗	交通	时空	家庭	物联网	社会综治视觉	网络运维	网络运维	文体	行业通话	金融风险	储能	司法

九天·网络大模型

构建网络AI大模型，实现从“网络+AI”向“AI+网络”转变，降低AI赋能网络的边际成本，指数级扩大赋能成效

为网络智慧内生提供AI核心基座，助力网络与AI全面、深度融合

- 2023年中国移动合作伙伴大会上发布网络大模型1.0，优先服务四大场景，驱动向“AI+网络”全面演进
- 基于网络大模型的网络运维AI助手正式上线中国移动MOA网络运维中心2个应用场景，端到端准确率达88%以上

九天·海算政务大模型

九天·海算政务大模型是中国移动基于近年来积累的丰富数字政府建设经验所打造的面向政务领域的行业大模型。

九天 · 海算政务大模型面向政务领域特殊性，融合了三大特色：深度行业智能、政务信息场、多元式交互

九天•海算政务大模型：已服务数字政府一网通办和一网统管

2023年世界人工智能大会上，发布了九天•海算政务大模型，已落地应用于黑龙江省数字政府项目政务智能客服、智能搜索、数字人、公文辅助写作等应用场景中的落地验证

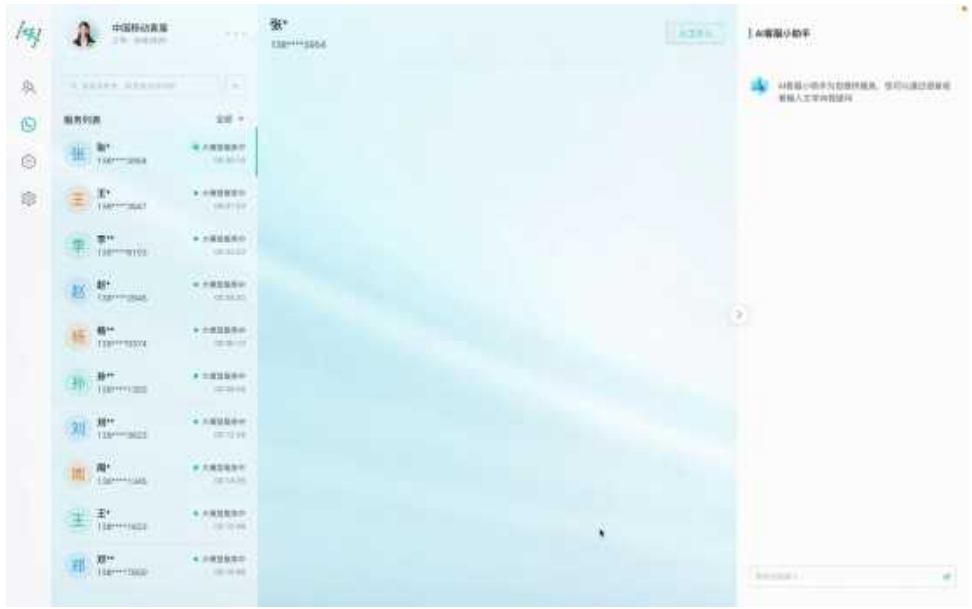
九天·客服大模型

面向企业级智能客服场景，基于10086服务近十亿客户的海量客服数据、业务知识和服务经验，打造九天·客服大模型，让每个用户都拥有24小时在线的“专属管家”，极大提升客服工作效率和用户体验

九天•客服大模型：率先服务大规模生产系统的行业大模型

2023年中国移动合作伙伴大会上，发布了首个基于客服行业大模型的智能客服产品，实现大模型客服领域应用破冰已在北京、广东等试点省份生产上线

中国移动10086客服
焕新升级



中国移动app
打造全新交互体验

构建“人工智能大平台”：成为通用智能时代的供给者、汇聚者和运营者

构建以平台和大模型为核心的智能基座，成为通用人工智能时代泛在智能的供给者、汇聚者和运营者，全面实现AI+转型升级

- 供给者：为产业提供智算、模型、平台等资源及服务
- 汇聚者：广泛汇聚国内外优质模型、数据、工具链和AI原生应用等
- 运营者：算、网、智等AI+应用的一体化服务及生态运营

建立面向生产服务的评估与审核体系，更好实现汇聚与承载

面向生产服务需求，汇聚业界优秀的通用和专用大模型及能力，建立“多层次-多维度-多任务-多指标-多模式”的大模型评估体系，确保汇聚的大模型安全、优质、高效，推动大模型产业规范化发展

模型汇聚与承载

汇聚对象

承载开源、业界领先的通专模型及工具

自研大模型	开源大模型	第三方大模型
小模型	工具链	AI 原生应用

入驻服务

构建承载平台，提供一体化、全流程的汇聚服务

入驻标准　模型测评　安全审核　大小模型端云协同

模型评测与安全审核

行业大模型业务评测维度	政务大模型 意图识别　答案有效 拟人程度	客服大模型 主观感受　意图识别 域内知识　域外幻觉	行业大模型 … …
通用大模型评测维度	语言大模型 理解　交互 生成　推理	视觉大模型 感知　认知 交互　推理	多模态大模型 多模态序列转换 …
评测指标	功能指标 任务支持度 场景支持度	性能指标 客观：准确性、鲁棒性 … 主观：准确性、安全性 …	服务成熟度 实时性　并发性 稳定性
安全审核	训练数据安全	输入问题的安全	模型结果安全
评测模式	自动评测 + 人工评测		

九天平台已汇聚开源模型20+个

百川-7B	百川-13B	百川2-13B	Yi-32B	Belle-7B	Belle-13B	Bloomz-3B	Bloomz-176B	ChatGLM-6B	ChatGLM2-6B
GLM-130B	Stable Diffusion	LLaMA-7B	LLaMA-13B	LLaMA-33B	LLaMA-65B	GPT-NeoX	Dolly	Falcon-40B	Moss

内容提纲

一、中国移动大模型布局及进展

二、中国移动大模型评估体系

三、九天客服大模型应用评估实践

面向生产的大模型评估体系

面向生产服务场景，建立语言大模型、行业大模型、多模态大模型、智能体应用、安全评测等**五大评测基准**，围绕评测数据、指标、方法与分析**三大建设方向**，高效开展综合全面的大模型评估评测。

大评

五大基准

- **语言大模型评测基准**：基础评测、专项评测、领域评测、体验评测
- **行业大模型评测基准**：客服大模型、政务大模型、网络运维大模型、其他行业模型
- **多模态大模型评测基准**：图文领域、音视频领域、混合多模态、数字人
- **智能体应用评测基准**：智能客服、知识管理、数据分析、智能助理
- **安全评测基准**：
  - 内容安全：政治敏感、违法违规
  - 伦理安全：价值观对齐、偏见歧视
  - 信息安全：个人隐私、商业机密
  - 模型安全

三大建设

- **评测数据建设**：2千+簇评测数据集
  - 基础数据
  - 行业数据
  - 专项数据
  - 安全数据
- **评测指标建设**：5大类评测指标
  - 功能指标
  - 性能指标
  - 业务指标
  - 服务成熟度
- **评测方法与分析建设**：可视化展示评测结果
  - 可视化报告
  - 数据看板
  - 自动化分析
  - ELO排行

自动评测 + 人工评测

技术支撑 ↑　　↓ 理论支撑

评测平台

评测平台支撑

- 模型注册
- 模型管理
- 评测任务管理
- 评测场景管理
- 评测指标汇聚
- 评测数据管理
- 评测报告分析
- 模型能力排行

面向生产的大模型评估体系-两阶段全生命周期评估

模型接入生产系统时，需要经过一系列“全面考验”，接入生产开始服务后，要开展“持续考验”，根据用户市场真实反馈，形成动态反馈机制持续优化提升大模型的落地成效

面向生产的模型评估

模型接入阶段

第一阶段 全面考验

行业大模型业务评测维度	政务大模型 意图识别 答案有效 拟人程度	客服大模型 主观感受 意图识别 域内知识 域外幻觉	行业大模型 … …
通用大模型评测维度	语言大模型 理解 交互 生成 推理	视觉大模型 感知 认知 交互 推理	多模态大模型 多模态序列转换 …
评测指标	功能指标 任务支持度 场景支持度	性能指标 客观：准确性、鲁棒性 … 主观：准确性、安全性 …	服务成熟度 实时性 并发性 稳定性
安全审核	训练数据安全	输入问题的安全	模型结果安全
评测模式	自动评测 + 人工评测		

模型服务阶段

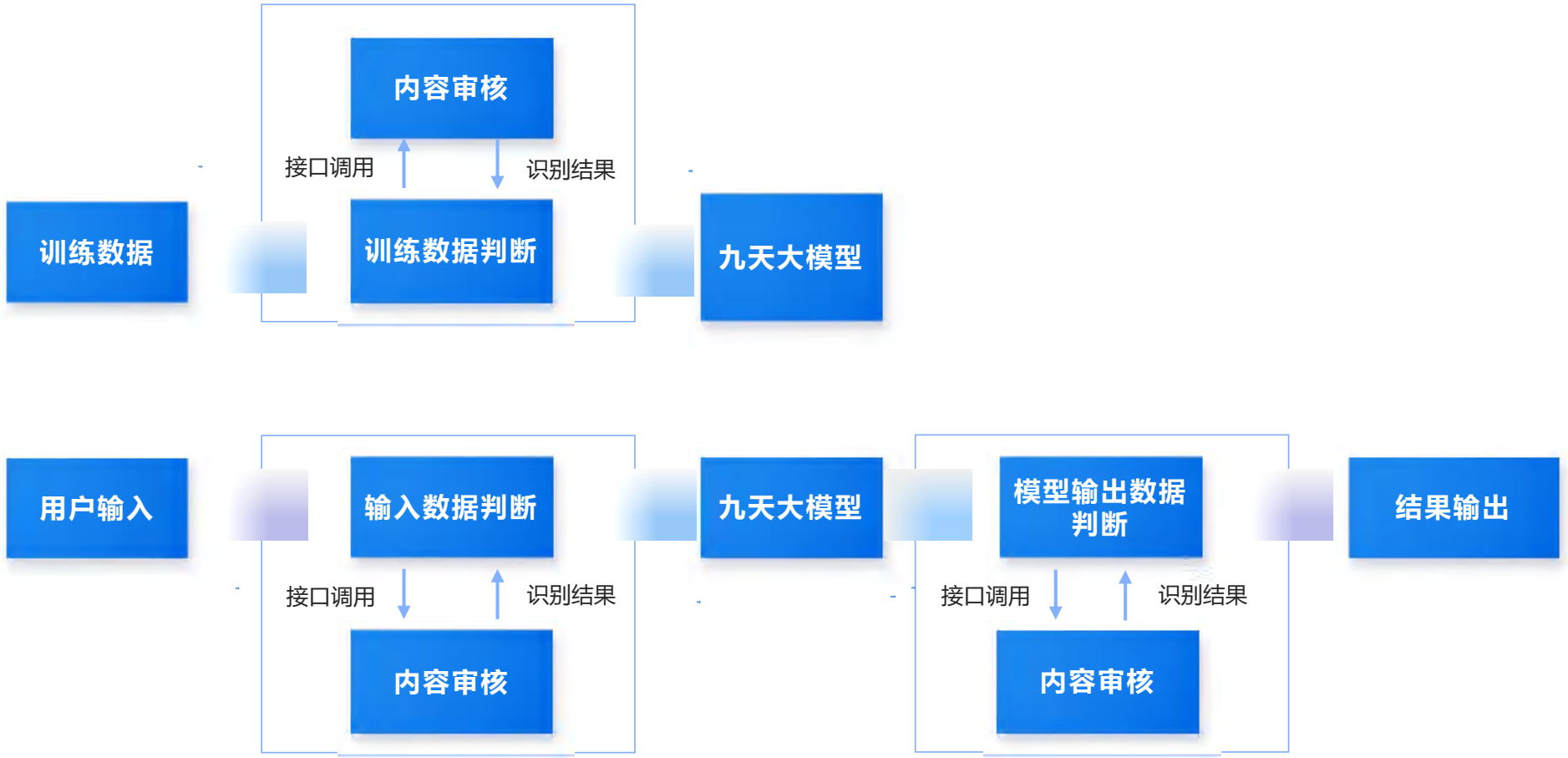
第二阶段 持续评测 迭代优化 持续考验

业务市场评测维度	用户评测 主观感受	专家评测 回复专业性	特定场景评测 问题解决率
评测数据	多样性、代表性	灵活性、适应性	准确性、应用性
评测指标	性能指标 客观：准确性、鲁棒性 … 主观：准确性、安全性 …	服务成熟度 实时性 并发性 稳定性	
持续评测	结果收集	结果分析	反馈优化
评测模式	自动评测 +用户评测+业务专家审核		

面向生产的大模型评估体系-安全审核

整体目标：构建覆盖训练数据、模型输入和输出的审核体系

审核机制：建立审核标签体系，对数据进行变体词识别等预处理后，通过多个模型标记审核标签，给出回答建议



面向生产的大模型评估体系-大模型评测平台

打造大模型评测平台，通过分层架构设计，增强其扩展性和灵活度，实现一键注册、快速评测、智能分析的大模型标准化评测流程

标准化评测流程

数据准备
准备评测数据集

覆盖基础评测、专项评测、领域评测、体验评测等4大评测维度、2000+个评测场景数据集

模型注册
准备模型信息

灵活快速接入业界多种类大模型，支持最大tokens数、并发线程、引导开关等配置项

模型评测
发起模型评测

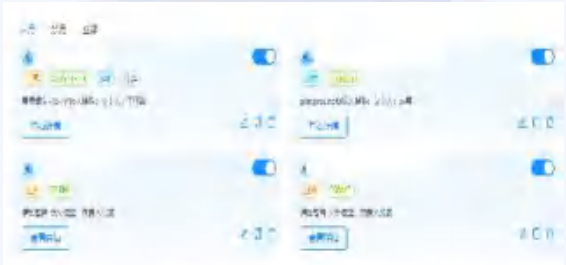
通过评测任务管理历史评测项。支持prompt模板配置、自动化打分、人工审核校验

查看报告
查看模型评估报告

基于准确性、鲁棒性、公平性、安全性等多维度量化打分，支持评测榜单快速查看

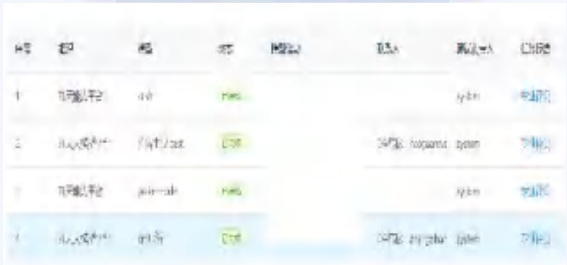
多功能支持

多模态支持



- 支持图生文
- 支持文生图
- 更多能力支持

对外服务

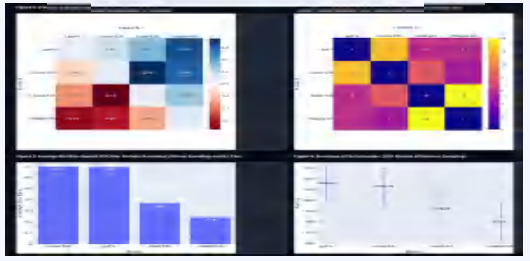


- 支持PaaS平台
- 支持MaaS平台
- 承接更多对外服务

交互评测

- 批量起聊
- 自动多轮对话
- 自动数据填充，自动场景打分

ELO排行榜



- 相对评估，消除差异
- 动态调整，确保准确性、适应性
- 可扩展性，适应不同数量和类型

面向生产的大模型评估体系-语言大模型评测基准

以语言大模型为出发点，建立评测基准，已形成4大评测维度、5大类指标、600+个评测场景、2000+簇评测数据集

评测维度

- **4大**评测维度：**基础评测、专项评测、领域评测、体验评测**
- **600+**评测场景：例如学术任务场景、创作写作、事实知识、计算、逻辑推理、闲聊、安全、自我认知等

评测数据

- **海量**评测数据集：**2000+簇**评测数据集
- 开源数据集：涵盖CMMU、CEVAL、AGI、GAOKAO、MMLU等
- 自建数据集：例如**安全类**数据集3万+条，**央企特色**数据2万+条

评测指标

- **5大**类评测指标：

  准确性、鲁棒性——着重指大模型的**功能**、**稳定性**表现

  安全性、公平性——着重指大模型的**非功能**表现

  高效性——着重指大模型的**响应时延、并发度**

面向生产的大模型评估体系-行业大模型评测基准

行业大模型评测与通用大模型评测不同，行业大模型更加**专注于行业领域知识和实际应用**，为此行业大模型评测应**深度融合行业特色**，评估大模型的高级理解、生成能力，如意图识别、意图改写和话术润色等，从而系统评估和分析行业模型的性能、准确性、适应性和实用性，确保模型满足行业标准和实际应用需求

内容提纲

一、中国移动大模型布局及进展

二、中国移动大模型评估体系

三、九天客服大模型业务应用实践

九天•客服大模型生产服务目标和技术要求

率先应用在10086全球最大的客服系统，驱动客服领域行业应用破冰。

服务目标

- **用户体验**
  Min（T1+T2+T3+T4）
- **服务效率**
  Max（工具和知识边界）

业务能力

- **拟人化**
- **强洞察**
- **多模态**

技术要求

- **稳健性与灵活性联合优化**
- **强系统集成**
- **多元多级高可控性**
- **开创人机协同新模式**

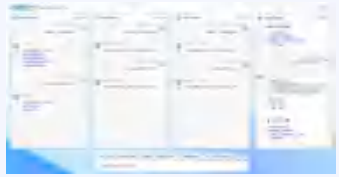
九天·客服大模型的评测与生产运营

- 面向10086智能客服系统生产级别上线要求，建立多维度、多层次的客服大模型评测体系，确保评测的完备性和合理性
- 针对真实客服场景中面临的安全可信问题，提出溯源信息场和一致性校验的方案，贯穿客服大模型的全流程，实现客服回复内容的可信响应，保证服务的安全可控

能力评测

构建大模型评测体系

搭建大模型测试工具



横向比较 → 对标九天客服大模型行业水平

纵向跟踪 → 九天客服大模型的迭代演进效果

六维度技术评测 ⟷ 六维度业务盲测

意图理解任务	平均响应时间
对话状态判断	非拒识交互占比
信息抽取任务	情绪识别任务

意图理解力	回答准确性
回答完整性	回答及时性
回答友好性	回答安全性

客服人员完成多轮业务评测和多轮技术评测。

安全可控

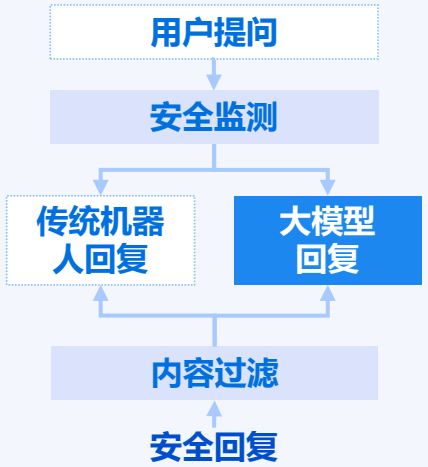
安全能力：6大维度，34个细项，80个细分小项

价值观	商业违法违规	侵犯他人合法权
歧视	行为攻击	特定服务安全需求

安全管控工具：解决不该答的不答问题

实现对用户表达、大模型生成内容进行双向安全管控

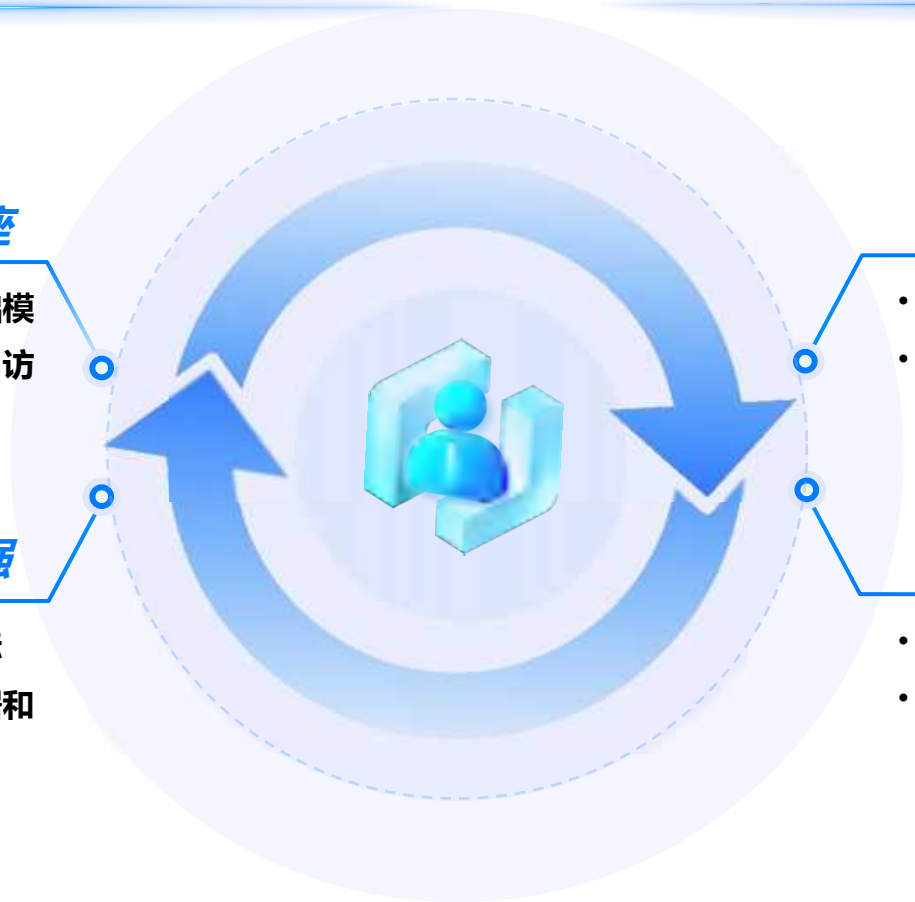
- 已对接智能在线客服敏感词库
- 支持运营人员自定义安全监测、内容过滤规则



技术目标：稳健性与灵活性联合优化

解决复杂系统智能化体系大而不稳的挑战。

系统和用户双驱动对话模式的灵活多样性。



预训练行业基座

- 客服知识融入基础模型、进阶式学习，访问增强

数据精标及增强

- 人人对话数据精标
- 动态用户属性数据和业务知识补充

主动服务

- 交互模式升级
- 单一界面 → 360度全方位提前想你所想

情感关怀

- 服务中双方情感流动，拟人化
- 思维链逻辑增强

技术目标：多元多级高可控性

业务运营工具化，与模型底座解耦合，通过大小模型协同、信息不出场、流程嵌入确保整体准确和可控。

技术目标：强系统集成

简单入口、现有系统重用和系统伴随，提供润物细无声的服务。

- **内部多层级系统集成**

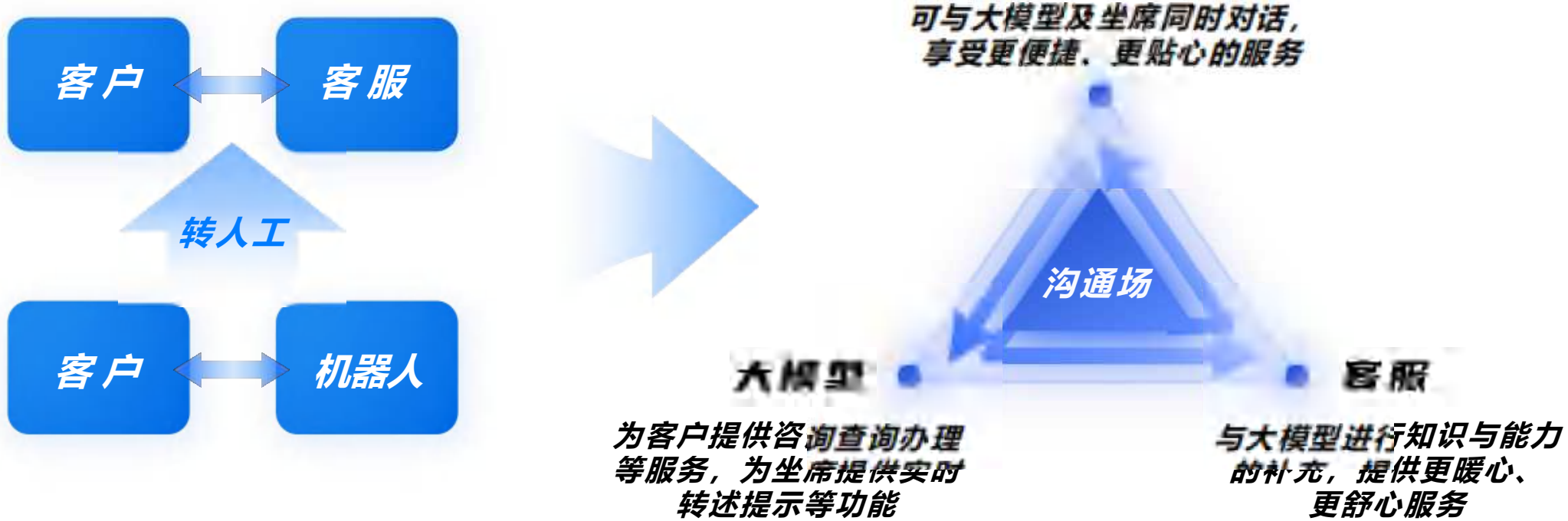
  从底层知识库、到接口级别的API、到系统界面级别的解析和集成

- **外部系统和工具集成**

  外部系统和工具集成，贯通客服咨询到运营分析全流程

技术目标：开创人机协同新模式

从客户-机器人、客户-客服的单点的沟通，升级为客户+大模型+客服三方协同交互。



AI人工智能产业链联盟

#每日为你摘取最重要的商业新闻#

更新 · 更快 · 更精彩

Zero

AI音乐创作人
水墨动漫联盟创始人
百脑共创联合创始人
人工智能产业链联盟创始人
中关村人才协会秘书长助理
河北北大企业家分会秘书长
墨玫星辰智能科技有限公司CEO
河北清华发展研究院智能机器人中心线上负责人
中关村人才协会数字体育与电子竞技专委会秘书长助理

主要业务:AI商业化答疑及课程应用场景探索，各类AI产品学习手册，答疑及课程

欢迎扫码交流

提供：学习手册/工具/资源链接/商业化案例/行业报告/行业最新资讯及动态

人工智能产业链联盟创始人

邀请你加入星球，一起学习

人工智能产业链联盟报告库

星主：人工智能产业链联盟创始人

每天仅需0.5元，即可拥有以下福利！
每周更新各类机构的最新研究成果。立志将人工智能产业链联盟打造成市面上最全的AI研究资料库，覆盖券商、产业公司、研究院所等...

知识星球

微信扫码加入星球