“弈衡”大模型评测体系研究及技术发展趋势浅析

中国移动研究院 中国移动技术能力评测中心
刘伟东

打造中国移动权威第三方评测机构

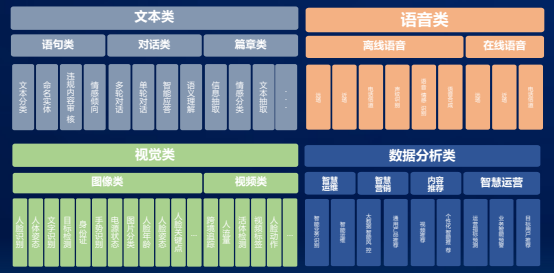
愿景：成为中国移动核心技术能力的度量衡和磨刀石

目标：以公司“一体五环”重点产品和能力评测为中心，打造中国移动权威、中立、客观的第三方技术能力评测机构

积淀全栈评测技术

涉及领域广

数据积累多



工具手段精

中国移动 China Mobile

弈衡 YI HENG

搭建中国移动技术能力评测图谱，全面助力公司打造优质产品

个人/新兴业务（CN）

云游戏	超级SIM	云渲染
云AR	中间号	视频直播
云VR	5G新通话	视频点播
云手机	5G消息	数字身份
云盘	移动认证	数字人
云魔百和	数字货币	视频彩铃
...		

家庭业务（H）

家庭连接	智能网关	高清机顶盒
家庭宽带	家庭安全	智能交互
视频通话	音频通信	健康养老
室内摄像头	教育加速	云魔百盒
实时音视频	游戏加速	语音交互
智能音箱	智能门锁	智能门窗磁
...		

行业应用（BGV）

工业	金融	位置	医疗	教育	农业	智慧城市
工业质检(螺钉检测仪表识别)	智能客服	室外定位	医疗影像云	课程推荐系统	智慧养殖猪圈整体估种	数字孪生
工业质检（光伏电池）	数字人客服平台	室内定位	急救产品（急诊分诊）	平安校园	智慧养殖猪只盘点	园区安防
工业网关（协议性能）	大数据风控模型	三维地图可视化引擎	AI辅诊产品	智慧体育	农机作业面积识别	跨境追踪
网关管理平台	风控平台	卫星遥感地图	急救产品（预案处置）	智慧考场		城管场景识别
工业质检(包装传送检测)	交通	地基增强	急救产品（紧急救援）	无人机	保密专网	大数据平台
	远程驾驶	星基增强	OneHealth智慧医疗云平台	中移凌云平台	密钥安全SIM卡...	城市运营管理
	自动驾驶	...		哈勃一号	5G保密专网管理平台	一网通办
	智慧港航					OneDB数据库
						城市AI平台

AI能力

文本类

语句类				对话类				篇章类			
文本分类	命名实体	违规内容审核	情感倾向	多轮对话	单轮对话	智能应答	语义理解	信息抽取	情感分类	文本抽取	对话情绪

语音类

离线语音						在线语音		
远场	近场	电话信道	声纹识别	语音情感识别	语音合成	远场	近场	电话信道

视觉类

图像类												视频类					
人脸识别	人体姿态	文字识别	目标检测	身份证	手势识别	电源状态	图片分类	人脸年龄	人脸姿态	人脸关键点	...	跨境追踪	人流量	活体检测	视频标签	人脸动作	...

数据分析类

智慧运维		智慧营销		内容推荐		智慧运营		
智能业务识别	智能运维	大数据智能风控	通用产品推荐	视频推荐	个性化智能推荐	运营指标预测	业务智能预警	目标用户推荐

通用大模型

行业大模型

成立CCIR大模型评测工作组

专委会副主任郭嘉丰向中国移动技术能力评测中心主任刘伟东授牌

促进国产大模型发展，建立评测体系是关键一环

以 评 促 研

积极促进自有大模型提升

以 测 选 优

遴选应用外部优质模型

初步构建“弈衡”体系，充分发挥评测价值

构建
完整体系

覆盖
主流模型

发挥
智库价值

推进
行业发展

“弈衡”大模型评测体系——“2-4-6”

2种 评测场景

基础任务


文本分类 目标检测

应用任务


智能客服 代码生成

4项 评测要素


评测方式


评测指标


评测数据


评测工具

6个 评测维度

功能性

安全性

准确性

应用性

可靠性

交互性

研究多维指标评测方法，推动评测体系落地应用

基于“弈衡”大模型评测体系，完成20+款主流模型评测

国外市场

ChatGPT3.5 OpenAI 1750 亿	GPT-4 OpenAI 约1万亿	Claude Anthropic 62亿	Bard 谷歌 5400亿	Vicuna-13B UC 伯克利 130 亿	Llama2 Mate 700 亿

……

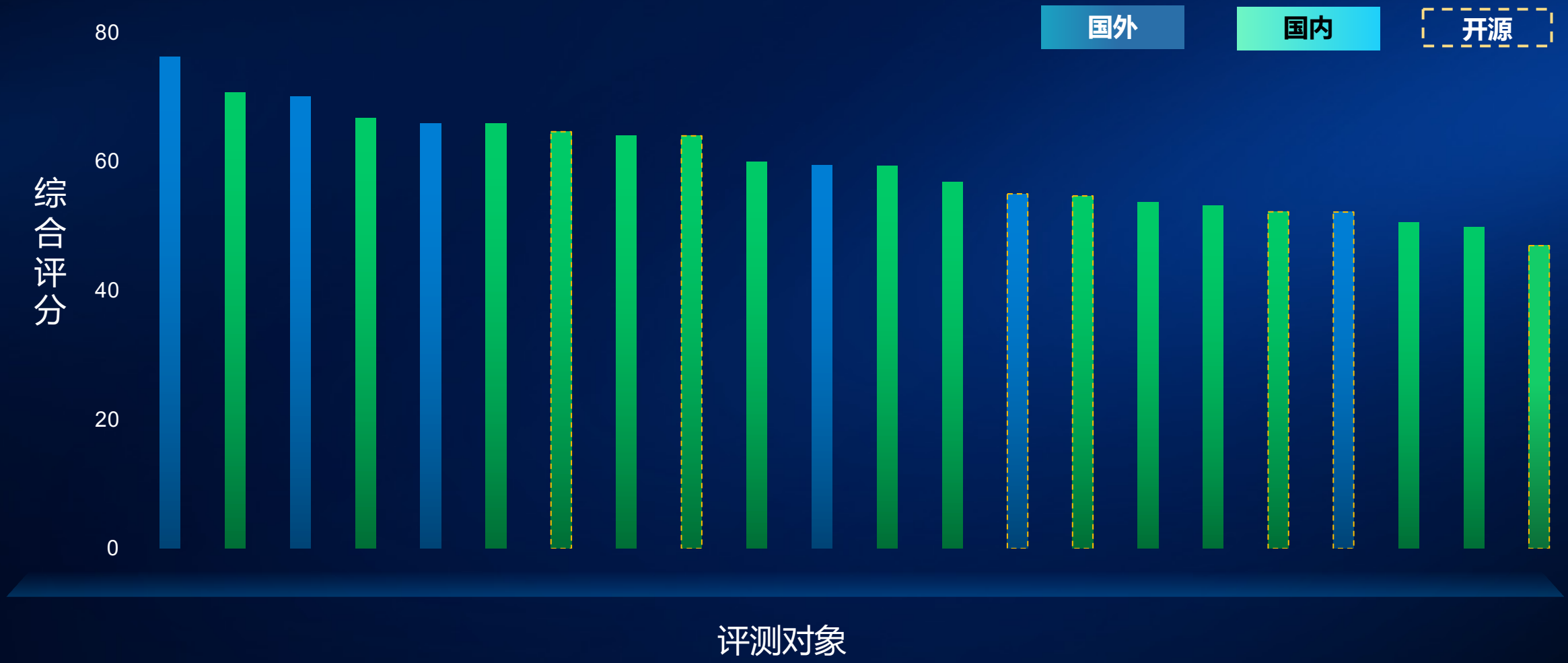
国内市场

文心一言 百度 1750亿	通义千问 阿里 未公开	星火认知 科大讯飞 500亿	百川 百川智能 530亿	混元 腾讯 1000亿	……	悟道3.0 智源 70亿	ChatGLM-130B 智谱AI 1300亿	紫东太初 中科院自动化所 1000亿
360智脑 奇虎科技 约千亿	SenseChat 商汤 1800亿	天工 昆仑万维 约千亿	ABAB MiniMax 约千亿	云雀 抖音 未公开		Moss 复旦大学 160亿	ChatGLM 2-6B 智谱AI 60亿	……

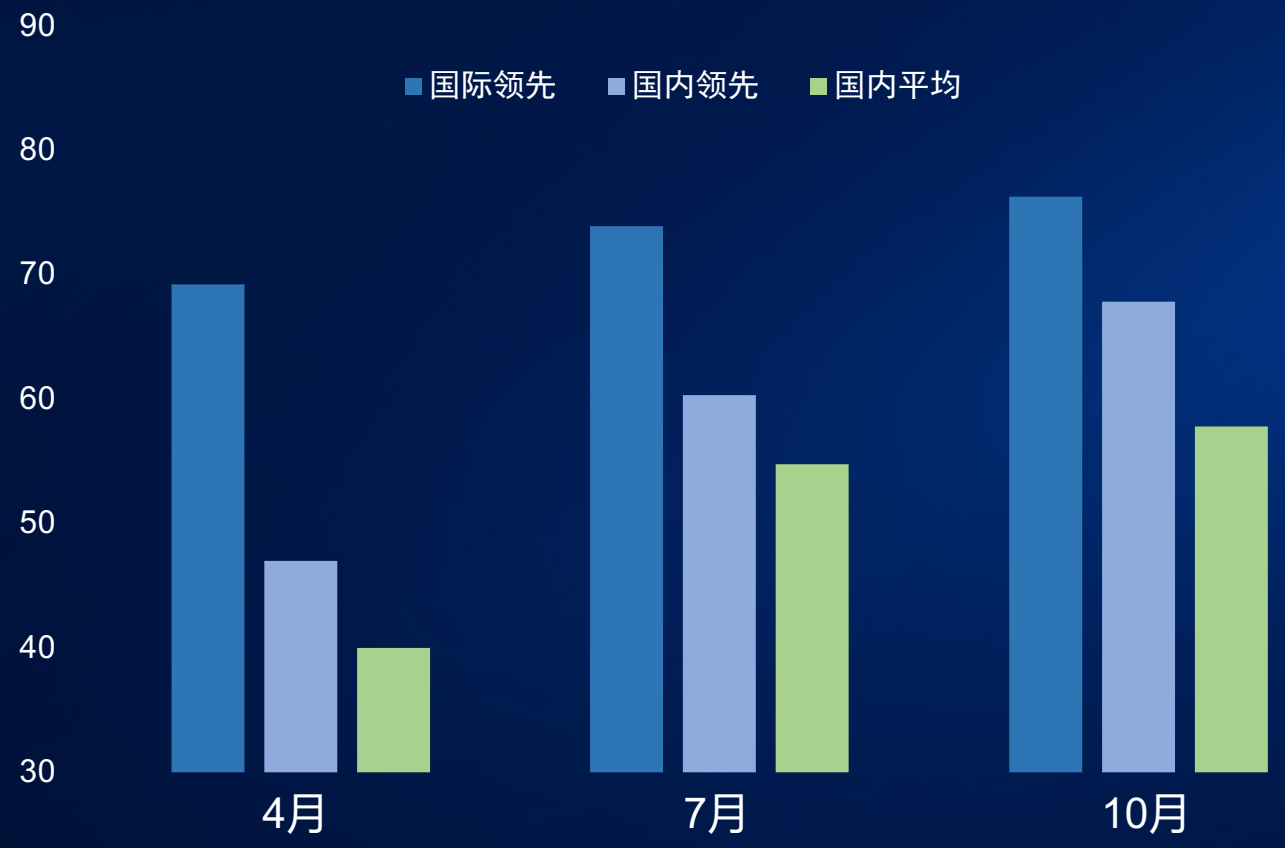
科技大厂　　科研机构

★ 开源模型

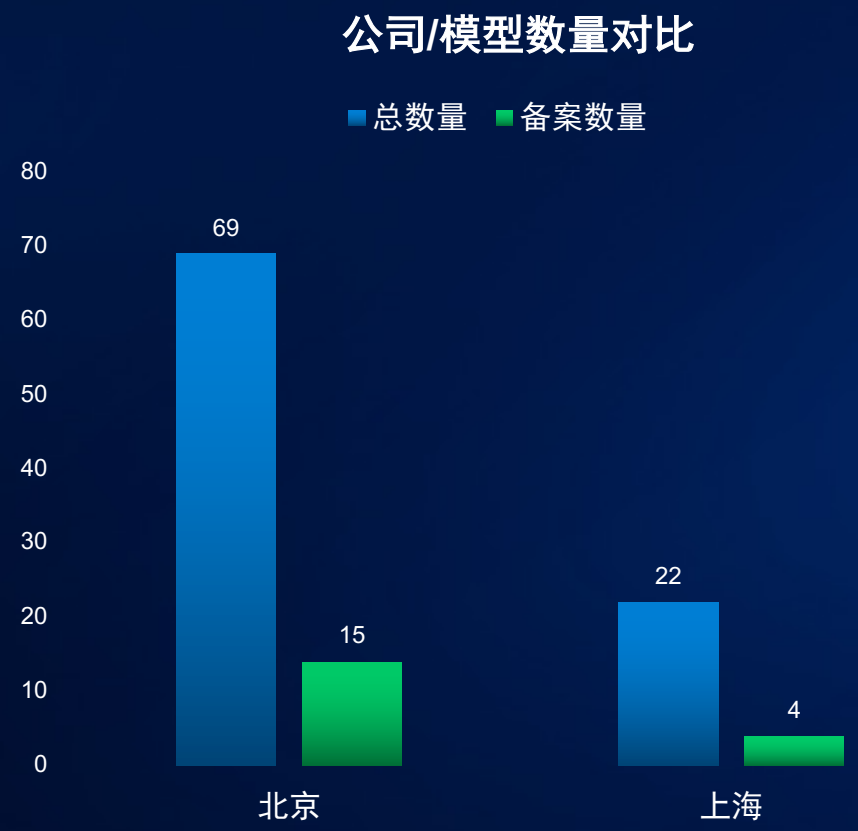
国内外主流大模型整体评测情况

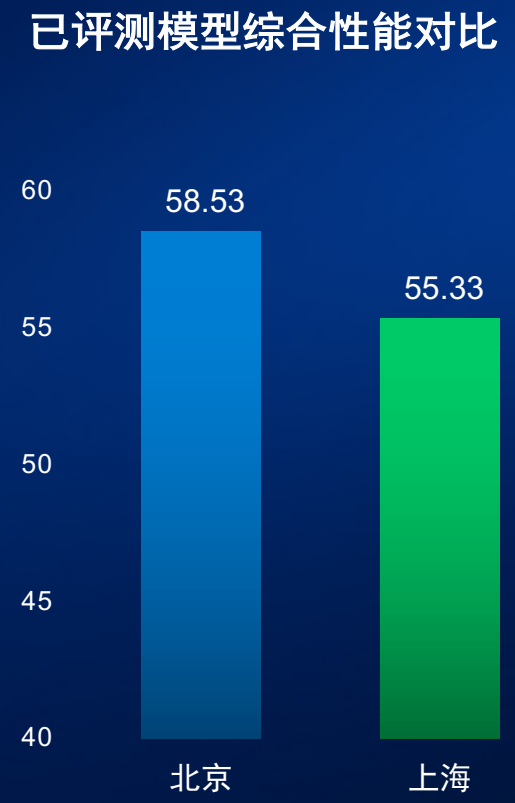


洞察1：国内总体水平与国际领先存在差距，头部企业追赶趋势明显

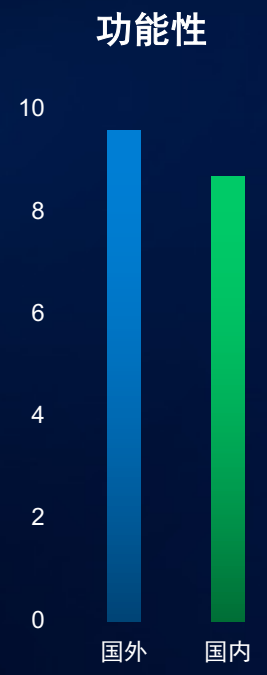


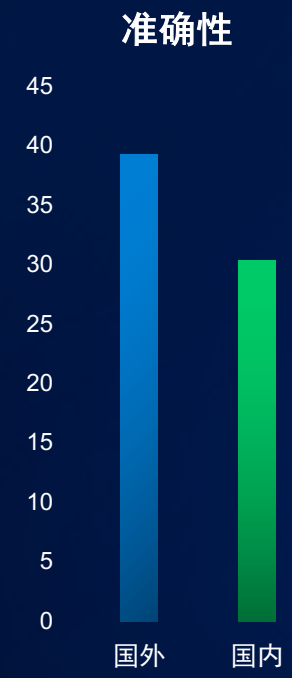
洞察2：北京上海成大模型高地，综合实力领先

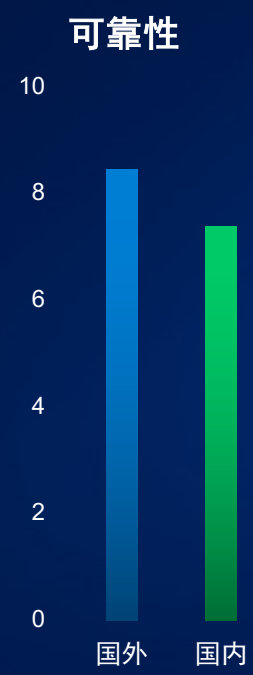


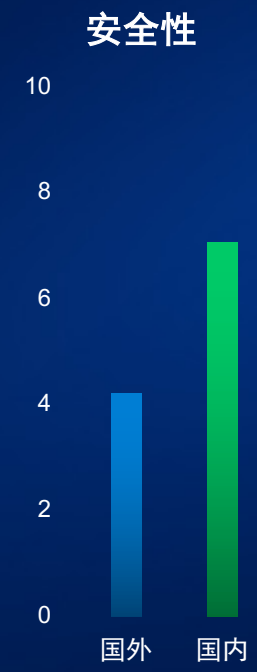


洞察3：国内外详细对比，准与稳才是硬道理









1 “弈衡”评测厚积薄发

2 大模型发展趋势洞察

3 下一步发展展望

展望未来：持续完善评测体系，携手推进产业发展

多模态评测能力

通用⇔行业大模型

智慧政务

智能客服

评测合作平台

中国移动联合业界筹备人工智能评测联盟

正在征集首批联盟单位，欢迎大家踊跃报名！

联系邮箱：zgydjsnlpczxfs@chinamobile.com

中国移动希望与产业界和学术界携手，

共同构建大模型评测合作生态，

推进国产大模型蓬勃发展！