智能体时代 AI 基础设施发展研究报告（2026）

中国电信研究院

行业与企业战略研究所

2026 年 9 月

目录

引言 1
一、2026 全球 AI 产业发展新变化 4
1、全球政策焦点转向主权 AI 基础设施建设及智能体应用 4
（一）美欧以规则管控持续完善 AI 基础设施治理体系，打造主权 AI 4
（二）日韩以巨额资本加码本土 AI 基础设施建设，重资产攻坚主权 AI 5
（三）我国以智能体应用为牵引，全面升级算力基础设施能力及服务水平 6
2、人工智能迈入“Agentic AI”新阶段，带动算力需求持续激增 8
（一）人工智能迈入 Agentic AI（代理式 AI）新阶段，全面开启智能体时代 8
（二）智能体加速落地，带动我国 Token 及算力需求激增 9
（三）AI Coding 成为主流 AI 应用场景之一，驱动软件生产范式重构 10
3、AI 基础设施竞争焦点转向系统全栈，市场空间持续扩大 10
（一）AI 基础设施竞争焦点由“算力/模型”单维能力比拼，转向“算力-模型-智能体-应用”系统性全栈布局 10
（二）生成式 AI 驱动云市场保持强劲增长 12
（三）AI 基础设施市场竞争主体多元，市场持续向头部厂商集中 13
4、AI 工程范式持续迭代，Harness 成为智能体核心底座 14
（一）AI 工程范式正经历从提示词工程、上下文工程到 Harness 工程的三阶段演进 14
（二）Harness 工程的核心价值在于弥补大模型原生能力无法实现的风险约束与流程治理短板 15
（三）头部科技厂商正加速将 Harness 工程产品化，构建面向生产环境的 Agent 托管服务 16
二、智能体时代 AI 基础设施发展及产业趋势预判 18
1、AI 基础设施总体视图：智能体重塑 AI 全栈能力和服务体系 18
2、算力：智能体重塑算力基础设施架构、规模及布局重心 19
（一）算力架构转向“CPU+GPU”协同主导，CPU 成为系统核心约束 19
（二）算力布局转向推理优先，推理基础设施持续向高效率、低成本演进升级 20
（三）超大规模高密度智算集群仍是 AI 产业刚需，超节点、十万/百万卡级集群持续深化建设 21
3、网络：算力需求牵引光网络成为构建高效 AI 基础设施的关键 22
（一）“光进铜退”成为超节点智算集群高密互联的技术刚需，LPO、NPO、CPO 多技术路线并行 22
（二）多数据中心协同正在成为 AI 集群规模扩张的关键手段，光网络是关键 22
（三）AI 算力扩张引爆光纤需求，供需失衡驱动主流厂商加速锁定光纤产能 23
4、云：云基础设施能力及产品服务面向智能体全面升级 24
（一）云基础设施能力全面升级，为智能体提供高效、稳定、安全运行环境 24
（二）面向 Agent 协同与应用优化产品能力，提升产品使用体验及易用性 25
5、模型：大模型能力迭代且供给格局分化，适配智能体规模化落地需求 27
（一）大模型面向智能体场景开展多维度技术能力升级 27

更多免费行业报告 并购家 www.ipoipo.net

（二）通用基座模型趋于收敛，垂直行业模型多元细分发展 28
6、智能体：智能体技术架构持续向高效化、开放化、解耦化方向演进，全面提升智能体运行效率与场景适配能力 29
（一）智能体通过记忆分层、协同互联技术实现高效记忆管理与跨系统协作29
（二） 智能体平台依托统一模型网关、可插拔架构实现多模型灵活调度，与模型解耦将成为行业标配 31
7、应用：业务价值优先，场景禀赋决定智能体落地节奏 32
（一）智能体应用将优先落地高价值、可闭环的数字化场景 32
（二）全生命周期治理与成本核算成为智能体生产准入硬性门槛 33
8、生态：多方协同与标准互通并行，兼顾智能体落地需求和行业发展引领 35
（一）企业生态重心转向智能体全栈体系构建和行业伙伴争夺 35
（二）以政策、倡议、标准等为指引推进智能体互联协同 36
三、我国 AI 及智能体产业区域性发展评估 36
1、智能体产业发展潜力评估模型框架 37
2、评估流程与方法 39
3、全国 31 省发展智能体产业的潜力评估 41
（一）综合评估结果：我国省域智能体产业发展呈现头部集聚与梯度分化并存的基本格局 41
（二）发展环境评估结果：东部地区在政策治理、经济基础和创新资源方面的综合优势较为突出 42
（三）基础设施评估结果：省域智能体基础设施已由单一算力供给转向算力、数据、网络和模型工具协同发展 44
（四）服务应用评估结果：智能体服务应用能力主要集聚于 AI 产业基础和场景转化能力较强的地区 46
（五）基础设施与服务应用的协同分析：基础设施与服务应用的协同程度，直接决定算力资源能否转化为智能体产业能力 48
（六）Top10 省份优劣势分析及发展建议 49
4、全国 31 省发展行业智能体的潜力评估 51
（一）评估方法 51
（二）评估结果 53
5、全国区域性智能体产业发展建议 56
（一）推动基础设施建设从资源扩张转向服务化和可调用化 56
（二）建设公共智能体平台，降低企业从试点到规模化的门槛 56
（三）以高价值、可验证、可闭环场景作为应用推广主线 56
（四）实施差异化行业布局，构建“一省一业多场景”推进体系 57
四、总结与展望 57
五、附录 58
1、智能体产业发展潜力评估模型的指标定义与数据依据 58
2、各省智能体产业发展评估结果 61
3、各省行业智能体发展评估结果 63
六、参考文献 66

引言

人工智能（AI）作为新一轮科技革命和产业变革的核心驱动力，正以前所未有的速度、广度与深度全面渗透人类生产生活，全方位重塑生产方式、产业形态和社会治理逻辑。我国高度重视AI创新发展，深入实施“AI+”行动，统筹推进全国一体化算力网建设，持续优化智算资源布局，推动AI技术与实体经济深度融合，加快构筑智能经济发展新格局。根据摩根士丹利数据[1]，2026年我国头部科技型企业AI的资本支出接近6000亿元，占全社会总投资比重超过10%；短期内（2026-2027年）可直接拉动我国GDP增长约0.3个百分点；中长期（到2035年）AI将成为我国经济增长的重要引擎，预计未来十年（2026-2035年）AI将拉动我国全要素生产率提升约3个百分点；到2035年，我国潜在GDP水平，相比没有AI赋能的情景，将高出约3.5个百分点。

当前人工智能正进入以智能体为主要形态的全新发展阶段，智能体让复杂的人工智能技术隐形化、普惠化，大众和企业无需掌握专业技术，借助智能体就能完成全流程的复杂任务。随着人工智能技术从“模型被动问答”向“智能体自主规划-决策-执行”升级演进，全球智能体数量爆发。根据IDC数据，全球活跃智能体数量将从2025年的2860万个跃升至2026年的7940万个(近8000万个)，预计2030年达到22.16亿个，2025-2030年的年复合增长率达139%[2]。随着人工智能产业发展重心从大模型和算力竞争转向智能体规模落地、应用变现，我国Token（词元）需求呈现爆发式增长。智能体发

更多免费行业报告 并购家 www.ipoipo.net

展将带动 Token 经济兴起，预计 2026 年我国 Token 年消耗量将达到 10 亿亿，到 2030 年将超过 3500 亿亿。智能体发展加速，将对算力、网络、平台等底层 AI 基础设施能力提出更高要求，如智能体复杂推理、多智能协同及长周期任务运行的业务特性，带来成本、效率等挑战，传统 AI 基础设施难以充分适配智能体时代的发展需求。

AI 基础设施是支撑智能体创新迭代与规模化落地的核心底座，涵盖算力、网络、存储、云、模型、服务平台等软硬协同全栈能力，直接决定智能体任务执行效率、系统稳定性与产业落地成本，是抢抓智能体发展机遇、释放人工智能新质生产力的前提基础。在政策引导与产业需求双重拉动下，我国 AI 基础设施建设加速推进，以智能算力为代表的基础设施规模和水平位居全球前列，截至 2026 年 6 月底，我国智能算力规模达 2185 EFLOPS（FP16），同比增长 177%，全国算力设施整体上架率达 71.4%。当前我国算力供给规模持续扩大，算网融合能力稳步提升，产业链上下游协同创新成效显著，已构建起覆盖算力、网络、存储、平台等的完整基础设施底座体系，但面向智能体规模落地、多智能体协同等新诉求，仍面临现实挑战。

本研究报告聚焦智能体时代 AI 基础设施发展主题，全面且系统梳理了 2026 年全球 AI 产业发展新动向新变化，包括国内外政策导向、智能体爆发驱动算力需求激增、AI 基础设施行业竞争焦点转移、AI 技术及工程范式迭代演进等，前瞻性研判了智能体时代 AI 基础设施发展及产业七大趋势，聚焦算力、网络、云、模型、

智能体、应用、生态等核心维度，结合智能体时代发展新要求，深入分析其底层基础设施供给新变化，以期为推动我国 AI 基础设施高质量建设、赋能智能体产业蓬勃发展提供有益参考。此外，该报告创新性构建了 AI 及智能体产业发展潜力评估模型，用于评估我国 31 省的产业发展潜力，以期为我国 AI 产业提质升级、AI 基础设施科学布局提供量化数据支撑。

更多免费行业报告 并购家 www.ipoipo.net

一、2026 全球 AI 产业发展新变化

1、全球政策焦点转向主权 AI 基础设施建设及智能体应用

（一）美欧以规则管控持续完善 AI 基础设施治理体系，打造主权 AI

美国持续构建 AI 基础设施安全管控体系，依托标准与出口管制巩固全球 AI 竞争优势。2 月 17 日，美国国家标准与技术研究院发布《AI 智能体标准倡议》，制定自主智能体的互操作性、安全边界、身份认证等标准，搭建行业协作平台，保障智能体可信、可控落地。3 月 26 日，众议院推动《芯片安全法案》落地，建立高端芯片全生命周期溯源管控机制，同时不断拓展管制边界，于 2026 年先后出台《远程访问安全法案》、《云安全法案》，将管制范围从实体硬件延伸至云端，封堵竞争对手借助云服务获取美国先进算力的渠道。6 月 2 日，美国总统签署《促进先进人工智能创新与安全》行政令，要求前沿 AI 模型在发布前向政府开放最长 30 天的网络安全测试窗口，接入联邦网络安全共享体系。8 月 4 日，白宫与多家美国主要人工智能企业举行闭门会议，研讨前沿 AI 模型自愿网络安全测评机制落地方案。

欧盟独立于中美技术生态，立法深化主权 AI 与安全。1 月 16 日，欧盟理事会发布《欧洲高性能计算条例修订案》，新增“AI 超级工厂”相关规定，为超大规模人工智能模型训练、运行搭建

标准化基础设施，从立法层面夯实欧盟 AI 算力底座，强化本土算力自主能力，支撑前沿 AI 技术研发。6 月 3 日，欧盟《技术主权一揽子方案》首次将云、AI、芯片、软件、能源等基础设施设为独立政策框架，通过监管规则、基础设施控制、产业应用牵引，投入万亿欧元资金支持，降低对海外供应商依赖，扩大数字技术主权。7 月 7 日，欧盟发布《网络安全与人工智能行动计划》，要求抵御先进人工智能脆弱环节，增强欧盟整体网络安全韧性与技术能力。7 月 30 日，欧盟发起“AI 超级工厂”建设招标，计划在欧洲设立最多 7 个“AI 超级工厂”，包括高端 AI 芯片、配套软件与云架构、高速互联网络、绿色节能数据中心等，目标提升欧洲人工智能自主计算能力。8 月 1 日，欧盟《人工智能法案》高风险 AI 相关条款正式生效，对可能造成系统性风险的先进通用人工智能模型，要求提供商还须履行额外义务，以防范大规模安全风险。

（二）日韩以巨额资本加码本土 AI 基础设施建设，重资产攻坚主权 AI

日本以主权 AI 为牵引，扩大智能体等 AI 基础设施投资及场景应用。4 月 13 日，日本经济产业省与新能源产业技术综合开发机构主导，软银、日本电气、本田、索尼等 4 家财团共同成立“日本 AI 基础模型开发公司”，主导一万亿日元国家级人工智能计划，形成从国产底层模型研发到实际应用的完整链条。7 月 14 日，日本内阁在仅 7 个月后通过第二期《人工智能基本计划》，将智能

更多免费行业报告 并购家 www.ipoipo.net

体式人工智能、垂直领域人工智能和物理人工智能列为重点应用方向，计划提出建立开放式人工智能主权，降低对单一国家、企业或基础模型的过度依赖。

韩国持续扩大 AI 基础设施资金投入规模，完善 AI 治理及应用。1 月 5 日，韩国科学技术信息通信部公布《国家 AI 战略》蓝图，2026 年 AI 年度预算提升至 10 万亿韩元，计划撬动公私部门合计 30 万亿韩元投资。1 月 22 日，韩国《关于人工智能发展和构建信赖基础的基本法》施行，旨在推动 AI 产业创新发展，同时防范深度伪造、虚假信息传播等潜在风险。2 月 25 日，《韩国人工智能行动计划 2026-2028》提出了至 2028 年的具体路线图，旨在将人工智能融入政府、产业和社会。6 月 29 日，韩国总统进一步宣布三位一体国家级 AI 战略，确立半导体、物理 AI、AI 数据中心为三大核心产业支柱，三星电子、SK 集团本土合计投资规模达 4700 万亿韩元，创下韩国最大规模本土投资纪录。8 月 28 日，韩国科学技术信息通信部正式公布“全民 AI”项目，是全球首个由政府出资向全体国民提供免费、无限量 Token 的 AI 公共服务。

（三）我国以智能体应用为牵引，全面升级算力基础设施能力及服务水平

一是首次明确智能体创新发展路径，牵引算力基础设施高质量配套升级。3 月 13 日，“十五五规划”明确鼓励包括智能体在内的技术创新，探索通用人工智能发展路径。5 月 8 日，国家网信办等三部门落地《智能体规范应用与创新发展实施意见》，首

次明确智能体创新发展路径，明确要求完善智能体研发、测试、部署、运维等工具链。6 月 10 日，工信部《“人工智能+信息通信”创新发展实施意见（2026—2028 年）》提出提升优化智算云服务，面向智能体的训练和推理算力需求，提升智算云服务水平，推动智能体云化部署。8 月 31 日，工信部《关于开展人工智能应用服务商培育专项行动的通知》提出探索通过首购首用、风险补偿等模式，加大大模型、智能体、Token 等服务采购力度，提升行业用模用智质效。9 月 7 日，工信部《信息通信行业发展“十五五”规划》要求加快发展智能体互联网络，打造智网融合的新一代互联网基础设施。

二是围绕算电协同与全国一体化算力网构建，持续推进算网互联与多网协同。3 月 13 日，《政府工作报告》首次明确提出“实施超大规模智算集群、算电协同等新基建工程，加强全国一体化算力监测调度，支持公共云发展”。4 月 8 日，国家发改委等印发《关于促进人工智能与能源双向赋能的行动方案》，提出统筹优化能源资源与算力布局、提高算力设施多元电力供给能力、提升算力设施能源供给质量。4 月 28 日，中共中央政治局将“算力网”写入“六大网络基础设施”建设部署，国家发改委表示将发挥“多网协同”作用，共同推动现代化基础设施体系优化布局结构并促进集成融合。6 月 18 日，国家发改委表示一体化算力网围绕 8 大国家算力枢纽、10 个国家算力集群和 3 个算电协同发展区域，已形成“8+10+3”空间布局。8 月 19 日，国家发改委在“六

更多免费行业报告 并购家 www.ipoipo.net

张网”重大项目协调调度机制会上，研究建立算力网、新型电网、新一代通信网“2+3+N”（2 家电网+3 家电信运营商+N 家算力相关企业）协调机制，加大统筹力度，形成工作合力。

2、人工智能迈入“Agentic AI”新阶段，带动算力需求持续激增

（一）人工智能迈入 Agentic AI（代理式 AI）新阶段，全面开启智能体时代

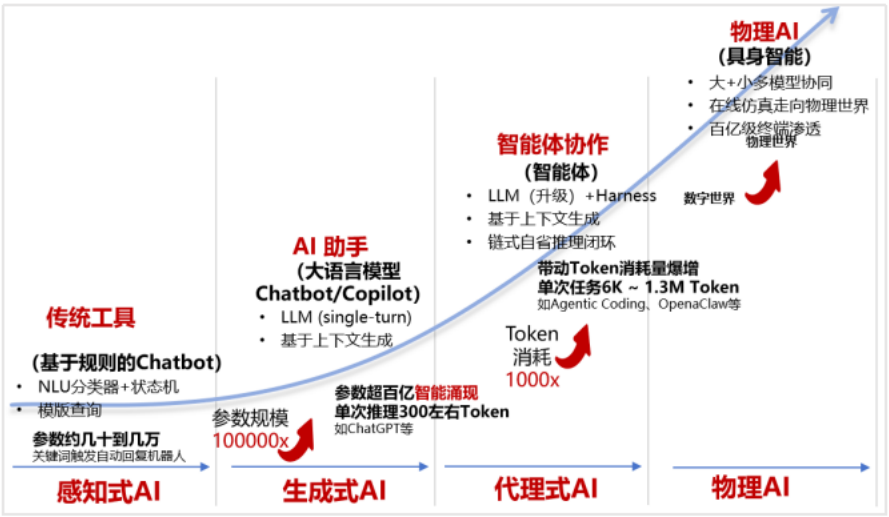


图 1 人工智能发展历程

全球 AI 发展演进可清晰划分四个阶段：一是感知式 AI（Perception AI），以图像识别等为代表，AI 可识别文本、图像和语音；二是生成式 AI（Generative AI），以大模型为代表，AI 能输出文本、图像和视频；三是代理式 AI（Agentic AI），以智能体为代表，AI 可思考、能推理、能调用工具；四是物理 AI（Physics AI），以具身智能为代表，AI 能理解物理规律、驱动实体机器人。当前全球人工智能正处于从生成式 AI 迈向代理式 AI 阶段的拐点，智能体成为人工智能产品及服务的重要形态。智

能体具备自主感知、记忆、决策、交互与执行能力，成为 AI 规模化落地的核心载体，正逐步融入产业、民生、治理等领域，把大模型能力组织成能对任务结果负责的系统嵌入场景应用，把大模型能力变成直接交付结果的稳定生产力，并加速与网络空间、物理世界深度融合，智能体时代全面开启。

（二）智能体加速落地，带动我国 Token 及算力需求激增

2026 年，智能体由概念验证加快进入生产环境，应用从客服、知识检索向数据分析、流程自动化、软件研发等核心业务扩展。智能体需要持续进行任务规划、长上下文读取、检索、工具调用、结果校验和失败重试，形成多轮模型调用。智能体单次任务的 Token 消耗可达普通对话数十倍，复杂场景更高。IDC 预计，到 2030 年，我国 Token 年消耗量将达 3547 亿亿[2]。AI 基础设施需求形态将由短时、单次推理转向长周期、高并发运行，在持续拉动 GPU 推理算力的同时，显著增加 CPU、内存、存储、网络及智能体运行平台需求，推动 AI 基础设施竞争转向单位 Token 成本、持续服务能力与异构协同效率。从算力需求来看，智能体的爆发将带动我国算力需求持续高速增长，预计未来 2-3 年，我国算力需求平均每年将增长近 10 倍。从算力需求结构来看，推理算力需求将远远超过训练算力，预计 2029 年我国推理算力市场占比将达到 80%。

更多免费行业报告 并购家 www.ipoipo.net

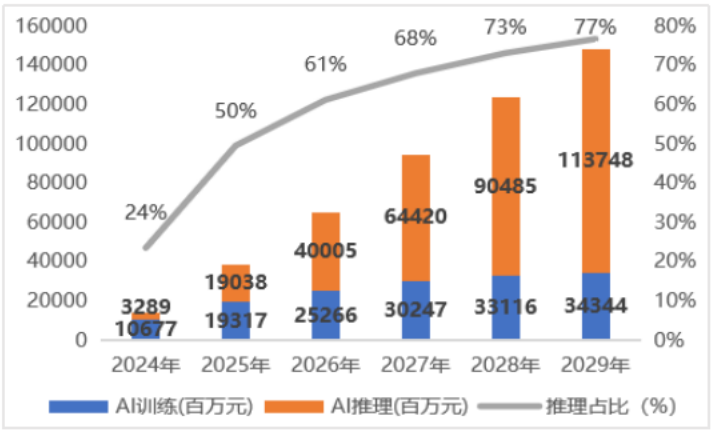


图 2 我国推理算力的市场占比[3]

（三）AI Coding 成为主流 AI 应用场景之一，驱动软件生产范式重构

AI Coding 流程高度数字化、任务边界清晰、结果可经编译和测试验证，是智能体最易形成业务闭环和量化价值的场景之一。2026 年，编码智能体正由代码补全和问答助手，向需求解析、架构设计、跨文件修改、测试验证、代码审查、持续集成和运维交付延伸，研发人员由直接编写代码转向定义任务、监督过程和验收结果。IDC 数据显示，2025 年中国 AI 编程市场规模为 3.99 亿元，预计 2026 年底增至 11.73 亿元，增长约 194%；阿里 Qoder 以 47.6%的份额居首[4]。全球市场同样加速，OpenAI Codex 周活跃用户已超过 400 万，GitHub 项目中的编码智能体活动占比较 2025 年末翻倍。

3、AI 基础设施竞争焦点转向系统全栈，市场空间持续扩大

（一）AI 基础设施竞争焦点由“算力/模型”单维能力比拼，转向“算力-模型-智能体-应用”系统性全栈布局

早期 AI 基础设施竞争以算力规模、集群性能、模型跑分等单一维度能力比拼为主。随着 AI 智能体加速落地，算力的商业价值必须经由模型能力封装、智能体自主任务编排等核心环节，最终下沉至行业应用才能形成可持续的商业闭环，算力、模型、智能体、应用之间的垂直耦合关系持续增强。全球头部科技企业纷纷立足自身资源优势与核心禀赋，差异化布局、系统性构建全栈 AI 基础设施体系，具体布局路径主要分为四大方向。

一是坚持端到端全栈自研，筑牢软硬件协同竞争壁垒。聚焦芯片、云平台、基础模型、智能体及智能体平台、应用等全链条，搭建完整自研闭环体系，最大化软硬件协同效能，夯实底层技术自主权，如谷歌依托自研 TPU、谷歌云底座、Gemini 模型、全域智能体与搜索生态等完成全链条布局；阿里依托平头哥自研芯片、阿里云、通义大模型、智能体平台及产业应用等构建完整自研栈。

二是通过“超级应用”场景牵引，抢占下一代 AI 交互入口。依托国民级超级应用的海量用户和场景优势，补齐底层算力、云、模型、智能体等基础设施能力，实现场景与技术的双向迭代，如字节跳动以豆包为核心智能交互入口，联动抖音、飞书等产品矩阵，持续拉动底层智算集群优化与大模型、智能体能力迭代升级；腾讯依托微信、腾讯会议等应用生态，以微信智能体、元宝 AI 为核心抓手，基于海量用户交互场景与产业服务需求，持续打磨并完善全栈 AI 基础设施底座。

三是打造中立开放平台，构建多元兼容的 AI 产业供给生态。

更多免费行业报告 并购家 www.ipoipo.net

以公有云与智能体工程化平台为核心载体，秉持多元兼容、开放共享的发展策略，不强制绑定自有技术产品，打破单一技术体系的局限，为行业客户提供多样化、可选择的 AI 技术与服务方案，如 AWS 依托自研 Trainium 芯片、安全可信云底座与 Bedrock 智能体平台，在落地自有技术体系的同时，全面兼容第三方主流大模型，打造开放中立的 AI 基础设施，适配客户的多样化技术需求。

四是推行“多芯多云”弹性布局，降低单一供应链依赖风险。以基础大模型研发、智能体技术迭代为核心目标，通过多品牌芯片选型、多云端平台部署的弹性架构，构建灵活可控的全栈 AI 服务能力，规避单一算力、单一云厂商绑定带来的供应链约束与运营风险，如 OpenAI、Anthropic 等头部 AI 企业，灵活组合英伟达 GPU、AMD 的 GPU、谷歌 TPU、亚马逊 Trainium 等各类优质算力资源，同时依托微软、谷歌、亚马逊等多云平台输出模型与智能体服务，搭建起高弹性、低风险的 AI 基础设施体系。

（二）生成式 AI 驱动云市场保持强劲增长

全球看，据 Synergy Research Group 数据，2026 年第一季度企业云基础设施服务支出达 1290 亿美元，同比增速 35%，已连续第九个季度上升[5]。高盛预测全球 AI 基础设施（含计算、数据中心及电力投资）2026 年-2031 年累计投入将达到 7.6 万亿美元，其中 2026 年支出约 7650 亿美元[6]。**国内看，**根据 IDC 数据，我国公有云 IaaS 市场同比增速恢复至 20.0%，公有云整体市场 2025 下半年总值突破 2000 亿元，GPU/MaaS 等 AI 相关领域贡献显著提

升[7]。AI 是公有云增长的主要驱动力。中国生成式 AI 的 I+M 市场规模预计 2029 年底达 2180 亿元，年复合增长率 72%，且在公有云占比提升迅速（6%→39%）[8]。

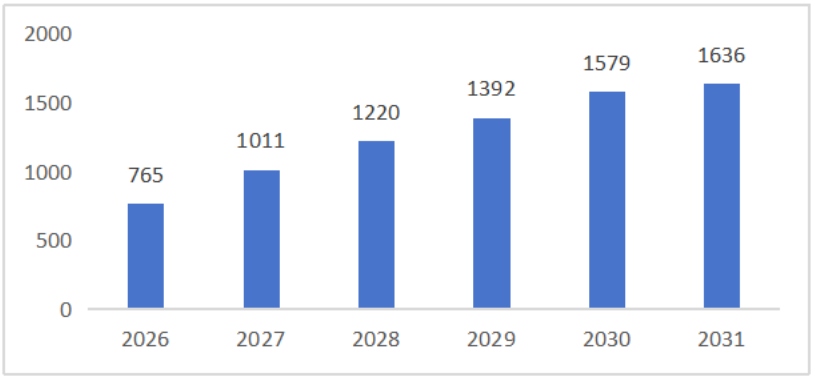


图 3 2026-2031 年全球 AI 基础设施投资预测（单位：千亿美元）

（三）AI 基础设施市场竞争主体多元，市场持续向头部厂商集中

一是除云商、运营商等算力服务商外，模型商也已着手布局 AI 基础设施服务能力。2026 年 7 月 1 日，Meta 被报道正在筹备云计算业务，将 AI 基础设施中的过剩算力对外出售[9]；WAIC 大会期间，商汤定位自己为“最懂大模型的 AI 基础设施服务商”，与 20 多家生态伙伴共同发起国产 AI 基础设施生态共建计划[10]。

二是 AI 基础设施市场寡头垄断趋势明显。头部云商具备智算、模型、平台等全栈 AI 服务能力，有利于形成高壁垒的行业 AI 解决方案，稳固其云市场份额。全球云市场中，前三名合计占据 67%的公有云市场份额；中国智算 IaaS 基础设施、大模型 PaaS 训推公有云平台、MaaS 公有云平台市场前三名份额总量均超 50%[11][12][13]。

更多免费行业报告 并购家 www.ipoipo.net

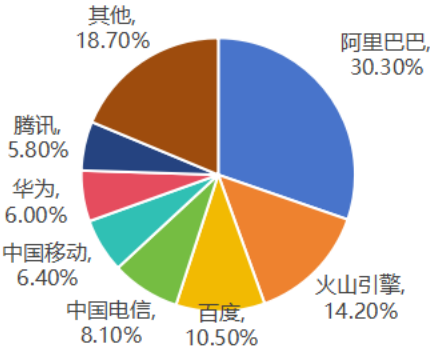


图 4 中国 GenAI IaaS 市场份额，2025H2

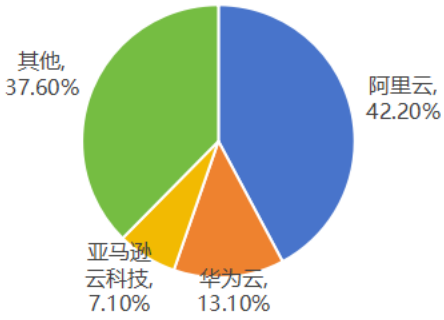


图 5 中国大模型训推公有云市场份额，2025

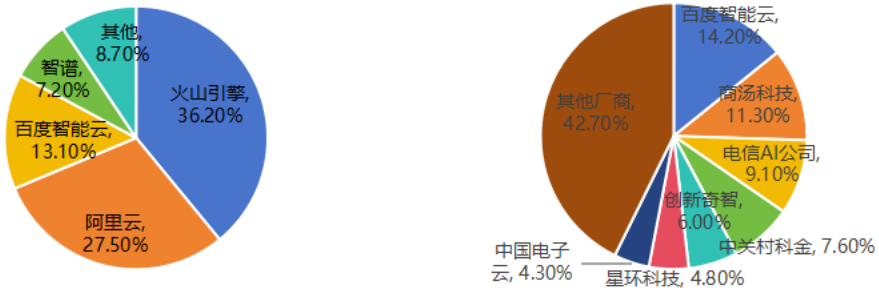


图 6 中国 MaaS 公有云（左）、私有化（右）市场份额，2025

4、AI 工程范式持续迭代，Harness 成为智能体核心底座

智能体从“对话”走向“行动”的能力跃迁，是驱动 Harness 工程成为基础设施底座的根源。当智能体具备工具调用能力，即相当于延伸出可对外操作的“手脚”，可联动的外部系统与工具种类不断丰富，行为边界持续扩张。运行过程因此更加难以管控，安全风险、流程失控、成本不可预测等挑战显著增加。Gartner 预测，到 2026 年底，40%的企业应用将嵌入 AI Agent，如何确保这些 Agent 在复杂环境中稳定、安全、可控地运行，正成为产业亟需解决的核心命题[14]。AI 工程化正是解决这一问题的关键核心。

（一）AI 工程范式正经历从提示词工程、上下文工程到 Harness 工程的三阶段演进

2022 至 2024 年，AI 领域聚焦提示词工程，核心在于优化单次模型调用的输入质量。2025 年起，伴随复杂长周期业务兴起，上下文工程成为新重心，开发者不再仅关注“如何写提示词”，而是思考“每一步该向模型提供什么上下文”，工作重心转向上下文管理与信息环境构建。进入 2026 年，伴随智能体自主执行能力的持续升级，AI 工程开始迈入 Harness 工程阶段。不同于前两阶段聚焦于“如何让模型说得更好”，Harness 工程关注的是“如何让智能体做得更稳”，即为自主 AI 智能体构建一套完整可控的运行环境，通过标准化约束、工具编排、校验闭环与反馈机制，保障智能体在生产环境中稳定执行任务。行业共识以比喻阐明其定位：“模型是引擎，Harness 是把引擎变成整车的工程——发动机决定上限，Harness 决定能不能跑、跑多远、跑多稳。”[15] 北京市专项政策明确将“实施驾驭层工程”列为强化智能体底层共性技术攻关的核心方向，围绕上下文工程优化、任务持久化、多智能体协作、系统可扩展等夯实智能体共性底座[16]。



图 7 智能体工程化演进各阶段关系示意图

（二）Harness 工程的核心价值在于弥补大模型原生能力无

更多免费行业报告 并购家 www.ipoipo.net

法实现的风险约束与流程治理短板

通过指令、上下文、工具、边界、反馈、治理六层构建标准化的运行框架，对智能体的工具调用、任务执行等建立标准化管控机制，弥补大模型原生能力无法实现的风险约束与流程治理短板，从根本上解决智能体运行不可控、成本不可预测、安全风险突出等生产级痛点。Anthropic 的对比实验直观验证了这一价值：同一个模型、同一句提示词，以简单方式运行 20 分钟耗费 9 美元，核心功能完全无效；而以完整 Harness 运行 6 小时耗费 200 美元，却交付了真正可用的产品[17]；据 Nate B Jones 研究，同一模型不变，仅优化 Harness，其编程任务成功率从 42%提升至 78%。以上实验表明，在底层大模型不变的前提下，Harness 架构能够从根本上保障智能体在真实业务场景中的执行质量。AI 竞争正从“大脑对决”转向“身体格斗”——模型负责思考，Harness 负责将思考转化为可靠的行动。



图 8 智能体架构示意图：模型与 Harness 的关系及六大组件布局[18]

（三）头部科技厂商正加速将 Harness 工程产品化，构建面向生产环境的 Agent 托管服务

一是云服务商将 Harness 定位为智能体时代云基础设施的核心底座。AWS 于 2026 年 6 月正式宣布 Amazon Bedrock AgentCore 中托管 Agent Harness 全面可用，无需编写编排代码，仅通过配置即可快速生成生产级智能体，内置代码沙箱、跨会话记忆、网页浏览等配套能力，Harness 实现与底层模型解耦，支持会话内无损切换多厂商模型，依托统一网关完成工具访问、身份权限、全链路观测一体化管控[19]；阿里云 AgentRun 定位为面向生产环境的一站式 Agentic AI 基础设施平台，支持模型无关接入、毫秒级沙箱隔离、MCP/Skills 双轨工具扩展，助力企业降低 60%总拥有成本（TCO）[20]。PPIO（派欧云）通过一体化 Agent Harness 整合沙箱、记忆、工具编排能力，拉长智能体任务循环时长，并落地 PPClaw、PPHermes 托管智能体，实现 7×24 小时长效稳定运行与自我修复[21]。

二是模型厂商，将 Harness 以托管 API 形式直接交付。Anthropic 于 2026 年 4 月正式发布 Claude Managed Agents，以托管 API 形式提供服务，将 Harness 与生产级基础设施（状态、记忆、权限、定时执行）打包交付，使开发者无需自建编排循环与安全沙箱，即可在数天内完成生产级智能体的构建与部署[22]。Claude Managed Agents 已在多个客户实践场景落地，取得显著业务成效。如 Notion 集成后原型制作时长从 12 小时缩短至 20 分钟，可同时启动 30 余个智能体处理编码、生成幻灯片等多种任务；Sentry 一名工程师在数周内完成了从 Bug 检测到可合并 Pull

更多免费行业报告 并购家 www.ipoipo.net

Request 的完整基础设施构建，而传统方式需要数月。

二、智能体时代 AI 基础设施发展及产业趋势预判

1、AI 基础设施总体视图：智能体重塑 AI 全栈能力和服务体系

智能体规模落地实际生产场景，本质是一项复杂系统性工程。智能体工程的核心价值，是让 AI 从“内容生成工具”转变为“任务执行系统”，真正成为驱动业务、降本增效、创造价值的生产力工具。智能体工程化的难点不只是大模型能力，而是如何将非确定性大模型系统转化为可控、可观测、可恢复的生产级应用，围绕业务目标、推理、工具、记忆、执行、评估和治理，构建一个可运行、可监控、可优化的 AI 任务执行体系。智能体时代，产业关注重点转向如何打造 AI 全栈引擎，构建高性能、低成本、高可靠、可复用的 AI 基础设施，支撑智能体工程的开发、运行、协同和治理全链路，推动智能体实现体系化建设、规模化应用和持续化运营。

智能体驱动 AI 全栈能力和服务体系重塑和跃迁。AI 全产业链呈现较明晰的三层架构协同发展生态，并且“自上而下（智能体应用驱动）”和“自下而上（基础设施底层支撑）”形成跨层级的动态、正反馈增强回路。最上层是 Agent 应用层，关键核心要素是智能体落地场景应用和集成交付，核心解决把智能嵌入到业务场景、集成到企业工作流；中间层是 Agent 基础设施层，关键核心要素是模型和智能体平台，重点围绕模型外部运行环境构

建系统能力，核心为支撑智能体开发、运行、协同和治理的基础体系；最底层是 AI 云基础设施层，关键核心要素包括云计算、算力和网络，云网基础设施跃迁重点是向 Agent 原生演进，打造低成本高效率 Token 工厂。

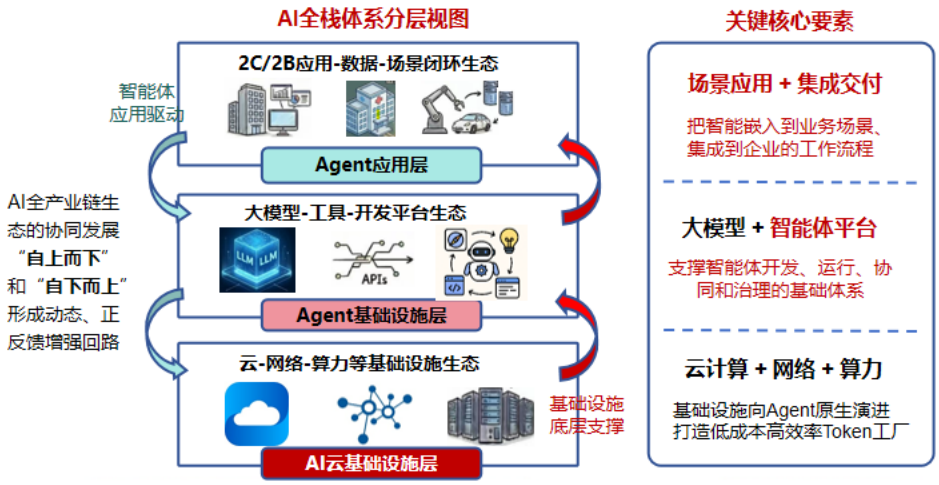


图 9 智能体时代 AI 全栈体系分层示意图和关键核心要素

2、算力：智能体重塑算力基础设施架构、规模及布局重心

（一）算力架构转向“CPU+GPU”协同主导，CPU 成为系统核心约束

一是 AI 智能体重塑算力负载结构，带动计算架构重心向 CPU 迁移。相比传统生成式 AI 偏向单轮请求与单次生成，智能体系统需要经历规划、检索、工具调用、执行、反馈、迭代等多步骤流程，对系统级调度与控制的要求显著提高。在智能体场景中，GPU 继续负责 token 生成与高并行矩阵运算，但任务切分、内存管理、I/O 调度、API 调用、长上下文状态维护、外部工具接入等控制层负载则适合由 CPU 承载。市场研究机构 TrendForce 预测，当前

更多免费行业报告 并购家 www.ipoipo.net

AI 数据中心的 CPU 与 GPU 配比约为 1:4-1:8，在智能体时代，这一比例将演变至 1:1-1:2[23]。**二是 CPU 正在成为智能体时代算力性能约束的核心瓶颈。**一方面，智能体系统延迟更多由 CPU 决定。根据佐治亚理工学院与英特尔研究，在主流智能体框架任务测试中，CPU 侧处理时延占比 40%-90%。在 Haystack RAG（检索增强生成）任务中，甚至超过 90%[24]。另一方面，CPU 正面临严重的产能短缺。主要原因在于先进制程和封装被高毛利 GPU 和 HBM 挤占、ABF 载板等上游关键材料紧缺等。英特尔、AMD 相继于 3 月、4 月启动涨价，服务器 CPU 涨幅达 10%-15%，交货周期拉长至 8-12 周，部分型号长达 6 个月。

（二）算力布局转向推理优先，推理基础设施持续向高效率、低成本演进升级

一是科技巨头持续加大算力投资，布局重心向推理转移。TrendForce 预计，2026 年亚马逊、谷歌、Meta、微软和甲骨文部署的英伟达 GB/VR 芯片，将使其 AI 训练总算力同比增长超 56%、AI 推理总算力同比增长约 122%[25]。德勤预测，到 2026 年，“推理”（即运行 AI 模型）将占据全球 AI 总计算能力的三分之二[26]。**二是推理基础设施持续向高效率、低成本演进升级，训推解耦、异构协同等多路线并行。**2026 年 4 月，谷歌发布第八代 TPU，首次将训练与推理任务拆分至独立芯片，包括专为模型训练打造的 TPU 8t 和专为模型推理打造的 TPU 8i，每瓦计算性能比第七代 TPU（Ironwood）提升了一倍。英伟达 Vera Rubin 依托 NVLink 6

高速互联、HBM4 高带宽内存等技术，较上一代 Blackwell，每瓦推理吞吐量提高 10 倍，单 token 成本降至 1/10；其 Groq LPX 推理加速机架（含 256 个 LPU）与 Vera Rubin 机架协同，进一步降低数据传输损耗，可使每兆瓦推理吞吐量最高提升 35 倍，高度适配智能体低延迟、长上下文推理需求。

（三）超大规模高密度智算集群仍是 AI 产业刚需，超节点、十万/百万卡级集群持续深化建设

一是超节点驱动算力基础设施竞争迈入芯片、网络、存储、能耗、软件等的系统级博弈新阶段。超节点是指通过高速互联网络、统一调度系统和高性能计算架构，将数十上百个计算节点整合为一个逻辑上高度协同的巨型计算单元，同时满足大带宽、低时延和内存统一编址，旨在解决单机算力不足、跨节点通信延迟高等问题。全球头部厂商纷纷推出超节点产品，如华为 CM384、英伟达 NVL72/144、阿里磐久 AL128 等。同时超节点集群规模加速由百卡向千卡级演进升级，如华为最新发布的 Atlas 950 SuperPod 已达 1024 卡规模，最大可扩展至 8192 卡。**二是 AI 算力集群持续向十万/百万卡演进。**2026 年 1 月，美国 xAI 投运全球最大 AI 训练集群 Colossus 2，拥有超 55 万颗英伟达 GPU，功耗达 1 吉瓦[27]。同年 4 月，微软宣布其 Fairwater AI 数据中心已提前上线，单个集群内整合数十万颗英伟达 Blackwell 架构 GB200 GPU。同年 7 月，中科曙光发布我国首台全国产 10 万卡 AI 超算集群曙光 8000，同步接入国家超算互联网郑州核心节点，宣告我国

更多免费行业报告 并购家 www.ipoipo.net

算力基建进入 10 万卡级规模化部署新阶段[28]。

3、网络：算力需求牵引光网络成为构建高效 AI 基础设施的关键

（一）“光进铜退”成为超节点智算集群高密互联的技术刚需，LPO、NPO、CPO 多技术路线并行

一是随着超节点规模持续扩大，铜缆逐渐逼近距离、带宽等物理极限。受信号衰减等固有物理特性限制，铜缆的有效传输距离存在刚性约束，如 224G 速率下，DAC（无源铜缆）有效传输距离小于 2 米；ACC/AEC（有源铜缆）通过集成信号调理芯片，以增加芯片功耗和成本为代价，可将铜缆有效传输距离延伸至 3-5 米[29]。同时传输速率越高，信号衰减越严重，从 224G 迈向 448G，提升铜缆的单通道速率异常困难，且多通道并行会导致严重串扰，并增加布线难度和成本，实用价值较低。**二是光互联成为超节点主流互联方案，LPO、NPO、CPO 多技术路线并行。**光互联具备高带宽、低时延、抗干扰等优势，且低损耗特性使其不受传输距离约束，高度适配大规模智算集群内卡间高速互联。头部企业加快布局光互联方案，如华为 Atlas 950 超节点已规模使用 800G LPO 光互联，并明确下一代 960 超节点将落地 NPO 光互联；阿里明确下一代超节点 UPN512 将使用 LPO/NPO 光互联代替铜互联；英伟达明确下一代超节点 NVL576 将采用 CPO 光互联代替铜互联。

（二）多数据中心协同正在成为 AI 集群规模扩张的关键手段，光网络是关键

受建设周期、许可、电力供应等客观条件的限制，传统只在单个数据中心进行大模型训练的方法面临性能和成本难平衡的难题。头部企业已经开始执行相关计划，将其大模型训练从一个站点扩展到多个数据中心园区，如微软通过高速光纤网络，把分布在不同州、相距超过1100公里的2个十万卡级大型AI数据中心连接成一个“虚拟的超级计算机”，主要用于前沿模型训练，旨在快速提升AI计算效率，同时降低单一地区能源压力。中国电信完成千亿参数大模型跨500公里、2个智算中心分布式训练试商用，性能达集中式的99%以上。谷歌自Gemini1.0开始，就已实现多个数据中心（最远距离几十公里）协同训练模型。

（三）AI算力扩张引爆光纤需求，供需失衡驱动主流厂商加速锁定光纤产能

一是AI成为光纤需求增长核心驱动力。以英伟达H100/H200 SuperPOD集群为例，单机柜光纤消耗量约是传统机柜的5-10倍以上。根据英国商品研究所（CRU）预测，全球AIDC光纤需求占比预计从2024年不足5%飙升至2027年30-35%[30]。**二是光纤供给持续紧缺，头部厂商加速锁定光纤产能。**光纤扩产周期较长，如上游光纤预制棒扩产周期长达18-24个月，且全球产能高度集中于我国。据CRU调研，我国已有13家成熟厂商制定了明确的光纤预制棒重启或扩产计划，预计最早2027年形成有效供给。英伟达、Meta等主流厂商加快通过签订长期协议等方式锁定光纤产能，如2026年5月，英伟达与康宁宣布达成长期合作，康宁承诺将其在美国的光连接制造能力提升10

更多免费行业报告 并购家 www.ipoipo.net

倍，光纤产能提升 50%以上，同时英伟达将投资 5 亿美元换取康宁认股权证。

4、云：云基础设施能力及产品服务面向智能体全面升级

（一）云基础设施能力全面升级，为智能体提供高效、稳定、安全运行环境

智能体应用带动云服务从“为人提供计算资源”变成“为智能体提供运行环境”。智能体运行具有长任务、有状态、不可预测、大规模并发、多组件协同等特点，传统的云原生服务聚焦算存资源供给，针对无状态、一次性调用的“响应请求”服务机制已无法满足智能体需求。

表 1 云原生与 Agent Native 云工作负载特征对比

	传统云工作负载	Agent Native 云工作负载
运行特点	以“响应请求”为中心，一次性调用，无状态	以“任务执行”为中心，持续运行，有状态，长上下文
生命周期	毫秒级	分钟~小时级、天级
资源消耗	相对平稳，内存峰值是平均值的 1~3 倍	突发，内存峰值可达平均值的 15.4 倍，CPU 利用率在 13%~175%波动

运行流程方面，相比于传统的云原生模式的函数调用，智能体运行以“任务执行”为中心，运行过程包括组装上下文、调用 LLM、调用工具/读写文件、回填上下文等多个操作，并可能形成多个循环（Loop），运行时间从毫秒级跃升至分钟/小时级，有些场景甚至需要天级；资源消耗方面，智能体所需资源波动明显，内存峰值可达平

均值的 15.4 倍，CPU 利用率在 13%-175%波动，而传统云负载内存峰值是平均值的 1-3 倍，CPU 利用率相比之下较为平稳[31]。

工作负载变化推动云平台能力升级。一是极致性能提升。主要表现在冷启动时延从秒级缩短到毫秒级，内存开销从 GB 级缩减至 MB 级，单节点支持数千沙箱并发运行。如腾讯云 CubeSandbox 冷启动时间＜60ms，实际内存开销＜5MB，单节点支持 2000 沙箱并发。**二是支持长链、有状态任务运行。**通过分层记忆架构实现记忆持久化，支持多 Agent 间状态认知一致、多轮对话上下文记忆一致。阿里云基于 Tair 的分层记忆架构使其成本降低 81.5%、延迟降低 80%。**三是隔离增强，风险可追溯。**从容器软隔离升级到物理硬隔离，并提供 Agent 身份认证、行为管控、skill 安全风险防范等更精细的安全治理能力。阿里云 Agent 安全中心支持统一管理 Agent 暴露面、身份和运行时安全，AI 安全护栏 2.0 支持实时动态行为监控。

（二）面向 Agent 协同与应用优化产品能力，提升产品使用体验及易用性

一是产品全面 Skill 化、MCP 化、CLI 化，简化被 Agent 调用难度。主流云厂商全面开展产品 Skill 化、MCP 化、CLI 化改造，核心目的通过标准化能力封装，将各类产品能力转化为可被上层智能体灵活调度、无缝调用的标准化工具服务，彻底打通产品与智能体的协同壁垒，实现产品能力的智能化复用与自动化调度。现阶段重点覆盖产品管理、业务运营、数据智能分析三大核心领

更多免费行业报告 并购家 www.ipoipo.net

域，适配企业数字化、智能化运营的核心需求，助力业务流程自动化、智能化升级。如阿里云已覆盖 2 万余个 API，为 300 余款产品提供 Skills 全生命周期管理；百炼核心能力已 CLI 化。阿里云、腾讯云均已上线 Skill 聚合平台，提供自有产品 Skill 以及开源 Skill 等。飞书、腾讯文档等多款主流应用均上线 Skill。

表 2 Skill 、MCP、CLI 能力对比

	Skill	MCP	CLI
作用	规定 Agent 内部执行方式	标准化连接 Agent 外部系统	Agent 在本地系统执行任务
运行位置	进程内	独立进程/网络服务	子进程/Shell 执行
适用场景	流程编排、约束智能体行为…	多 Agent 协同、调用外部数据/能力…	本地文件处理、执行本地命令…

二是将 Agent 融入产品，提高产品易用性。一方面，将 Agent 深度集成到 SaaS 产品中，围绕产品原生业务场景进行能力延伸与功能拓展，依托智能体赋能重塑业务交互模式，丰富产品价值边界，提升场景化业务处理能力。如腾讯 IMA 新增 copilot 功能，支持用户创建专属 Agent 作为专属知识伙伴，基于用户数据自主扩充知识体系，沉淀并支持自定义拓展 Skills。另一方面，立足原有核心业务体系，引入 Agent 能力重构与优化现有业务流程，全面提升业务运行效率与整体运营效能。阿里云瑶池 Data Agent 支持自然语言输入，支持复杂、多步骤的数据分析任务执行，在全球权威数据区评测基准 BIRD-CRITIC 中登顶，已服务超 6000 家企业。

5、模型：大模型能力迭代且供给格局分化，适配智能体规模化落地需求

（一）大模型面向智能体场景开展多维度技术能力升级

大模型产业发展重心从单次问答转向多步骤自主任务执行，行业评判大模型价值的核心标尺同步切换为智能体自治水平。依据 Artificial Analysis 2026 评测体系，工具调度、超长上下文管理、Token 精细化成本管控、输出结果复核等长流程能力成为四大核心考核指标，倒逼头部厂商围绕推理、上下文、工具调用、成本控制四大模块完成技术迭代。

在推理层面，厂商落地可控混合推理、可验证编码能力，引入思考过程自主调控机制，动态平衡算力投入、运行速度与任务完成质量。例如 Anthropic 在 Claude Opus 4.8 中引入"投入控制"机制，允许用户直接控制思考资源投入。

在上下文层面，实现 200M 超长上下文承载，支撑复杂长周期业务会话。Google Gemini 3.1 Pro 通过模型架构优化、分层记忆工程、超算集群底座的综合工程成果，实现了 200M Token 海量上下文低损耗、高精准的稳定推理。

在工具调用层面，原生工具与代码调用能力深度内置，打通模型与外部系统的自动化交互链路，有效降低传统外挂自动化方案的调用复杂度与工程损耗。Open AI GPT 5.4 则通过"截图→视觉分析→操作决策→结果验证"的端到端推理，在单次推理内直接输出点击、输入等电脑操作指令，降低传统外挂自动化方案的调

更多免费行业报告 并购家 www.ipoipo.net

用复杂度与工程损耗，大幅简化执行链路。

在 Token 消耗层面，行业持续压缩 Token 消耗，主流新一代模型 Token 使用量较前代降低近五成，从底层解决规模化落地成本高企痛点。Open AI GPT-5.5 原生搭载动态视觉 Token 压缩算法，针对 4K 高清图像场景，在完全保留语义精度的前提下，智能剔除冗余像素 Token，可将图像类 Token 体积压缩 35%以上。

整套技术体系围绕“自主执行”构建，呈现向 Harness 六层标准化工程架构理念演进的趋势，与 Harness 共同补齐智能体全流程可观测、故障自愈、算力弹性调度等基础设施能力，支撑智能体从简单测试走向大规模生产级落地。



图 10 大模型能力升级主要方向

（二）通用基座模型趋于收敛，垂直行业模型多元细分发展

模型市场格局呈现两极分化演进态势，与智能体基础设施商业化路径高度匹配。**一是通用大模型赛道资源向头部集中，**厂商依托 MoE 稀疏架构、轻量化底座打造标准化普惠模型，以平价规模化路线抢占通用智能体、海外增量市场，依靠海量调用量摊薄研发与推理成本；头部通用模型厂商同步布局高端增值服务，面向复杂推理场景推出高毛利专属版本，依靠持续迭代与稳定 SLA 服务获取溢价收益。**二是政务、工业、金融等垂直领域催生大量**

定制化行业模型，形成多元细分供给格局。行业模型高度适配本地业务流程、数据规范与安全合规要求，对中立多模型网关、政务级安全沙箱、行业智能体模板等基础设施存在刚性需求。 供给侧分化也催生全新商业变现模式：通用模型厂商主打标准化 API 订阅收入，行业垂直模型依托开放生态分成、平台托管交付实现收益转化，中立 AI 基础设施平台（TokenHub、智能体引擎）作为中间载体，打通通用基座与行业定制模型的统一调度、计量分账链路，承接未来四年国内 AI 市场 25 倍规模增长红利[32]，化解各厂商算力成本高、开源流量难以变现的经营困境。

6、智能体：智能体技术架构持续向高效化、开放化、解耦化方向演进，全面提升智能体运行效率与场景适配能力

（一）智能体通过记忆分层、协同互联技术实现高效记忆管理与跨系统协作

一是记忆机制持续突破，从单层存储向分层解耦架构演进。传统记忆系统长期困于碎片化存储与粗粒度摘要两类极端方案之间，前者保留细节但孤立脆弱，后者压缩成本但丢失关键信息。记忆机制从架构层面，进行存储与检索的解耦以及信息分层，使检索效率与细节保真不再相互制约。如微软 Memora 通过将“存储内容”（丰富记忆）与“检索方式”（轻量级抽象与线索锚点）解耦，解决了传统记忆系统在细节保留与高效组织之间的两难困境。在 LoCoMo 和 LongMemEval 基准测试中，Memora 创下了新的

更多免费行业报告 并购家 www.ipoipo.net

最优性能，相比全上下文推理可节省高达 98%的 Token 消耗[33]。腾讯云开源的一套面向 AI Agent 的分层记忆引擎 TencentDB Agent Memory 如下图所示，通过 L0 原始对话层、L1 原子记忆层、L2 场景归纳层、L3 用户画像层的四层架构，将不同粒度的信息分层存储，化解单次长任务的信息过载。在生产环境验证中，该记忆系统作为 OpenClaw 插件接入后，最高可节省 61.38%的 Token 消耗，任务通过率相对提升 51.52%，长期记忆 PersonaMem 准确率从 48%提升至 76%[34]。



图 11 TencentDB Agent Memory 四层渐进式记忆架构

二是多智能体协同正从框架内协作走向跨系统互联。复杂问题的解决日益依赖多智能体协同，主流采用中心化分工模式，调度协调智能体负责目标拆解与任务路由，各专家智能体承担检索、代码生成、数据分析等专项工作，多个专业化智能体并行执行、交叉校验，显著提升复杂任务完成质量与运行效率。AutoGen 框架的 GroupChatManager 和 MetaGPT 的角色分工体系是实现这一模

式的典型代表[35]。在此基础上，MCP、A2A 等通信协议的标准化正将多智能体协同从“同一框架内的协作”推向“跨平台、跨系统的互联互通”。智源研究院在《2026 十大 AI 技术趋势》中指出，随着 MCP、A2A 等通信协议趋于标准化，智能体间拥有了通用“语言”，多智能体系统将突破单体智能天花板，在科研、工业等复杂工作流中成为关键基础设施[36]。

（二）智能体平台依托统一模型网关、可插拔架构实现多模型灵活调度，与模型解耦将成为行业标配

企业不再满足于绑定单一模型供应商，智能体平台在多模型之间自由切换正成为显著趋势。Vercel CEO 指出“AI 时代的关键之争在于模型与智能体是否解耦”；Omdia 将 2026 年 AI 云架构重新定义为基础设施、MaaS、AaaS 三个独立层级，智能体平台与模型分离已成为产业共识。AWS AgentCore Harness 明确将模型解耦作为核心设计原则，“客户可以选择任意模型，并在会话中途切换提供商而不丢失上下文”，支持 Bedrock、OpenAI、Gemini 及任何 OpenAI 兼容模型，实现“一套 Harness、任意模型”[37]；阿里云多智能体治理平台 AgentTeams 支持接入多个 AI 模型供应商，在不同 Worker 间灵活切换模型；Microsoft Agent Framework 允许开发者无论使用哪个底层模型提供商，都能获得一致的接口，提供统一的开发体验。

更多免费行业报告 并购家 www.ipoipo.net



图 12 Agent 开发与托管平台架构

7、应用：业务价值优先，场景禀赋决定智能体落地节奏

（一）智能体应用将优先落地高价值、可闭环的数字化场景

一是业务价值与结果可验证性决定智能体应用优先级。企业部署智能体的评价标准将由“是否使用前沿模型、是否具备自主执行能力”，转向“能否形成可量化的业务净收益”。任务频次高、流程边界清晰、数据高度数字化、结果能够验收、异常可以回退的场景，更容易建立“执行-校验-反馈-优化”闭环，将率先从辅助工具升级为业务执行单元。麦肯锡 2025 年全球调查显示，23%的受访企业已经在至少一个职能中规模化部署智能体，另有39%处于试验阶段，但在任一具体职能中，规模化部署比例均未超过 10%，反映出企业仍主要围绕少数价值明确的流程集中投入，而非全面铺开[38]。Anthropic 披露的 Notion 客户案例显示，编码智能体将原型制作时间由 12 小时缩短至约 20 分钟，并可同时运行 30 余项智能体任务，原型制作时间下降 97%，说明软件研发等数字原生、结果易验证场景已具备较为清晰的价值闭环[39]。

二是数据、系统和风险禀赋决定智能体由试点走向规模化的

节奏。智能体不仅需要调用模型，还要持续访问业务数据、操作企业系统并承担任务结果，因此，数据质量、系统接口、权限配置、责任边界和人工兜底机制越成熟，智能体可获得的自主权限和应用规模越高。预计未来 1-2 年，客户服务、企业知识管理、软件研发和数据分析等数字化场景将率先规模化；财务、法务、合规、人力资源等专业职能将主要采用“智能体执行、人工审批”模式；生产制造、供应链、采购及设备运维等连接物理设备的场景，则需先从局部流程和低风险任务切入，待实时数据、设备接口和责任机制成熟后逐步扩权[40]。Gartner 此前预测，到 2026 年，缺乏 AI 就绪数据支撑的 AI 项目中约 60%将被放弃，表明场景数据基础将直接影响智能体项目的存续率和规模化进度。

（二）全生命周期治理与成本核算成为智能体生产准入硬性门槛

一是智能体生产准入将由“模型可用”升级为“全链路可治理”。智能体能够自主调用工具、修改数据和操作业务系统，单纯依靠上线前模型评测已无法覆盖其运行风险，治理机制需要贯穿场景设计、开发测试、上线审批、运行监控、版本变更和退出处置全生命周期。上线前应完成场景风险分级、数据授权、工具白名单、权限边界和压力测试；运行中应实行智能体身份认证、最小权限、关键动作人工审批、异常熔断和持续监控；任务结束后应保留模型版本、提示词、上下文、工具调用、数据读写和审批记录，支持责任追溯、结果复现和故障回滚。IBM 联合 Oxford Economics 于 2026 年对 2000 名技术高管的调查显示，77%的受访机构认为

更多免费行业报告 并购家 www.ipoipo.net

AI 应用速度已经超过现有治理能力；受访机构过去一年平均发生 54 起需要人工纠正的智能体事件，其中 17%属于高严重度事件[41]。智能体治理将成为其能否获得业务系统访问权和自主执行权的前置条件。

二是成本核算将由“单次 Token 价格”升级为“单位成功任务成本和风险调整后 ROI”。智能体需要多轮规划、检索、工具调用、结果校验和失败重试，单次模型调用价格难以反映真实成本。企业应将模型与算力、上下文和数据检索、API 及软件工具、沙箱与存储网络、人工复核、失败重做和预期风险损失统一纳入核算，重点考核单位成功任务成本，并同步设置单任务预算上限、最大执行步数、超时熔断、低成本模型路由、缓存复用和人工接管条件。2026 年一项针对智能体编程任务的研究发现，此类任务的 Token 消耗可达到推理和对话任务的约 1000 倍，同一任务不同运行之间的 Token 消耗最多相差 30 倍，且消耗更多 Token 并不必然带来更高准确率[42]。另一项对 30 个多智能体软件开发任务的分析显示，代码审查与迭代环节平均占总 Token 消耗的 59.4%，说明校验、重试和修正可能成为智能体的主要成本来源[43]。IBM 调查同时显示，85%的受访技术高管尚不具备完整的 AI 实时支出可视能力；Gartner 则预测，到 2027 年底，超过 40%的智能体项目可能因成本持续上升、业务价值不清或风险控制不足而被取消[44]。

8、生态：多方协同与标准互通并行，兼顾智能体落地需求和行业发展引领

（一）企业生态重心转向智能体全栈体系构建和行业伙伴争夺

一是搭建全栈体系，贯通从技术底座到业务结果的完整链路，加速智能体进入企业生产环境。AWS 发布 Agentic AI 五层技术栈全景地图，覆盖 AI 基础设施、模型、数据与知识、Agentic 平台及智能体应用，并将安全、效果、性能和成本贯穿始终，以支持业务产出和价值。谷歌云发布智能体企业（Agentic Enterprise）蓝图，定义企业 AI 五层架构，即 Agentic 超级计算机、Agentic 防护、Agentic 数据云、Agentic 平台和模型、Agentic 小组任务，以一整套垂直优化的技术栈赋能企业将智能转化为增长引擎。阿里云发布智能体原生云（Agent Native Cloud），包括提供运行基础的 Infra 基建、决定规模化的 Platform 平台和连接业务的 Desktop 桌面等三层，旨在将智能体变成可控、可复用、可协作、会进化的组织资产，为客户释放生产力。

二是联合行业伙伴，以“技术+业务”双轮驱动，实现智能体从“可用”到“专业”的关键突破。阿里云携手近百家金融 AI 联盟伙伴推出“超级智能体计划”，旨在为金融行业构建通用智能体。“超级智能体计划”包括两大板块，一是基于点金通用智能体的技术底座，内置金融场景所需的核心能力，帮助金融机构快速搭建 AI 应用；二是行业共享 Skills，整合阿里云及伙伴在信贷、理赔、投研等领域的实践经验，提升智能体在金融场景的

更多免费行业报告 并购家 www.ipoipo.net

实际应用能力。同时，根据智能体时代对行业 ISV 伙伴的需求，阿里云通过专门技术团队支持、成熟产品支持、多端口流量支持和区域销售团队支持等，帮助合作伙伴转型升级。

（二）以政策、倡议、标准等为指引推进智能体互联协同

国家网信办、国家发展改革委、工业和信息化部联合印发《智能体规范应用与创新发展实施意见》，要求“加强智能体互联协议（AIP）等智能体互联关键国家标准、行业标准的推广应用”。中国信通院联合政产学研多方代表正式发布《智能体互联互通协同发展与治理倡议》，从标准连接、规则治理、生态共融、互联互通等方面提出举措，构建 “可连接—敢协同—可持续” 的完整发展路径。《人工智能 智能体互联》系列 7 项国家标准建立起统一的技术框架和标准依据，规范智能体互联的关键环节，提升互操作性、可组合性和工程化应用能力。

三、我国 AI 及智能体产业区域性发展评估

智能体已成为 AI 产业新质生产力的核心载体，各省在政策供给、算力网络、数据要素、行业应用层面积累不同发展根基，区域发展差距逐步显现。本章创新构建 AIDP-EM 省域智能体产业潜力评估模型，围绕发展环境、基础设施、服务应用搭建可量化指标体系，完成全国 31 省份综合指数测算与梯队划分，并进一步开展分行业适配潜力评估，剖析各地算力资源与场景落地的协同短板，梳理不同区域适配主导赛道，形成分区域、分行业的差异化发展指引，为各地精准布局智能体产业提供数据支撑与路径参考。

1、智能体产业发展潜力评估模型框架

智能体产业发展潜力评估模型（Agentic Industry Development Potential Evaluation Model，AIDP-EM），从发展环境、基础设施、服务应用 3 大维度构建评估体系，包含 10 项二级指标、29 项三级指标，用于量化分析省域智能体产业发展的资源基础、平台能力、应用转化和生态成熟度。详细指标定义、数据来源见附录。

与上一年度智算评估框架相比，新模型主要进行了四方面扩展：一是将政策环境由“政策数量”升级为政策强度，并单列智能体专项政策；二是将基础设施从算力和网络拓展至存力与数据、模型与智能体能力底座；三是在服务应用中增加智能体示范场景、场景多样性和平台化率；四是统一采用 2025 年完整年度数据和 2026 年最新政策、试点名单，并引入数据质量调整机制。



图 13 智能体产业发展潜力评估指标体系

更多免费行业报告 并购家 www.ipoipo.net

表 3 AIDP-EM 三级指标体系

一级指标	二级指标	三级指标	二级权重
发展环境	政策与治理环境	• AI/算力政策强度 • 智能体专项政策强度 • 公共数据开放综合指数	9%
	经济与国家布局环境	• 地区生产总值（GDP） • GDP 增速 • 国家算力枢纽/AI 试验区覆盖数	9%
	绿色能源环境	• 国家绿色算力设施数量 • 绿色算力设施密度	7%
基础设施	算力基础设施	• 三大运营商对外服务机架估算 • 可调度智能算力估算 • 公共算力服务与调度指数	14%
	存力与数据基础设施	• 高质量数据集典型案例数 • 可信数据空间创新试点数 • 公共数据开发利用指数	11%
	网络基础设施	• 每万人 5G 基站数 • 每万人千兆宽带用户数 • 完成万兆光网试点项目数	10%
	模型与智能体能力底座	• 生成式 AI 服务备案数 • 模型/工具平台典型案例数 • 智能体开发/编排平台案例数	10%
服务应用	AI 产业创新供给	• AI 企业存量 • AI 专利存量 • AI 高校/科研机构供给	10%
	AI 应用与产业载体	• 国家级 AI 典型应用案例数 • AI 应用密度 • AI 产业园/创新载体存量	10%
	智能原生应用	• 智能体示范场景/案例数 • 智能体场景多样性 • 智能体平台化率	10%

2、评估流程与方法

(1) **指标构建：**基于 AIDP-EM 模型建立三级指标体系。一级指标保持“发展环境—基础设施—服务应用”的总体逻辑，二级指标对应智能体产业的关键能力环节，三级指标强调可量化、可比较和可持续更新。

(2) **数据赋值：**评估时点为 2026 年 7 月，基础统计以 2025 年完整年度数据为主，政策文件、绿色算力设施、可信数据空间和万兆光网等名单更新至 2026 年最新可得时点。

(3) **权重设计：**一级指标权重分别为发展环境 25%、基础设施 45%、服务应用 30%；二级指标根据产业逻辑配置权重；三级指标依据重要性、可得性和避免重复计权原则分配，合计 100%。

(4) **数据质量分级：**A 级为官方直接统计或公告，B 级为正式名单或统一研究数据，C 级为使用可核验数据形成的透明代理。质量调整系数分别为 1.00、0.90 和 0.75。

(5) **数据处理：**对 GDP、算力、备案、企业、专利和案例数等指标进行标准化处理，标准范围在 0-100 分。

(6) **指数计算：**分别生成综合评估指数、发展环境指数、基础设施指数、服务应用指数及 10 项二级指标指数。综合排名以质量调整得分为主，以降低估算指标和历史存量指标对结果的影响。

(7) **行业适配层：**在省域综合潜力评价之外，构建 31 省×11 行业评估矩阵。最终潜力得分由本地行业机会度 55%、任务适配度 25%和省域共性支撑 20%构成。本地行业机会度包括行业规模 15%、专

更多免费行业报告 并购家 www.ipoipo.net

业化 10%、数字化准备 10%、数据准备 8%、案例证据 7%、政策/集群 5%。对已完成指标进行归一化处理，并继续采用 A/B/C 数据质量调整系数。

表 4 数据质量分级与处理规则

等级	数据类型	质量系数	典型指标
A	官方直接统计或公告	1.00	2025 年 GDP、生成式 AI 服务备案、万兆光网完成项目
B	正式名单或统一研究数据	0.90	绿色算力设施、可信数据空间、智能体统一样本案例
C	透明公式或存量数据	0.75	省级算力校准值、AI 企业/专利存量、应用密度

3、全国 31 省发展智能体产业的潜力评估

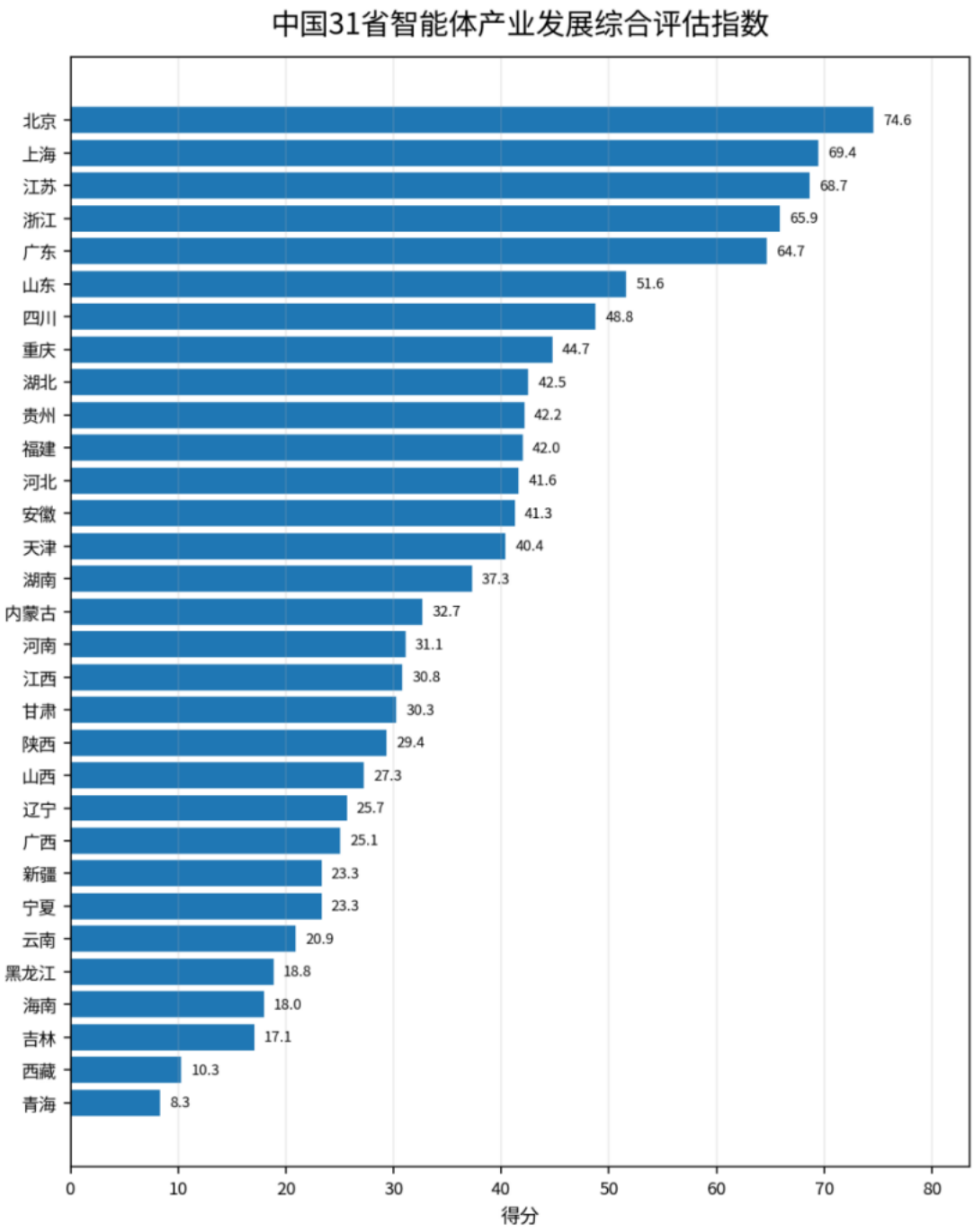


图 14 中国 31 省智能体产业发展综合评估指数

（一）综合评估结果：我国省域智能体产业发展呈现头部集聚与梯度分化并存的基本格局

更多免费行业报告 并购家 www.ipoipo.net

从质量调整得分看，全国省域智能体产业发展形成四个梯队。领先梯队平均得分为 68.6 分，较强梯队为 44.8 分，中等梯队为 32.6 分，追赶梯队为 18.3 分。政策、数据、平台和产业生态正在向少数头部地区加速集聚。

表 5 省域智能体产业发展梯队划分

梯队	代表省份	得分区间	主要特征
领先梯队	北京、上海、江苏、浙江、广东	64.7—74.6	政策、数据、算力、模型平台和产业应用多维领先，具备全国辐射能力。
较强梯队	山东、四川、重庆、湖北、贵州、福建、河北	41.6—51.6	在算力、数据、网络或行业应用方面具有特色，但存在模型平台或智能原生应用短板。
中等梯队	安徽、天津、湖南、内蒙古、河南、江西、甘肃、陕西、山西、辽宁	25.7—41.3	具备一定资源和产业基础，综合能力尚未形成闭环，需要聚焦优势场景突破。
追赶梯队	广西、新疆、宁夏、云南、黑龙江、海南、吉林、西藏、青海	8.3—25.1	多以绿色能源、网络或单点项目为优势，数据、平台和应用生态仍需补齐。

区域格局呈现“京沪创新策源、长三角均衡集群、粤鲁产业应用、川渝黔算力节点”的基本特征。领先梯队全部位于东部地区，说明智能体产业并非单纯由低成本算力驱动，而是更依赖高质量数据、模型工具平台、人才和高价值应用场景的综合集聚。中西部地区在绿色能源和算力设施方面具备比较优势，但其资源优势能否转化为智能体产业优势，取决于数据流通、行业工具链和服务应用生态。

（二）发展环境评估结果：东部地区在政策治理、经济基础和创新资源方面的综合优势较为突出

浙江、上海、江苏、广东和北京在发展环境维度评估得分位居前五。浙江依托高强度政策、较快经济增长和完善的公共数据开放体系获得最高分；上海政策治理得分居首，形成较强的制度供给和平台集聚能力；江苏在经济基础和国家绿色算力设施数量方面表现突出。

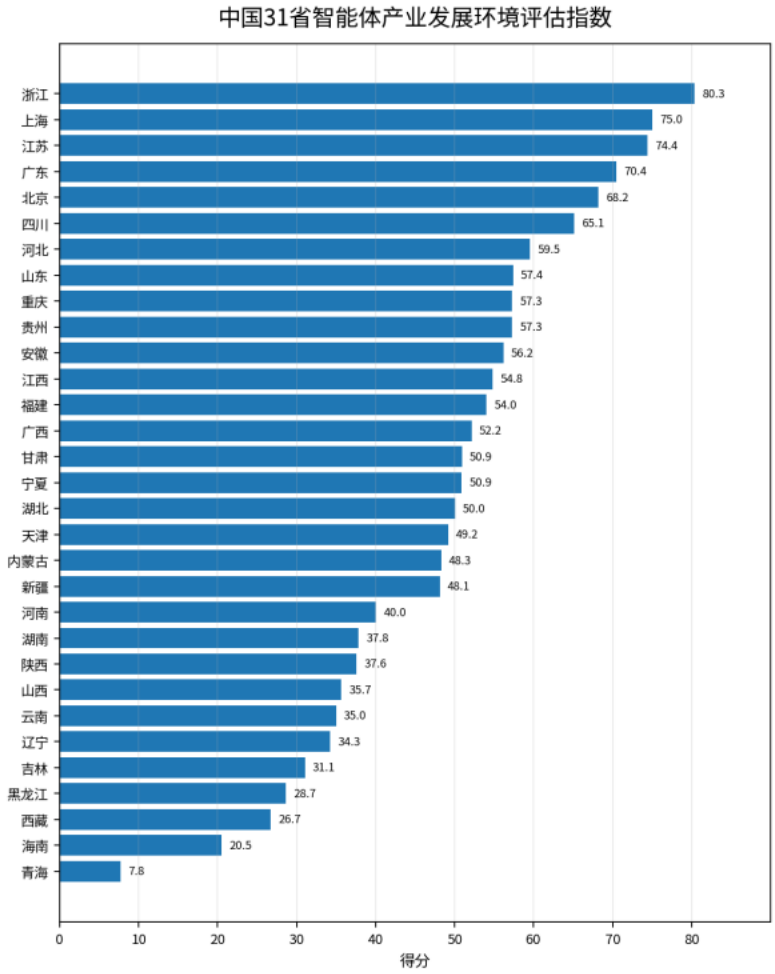


图 15 中国 31 省智能体产业发展环境评估指数

绿色能源维度呈现与综合排名差异化空间格局。新疆、宁夏、江西、江苏等地区依托绿色算力设施数量或密度取得较高分，表明绿色能源禀赋可为训练、推理和数据处理等高耗能环节提供成本优势。但绿色算力设施数量并不直接等同于智能体产业规模，仍需通过跨区域

更多免费行业报告 并购家 www.ipoipo.net

调度、行业数据集和模型服务把能源优势转化为可交易、可调用的智能服务。

（三）基础设施评估结果：省域智能体基础设施已由单一算力供给转向算力、数据、网络和模型工具协同发展

基础设施指数覆盖算力、存力与数据、网络、模型与智能体能力底座。北京、上海、江苏、广东、浙江位居前五。截至 2025 年底，全国 5G 基站达到 483.8 万个，千兆宽带用户达到 2.38 亿户，三大运营商对外服务机架数达到 93.8 万架，可调度智能算力超过 94.4 EFlops[45]。2026 年，工业和信息化部公布 136 个完成万兆光网试点任务的项目，进一步推动模型训练、推理和超高清视频等高带宽业务落地[46]。

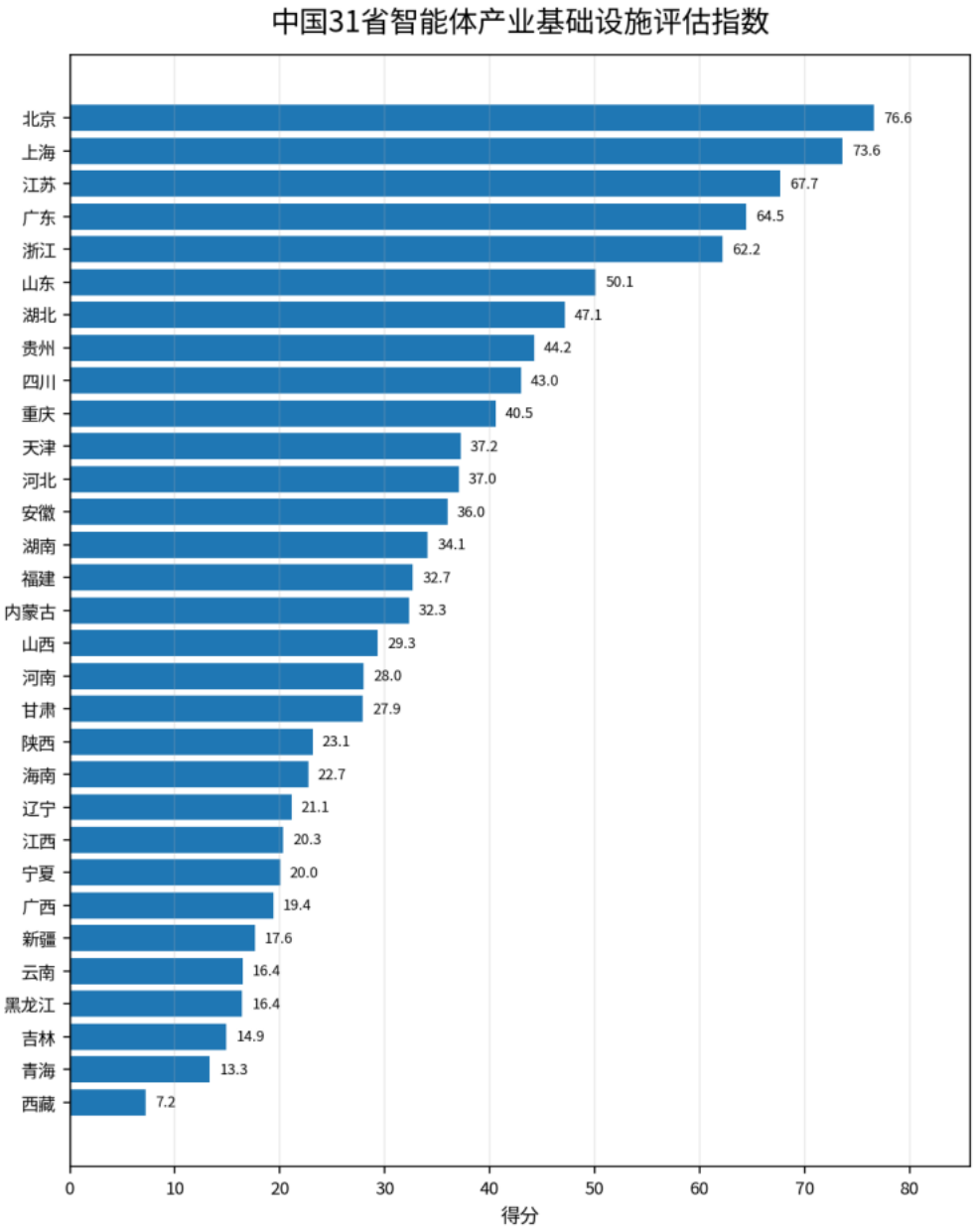


图 16 中国 31 省智能体产业基础设施评估指数

北京的优势主要来自存力与数据、模型与智能体底座；上海在算力基础设施和数据流通方面较为均衡；江苏兼具算力规模与万兆网络项目优势；广东算力基础设施得分最高，但数据和智能体平台环节仍有提升空间。贵州则呈现典型的“算力强、应用弱”特征，算力基础设施得分较高，但存力数据、模型工具和智能原生应用尚未形成同等

更多免费行业报告 并购家 www.ipoipo.net

规模。由此看，省域智能体基础设施已从“建算力”转向“算力、数据、网络、模型工具协同”。

（四）服务应用评估结果：智能体服务应用能力主要集聚于 AI 产业基础和场景转化能力较强的地区

服务应用指数包含 AI 产业创新供给、AI 应用与产业载体、智能原生应用。北京、江苏、广东、浙江、上海位居前五。服务应用得分的省际标准差约为 20.4，高于发展环境和基础设施，说明应用侧的**头部集中度最为明显**。

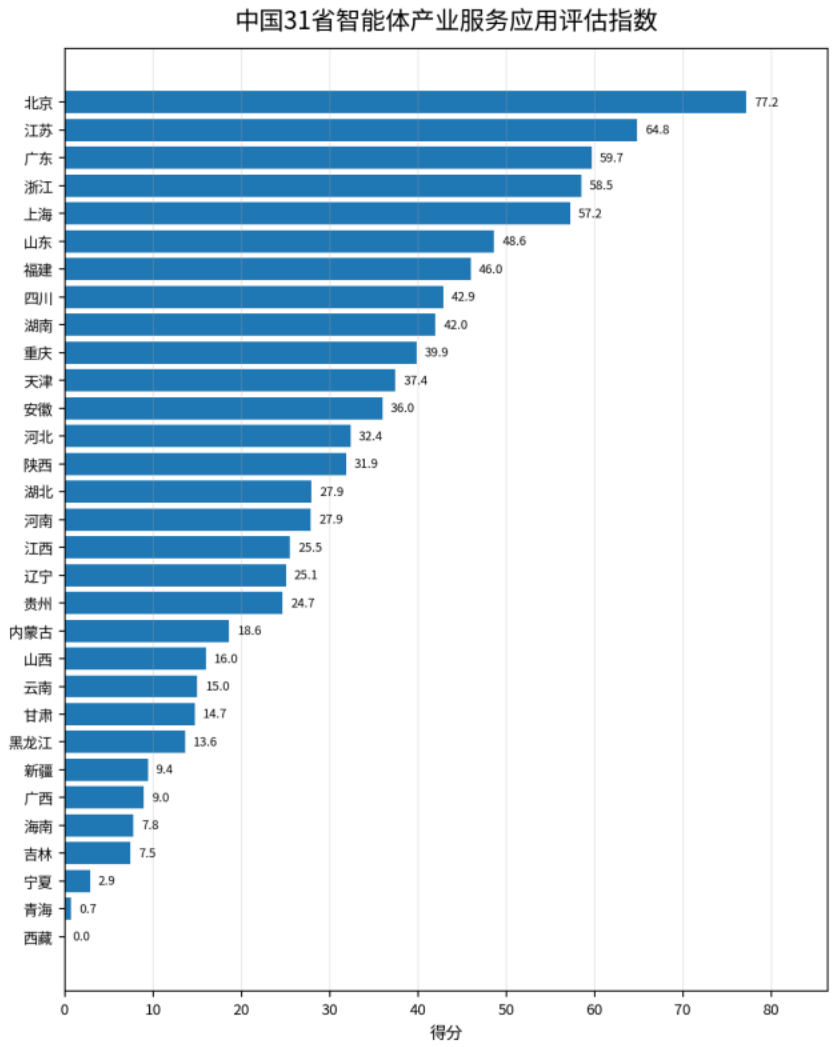


图 17 中国 31 省智能体产业服务应用评估指数

北京凭借 AI 企业、科研机构、模型服务、智能体案例和平台型应用的综合优势领先；江苏在工业应用和多场景案例方面表现突出；广东国家级 AI 典型应用案例较多，产业转化能力较强；浙江在模型工具平台和应用场景之间形成较好衔接。相比之下，部分算力节点省份服务应用得分偏低，反映出算力供给与行业需求之间仍存在接口、数据、人才和商业模式障碍。

北京、江苏、浙江等地区在智能体案例数量、场景覆盖和平台化能力方面处于相对领先位置。智能原生应用指数由智能体示范场景/案例数、场景多样性和平台化率构成。北京以 79.8 分显著领先，江苏、浙江分列第二和第三，福建、重庆并列其后。在统一样本中[47]，北京拥有 33 个智能体案例，覆盖 9 类主要场景，其中含 9 个平台型案例；江苏案例数量和场景覆盖较多，但样本中平台型案例相对不足；浙江则在案例数量和平台化之间较为均衡。

更多免费行业报告 并购家 www.ipoipo.net

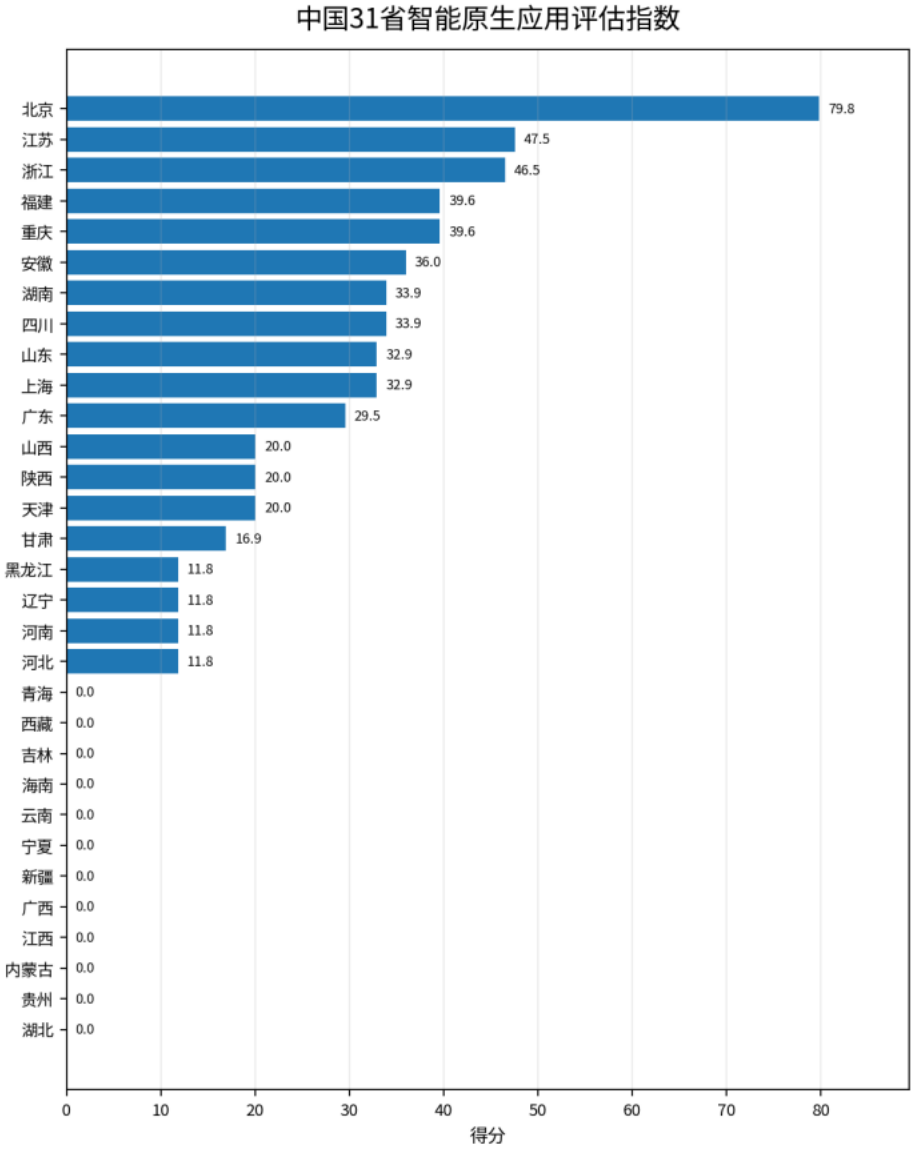


图 18 中国 31 省智能原生应用评估指数

（五）基础设施与服务应用的协同分析：基础设施与服务应用的协同程度，直接决定算力资源能否转化为智能体产业能力

以 31 省基础设施和服务应用平均得分为界，可形成四类发展状态。北京、上海、江苏、浙江、广东、山东、四川、重庆、河北、安徽、天津位于“高基础设施—高应用”象限；湖北、贵州位于“高基础设施—低应用”象限；福建、湖南、陕西位于“低基础设施—高应用”象限；其余地区主要处于“双低”象限。

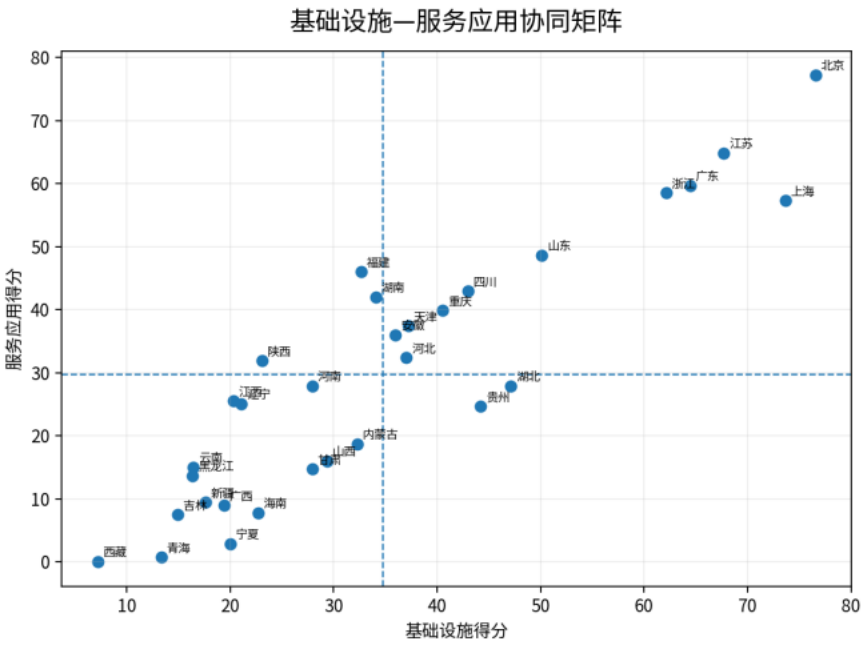


图 19 基础设施—服务应用协同矩阵（AIDP-EM 模型测算；虚线为 31 省平均水平）

“高基础设施-低应用”地区应把发展重点从继续扩张资源规模转向建设公共模型服务、行业工具链、可信数据空间和标杆场景；“低基础设施—高应用”地区则可依托全国一体化算力网、云服务和边缘节点补足本地资源，避免重复建设。该结果说明，智能体产业政策需要从单一基础设施投资转向“资源供给—平台能力—场景闭环”的协同治理。

（六）Top10 省份优劣势分析及发展建议

综合潜力 Top10 省份呈现差异化竞争格局，各省在政策治理、算力基础设施、数据要素、模型智能底座、AI 产业供给、行业场景六大链条（最高评估得分在 80 及以上）上优势分布不均，短板特征清晰。建议各省结合自身优劣势推动智能体各产业链条协同发力。

更多免费行业报告 并购家 www.ipoipo.net

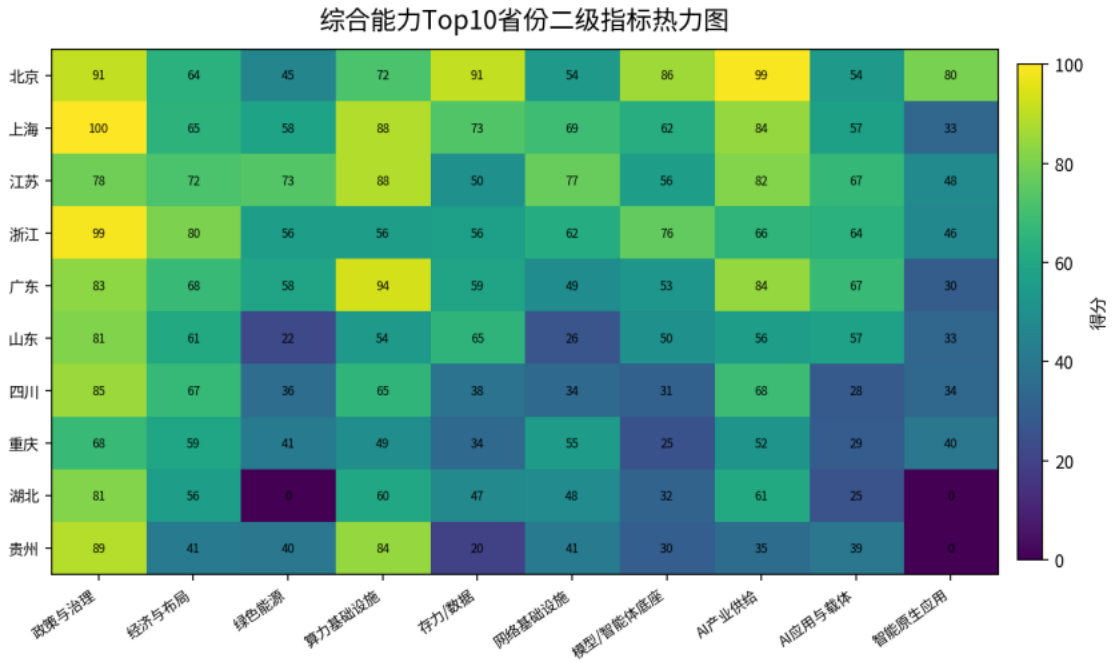


图 20 综合能力 Top10 省份二级指标热力图

表 6 综合能力 Top10 省份优势、短板与发展方向

省份	主要优势	主要短板	建议发展方向
北京	数据、模型与智能体底座、AI 产业供给、智能原生应用	绿色能源、网络基础设施相对弱	强化绿色算力协同和跨区域资源调度，输出智能体平台、标准和高端服务。
上海	政策治理、算力基础设施、存力数据	智能原生应用和应用转化不足	发挥金融、航运、制造等场景优势，提升智能体案例规模和平台化能力。
江苏	算力、万兆网络、AI 产业供给和应用案例	存力数据和智能体平台型案例不足	建设工业智能体公共平台，打通制造数据、工具和执行系统。
浙江	政策治理、经济活力、模型工具平台、应用场景	算力规模和智能原生应用仍有提升空间	发挥平台经济和民营企业优势，推动智能体从应用插件走向企业流程平台。
广东	算力基础设施、AI 产业供给、国家级应用案例	数据流通、网络协同和智能体平台化不足	围绕制造、终端和互联网生态建设跨企业智能体协作网络。

山东	存力数据、工业应用和产业基础	绿色能源、网络和模型底座偏弱	以工业数据空间和行业智能体为抓手，强化模型服务与网络支撑。
四川	政策环境、经济布局、AI 产业供给	模型工具和应用转化不足	依托成渝算力与科研资源，培育公共智能体平台和优势行业场景。
重庆	政策、网络和部分平台型案例	数据、模型底座和应用规模不足	聚焦汽车、装备、物流等流程闭环，提升案例数量和跨系统执行能力。
湖北	政策、网络、数据和科研供给	绿色设施和智能原生发展较弱	依托科教和产业集群扩大智能体示范场景，形成平台化复制能力。
贵州	政策与算力基础设施	数据利用、模型平台、服务应用和智能体案例不足	从算力承载向数据加工、模型推理和行业智能体运营延伸。

4、全国 31 省发展行业智能体的潜力评估

（一）评估方法

为了进一步分析我国 31 省发展行业智能体的潜力，本报告在省域智能体产业发展潜力评估模型（AIDP-EM）基础上增设行业适配评估。该评估综合考虑省域共性支撑、行业发展基础和智能体任务特征，形成 31 省与 11 个重点行业的交叉评价矩阵，为各省确定行业智能体的重点发展方向和应用优先顺序提供参考。

(1) 模型分析

本报告从本地行业机会、任务适配和省域共性支撑三个方面对行业适配模型进行分析。**本地行业机会**反映各省相关行业的规模、专业化程度、数字化基础、数据资源、应用案例及政策集群条件；**任务适配**反映智能体在特定行业中的业务价值、流程适配、技术可行、数据可得、风险可控和复制扩展能力；**省域共性支撑**则沿用 AIDP-EM 对

更多免费行业报告 并购家 www.ipoipo.net

发展环境、基础设施和服务应用的评价结果。

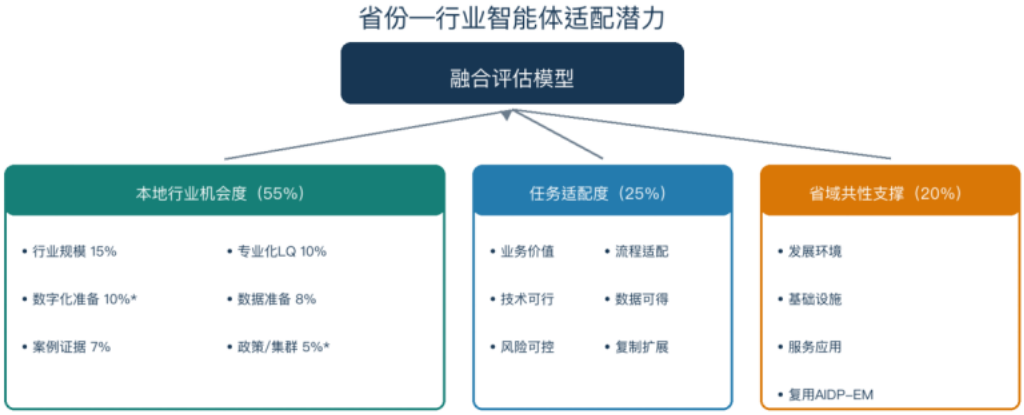


图 21 省份-行业智能体适配评估框架

(2) 评估方案

指标构建：依据行业适配模型，构建由本地行业机会、任务适配和省域共性支撑构成的评估指标体系。本地行业机会包括行业规模、专业化区位商[1]、数字化准备、数据准备、案例证据和政策集群六项指标。

指标赋值：法人单位规模和专业化区位商主要依据《中国统计年鉴 2025》进行赋值；数据准备和案例证据依据国家数据局正式名单及现有智能体案例样本进行赋值；数字化准备和政策集群指标将依据行业主管部门及省级政府公开资料进一步补充。

权重确定：行业适配综合得分中，本地行业机会、任务适配和省域共性支撑的权重分别为 55%、25%和 20%。

评估指数结果：根据指标得分、指标权重和数据质量调整系数，计算各省在 11 个重点行业中的适配得分，并形成行业任务适配度、

1 区位商（Location quotient），又称专门化率，是区域经济学中衡量地区产业专业化程度的指标

省份—行业综合适配指数和省域行业适配空间格局等评价结果。

（二）评估结果

基于行业适配评估方法确定的指标、权重和阶段性数据，本报告从行业任务适配度和省域行业适配布局两方面，对 31 省发展行业智能体的重点方向进行分析，31 省×11 行业智能体适配阶段性得分热力图见附录（三）。

(1) 行业任务适配度

数字化程度高、跨系统协同需求强的行业，智能体任务适配度相对较高。

如图 10 所示，在暂不考虑省际差异的情况下，信息通信与软件的任务适配度为 96.7 分，位居 11 个行业首位；商贸零售、金融服务、交通物流、能源电力和工业制造的得分均超过 83 分。上述行业的业务流程数字化程度较高，跨系统协同需求较强，任务结果相对易于验证，具备规模化复制条件。医疗健康和农业农村受到高风险决策、敏感数据治理、设备联网水平和县域数据颗粒度等因素约束，现阶段可优先发展辅助服务、运营管理及保留人工复核的应用场景。

更多免费行业报告 并购家 www.ipoipo.net

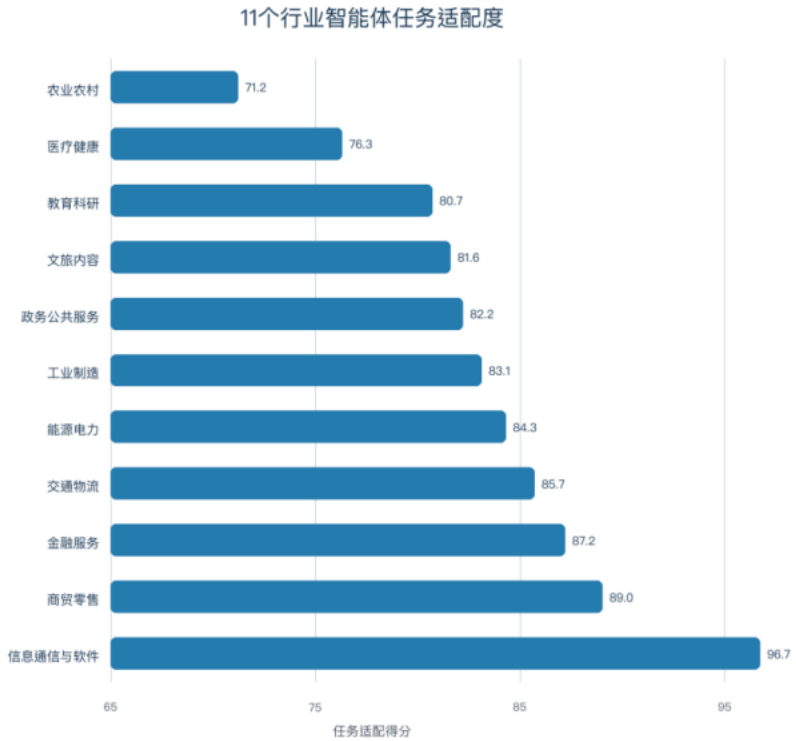


图 22 11 个行业智能体任务适配度

(2) 省域行业适配布局

东部地区的平台型行业优势较为突出，中西部地区在特色产业场景中具有比较优势。

如下图所示，农业农村在 9 个省份处于阶段性首位，政务公共服务在 6 个省份处于首位，能源电力在 3 个省份处于首位；工业制造、交通物流、金融服务、医疗健康和信息通信与软件则分别在部分省份形成相对优势。该结果表明，行业适配格局与省域综合潜力排序并非简单对应。综合潜力较高地区具备建设平台型智能体、跨行业工具链和开展全国性服务输出的条件；产业基础较强地区可围绕制造、物流和商贸等高频业务流程构建行业级平台；资源型和农业地区可依托能源、农情及政务服务等数据资源，形成专业化的任务闭环。

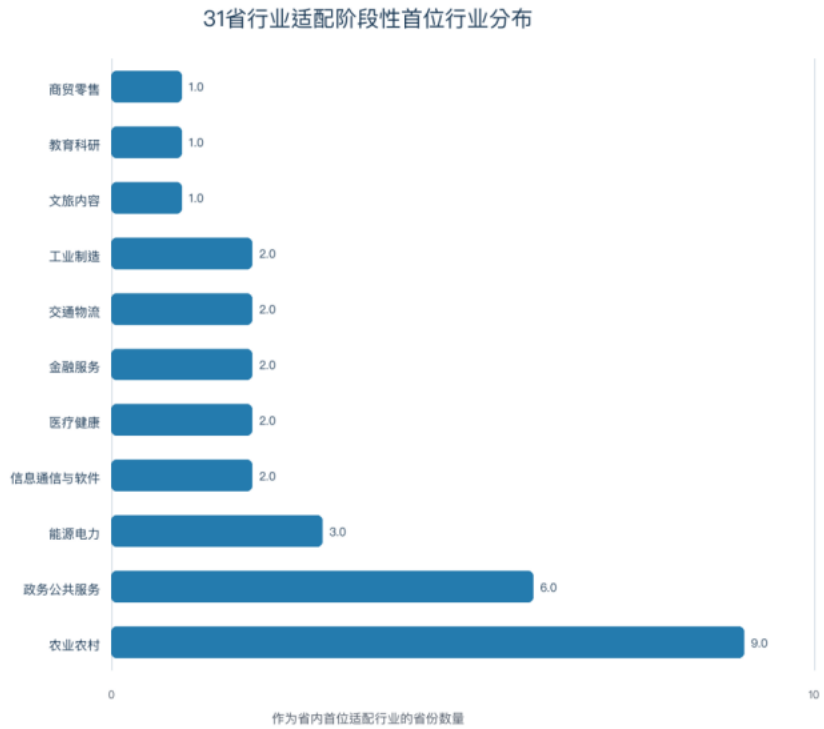


图 23 31 省行业适配阶段性首位行业分布

结合省域共性支撑得分及各省 Top3 行业（评估得分见附录）匹配特点，可将国内省份划分为平台输出型、产业闭环型、科教医疗型、资源场景型四类，不同类型地区适配的重点行业存在明显差异。各地应立足自身产业基础，差异化推进行业智能体建设，找准平台输出、工业打通、公共服务、场景闭环等不同发展侧重点，详情如下表所示。

表 7 省域类型与行业智能体推进重点

类型	代表地区	重点适配行业	推进重点
平台输出型	北京、上海、浙江、广东	信息软件、金融、商贸、文旅、交通	建设模型工具、智能体平台和跨区域服务输出能力
产业闭环型	江苏、山东、河北、重庆、安徽	制造、物流、农业、能源	打通工业软件、设备、供应链和数据空间

更多免费行业报告 并购家 www.ipoipo.net

科教医疗型	湖北、四川、陕西、辽宁	医疗、教育科研、政务、文旅	优先非诊断辅助、科研知识与公共服务场景
资源场景型	山西、内蒙古、江西、新疆、宁夏、青海	能源、农业、政务	以运维、调度、农技和公共服务形成可验证闭环

5、全国区域性智能体产业发展建议

（一）推动基础设施建设从资源扩张转向服务化和可调用化

对湖北、贵州等基础设施能力较强但应用得分偏低的地区，应减少单纯追求机架和算力规模的倾向，重点提升公共算力平台利用率、模型 API 服务、工具接入、数据治理和行业客户运营能力。对算力资源相对不足但应用活跃的地区，可通过跨区域调度、云边协同和万兆光网降低本地重资产建设压力。

（二）建设公共智能体平台，降低企业从试点到规模化的门槛

公共智能体平台应提供模型选择、工作流编排、工具注册、长期记忆、权限管理、身份认证、评测监控和人工兜底等通用能力。地方政府、运营商和云服务商可共同建设行业智能体开发环境，以标准化接口连接政务系统、工业软件、CRM、ERP、知识库和物联网设备，避免企业重复开发底层能力。

（三）以高价值、可验证、可闭环场景作为应用推广主线

智能体场景应优先选择业务价值高、流程适配度高、技术可行性强且风险可控的任务，重点关注流程重复、跨系统操作、多步骤判断和结果可验证的业务。可从客服、运维、软件开发、供应链、财务对

账、合同审核和报告生成等场景切入，按照“辅助建议—人工审批—受控执行—闭环自治”分阶段推进。

（四）实施差异化行业布局，构建“一省一业多场景”推进体系

各地区应避免同质化的全行业布局，可依据前三位适配行业形成“一个主导行业、若干可复用场景”的项目组合。综合潜力领先地区应重点提升智能体平台、技术标准和工具生态的外溢能力；制造业与商贸业基础较强地区应强化跨企业供应链协同和生产执行闭环；能源与农业资源型地区应重点建设行业数据空间、远程运维及县域服务智能体；医疗、金融等高风险行业应优先部署辅助决策和运营管理应用，并持续保留人工审批、责任界定和全链路留痕机制。

四、总结与展望

2026 年迎来智能体规模化爆发的关键之年，通用智能体、行业智能体、多智能体协同系统加速从概念验证走向常态化部署，AI 发展正式迈入智能体时代。智能体具备自主感知、持续推理、任务规划、长期交互与多轮执行能力，彻底重塑 AI 算力调用模式、资源调度逻辑与基础设施架构，推动 AI 基础设施从支撑大模型训练推理的基础载体，转向承载智能体持续运行、自主决策、动态协同的新型运行底座，带动全产业链迎来新一轮结构性变革。

在此背景下，AI 基础设施需求呈现全新特征：不再是单一静态算力供给，而是面向智能体 7×24 小时持续运行、长短任务混合调度、多智能体互联协同的动态弹性资源需求。从底层算力硬件、高速互联网络、异构算力调度平台，到中间件、智能体运行环境、资源运营服

更多免费行业报告 并购家 www.ipoipo.net

务、行业解决方案，整条 AI 基础设施产业链迎来价值重构机遇，软硬件厂商、云服务商、行业客户围绕智能体落地开展深度协同，催生全新市场空间。

与此同时，面向智能体时代的 AI 基础设施建设仍存在诸多突出挑战。如异构算力资源调度体系难以适配智能体碎片化、突发性、长周期混合任务特征；国产算力、框架、智能体开发工具链全栈适配生态仍需持续完善；传统以 GPU 硬件、基础算力租赁为主的商业模式，无法匹配智能体“算力+运行环境+调度服务+生态工具”一体化交付的客户诉求，面向多智能体场景的分层价值变现模式亟待探索。

立足智能体产业发展机遇与现实挑战，我们倡议产业各方深化协同创新、共建开放生态：推动产学研用联合攻关适配智能体场景的新型算力架构、高速互联技术与动态资源调度体系；打通软硬件全栈适配链条，完善国产智能体开发与运行底座；探索面向智能体业务的分层化、可持续运营商业模式；联合打造多行业智能体标杆落地场景。以开放联动、务实共建的理念，推动 AI 基础设施向弹性化、自治化、服务化持续演进，助力智能体产业健康有序发展，共建供需均衡、自主可控、价值闭环的智能体时代 AI 基础设施产业生态。

五、附录

1、智能体产业发展潜力评估模型的指标定义与数据依据

编码	一级指标	二级指标	三级指标	权重	口径与来源
1	发展	政策与治	AI/算力政策	3%	算力政策与 AI 政策两项各占 2.5 分；沿用

编码	一级指标	二级指标	三级指标	权重	口径与来源
	环境	理环境	强度		前版省级正式政策核验结果。
2	发展环境	政策与治理环境	智能体专项政策强度	3%	省级正式政策明确部署智能体记 5 分，否则记 0 分；逐省文件与证据保留在政策审计页。
3	发展环境	政策与治理环境	公共数据开放综合指数	3%	采用 2025 中国开放数林指数省域/直辖市综合指数。
4	发展环境	经济与国家布局环境	地区生产总值（GDP）	3%	各省 2025 年地区生产总值初步核算值。
5	发展环境	经济与国家布局环境	GDP 增速	3%	各省 2025 年 GDP 不变价增速。
6	发展环境	经济与国家布局环境	国家算力枢纽/AI 试验区覆盖数	3%	国家算力枢纽覆盖标记加国家新一代人工智能创新发展试验区城市数。
7	发展环境	绿色能源环境	国家绿色算力设施数量	4%	2025 年度国家绿色算力设施名单按所在地归集，共 60 项。
8	发展环境	绿色能源环境	绿色算力设施密度	3%	国家绿色算力设施数量÷2025 年常住人口×1000（人口单位：万人）。
9	基础设施	算力基础设施	三大运营商对外服务机架估算	5%	以原表省级 IDC 机架分布为权重，将工信部 2025 年 93.8 万架全国总量分配到各省。
10	基础设施	算力基础设施	可调度智能算力估算	6%	以原表省级智能算力分布为权重，将 2025 年三大运营商 94.4EFlops 全国总量分配到各省。
11	基础设施	算力基础设施	公共算力服务与调度指数	3%	智能算力非零、智能算力高于中位数、机架高于中位数、国家布局覆盖、万兆项目覆盖五项各 1 分。
12	基础设施	存力与数据基础设	高质量数据集典型案例	3.5%	国家数据局 2025 高质量数据集典型案例中地方推荐案例按属地归集。

更多免费行业报告 并购家 www.ipoipo.net

编码	一级指标	二级指标	三级指标	权重	口径与来源
		施	数		
13	基础设施	存力与数据基础设施	可信数据空间创新试点数	3.5%	国家数据局 2025 可信数据空间创新发展试点按申报单位属地归集，共 63 项。
14	基础设施	存力与数据基础设施	公共数据开发利用指数	4%	采用 2025 开放数林指数利用层分值；直辖市分项缺失时使用统一比例估算。
15	基础设施	网络基础设施	每万人 5G 基站数	4%	以原表省级 5G 基站分布为权重，将 2025 年全国 483.8 万站分配后除以 2025 人口。
16	基础设施	网络基础设施	每万人千兆宽带用户数	3%	以原表光宽用户分布为权重，将 2025 年全国 2.38 亿千兆用户分配后除以 2025 人口。
17	基础设施	网络基础设施	完成万兆光网试点项目数	3%	工信部 2026 年完成万兆光网试点任务的 136 个项目按所在地归集。
18	基础设施	模型与智能体能力底座	生成式 AI 服务备案数	4%	国家网信办累计完成备案的生成式 AI 服务按地区归集。
19	基础设施	模型与智能体能力底座	模型/工具平台典型案例数	3%	人工智能赋能新型工业化典型案例中的技术底座方向，按地方推荐主体归集。
20	基础设施	模型与智能体能力底座	智能体开发/编排平台案例数	3%	从中国互联网协会 80 个统一样本案例中筛选平台、编排、聚合、开发与生态类案例。
21	服务应用	AI 产业创新供给	AI 企业存量代理	4%	沿用用户原始底表的省级 AI 企业数量，作为产业主体存量代理。
22	服务应用	AI 产业创新供给	AI 专利存量代理	3%	沿用用户原始底表的 AI 专利数量，作为创新产出存量代理。
23	服务应用	AI 产业创新供给	AI 高校/科研机构供给	3%	沿用用户原始底表的 AI 相关高校、科研机构数量。

编码	一级指标	二级指标	三级指标	权重	口径与来源
24	服务应用	AI 应用与产业载体	国家级 AI 典型应用案例数	4%	人工智能赋能新型工业化典型应用案例，剔除已计入 I42 的技术底座案例。
25	服务应用	AI 应用与产业载体	AI 应用密度	3%	国家级 AI 典型应用案例数÷AI 企业存量代理×1000。
26	服务应用	AI 应用与产业载体	AI 产业园/创新载体存量	3%	沿用用户原始底表的 AI 产业园、孵化器和创新载体数量。
27	服务应用	智能原生应用	智能体示范场景/案例数	4%	中国互联网协会《智能体应用发展报告（2025）》80 案例按申报单位属地归集。
28	服务应用	智能原生应用	智能体场景多样性	3%	将 80 案例归入网络运维、政务公共服务、工业制造、金融、医疗、内容、司法安全、客服等类别，统计各省覆盖类别数。
29	服务应用	智能原生应用	智能体平台化率	3%	智能体开发/编排平台案例数÷智能体案例数×100；无案例省份记 0。

2、各省智能体产业发展评估结果

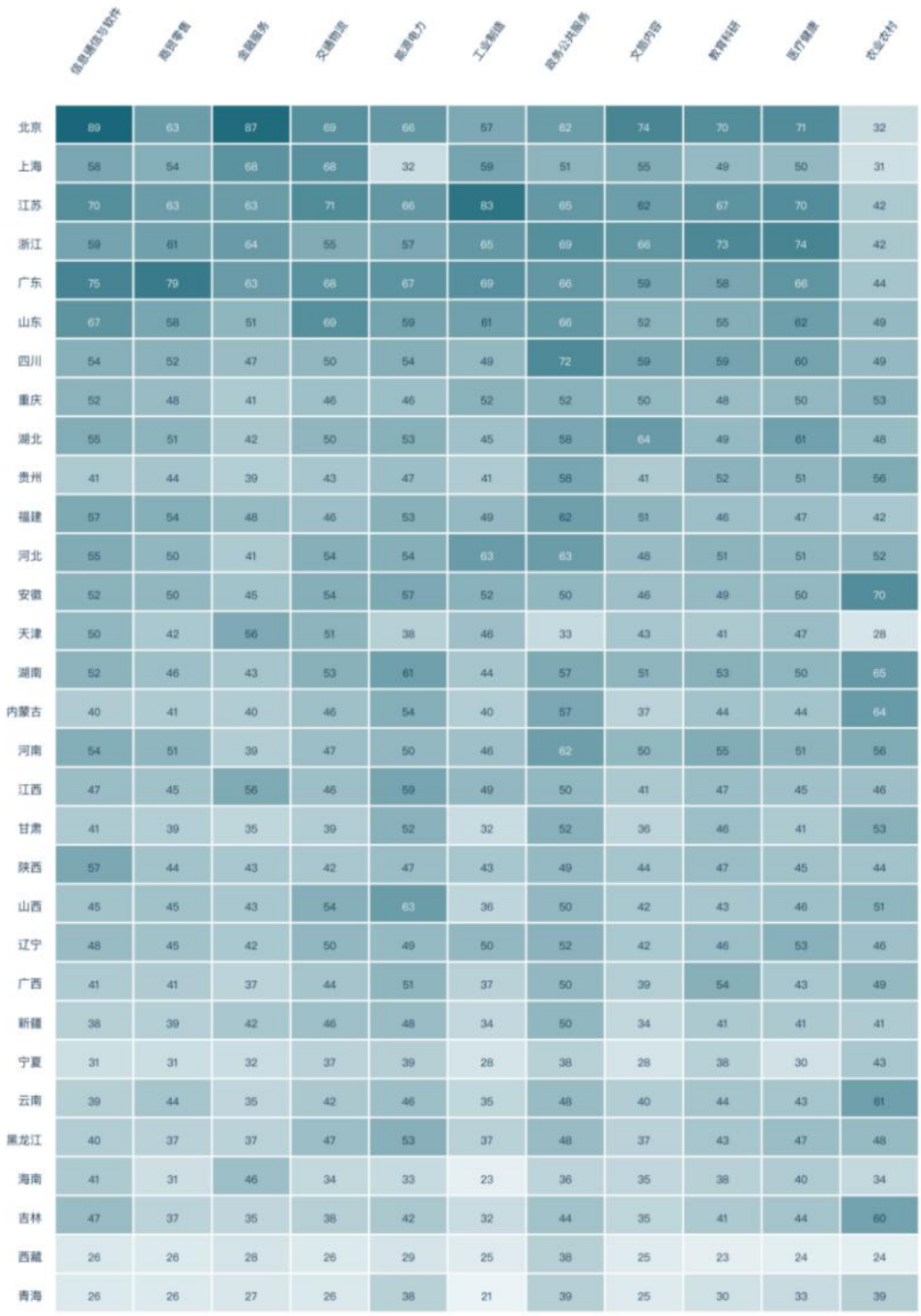
排名	省份	发展环境	基础设施	服务应用	质量调整得分	梯队
1	北京	68.2	76.6	77.2	74.6	领先
2	上海	75.0	73.6	57.2	69.4	领先
3	江苏	74.4	67.7	64.8	68.7	领先
4	浙江	80.3	62.2	58.5	65.9	领先
5	广东	70.4	64.5	59.7	64.7	领先
6	山东	57.4	50.1	48.6	51.6	较强
7	四川	65.1	43.0	42.9	48.8	较强
8	重庆	57.3	40.5	39.9	44.7	较强
9	湖北	50.0	47.1	27.9	42.5	较强

更多免费行业报告 并购家 www.ipoipo.net

排名	省份	发展环境	基础设施	服务应用	质量调整得分	梯队
10	贵州	57.3	44.2	24.7	42.2	较强
11	福建	54.0	32.7	46.0	42.0	较强
12	河北	59.5	37.0	32.4	41.6	较强
13	安徽	56.2	36.0	36.0	41.3	中等
14	天津	49.2	37.2	37.4	40.4	中等
15	湖南	37.8	34.1	42.0	37.3	中等
16	内蒙古	48.3	32.3	18.6	32.7	中等
17	河南	40.0	28.0	27.9	31.1	中等
18	江西	54.8	20.3	25.5	30.8	中等
19	甘肃	50.9	27.9	14.7	30.3	中等
20	陕西	37.6	23.1	31.9	29.4	中等
21	山西	35.7	29.3	16.0	27.3	中等
22	辽宁	34.3	21.1	25.1	25.7	中等
23	广西	52.2	19.4	9.0	25.1	追赶
24	新疆	48.1	17.6	9.4	23.3	追赶
25	宁夏	50.9	20.0	2.9	23.3	追赶
26	云南	35.0	16.4	15.0	20.9	追赶
27	黑龙江	28.7	16.4	13.6	18.8	追赶
28	海南	20.5	22.7	7.8	18.0	追赶
29	吉林	31.1	14.9	7.5	17.1	追赶
30	西藏	26.7	7.2	0.0	10.3	追赶
31	青海	7.8	13.3	0.7	8.3	追赶

3、各省行业智能体发展评估结果

31省×11行业智能体适配阶段性得分热力图



更多免费行业报告 并购家 www.ipoipo.net

图 24 31 省×11 行业智能体适配阶段性得分热力图

模型阶段性质量调整得分；颜色越深表示当前可得数据下的相对适配度越高。

表 8 各省 TOp3 适配行业

省份	第一位行业	第二位行业	第三位行业	说明
北京	信息通信与软件（88.8）	金融服务（87.0）	文旅内容（73.9）	阶段性
上海	交通物流（68.3）	金融服务（67.5）	工业制造（59.1）	阶段性
江苏	工业制造（82.6）	交通物流（71.1）	医疗健康（70.3）	阶段性
浙江	医疗健康（74.3）	教育科研（72.7）	政务公共服务（69.0）	阶段性
广东	商贸零售（78.9）	信息通信与软件（74.7）	工业制造（69.3）	阶段性
山东	交通物流（69.2）	信息通信与软件（67.3）	政务公共服务（65.8）	阶段性
四川	政务公共服务（71.9）	医疗健康（59.5）	教育科研（59.5）	阶段性
重庆	农业农村（53.1）	政务公共服务（52.5）	工业制造（52.2）	阶段性
湖北	文旅内容（63.5）	医疗健康（60.9）	政务公共服务（57.8）	阶段性
贵州	政务公共服务（58.0）	农业农村（56.0）	教育科研（51.9）	阶段性
福建	政务公共服务（61.8）	信息通信与软件（56.5）	商贸零售（54.2）	阶段性
河北	工业制造（63.4）	政务公共服务（62.6）	信息通信与软件（55.4）	阶段性
安徽	农业农村（70.4）	能源电力（56.8）	交通物流（53.8）	阶段性
天津	金融服务（56.2）	交通物流（51.2）	信息通信与软件（50.2）	阶段性
湖南	农业农村（65.1）	能源电力（61.5）	政务公共服务（56.5）	阶段性
内蒙古	农业农村（63.9）	政务公共服务（57.3）	能源电力（54.0）	阶段性
河南	政务公共服务（62.0）	农业农村（55.7）	教育科研（55.2）	阶段性
江西	能源电力（58.7）	金融服务（56.1）	政务公共服务（49.8）	阶段性

省份	第一位行业	第二位行业	第三位行业	说明
甘肃	农业农村（53.2）	政务公共服务（52.4）	能源电力（52.4）	阶段性
陕西	信息通信与软件（57.3）	政务公共服务（48.5）	教育科研（47.1）	阶段性
山西	能源电力（62.9）	交通物流（53.8）	农业农村（50.5）	阶段性
辽宁	医疗健康（52.9）	政务公共服务（52.0）	交通物流（50.1）	阶段性
广西	教育科研（53.7）	能源电力（50.9）	政务公共服务（49.8）	阶段性
新疆	政务公共服务（50.3）	能源电力（47.7）	交通物流（45.7）	阶段性
宁夏	农业农村（42.6）	能源电力（38.7）	政务公共服务（37.6）	阶段性
云南	农业农村（61.4）	政务公共服务（47.7）	能源电力（46.4）	阶段性
黑龙江	能源电力（53.3）	农业农村（48.4）	政务公共服务（48.1）	阶段性
海南	金融服务（46.4）	信息通信与软件（41.2）	医疗健康（40.3）	阶段性
吉林	农业农村（60.0）	信息通信与软件（47.3）	政务公共服务（44.4）	阶段性
西藏	政务公共服务（38.0）	能源电力（28.6）	金融服务（28.3）	阶段性
青海	农业农村（39.1）	政务公共服务（38.5）	能源电力（37.8）	阶段性

更多免费行业报告 并购家 www.ipoipo.net

六、参考文献

[1] 摩根士丹利,中国 AI 2.0：进入新阶段[R],2026.
[2] https://www.idc.com/resource-center/blog.
[3] IDC,中国 GenAI IaaS 市场跟踪与预测（2024－2029）[R]，2025.
[4] IDC,中国 AI 编程市场份额（2025）[R]，2026.
[5] http://www.iccsz.com/site/cn/News/2026/04/30/20260430013821788015.htm
[6] Goldman Sachs. Tracking Trillions: The Assumptions Shaping the Scale of the AI Build-Out[R/OL]. (2026-05-01).
[7] IDC. 中国公有云服务市场跟踪研究，2025H2 [R/OL]. (2026-04-30).
[8] https://mp.weixin.qq.com/s/qU2yH6nj0tUIzJ-G_8vEKQ
[9] https://mp.weixin.qq.com/s/HjTUkRf_ZXdMfBOGkIUeyw
[10] https://mp.weixin.qq.com/s/dtPX73vX70Nsjf2SHQatiw
[11] IDC. 2025H2 中国 AI 专业服务市场跟踪报告 [R/OL]. (2026-05).
[12] IDC. 中国智算云服务市场（2025 下半年）跟踪 [R/OL]. (2026-06).
[13] IDC. 中国 AI 软件市场半年度追踪，2025H2 [R/OL]. (2026-06).
[14] Gartner. Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5% in 2025[EB/OL]. Gartner, 2025-08-26.
[15] 锌财经. Token 是风口，Harness 是风向[EB/OL]. 创业邦, 2026-05-02.
[16] 北京市发展和改革委员会，中共北京市委网信办，北京市科委中关村管委会，等。北京市关于加快智能体引领发展的若干措施：京发改〔2026〕1185 号[EB/OL]. (2026-07-23)[2026-07-24].
[17] 新智元. AI 成功率从 20%飙到 100%！只需一个 Harness 文件[EB/OL]. 新智元, 2026-05-25.
[18] https://www.waytoagi.com/zh/blog/news-20260401.
[19] Amazon Web Services. AgentCore harness is now generally available[EB/OL]. (2026-06-17).
[20] 阿里云开发者社区. 托管 Agent 执行循环只是起点，AgentRun 托管的更是企业 AI 生产全链路[EB/OL]. (2026-05-11).
[21] 姚欣(PPIO 联合创始人). 2026 世界人工智能大会(WAIC)主题演讲[EB/OL]. (2026-07).
[22] Anthropic. Claude Managed Agents API[EB/OL]. (2026-04-08).
[23] https://www.trendforce.com/insights/ai-inference-drives-memory-demand
[24] Raj R, Wang H, Krishna T, et al. Towards Understanding, Analyzing, and Optimizing Agentic AI Execution: A CPU-Centric Perspective[EB/OL]. (2025-11-01).
[25] https://mp.weixin.qq.com/s/AhKZv5puFuxKpwWRATvtjA
[26] https://mp.weixin.qq.com/s/ppKgpMccoNlpDgl-Z7rcGg
[27] https://mp.weixin.qq.com/s/Y9FKyAUULk9o037DmaDzGQ
[28] https://mp.weixin.qq.com/s/cbLBnevJsJzMjzIOkFMw5A
[29] https://mp.weixin.qq.com/s/488qHtB4C99KX4zlf1xsPA

[30] https://mp.weixin.qq.com/s/A9gTGLV6iwvIvHYAXSYOKg
[31] Zheng Y, Fan J, Fu Q, et al. AgentCgroup: Understanding and Controlling OS Resources of AI Agents[EB/OL]. (2026-07-22).
[32] 高盛. 高盛深度报告：谁将成为中国 AI 大模型行业的长期赢家？[EB/OL]. 腾讯新闻，2026-07-10.
[33] Microsoft Research. Memora: A Harmonic Memory Representation Balancing Abstraction and Specificity[EB/OL]. (2026-07-27).
[34] 腾讯云. TencentDB Agent Memory 正式开源：让 Agent 沉淀经验，让人专注创造[EB/OL]. (2026-05-13).
[35] HONG S R, ZHUGE M C, CHEN J, et al. MetaGPT: Meta Programming for a Multi- Agent Collaborative Framework[EB/OL]. (2023-08-01).
[36] 沈春蕾.智源研究院发布《2026 十大 AI 技术趋势》[EB/OL]. 2026-01-09.
[37] AWS. AgentCore 框架现已全面推出 [EB/OL]. 2026-06-17.
[38] McKinsey & Company. The State of AI: Global Survey 2025. 2025.
[39] Anthropic. Claude Enterprise—Managed Agents 客户案例. 2026.
[40] Gartner. Lack of AI-Ready Data Puts AI Projects at Risk. 2025.
[41] IBM Institute for Business Value, Oxford Economics. Growing AI Control Gap as Enterprise Deployment Scales. 2026.
[42] Bai L., Huang Z., Wang X., et al. How Do AI Agents Spend Your Money? Analyzing and Predicting Token Consumption in Agentic Coding Tasks. arXiv:2604.22750, 2026.
[43] Salim M., et al. Tokenomics: Quantifying Where Tokens Are Used in Agentic Software Engineering. arXiv:2601.14470, 2026.
[44] Gartner. Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027. 2025.
[45] 工业和信息化部. 2025 年通信业统计公报及解读 [R/OL]. (2026-01-28).
[46] 工业和信息化部办公厅. 工业和信息化部办公厅关于公布万兆光网试点项目完成情况的通知 [EB/OL]. (2026-04-23).
[47] 中国互联网协会. 智能体应用发展报告（2025）[R/OL]. 2025-12

更多免费行业报告 并购家 www.ipoipo.net