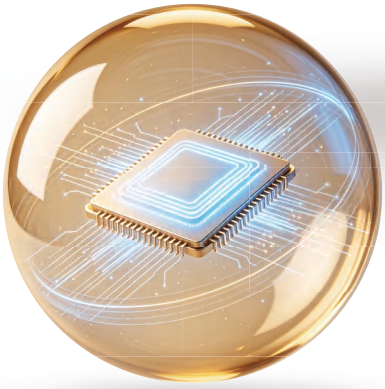
HUAWEI











智能世界 2035

从愿景到行动

构建万物互联的智能世界

序言



高文

近年来，大模型、生成式人工智能快速发展，人工智能从相对单一的感知、识别和预测，逐步走向内容生成、知识处理和复杂任务协同，并与计算、网络、数据、机器人及各类产业技术深度融合，正在从数字空间走向物理世界。这意味着，人工智能带来的变化，不能简单理解为又出现了一项新的信息技术，而是生产力体系的一次深刻变革。

其中一个关键变化，是“智能”正在成为一种可以被计算、被传输、被调用、被规模化使用的生产力要素。过去，知识和经验主要被劳动者掌握；现在，越来越多的知识可以通过数据和模型沉淀下来，借助算力在不同场景中快速调用。当这种能力进入研发、设计、制造、交通、医疗、教育等生产活动，人工智能就不再只是辅助工具，而会逐步成为新的生产要素和生产工具。

因此，真正值得关注的，不是模型有多少参数、回答问题有多快，而是人工智能能在多大范围内进入真实生产过程、创造多少新的价值。也要清醒看到，当前人工智能仍处于快速发展的早期阶段：在复杂推理、持续学习、因果理解、长期规划以及真实世界交互等方面，还有许多基础问题有待解决。技术发展越快，越要尊重规律，不能把阶段性突破等同于最终成熟。

人工智能进一步发展，越来越与基础条件紧密绑定，主要有三个。

第一是算力。大模型使算力从信息化建设的基础资源，变成人工智能时代的重要生产力，但算力资源分布不均、利用效率不高、供需匹配不足的问题仍然突出。这与电力的早期阶段相似：最初电力是哪里生产就哪里使用，后来才逐步形成覆盖全局的电网体系，算力也要经历类似的过程。我们推进鹏城云脑和中国算力网建设，就是要让不同地区、不同类型的算力高效连接、统一调度，使算力从相对孤立的资源变成便捷可用的公共基础能力——未来，算力的使用应该像用电一样方便。

更多免费行业报告 并购家 www.ipoipo.net

同时，随着人工智能规模化发展，算力与电力的关系越来越紧密，算力布局要统筹考虑能源供给和利用效率，推动算电协同。

第二是数据。大模型的能力离不开数据，但真正有价值的数据不只是规模大，更要质量高、样本分布合理、标准统一、反映真实业务。当前产业数据分散在不同系统和企业之中，质量、安全、标准、产权问题交织。如何把行业数据和知识转化为机器可理解、可学习、可利用的资源，是人工智能向纵深发展的基础。

第三是软硬件体系。人工智能已经形成算法、模型、芯片、计算系统、网络和数据相互耦合的技术体系，单靠一个环节的突破，难以形成整体优势。既要加强基础理论研究，也要夯实芯片、框架、工具链和训练系统等工程能力，构建完整的技术和产业生态。

人工智能最终要靠应用走向成熟。要推动人工智能与制造、交通、能源、医疗、教育等领域深度融合，让技术进入真实生产流程，在应用中发现问题、持续迭代，形成“技术研发—场景应用—持续迭代—产业升级”的良性循环。与此同时，人工智能越发展，越要处理好创新与规范、发展与安全的关系，同步建立法律、伦理和安全保障体系，使人工智能始终服务人的发展。

展望 2035 年——那也是我国基本实现社会主义现代化的时间坐标——我们很难准确预测智能社会发展到什么程度，但可以预见，智能将越来越深地融入经济社会运行，算力、数据、模型和网络将成为重要的智能基础设施。真正值得期待的智能世界，不是机器简单地替代人，而是人与智能

系统高效协同，让技术拓展人的能力边界。所以，比“未来的智能世界是什么样”更重要的问题是：今天做什么，才能走向这样的未来。未来不是等来的，是干出来的。

《智能世界 2035：从愿景到行动》正是沿着这个思路展开的。报告以 2035 年为坐标，用“变局—命题—行动”的框架，把对未来的判断落实为面向现实的行动方案：既分析人工智能给技术和产业带来的关键变化，也提出迈向智能世界的十个关键命题，进而围绕智能体时代的需求，给出构建智能基础设施、重塑用户体验和生产模式的具体路径。这种把愿景和行动连接起来的努力，正是当下最需要的。

愿这份报告带给读者启发，也期待我们共同以更长远的眼光认识人工智能，以更扎实的行动走向智能时代。



鹏城实验室主任
北京大学博雅讲席教授

更多免费行业报告 并购家 www.ipoipo.net



Otthein Herzog

所有实用的 IT 应用都应建立在对特定问题解决方案的理解之上。这一理解所需的知识既可嵌入程序代码中，也可通过知识描述进行详细记录。AI 领域历来高度重视应用中的知识相关要素。自 1956 年 AI 概念提出以来，该领域经历了深刻的变革。知识描述方法从早期依赖 AI 编程语言（如 LISP（列表处理）和 PROLOG（逻辑编程）），逐步发展为基于规则的系统，并进一步演进到统计学习。基于数理逻辑的方法（包括规则系统、本体和知识图谱）与亚符号神经网络在知识模型领域长期并存。直到近三年来，基于概率的大规模生成式模型才逐渐成为知识模型研究领域的主导范式。

“模型”是贯穿这一演进历程的核心概念，是对知识、规律或函数的结构化表征。模型得以在解决问题场景中使用的前提，是其能够对现实世界各个组成部分进行编码。模型采用的方法可分为两类：显式方法（如由知识工程师构建知识图谱）和隐式方法（如通过神经网络从数据中自主学习）。隐式知识模型的创建过程，是将工程重心从繁重的人工知识模型构建，转向大规模数据样本集的收集。而知识模型对应用系统能力和局限性的影响是该领域的一个关键考量因素。

从广义上看，AI 模型是一种数学或计算架构。AI 模型可以配置为基于公式执行简单的计算（如销售业绩预测），也可配置为深度神经网络（如生成逼真的人类语言）。作为对现实世界的抽象，模型旨在捕捉数据中概念、规律和结构之间的关系。此类知识关系有助于将模型用于分类、预测和决策。从符号模型到数据驱动型统计模型的迁移是近期 AI 发展的标志。以生成式对抗网络和大语言模型为代表的生成式模型，已展现出基于习得的概率分布生成多模态内容（如文本、图像、音频、

视频等）的能力。这些深度学习模型在处理高维数据以及结构化和非结构化数据（包括自然语言）任务中展现出强大的能力。其主要优势在于能够从原始数据中自动提取特征：模型可直接从样本中学习知识，而无需依赖人工设计的特征，从而挖掘出人类难以察觉的规律。

大语言模型是深度学习模型的子集，通常具备数十亿级别的参数规模，并在海量数据集上进行训练，无需针对特定任务进行训练即可具备执行各类任务的能力。大语言模型的能力包括编写代码、输出文档摘要、问答以及自然语言对话等。此外，大语言模型还可用于实现工作流程、辅助决策以及利用外部工具来扩展自身能力。这些能力是行业应用创新的基础。

大语言模型还展现出仅需极少额外训练即可适应新任务的能力，并具备多模态输入和输出能力。这一范式转变克服了最初针对特定任务设计的狭义 AI 系统存在的局限，向具备灵活推理与决策能力的通用系统演进。然而，这些模型的质量取决于三个因素：模型训练数据、模型架构以及优化过程。

大语言模型的应用近期显著增长，在通信和生产等多个领域支持并实现决策流程的自动化（如聊天机器人和工业系统）。在制造领域，预测模型有助于管理工厂的资源分配、早期故障检测以及增强其适应不同环境的能力。在交通领域，大语言模型在自动驾驶车辆控制以及交通和物流优化等方面也发挥着重要作用。

更多免费行业报告 并购家 www.ipoipo.net

然而，在将大语言模型融入到决策流程的过程中，我们需要避免放大偏见，保证透明度，确保系统作出的决策经过人工评审。在现实场景中部署 AI 模型时，确保公平性、问责制和可解释性至关重要。

AI 模型已从单纯的分析和预测工具演变为重要的生产力引擎。生成式模型的运用有望推动工业设计等领域的创意开发。大语言模型可用于探索新的形态、助力科学家生成假设，辅助工程师设计复杂系统。

大语言模型的开发需要获取大量高质量的数据集，而从各行业真实场景获取这些数据集通常非常困难或获取成本高昂。此外，训练大语言模型需要耗费大量计算资源和电力，加剧环境问题。因此，在提供行业基础设施时应考虑这些要素，这一点至关重要。

正如本报告所指出的，将多模态大语言模型应用于智能行业基础设施，对于行业系统的发展至关重要，有助于跨越数据、信息与知识之间的鸿沟，从而实现自主决策。近期研究主要聚焦以下方向：

- 采用小规模、本地运行、高效能的大语言模型，实现高性能行业应用。
- 推动学习模型与其具体应用场景的动态适配。

- 探索分布式智能体模型架构，支持自主规划、工具调用以及长周期任务执行。

- 制定完善的全球治理体系、伦理规范和负责任部署的框架。

知识模型驱动的系统的出现，正在加速重塑各行业的作业流程、工具与方法。随着这些系统的部署，企业有望构建起贯穿其各个领域的以知识为中心的基础设施，驱动其业务、生产和物流流程实现前所未有的优化。



中国工程院外籍院士
德国国家工程院院士
同济大学教授
德国不来梅大学教授

更多免费行业报告 并购家 www.ipoipo.net



周志华

与人工智能结缘三十余年，我愈发确信：人工智能并非一项孤立的技术，而是一种具有广泛“赋能”意义的基础力量。它正在从底层逻辑上改写所有行业的生产方式与创新范式。

正因如此，当看到华为《智能世界 2035：从愿景到行动》这份报告时，我感到由衷的欣慰。它没有停留于泛泛的趋势宣言，而是以工程思维将战略远见转化为清晰的行动框架。报告对产业变革的关键判断与核心挑战的精准定义，展现了难得的系统性与务实精神。尤其是，报告直面当前 AI 发展中的一项深层矛盾——数据驱动与知识驱动的融合难题。近年来，基于海量数据的机器学习取得了举世瞩目的突破，但“数据驱动”路径在可信性、可解释性方面的短板日益凸显；而“知识驱动”路径虽能提供逻辑保证，却在数据利用效率上天然受限。二者如何互补共生，是 AI 基础研究的重大命题，我本人在这方面做了很多探索，报告中对“数据向知识和记忆转化”的阐述，以及对“结果准确、可信与可追溯”的深入探讨，正是对这一命题的积极回应，方向值得肯定。

作为一名长期从事机器学习基础研究的科研工作者，我也想借此机会强调一个容易被热潮淹没的观点：没有任何一个算法模型能“包打天下”、解决一切问题。大模型能力强大，但绝非“万能钥匙”。我们必须纠正盲目跟风“大模型解决一切”的倾向，将更多资源投向 AI 算法的基础研究，鼓励多元路径的探索。在产业热情高涨的当下，保持科学家的清醒尤为重要。报告提出的解决方案——从技术架构到产业落地——恰恰体现了“具体问题具体分析”的原则，而非试图用一个通用模板覆盖所有场景，这种务实态度值得倡导。

我还注意到，报告将“智能型企业走向新四化”“组织液态化”作为重要趋势，揭示了一个深层逻辑：未来的竞争不仅是技术的竞争，更是人才与组织能力的竞争。作为教育工作者，我深感人

才培养必须从源头抓起。要真正释放 AI 的潜力，关键之一在于培养一批既深谙领域知识、又精通前沿 AI 技术的“双语”科学家——他们能跨越技术与产业的鸿沟，成为人工智能赋能的中坚力量。

至于报告提出的“五大倡议”——以人为本、科技向善、开放合作、系统思维、可持续发展，我完全赞成。AI 发展的根本目标是服务人类、赋能社会，而非取代人。我们始终应保持对技术边界的审慎，对“人类无法控制的人工智能体研究”暂不宜开展，这是科学伦理的底线。

回望 AI 七十余年的范式演进，每一次跃迁都源于基础研究的突破。今天，我们正站在人工智能技术从虚拟数字世界走向现实物理世界的新节点上。将愿景变为现实，需要产业界、学术界和政策制定者以冷静的头脑、开放的胸怀和务实的行动协同共进。我期待更多资源投向基础研究、人才培养和多元路径探索，让 AI 真正成为服务人类进步的有益力量。



中国科学院院士
南京大学教授、副校长
国际人工智能联合会理事会主席

更多免费行业报告 并购家 www.ipoipo.net



汪涛

我们正站在一个历史拐点上。历次工业革命解放了人类的体力，信息革命延伸了人类的感官，而今天，人工智能开始触及人类最核心的能力——认知与决策。AI 已经走出实验室，成为驱动科技创新与产业进步的新引擎。

一年前，我们发布《智能世界 2035：探索未知，跃见未来》，勾勒了面向未来十年的技术图景——算力增长十万倍、存储需求增长五百倍、数千亿智能体广泛联接。这些愿景并非简单的预测，而是基于产业与技术长期积累下的战略判断。

过去一年，人工智能以 Agent 的形态走进代码仓库、走进客服窗口、走进工厂产线。企业的追问，已从“要不要拥抱 AI”转向“AI 能替我们做成什么”。方向已然清晰，但愿景落地，还需坚实的行动。这正是今年《智能世界 2035：从愿景到行动》聚焦的方向：从当前 AI 革命带来的关键变化切入，剖析变化背后的核心命题，最终提出面向 Agentic AI 时代的行动建议——构建智能基础设施，重塑用户体验与生产模式。

跳出单一企业的视角，放眼全局：AI 正加速走进千行万业，成为驱动社会前行的加速轮。对每一个组织、每一个个体而言，真正的挑战已不是方向选择，而是速度——能否在指数级变化中踩准节奏，找到属于自己的增长曲线。

AI 对世界运行规则的改写，将给人类社会和经济发展带来前所未有的改变：Token 走向普惠供给，将重构企业的竞争基石；物理 AI 开始重塑万亿级实体产业；终端则成为 AI Agent 的感知与行动载体。这些变化，将从根本上重写产业的底层逻辑。

然而，AI 若要重构生产方式、竞争要素及产业形态，绕不开几道关键命题：如何构建具备自主认知与全域通用能力的 AGI？如何让算力集群规模百倍跃升的同时，使智能体任务成本千倍下降？如何构建以体验为中心的新一代智能通信网络？面向自主进化的 Agent，又如何构建安全与隐私保护的机制？这些命题，不会因等待而自行解答，唯有行动，才能一步步逼近答案。

行动从何处着手？我们看到，Agentic AI 正推动数字世界从“以应用为中心”走向“以智能体为中心”。这一转变的根本在于智能体的运行逻辑截然不同：它需要持续感知环境、自主推理决策、动态调用工具、实时与人交互——每一步都伴随着 Token 的大规模生成与高频流转。作为这一过程的载体，Token 的生产、运载与应用，对算力、通信网络和云提出了远超传统互联网应用的要求：长程任务更难、交互时延更低、并发规模更大、协同更复杂。归根结底，智能体的能力上限，不取决于单一环节的优化，而取决于整条链路的系统工程能力。

因此，面向 Agentic AI 时代，最关键的行动是围绕 Token 构建“生产—运载—应用—防护和运维”的全链路智能基础设施体系。在 Token 生产侧，超节点架构创新为大规模、高性能计算场景提供澎湃算力；在 Token 运载侧，以体验为中心的智能联接底座保障高速、无损的 Token 流转；在 Token 应用侧，跨端异构算力协同的终端设备将智能送达千家万户、千行万业。三者环环相扣，最终将智能基础设施的能力，转化为更自然的用户体验与更高效的企业生产力。

愿景不会因描绘而自动实现，行动才是丈量未来的尺度。“从愿景到行动”，构建智能基础设施，让创新有处生根；重塑用户体验与生产模式，让每一个人、每一家企业都能在这场变革中真切感受到 AI 带来的价值。

华为愿携手全球客户与伙伴，脚踏实地，共同探索，让今天的每一次实践，都成为通往智能世界 2035 的坚实一步。



华为副董事长、轮值董事长

更多免费行业报告 并购家 www.ipoipo.net

前言

随着人工智能迈入自主进化全新阶段，依托大模型与智能体的自主学习、持续迭代能力，彻底摆脱对人工输入与调优的深度依赖，突破人类认知与效率上限，形成自我驱动、正向增益的指数增长螺旋，推动科研、生产、开发、商业四大范式全面变革，正式开启 AI 产业超级周期。

科研范式重构

AI 跃升为继实验、理论、计算之后的第四大科研创新范式。依托海量数据与文献挖掘能力，可精准捕捉隐性关联、推演潜在规律、求解最优方案，将科研人员从低效重复的试错工作中解放，聚焦原创突破与前沿探索。这一变革让科研发现从偶然探索变为可推演、可落地的确定性创新，大幅压缩技术攻关周期，实现科研效率量级跃升。

生产范式革新

AI 智能体彻底重塑人机协作的生产力形态。智能体可自主拆解任务、调用工具、闭环执行、动态优化，全面替代标准化、流程化的重复劳动。人类得以剥离繁琐执行工作，专注顶层设计、复杂决策与创新突破。传统固化的线性生产模式，迭代为灵活高效的并行智能协作模式，实现生产组织轻量化、响应时效极速化，完成从 “人主导执行” 到 “人定策略、智能落地” 的生产变革。

开发范式变革

AI 打破软件开发的精英技术垄断，实现全民开发平权。自然语言替代专业代码门槛，各行各业从业者均可直接落地业务需求，开发者群体从专业从业者拓展至大众用户。软件交付从按月迭代压

缩至分钟级实时更新，颠覆 “先产后用” 传统模式。开发成本大幅降低，催生海量个性化需求，推动软件供给从中心化垄断转向分布式自产，产业价值升级为 “人人可生产、按需即消费” 的全新形态。随着 AI 技术的演进，自动化开发向硬件设计，硬件制造全流程扩展，进而开发进步速度发生结构性跃迁。

商业范式重塑

AI 深度重构传统商业体系与市场竞争逻辑。依托全链路智能匹配、精准运营与动态迭代，企业经营从被动响应市场，转向主动预判需求、前置创造价值。传统规模化、中心化的粗放经营模式，逐步迭代为轻量化、高灵活、个性化的智能商业形态。企业核心竞争力也从人力、渠道、规模优势，转变为智能迭代、数据配置与模型创新的核心能力。

四大范式的系统性变革，正深刻重塑人类社会的发展形态。全面智能化已不再是遥远的科技愿景，而是全产业深度落地、协同推进的共同行动。在技术迭代持续加速的背景下，智能化演进过程中的底层技术瓶颈与核心命题，亟待持续攻坚突破。当前 AI 加速从数字世界赋能物理世界，联接技术与计算技术双向激荡、协同演进，计算架构在通用算力与专用芯片之间动态钟摆、交替迭代。半导体韬定律、生成率 J 曲线、杰文斯悖论等产业规律，有效破除了行业发展的认知迷茫，为 AI 技术演进与产业升级提供了科学指引与底层支撑。

对于各类组织与个体而言，拥抱 AI 已是必然趋势。唯有顺应 AI 指数级增长浪潮、踩准智能变革的节奏，才能在时代变局中把握机遇，构筑专属的长期增长曲线。

更多免费行业报告 并购家 www.ipoipo.net

目录

CONTENTS

01 变局：AI 成为驱动经济发展与社会进步的加速轮 01

四大引擎驱动智能经济增长，迎接 Token 普惠时代 04
企业竞争要素变迁，智能型企业走向新四化 06
从执行工具到数字伙伴，智能体加速数智生产力 08
从比特到原子，物理 AI 重塑万亿实体产业新基建 09
从触控到多模交互，终端成为 AI 智能体全时空载体 10
智能体流量占比超九成，钟摆效应驱动计算联接交替爆发 12
数据决定 AI 能力上限，存储夯实智能产业根基 14
从“几何缩微”到“时间缩微”，新原则指导半导体产业持续演进 16
电力决定算力，AIDC 走向算电网储协同 17

02 命题：通往智能世界的十大关键命题与研究方向 19

构建具备自主认知与全域通用能力的 AGI 21
算力集群规模百倍提升，智能体任务成本千倍下降 25
构建因果准确、溯源可信的新型存储与记忆系统 27
AI 全面融入通信系统，构建体验为中心的智能联接底座 28
深耕韬定律下芯片设计方法和工具，支撑时延与能效协同优化 30
从驾驭环境到自主演进，Agent OS 构建智能体原生技术底座 32
定义无感适配、全域随行的终端智能新形态 35
实现智能驾驶的自我学习与极致安全 37
攻克供电与散热难题，支撑 AIDC 走向更大算力规模 39
面向自主 Agent 的安全与隐私保护，为智能世界保驾护航 41





03 行动：面向 Agentic AI 时代，构建坚实的智能基础设施，重塑用户体验和生产模式 43

计算系统与软件：
超节点架构创新与开放的软件生态，实现 Token 的高效生产 46

通信系统与网络：
构筑以体验为中心的智能联接底座，实现 Token 的高效运载 59

云基础设施与软件：
打造 Agentic 原生的云基础设施与应用平台，加速 AI 在行业深化落地 75

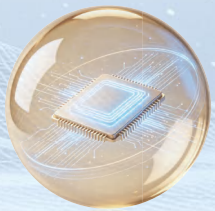
智能的应用和落地：
重塑个人用户体验和企业生产模式，实现 Token 的高效应用 82

04 共识：构建人与 AI 和谐共生的产业生态倡议 99

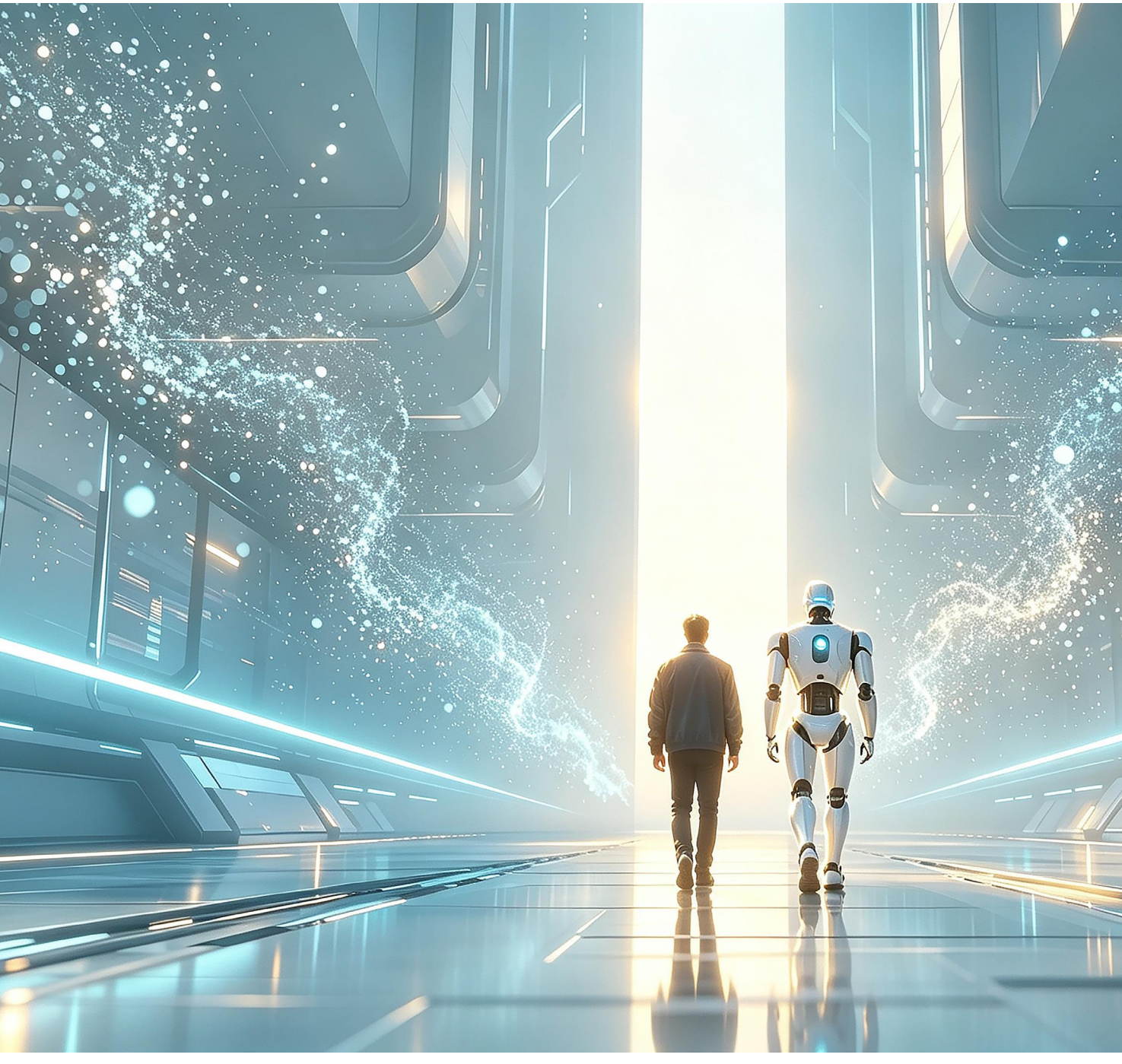


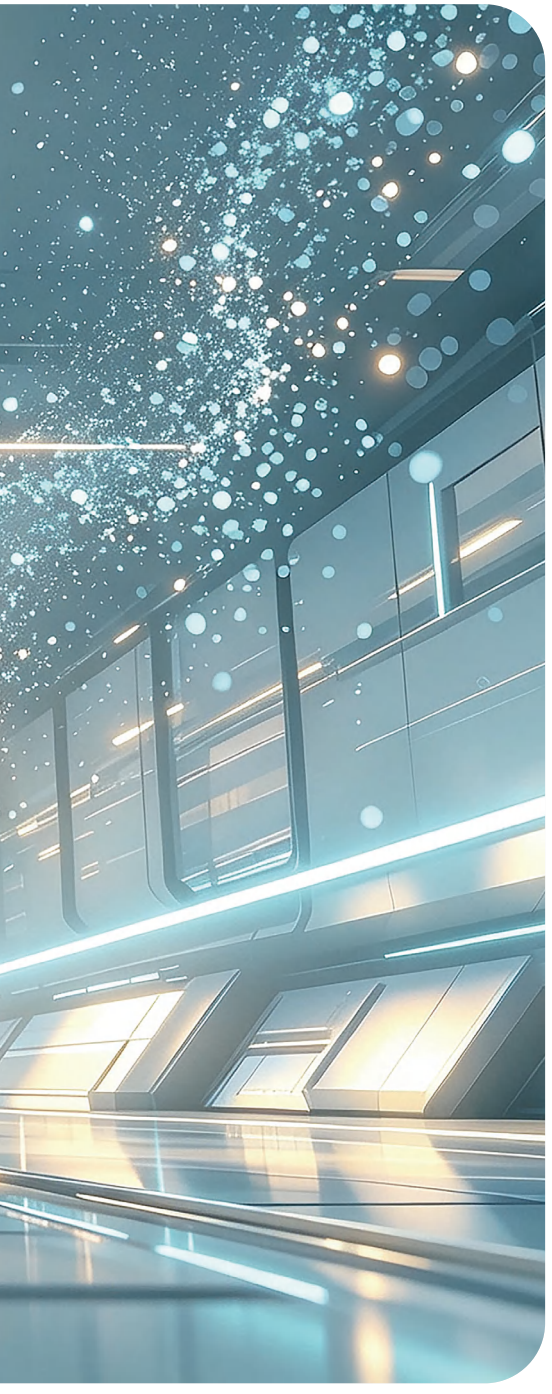






更多免费行业报告 并购家 www.ipoipo.net





变局

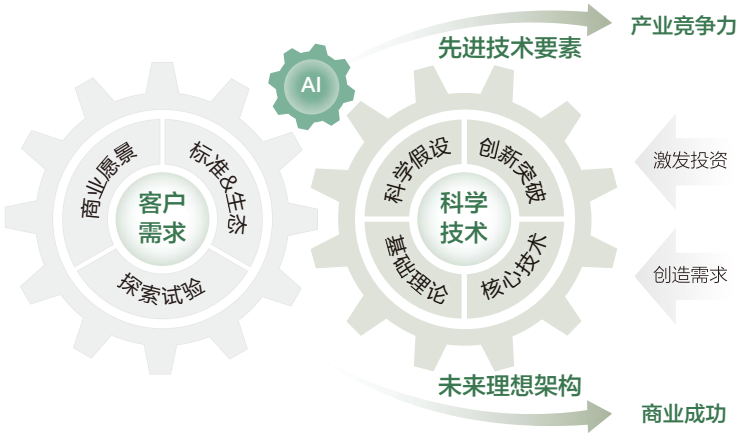
AI 成为驱动经济发展与
社会进步的加速轮

更多免费行业报告 并购家 www.ipoipo.net



我们正身处一场从信息时代向智能时代跃迁的结构性变局之中。这场变革的广度与深度，远超此前任何一次技术革命：它不再是单一技术的线性突破，而是 Token 经济、智能体、物理 AI、新型终端与基础设施的协同爆发，共同改写着产业底层逻辑。

Token 正从数字世界的价值载体蜕变为智能经济的“数字通货”，Token 调用量近两年增加了千倍以上，这一指数级扩张昭示着经济活动的计量单元与运行规则正在被重新定义。与此同时，AI 正从“执行工具”全面进化为“数字伙伴”，具备长程记忆与自主决策能力的智能体将占据未来网络流量的九成以上，催生出万亿级智能体互联的全新生态。更为深远的是，AI 的能力边界正突破比特世界的藩篱，物理 AI 开始直接改造原子世界，万亿级实体产业迎来智能化重构的历史机遇。



AI 成为经济增长加速轮

四大引擎驱动智能经济增长，迎接 Token 普惠时代

随着数字经济的不断深化，Token 经济作为新兴的数字资产与价值载体，正逐步成为推动产业升级和创新发展的核心引擎。截至 2026 年 3 月，中国日均 Token 调用量相比 2024 年初增长了 1400 倍。这已经不再是渐进式的技术普及，而是一场指数级的产业重构。

四大引擎驱动 Token 经济增长

Token 经济的增长并非偶然，而是规模、效率、利用率与价值四大引擎协同驱动的结果，共同构成了一套精密、高效、持续运转的增长机制。

首先，规模引擎决定了供给的天花板。规模引擎反映了 Token 发行和流通的基础容量，决定了整个经济体系中 Token 的最大供应量。合理的规模设计不仅保障了 Token 的稀缺性和稳定性，也为后续的经济活动提供了充足的流动性基础。

其次，效率引擎决定了产出的加速度。效率引擎强调生产和流通环节的优化，通过技术创新和流程改进，可以提高 Token 的生产效率，缩短交易确认时间，降低系统延迟。

第三，利用率引擎决定了价值的兑现度。Token 的价值不在“拥有多少”，而在“用了多久、用得多好”。利用率引擎关注 Token 在实际应用中的活跃度和使用频率，推动 Token 从静态持有向动态流转转变，增强 Token 的经济效用。

最后，价值引擎决定了溢价的护城河。不同 Token 因其背后的资产支持、应用场景和社区共识而具备不同的价值溢价能力。价值引擎可通过创新设计和市场机制，提升 Token 的独特性和竞争力，促进优质 Token 的价值发现和资本增值。

构建 Token 经济三层框架

Token 经济的复杂性决定了其必须依托一个多层次、一体化的框架体系，才能实现高效运转和持续发展。该体系主要包括 Token 消费、Token 调度与分发以及 Token 工厂三层。

Token 消费层是 Token 经济的终端应用场景，直面各类用户，涵盖企业、开发者、个人、智能硬件、机器人等，实现 Token 按需调用、弹性供给。消费层的活跃度将直接影响 Token 的流通速度和市场需求，是实现 Token 经济发展的关键环节。

Token 调度与分发层是资源流动的中枢，承担智能路由、交易撮合、合规清算、跨域调配等核心职能，实现 Token 的精准分发和调度，保障资源的最优配置，避免浪费和滞留，提升整体系统的响应能力。

Token 工厂层是 Token 生产的源头，传统智算中心正在全面升级为具备规模化生产能力的算力底座。未来将依托算网协同、算电协同，并采取“中心 - 边缘 - 现场”立体化的布局，实现 Token 规模化、低成本、绿色化的量产，为智能社会输送源源不断的动力。

这三层相互依托、协同运作，形成一个闭环的生产网络，推动整个智能产业从“人力驱动”转向“人机协同驱动”，生产效率与组织形态将迎来根本性变革。

更多免费行业报告 并购家 www.ipoipo.net

高效 Token 工厂需四要素协同

构建高效的 Token 工厂，离不开 ICT（联接与计算）、能源、模型和数据四大关键要素的协同优化。

ICT 作为基础设施，提供了高速、安全、可靠的网络环境和计算能力，是 Token 生产和流通的技术保障。能源则是 Token 工厂可持续运转的物理基础，绿色低碳的能源不仅降低了生产成本，也符合可持续发展的全球趋势。模型决定了 Token 工厂的能力上限，需要依托架构创新，持续增强逻辑推理、多模态理解、精准生成的能力。数据是提高 Token 价值的核心原料，高质量的数据集直接映射为 Token 的信息密度、准确性和价值性。

上述四要素并非简单物理堆叠，需要深度融合和协同发展，Token 工厂的全链条转化效率将成为衡量智能经济新型基础设施成熟度的核心标尺。



四大引擎驱动 Token 经济增长

Token 普惠时代必将到来

展望未来十年，Token 经济的稳态规模有望突破百万亿元，并将迎来普惠时代。我们预测，全球年度 Token 总消耗还将增长约 10 万倍，成本每年下降约 50%。

“规模在扩张，门槛在坍塌”，这就是普惠时代的核心逻辑。当更多个人与企业得以平等参与 Token 经济，资源分配将更公平，产业链协同将更高效，新兴产业的活力也将被真正激发。一个更开放、更包容、更可持续的智能经济生态，正在加速成形。

企业竞争要素变迁，智能型企业走向新四化

工业时代，传统企业优势主要建立在泰勒制、标准化流程、资产规模和运营效率之上，价值创造总体呈线性关系：增加资本、人员和渠道，获得可预测的产出增长。数字化时代，平台、数据和用户网络成为新的增长杠杆，网络效应使价值创造由线性走向平方级放大。智能型企业则把 AI 原生智能、群智协同和闭环自治嵌入企业运行机制，使企业从“依靠组织推动业务”转向“依靠智能系统持续感知、决策、执行、验证和进化”。其竞争优势不再主要取决于拥有多少静态资源，而取决于能否形成高频、低成本、可验证的智能闭环，由此形成新的智能涌现效应。

智能化优势的要素可以总结为智能型企业的“新四化”：流程 Agent 化、能力组件化、系统智能化和组织液态化。流程 Agent 化，是将大量标准化、重复性乃至部分复杂业务流程，从“人驱动流程”转变为 Agent 围绕目标端到端自主执行；能力组件化，是将原本依附于个人和部门的专业知识、经验和方法，沉淀为可调用、可组合、可评估和持续升级的能力组件，使企业能力能够像软件组件一样复用；系统智能化，是让数据、模型、算力和业务系统具备自主感知、动态调度、故障自愈和资源优化能力，为 Agent 规模化运行提供智能企业底座；组织液态化，则是让人与 Agent 突破固定岗位和部门边界，围绕任务动态组合、按目标协作。之所以说“新四化”构成智能型企业的核心要素，是因为它们分别重构了企业的流程、能力、底座与组织，共同将企业从依赖层级管理和人工推动的组织，转变为目标驱动和持续自主演进的智能系统。

与之对应，企业护城河也正在被重新定义。智能型企业真正难以复制的，是由“专有数据 × 强化学习环境 × 协同网络 × 验证机制”形成的复合优势。专有数据记录企业真实客户、业务状态、历史决策与运行结果，通过本体（Ontology）进一步将数据、逻辑和行动



更多免费行业报告 并购家 www.ipoipo.net

组织为具有统一语义的业务实体，形成“以决策为中心”的智能系统的企业世界模型；强化学习、仿真和数字孪生环境，使 Agent 能够低成本试错、探索并不断优化策略；协同网络把员工、Agent、客户和生态伙伴连接起来，使个体智能进一步形成群体智能；验证机制则通过评估、规则、证据和反馈判断一次行动究竟是否真正有效，并决定哪些经验能够进入下一轮闭环。单独拥有一个大模型并不能构成长期壁垒，因为竞争者同样可以买到模型；真正难以复制的是企业多年运行形成的数据、环境、协作关系和评价标准，以及它们共同构成的持续进化闭环。

这进一步推动智能型企业的价值创造逻辑，从传统的线性价值链走向“交付—反馈—进化”的飞轮。过去企业往往按照研发、生产、销售、交付的顺序创造价值，一次交付意味着一轮业务基本结束，经验还需要经过人工总结、管理决策和流程调整才能进入下一轮。智能型企业则让每一次交付成为下一轮进化的起点：产品和服务交付产生真实业务结果与客户数据，形成反馈；系统通过 Agent 执行轨迹、客户行为和验证结果识别成功经验与失败模式；有效经验进一步沉淀和更新为新的数据、组件、策略、模型和流程，完成一次进化；进化后的系统再进入下一轮交付，并产生更高质量的反馈。于是，企业运行不再只是重复业务，而是在每一次业务运行中同步进化和提升自身能力。

因此，智能型企业真正重要的变化，是企业开始拥有一个可以持续自我进化的价值创造闭环。“新四化”决定智能闭环能否运行，“专有数据 × 强化学习环境 × 协同网络 × 验证机制”决定闭环是否形成难以复制的护城河，而“交付—反馈—进化”飞轮则决定这种优势能否持续形成复利，从而在智能世界形成持续扩大的结构性复利竞争优势。



企业竞争要素的演进：从效率规模、网络效应到智能涌现

从执行工具到数字伙伴，智能体加速数智生产力

智能体是以大语言模型为认知内核，具备自主感知、推理决策、规划执行与持续学习能力的智能系统——它能够自主拆解任务、调用外部工具、与环境交互，并根据反馈动态调整策略，直至完成目标。其核心能力可归纳为四层递进的工程范式：Prompt Engineering（清晰表达意图）、Context Engineering（管理工作记忆）、Harness Engineering（装配工具与执行框架）、Loop Engineering（在“感知 - 决策 - 行动 - 反思”中持续迭代）。四层能力的成熟度，共同决定了智能体能够处理任务的复杂度上限。

当前，智能体正经历一次本质性跃迁——从被动执行任务的工具，全面进化为能够自主闭环、提供专业指导的“数字伙伴”。这背后是一条清晰的能力提升路径：从 L1 功能级——AI 作为工具被人类调用，到 L2 任务级、L3 协作级，直至 L4 指导级——AI 提供专家级服务，人类从“操作者”转变为“授权者”，再到“监督者”。这是基于 Agent 能力提升，带来的人机协作关系的根本重构。IDC 预测，2030 年 22 亿 AI Agent 将作为“新数字劳动力”席卷全球 [1]。

这一跃迁背后是Agent三大核心能力的质变：

长程记忆能力的持久化：Agent 正逐步具备月级乃至年级的持续工作能力，这意味着它不再是“一问一答”式的工具，而是能够在长周期任务中保持上下文、持续积累经验的协作者。

协作方式的双轨化：Agent 正通过一系列开放协议，实现从内部工具连接到外部智能体协同的全链路标准化，一端打通与内部系统、工具的标准化调用，另一端实现跨组织、跨平台智能体之间的身份认证与协同交互。这些协议正在共同构建起智能体之间“可对话、可协作”的基础设施，让 Agent 不再是孤立的功能节点，而是能够在开放网络中自由连接、协同工作的“社会化”个体。

训练范式的开放化：Agent 的训练正从依赖人工标注的监督学习，转向基于强化学习与环境反馈的自主探索范式，通过与真实任务场景的持续交互获得反馈信号，Agent 得以在实践中自我修正、自我进化，迭代速度也随之显著提升。

我们看到 Agent 企业级标准平台正在加速成形，感知、规划、执行、反思的四层闭环已成为标准技术架构。未来，企业将构建由专业 Agent 组成的“数字员工矩阵”，并通过打通 ERP、CRM 等核心系统，实现 Agent 托管 & 治理的规模化工程落地。

更为关键的是，Agent 正在重塑企业的数据资产，行业深度经验与核心技术的长期积累，将构筑起难以复制的竞争护城河。数据、模型与反馈的闭环质量，直接决定了 Agent 落地效果的天花板。

简言之：AI Agent 不仅是技术的升级，更是企业数智生产力的重塑。未来企业竞争的胜负手，不再只是“是否用了 AI”，而是“能否把 AI 训练成真正懂业务的数字伙伴”。

[1]《IDC 预测，2030 年 22 亿 AI Agent 将作为“新数字劳动力”席卷全球》

更多免费行业报告 并购家 www.ipoipo.net

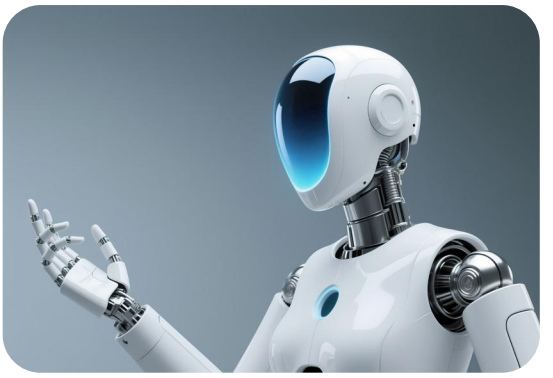
从比特到原子，物理 AI 重塑万亿实体产业新基建

当下 AI 扎根数字世界，依托文本、图像、音视频等比特数据完成感知与生成，但单纯虚拟智能的增长边界日益清晰。一场从比特走向原子的范式变革正在开启：物理 AI，也就是能够感知、理解、干预现实物理世界的智能体系，将成为重塑实体经济的新一代核心基建。不同于对话大模型、文生图等纯数字应用，物理 AI 打通数字大脑与实体世界，推动智能从屏幕走向产线、矿山、交通场景，开启智能技术产业化的全新周期。

第一，物理 AI 是 AI 从数字世界走向物理实体的范式革命。AI 不能永久局限在虚拟空间。依托互联网文本、图像数据训练的数字智能，仅能完成信息识别、内容生成，缺少对物理规则、空间交互、实体运动的理解，能力增长逐步触顶。若要迈向通用人工智能，AI 必须进入物理世界获取运动、受力、空间交互等高维感知数据，同时输出动作指令、作用于真实物理对象。未来，物理 AI 补齐现有智能“只能思考、无法行动”的短板，实现感知、决策、执行完整闭环，是人工智能进化的必经阶段。

第二，物理 AI 是万亿级蓝海赛道，长期增速有望超越数字 AI 产业。数字互联网改造的是信息流通环节，而物理 AI 直面制造、物流、能源、交通等体量庞大的实体经济。长期来看，物理 AI 改造实体经济的市场空间，量级远超过去二十年互联网数字化浪潮，将孕育规模达数十万亿的全新产业集群。

第三，比特的运算速度再快，也要尊重原子的惯性。物理世界存在动力学约束、传输时延、实体运动延迟，单纯依靠云端集中运算无法满足实时控制需求。因此物理 AI 的标准架构必然是云边端协同体系：云端承载大规模世界



模型训练、高保真环境仿真；边缘节点负责多设备协同调度、区域性决策；确定性工业网络保障数据低时延、高可靠传输；端侧硬件完成实时感知与动作执行，并持续把真实环境反馈回训练体系，形成“仿真推演—实体执行—数据回流”的高效闭环。世界模型 + 仿真训练正成为关键路径，Sim-to-Real（仿真迁移现实）是核心瓶颈。

第四，可靠性是商用起点，微小精度提升蕴藏千亿价值。通用能力决定物理智能发展上限，但安全可靠是商业化入场券。数字大模型出现幻觉大多仅影响内容质量，而物理 AI 一旦决策出错，可能引发设备损毁、生产事故、人身危险。当系统可用度从 99.9% 提升至 99.99%，业务中断时间大幅下降，对应的产业损失规避、运维成本节约价值可达千亿级别，可靠性工程将贯穿硬件、算法、控制系统全流程。

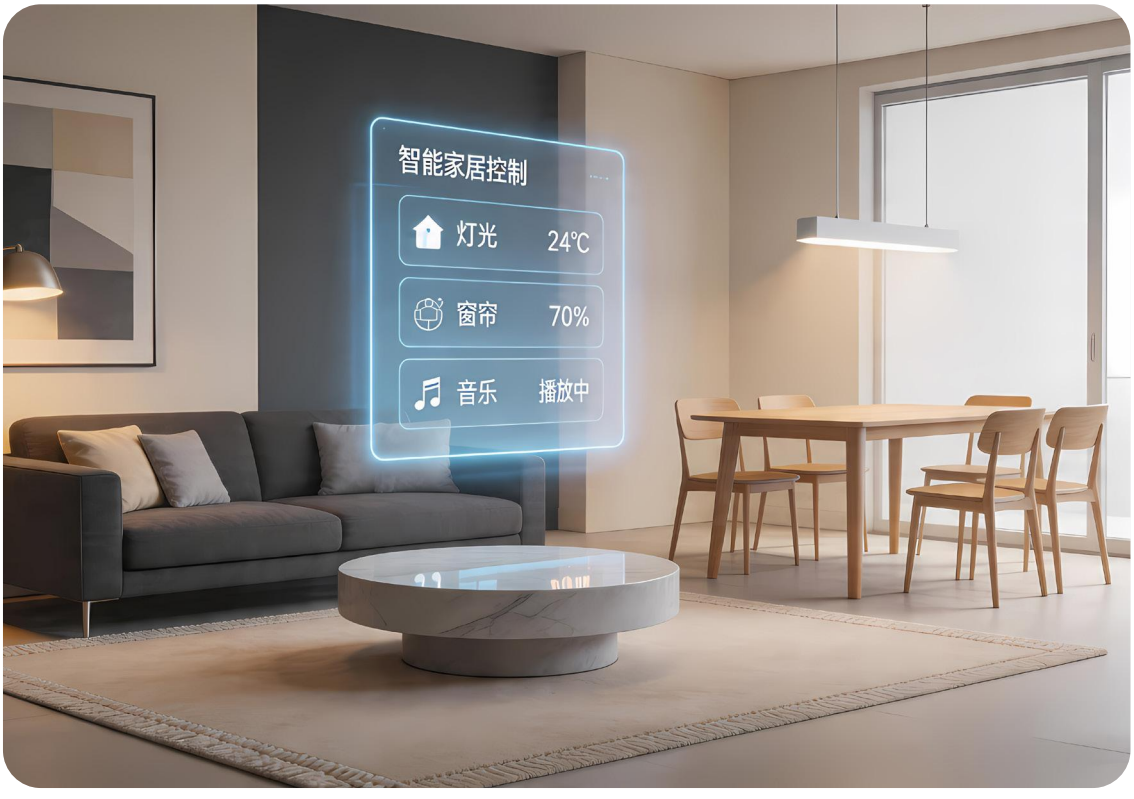
整体来看，数字 AI 正在推动信息智能化，而物理 AI 将推动生产力智能化。从比特到原子，智能技术不再只是信息工具，而是能够直接改造物质生产过程的基础设施。未来十年，围绕云边端算力底座、世界模型、仿真平台、智能执行硬件的产业链将持续成熟，物理 AI 将深度渗透实体经济各个环节，掀起新一轮产业变革。

从触控到多模交互，终端成为 AI 智能体全时空载体

过去三十年，个人计算的核心逻辑是将更多功能装入一台随身设备，但人需要主动唤醒终端、寻找应用并适应界面。AI 革命正在改变这一切：智能将不再被封装于某台设备，而是围绕人的身份、意图和所处场景，在不同终端间持续流动。Mark Weiser 早在 1991 年提出，最深刻的技术会融入日常生活，直至隐入人的注意力之外 [1]。到 2035 年，个人计算将从“设备随身”迈向“能力随人”，终端则由孤立工具演化为智能共生系统。

通过分体组合形态与端侧异构算力、云端智算服务协同，终端将提供无感泛在、基于场景的服务伴随。计算、显示、影像、传感等模块将逐步解耦，手机、眼镜、耳机、汽车、家居和机器人不再各自为战，而是根据任务动态组合为个人终端集群。端侧模型承担实时响应、隐私敏感计算与离线自治，边缘节点负责跨设备协同，云端模型提供复杂推理和知识扩展，形成端—边—云动态分工。三者相互配合，将基础模型的通用能力与终端的长期记忆、个人习惯与实时情境相结合，并转化为面向个人的主动服务。

五感融合将缔造更具沉浸感的临境体验。未来终端不仅“看得见、听得到”，还将通过触觉、嗅觉和味觉等双向接口感知并反馈物理世界；结合空间定位、环境理解和连续时间建模，形成对三维空间及其动态变化的“时空全景”。学界已有一些研究报道了柔性微型无线



更多免费行业报告 并购家 www.ipoipo.net

嗅觉接口、可远程感知和复现味觉系统等[2][3]，这些探索表明，终端正在由视听窗口演化为人类感官的物理延伸，成为数字世界与物理世界互通的桥梁、碳基生命与硅基智能交融的重要入口。

交互将走向空间多模态（SMUI），人机协同范式也将从过去“机器被动执行指令”，演进为围绕人的意图，由人与 AI 共同思考、完善、策划与分解，并协同完成任务。触控不再是唯一入口，语音、视觉、手势、眼动、姿态等模态将共同表达用户意图；系统不再是被动等待指令，而是能够理解场景上下文、主动预测需求，与用户协同推演并规划复杂任务。UI 将由固定菜单和层级页面，转变为依据人物、环境和任务即时生成的交互空间。未来界面将不再要求用户寻找功能，而是围绕当前意图组合最必要的信息与操作；硬件和 UI 将进一步融合，在提升能力的同时走向美观简约。

上述转变对终端操作系统提出了根本性挑战。传统 OS 以应用为中心、以设备为边界，难以支撑动态组合的终端集群、跨模态的自然交互，以及主动预测与协同推演。当智能不再依附于单一设备，而是围绕人的意图持续流动，操作系统必须从管理硬件资源与应用程序，转向管理智能体、上下文与协作关系。这正是 Agent OS 诞生的动因。以 Agent OS 为底座，海量智能体与元服务将形成相互连接的新生态，支撑人机共生体的高效运转。应用将逐渐由独立入口转化为可被智能体调用的能力模块，不同设备和服务中的专业智能体可以围绕与用户共同确立的目标分工协作，突破单一智能体的知识与能力边界。终端服务入口将由“用户选择哪个应用”，转向“系统如何组织能力实现用户意图”。

为用户提供无感、物理级的数据主权保护，将成为智能共生的信任底座。终端持续感知环境，智能体又能读取数据、调用服务甚至完成交易，隐私安全因而不能再作为合规补丁，而必须贯穿数据采集、存储、传输、使用、共享和销毁全生命周期。“物理级绝对数据主权”并非承诺不存在风险，而是让敏感数据本地优先、采集状态可感知、调用权限可控制、关键行为可审计和可撤销，并由安全芯片、可信执行环境、硬件隔离及机密计算提供底层保障。真正的“无感”不是无边界采集和不可见决策，而是在降低操作负担的同时，让用户始终拥有最终知情权、决定权与中止权。

到 2035 年，手机仍将是重要节点，但不再是个人智能的唯一中心。分体组合、智力可扩、临境传感、时空全景、安全可靠与美观简约将共同构成新终端的基本特征：设备形态逐渐隐入生活，智能跨越时间、空间和终端持续伴随。最终胜出的不再是功能最多的单台设备，而是最能自然感知、准确理解、安全调度并长期服务于人的智能共生系统。

---

[1] Weiser, M. The Computer for the 21st Century. Scientific American, 265(3), 1991.

[2] Liu, Y. et al. Soft, Miniaturized, Wireless Olfactory Interface for Virtual Reality. Nature Communications, 14, 2297, 2023.

[3] Chen, S. et al. A Sensor-Actuator–Coupled Gustatory Interface Chemically Connecting Virtual and Real Environments for Remote Tasting. Science Advances, 11, eadr4797, 2025.

智能体流量占比超九成，钟摆效应驱动计算联接交替爆发

随着人工智能（AI）技术的飞速发展，互联网正迎来一次深刻变革。“智能体互联”将成为下一个互联网的新范式，并推动通信网络从“连接百亿人”向“连接千亿智能体”演进。

智能体互联将重塑网络基础设施

随着AI技术的普及和智能体（Agentic AI、Physical AI等）的广泛应用，网络流量结构正在发生根本性变化。我们预测，到2035年，智能体的相关流量将实现100倍以上的增长，占整体网络流量的90%以上。这种指数级的流量膨胀，将催生以智能体为中心的新型网络架构。

智能体不仅仅是简单的终端设备，更是具备感知、决策和执行能力的自主实体。它们之间的互联，不仅涉及数据传输，更涵盖智能协作、意图交互和任务分配。智能体互联将重塑互联网的基础设施，推动通信网络从单纯的信息通道，转变为智能协同的生态系统。这一范式的转变，要求网络具备更高的智能化水平、更强的实时响应能力和更完善的安全保障。

端边云网协同是智能体规模化的必由之路

智能体的广泛部署和复杂应用，对计算资源和网络时延提出了极高要求。“端边云网”协同是实现智能体规模化的关键：端侧设备负责感知和初步处理，边缘节点承担实时推理和局部决策，云端则提供强大的算力支持和全局优化，而网络则是串起端边云，并产生系统价值的关键链条。面向智能体互联需要研究新的网络协议，用于解决智能体的身份发现与互操作问题，确保不同智能体之间能够安全、可靠地识别和通信；同时还需要构建网络信任与安全防护机制，防止恶意攻击和数据泄露，保障智能体生态的健康发展。一个多方参与、共生共赢、端边云网协同的智能体生态系统，将推动智能体互联网从概念走向规模化应用。



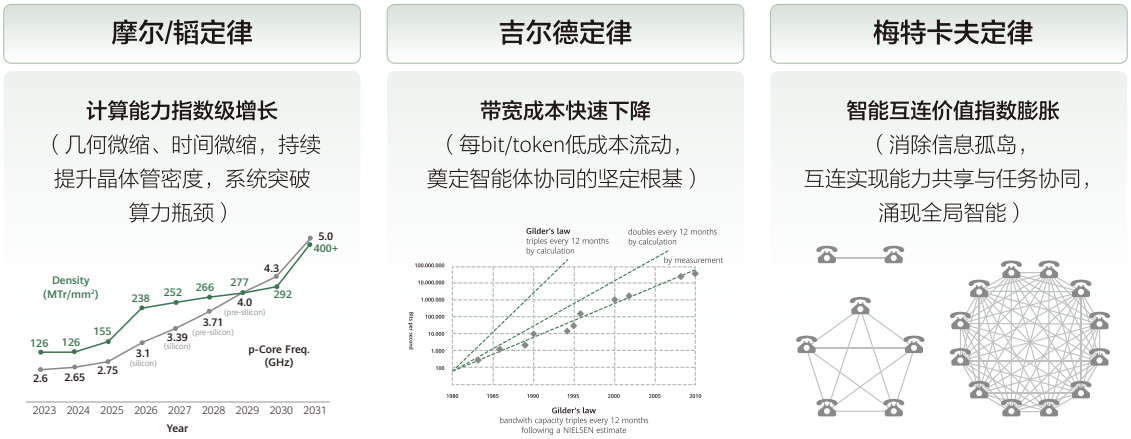
更多免费行业报告 并购家 www.ipoipo.net

三大定律驱动计算与联接将保持钟摆式发展规律

指导互联网数十年发展的三大定律在智能体时代仍将发挥关键作用。梅特卡夫定律（用户价值）和吉尔德定律（带宽能力）共同决定了网络的规模效应，而摩尔 / 韬定律（硬件性能）则定义了技术底座的实现路径。它们共同组成了互联网那只看不见的“流量之手”，指挥计算与联接、通用与专用、集中与分布等实现动态平衡与相互驱动。



三定律一效应 共同驱动智能体互联钟摆式交替发展



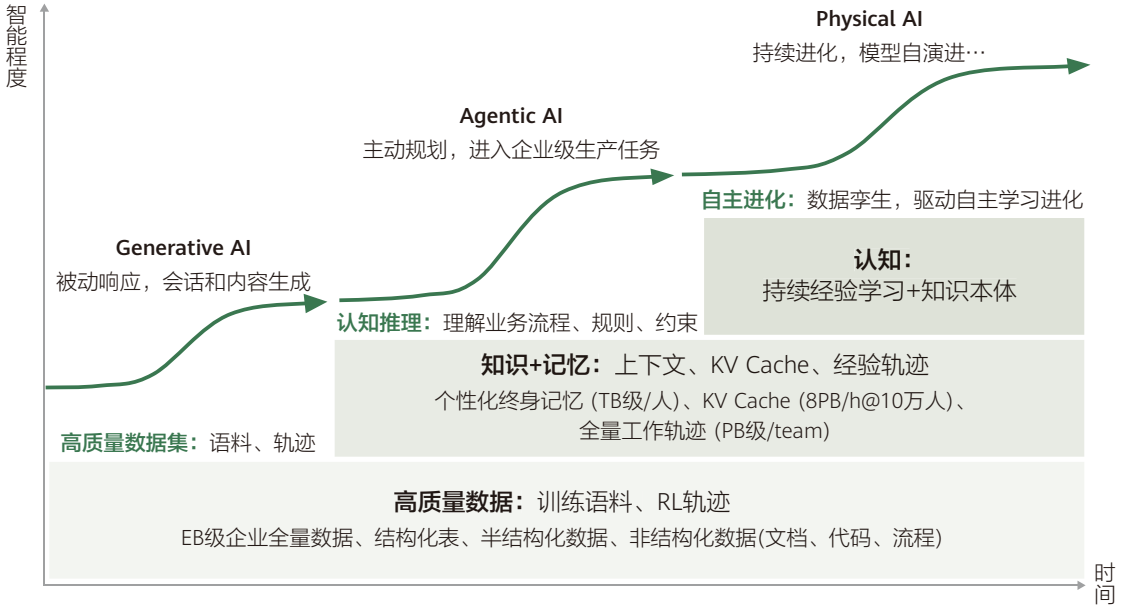
三大定律驱动计算与联接保持钟摆式发展规律

算力需求的爆发将推动网络带宽需求激增，而网络的升级又反过来促进计算资源的优化配置。通用计算与专用加速器的交替与结合，可实现性能与效率的最佳平衡。集中式数据中心与分布式边缘计算则相辅相成，可满足不同场景的需求。这种数据洪流驱动下的钟摆式互动关系，将构成智能体互联网演化的内在动力，并推动技术不断突破，生态不断壮大。

智能体互联网的价值将与其节点（智能体）数量的平方成正比。每接入一个优秀的新智能体，都可能为整个网络所有其他智能体带来新的协作可能和能力增强。我们预测，2035 年全球智能体数量将突破 9000 亿，超过人口 100 倍，当数千亿的专业智能体联接在一个协议互通的网络上时，其能组合出的解决方案数量将是天文数字，并催生出全新业态。

数据决定 AI 能力上限，存储夯实智能产业根基

在人工智能发展的长河中，我们正经历一场深刻的范式转移——从“算法驱动”走向“数据驱动”，从“被动响应”迈向“主动创造”。这一转变的核心，是数据、知识与记忆的协同作用，也将成为人类探索未知边界的重要支撑。



智能程度沿 S 型曲线不断攀升，其背后是底层要素的系统性重构

正如图中所示，AI 的智能程度沿 S 型曲线将不断攀升，其背后并非单一技术突破，而是底层要素的系统性重构。

范式转移：数据成为能力上限的决定性变量

过去十年，AI 的发展主要依赖算法架构的创新：从 CNN 到 Transformer，从 RNN 到 Attention 机制，每一次架构升级都带来性能跃升。然而，随着边际收益递减，单纯依靠算法优化已难以突破瓶颈。当前，模型发展的重心正悄然转向“数据竞争”。

据 Epoch AI 估计，前沿大模型的数据集每年增长 2.7 倍，且增长速度仍在加快 [1]。这意味着，谁拥有更高质量、更结构化、更具领域深度的数据，谁就掌握了定义 AI 能力天花板的关键钥匙。未来 AI 的竞争，不仅是“谁的模型更大”，也是“谁的数据更精、更准、更专”。

[1] David Owen ‘What will AI look like in 2030?’. Epoch AI, 2025.

更多免费行业报告 并购家 www.ipoipo.net

知识增强与记忆固化：从原始信息到领域专精，构建具备持续进化能力的长效智能体

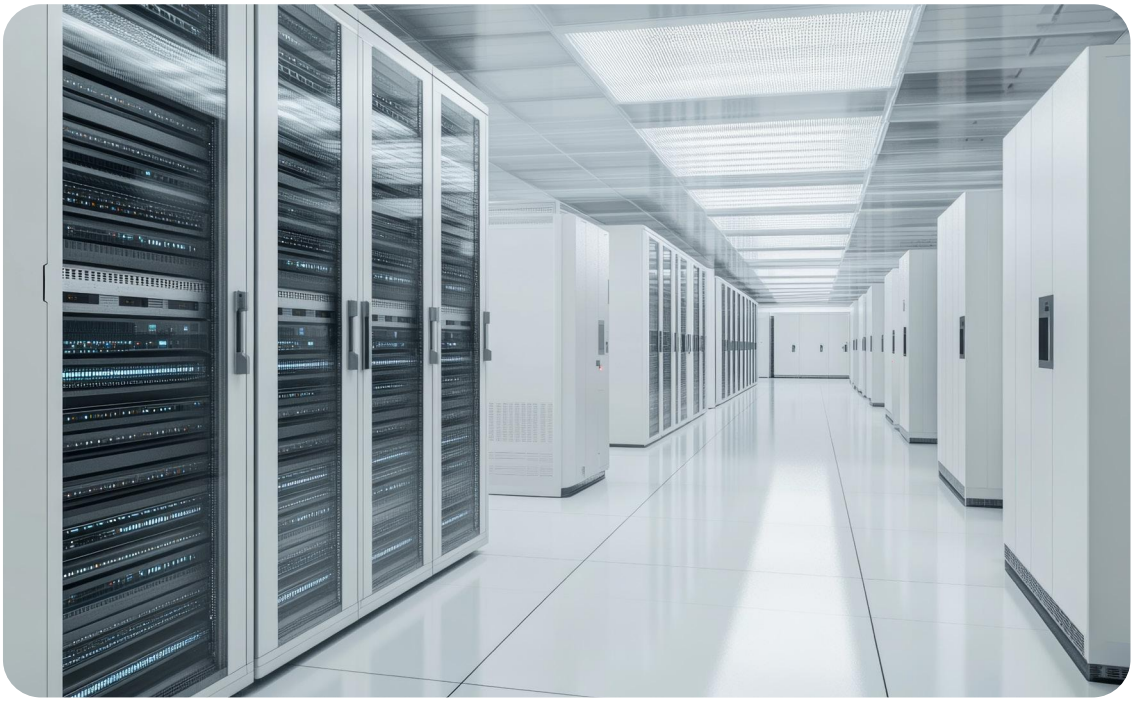
数据本身只是原材料，唯有转化为“知识”，才能赋予 AI 真正的理解力和推理力。传统的语言模型本质上是基于符号概率的统计拟合，难以真正理解现实世界的概念与逻辑。真正的知识增强，是在数据清洗之上，基于本体论（ontology）对特定领域内的概念实体、分类层级以及因果规则建立一套严密的形式化映射。

传统 AI 模型是“无状态”的——每次交互都是独立事件，无法积累经验、形成个性。而新一代 AI 正通过“记忆固化”技术，实现状态保持与跨期学习，打造真正的“成长型 AI”。

存储进化：打造具备知识和记忆能力的数据基础设施，支撑智能体持续成长

随着智能体对实时性和复杂性要求的提升，存储正经历深刻转型，从被动存放数据的静态“仓库”，跃迁为主动感知算力需求、调度数据、供给知识的“主干道”，也成为同时具备“高性能、低延迟、介质多元、架构网络化及弹性可扩展”的核心基础设施。这种跃迁不仅是技术底座的升级，更是 AI 智能体实现自我进化的先决条件。

从 Generative AI 到 Agentic AI，每一次跃迁都建立在数据、知识与记忆的协同进化之上。高质量数据提供依据，知识图谱赋予深度，记忆系统实现成长。而这一“智能引擎”的高效运转，离不开存储技术的持续革新与介质供应的多元化。存储作为承载数据、知识和记忆的关键载体，已从后台支撑走向前台核心，其技术进步降低了数据边际成本，其供应体系确保了算力扩展的可持续性。未来，筑牢智能产业根基的，不仅是更聪明的算法，更是先进、可靠且具弹性的存储技术与介质供应体系。



从“几何缩微”到“时间缩微”，新原则指导半导体产业持续演进

当我们谈论 2035 年的智能世界——9000 亿智能体互联，AI 流量远超人类流量时，一个根本性的问题浮出水面：支撑这一切的半导体技术，还能持续演进吗？

主导半导体产业半个多世纪的摩尔定律，正从“几何缩微单一驱动”的黄金时代，走向晶体管几何缩微放缓、成本红利消退的双重瓶颈期。半导体演进的逻辑，正从单纯依赖工艺制程微缩，转向亟需寻找全新可持续路径的关键路口——韬（τ）定律正是为这一转向给出的答案。

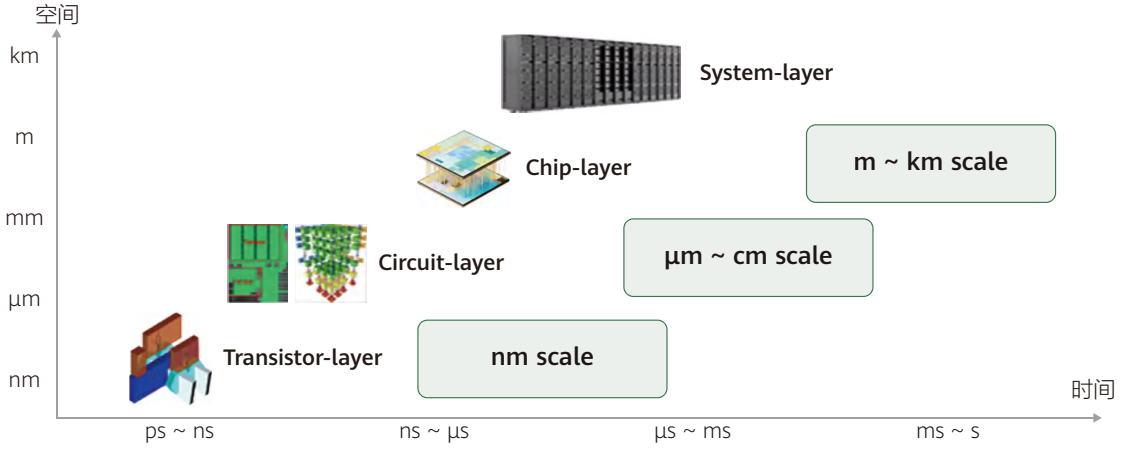
韬（τ）定律提出以“时间（τ）缩微”替代“几何缩微”作为半导体与电子系统演进的新指导原则——通过逻辑折叠等创新技术，持续压缩信号传播时延，不断提升晶体管密度，从而实现半导体与电子系统的持续演进。

韬定律将时间常数 τ 的优化分为四个层次：

在器件层，通过优化晶体管和互连电阻，从物理底层最大限度缩微器件级时间常数 τ；在电路层，通过逻辑折叠技术突破传统平面布局的物理边界，显著缩短关键路径的走线长度并有效降低信号传播的电阻和电容，实现晶体管密度和电路性能大幅提升；在芯片层，通过“软件、架构、芯片”的全栈软硬芯协同设计，基于实际工作负载实现指令流和数据流的细粒度控制，提高系统级并行度和效率，大幅降低端到端执行时间；在系统层，定义统一总线，重构计算系统互联协议，实现超节点的统一内存编址和原生内存语义，大幅降低系统通信时延。

四个层面从下到上，从物理到系统，每一层都在压缩 τ，每一层都在贡献乘积因子。这不是某一层突破就能实现百倍的线性叠加，而是四层乘积效应的指数级释放。

摩尔定律给了半导体产业第一个五十年，韬定律则揭示了下一个阶段的演进逻辑：当几何缩微遇到物理墙，时间缩微将打开新的优化空间，支撑性能、功耗、密度、成本的可持续演进。



τ 缩微跨越时空 12 个数量级，覆盖晶体管至系统四层

更多免费行业报告 并购家 www.ipoipo.net



电力决定算力，AIDC 走向算电网储协同

随着 AIDC 集群建设规模持续扩张，单纯依靠 GPU 硬件堆叠的增长模式已遭遇瓶颈，难以适配大规模智能算力的长期发展需求。未来 AIDC 的商业化落地与高效运营，不再取决于单一算力硬件规模，而是由电网供电容量、高压输电能力、供电稳定性、绿电消纳水平及大容量储能配套等多维度要素共同决定。在此背景下，AIDC 正逐步摆脱传统独立算力机房的单一形态，加速迭代为算力、电力、网络、储能深度耦合、协同联动的新一代分布式智能基础设施，实现算力供给与能源调度的一体化、常态化协同运行。

随着 AIDC 集群规模持续迈向 GW 级乃至超大规模，电力供给将决定算力扩容上限

当前 AIDC 单机柜功率已攀升至数百千瓦级别，集群耗电迈入 GW 级及以上规模，电力系统已经彻底摆脱传统配套支撑属性，成为制约算力规模化扩容的关键要素。未来，即便算力芯片硬件仍能持续迭代升级，但若电力供给无法同步适配，大规模智能算力部署也将面临落地瓶颈。据 IEA、彭博新能源财经等权威机构预测，2035 年全球数据中心总耗电量将达 1600TWh，用电占比将由 2025 年的 1.5% 攀升至 4.5%–10%。算力用电需求的爆发式增长，将倒逼 AIDC 电力供给体系与运营模式全方位重构，彻底摒弃传统就近取电、单向供电的粗放发展模式。

AIDC 将逐步向“算、电、网、储”统一闭环控制系统演进

AIDC 算力负荷具备高波动、大摆幅的典型特征，而新能源发电具有间歇性、随机性波动的特点，二者叠加对传统“源随荷动”的单向供电模式提出了严峻挑战。这将倒逼 AIDC 摆脱单一用电主体的定位，向“算电网储”四位一体的协同运营范式深度演进。

算力侧：AIDC 算力负荷将不再是电网刚性耗电负荷，未来可主动参与电网协同调节，具备灵活负荷调节与反向馈电能力，可有效平抑电网功率波动，提升新型电力系统的运行稳定性。



电源侧：依托风、光、水、火、核多能源互补协同，为大规模 AIDC 算力集群提供充足、稳定、低碳的电力支撑，持续降低算力产业碳排放，实现算力绿色低碳发展。

电网侧：实现电网与算力网络双向感知、动态联动调度，算力负荷可随电力资源富余态势灵活迁移；电网依托高压直流供电技术保障算力稳定供能，同时 AIDC 可反向参与电网调峰补缺，平抑新能源出力波动，实现网算双向赋能。

储能侧：融合电容、锂电、钠电、氢能、甲醇、液氨等多元储能技术，构建短时、长时分层协同的复合型储能体系，可覆盖从毫秒级、小时级到跨日、跨周的全周期功率支撑需求。储能系统也从传统备用电源，升级为集保供稳供、电网调峰、能源套利于一体的核心经营性资产。

多元绿色能源是支撑 AI 持续演进的核心基石

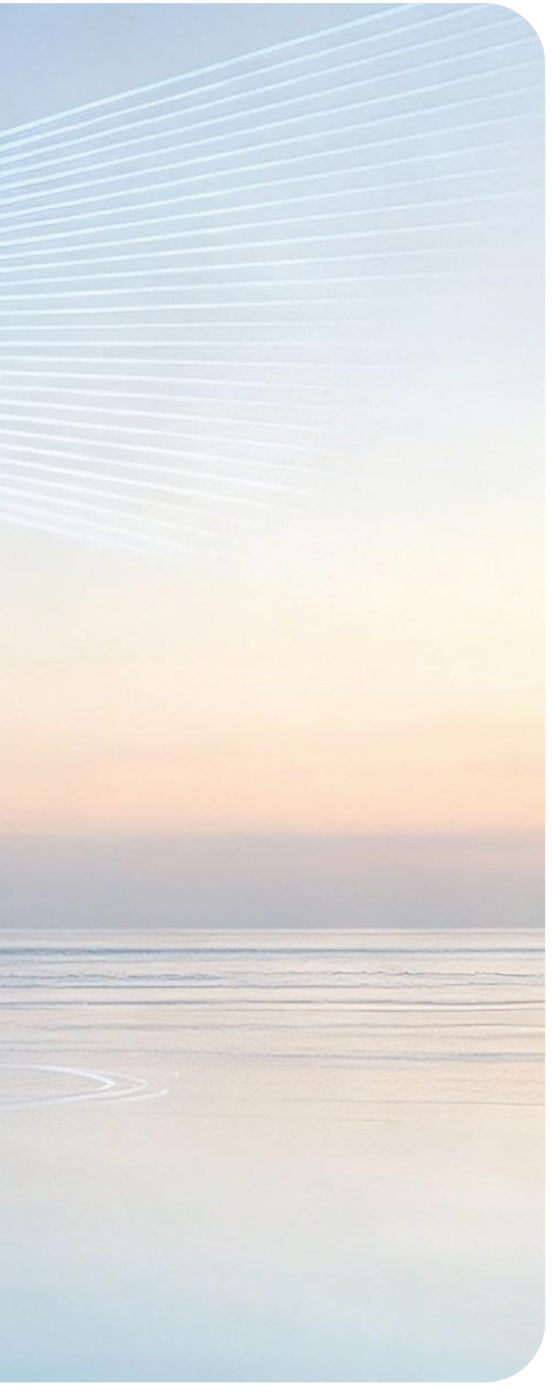
未来多元综合绿色能源体系的全面成熟，将彻底破除电力供给对 AI 算力扩张的制约，为人工智能长期高速演进提供充足、稳定、可持续的能源保障，算力发展无需再受电力资源瓶颈约束。未来风光储一体化综合能源体系将全面普及，以风电、光伏为主体、多元储能为配套的绿色能源架构，将成为 GW 级 AIDC 集群规模化、持续性扩张的核心底座。针对风光出力间歇性的天然短板，通过多形态储能系统削峰填谷、动态调谐，可精准匹配算力负荷的大幅波动，保障大规模智算集群全天候稳定连续运行。

在此基础上，生物质、绿氢、潮汐、地热等零碳能源形成多源互补的地面能源体系，构建高弹性、低碳化、可兜底的综合供能底座。同时，以太空光伏为代表的太空新能源范式快速兴起，依托轨道无间歇光照、超高光伏利用率、太空集中供能等独特优势，突破地面土地、光照、电网容量的物理限制，形成天地一体的新型绿色能源供给格局。

天地多维度绿色能源协同赋能、长短时储能体系互补兜底，彻底解决大规模 AIDC 与超级 AI 算力的用能难题，为通用人工智能（AGI）时代算力的持续扩容、长期迭代提供源源不断的绿色动力。

更多免费行业报告 并购家 www.ipoipo.net





命题

通往智能世界的
十大关键命题与研究方向

更多免费行业报告 并购家 www.ipoipo.net



在 AI 从技术变量向文明基石加速演进的进程中，通往智能世界的道路绝非坦途。唯有直面底层挑战、系统性地定义并攻克核心难题，才能将愿景真正转化为可落地的现实。围绕这一认识，我们凝练出十大关键命题，它们共同构成了智能世界必须跨越的“技术隘口”。这十大关键命题不是孤立的技术点，而是一张相互耦合的攻关图谱——唯有逐一破题、协同推进，方能为 Agentic AI 的规模化落地铺就坚实底座，让智能真正融入物理环境与生产流程，实现万物互联的智能世界蓝图。

构建具备自主认知与全域通用能力的 AGI

AI 技术高速迭代发展，对于何时真正实现通用人工智能（AGI），各个组织有不同的判断，跨度从 2027 年到 2035 年，或者更长。分歧主要是因为各个组织对 AGI 有不同的定义，不同的定义将决定走向 AGI 的路径和实现的时间点。本文采用维基百科的一个定义：“AGI 是一种假想的智能体，它能够学习并执行人或其他动物所能完成的任何智力任务”。其核心在于实现通用性、自主性与创新性：能够在新环境中理解、学习并完成任何人类可及的智力任务，胜任跨场景推理、复杂问题解决，并展现出类人级别的灵活适应与创造能力，我们认为走向物理世界是 AGI 形成的必由之路。所以，作为人工智能发展的目标，AGI 在未来十年的发展之路上，仍面临诸多核心挑战。

实现 AGI 面临的关键挑战

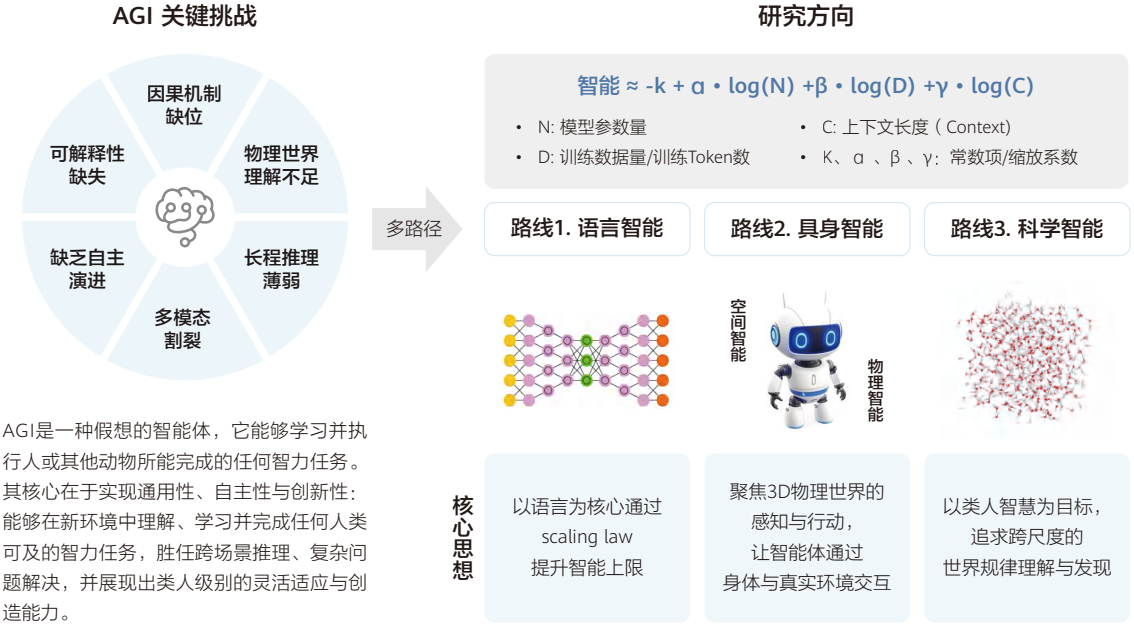
当前 AI 发展的主流路径是基于 Transformer 架构下概率统计为基础的相似注意力机制构建，存在六重结构性缺陷和挑战。

可解释性原生缺失，决策链路难以审计溯源

Transformer 的智能表征承载于海量参数的高维概率分布之中，模型输出结果源于参数空间的统计拟合与概率采样，不存在显式、可固化的推理逻辑链条。学界已多次验证，自注意力权重仅能反映单元间的关联热度，无法等价于模型真实的推理依据与决策逻辑，传统注意力可视化方法不具备可靠的归因解释能力。当前大模型整体处于典型“黑盒决策”状态，输出结果的生成逻辑、依据来源、误差成因难以追溯与量化审计。这一结构性缺陷，导致模型幻觉治理、安全价值对齐、风险溯源管控缺乏底层技术支撑，无法满足金融、医疗、工业控制等高风险垂直领域对决策可解释、可追溯、可审计的刚性落地要求。

具身物理认知匮乏，统计表征不等同于世界理解

现有大模型的知识习得完全依托海量数字化文本、图像等静态数据的监督与自监督学习，本质是对人类语言描述、视觉表象的统计模式复刻，并非基于真实物理世界的感知、交互与试错习得。其学到的“物理常识”仅是数据中的文本与图像统计关联，而非对真实世界运行规则、因果约束、空间拓扑、运动逻辑的底层认知，无法构建贴合物理现实的世界模型，进而难以支撑机器人操控、工业仿真、复杂场景决策等需要真实物理认知的高阶任务。



AGI是一种假想的智能体，它能够学习并执行人或其他动物所能完成的任何智力任务。其核心在于实现通用性、自主性与创新性：能够在新环境中理解、学习并完成任何人类可及的智力任务，胜任跨场景推理、复杂问题解决，并展现出类人级别的灵活适应与创造能力。

构建具备自主认知与全域通用能力的 AGI

更多免费行业报告 并购家 www.ipoipo.net

因果推理机制缺位，相关性拟合制约泛化能力

自注意力机制的核心运算逻辑是基于特征相似度的关联拟合与权重分配，本质捕捉的是数据维度的统计相关性，而非事物背后的内在因果机制。模型仅能基于历史数据的统计分布完成拟合生成，无法开展主动干预推理、反事实推演与因果归因。这是大模型产生事实幻觉、逻辑谬误的核心根源，同时直接导致模型在分布外场景、全新场景、小样本场景下泛化能力大幅衰减，难以适配真实复杂、动态变化的产业场景。

长程链式推理脆弱，误差累积导致复杂任务失效

当前主流大模型采用自回归逐 Token 生成范式，生成过程为串行递进式输出，缺乏全局前置规划、层级化任务拆解与实时误差校正机制。在长链条逻辑推理、复杂工程推演、多步骤任务规划等场景中，前置生成的微小误差会随推理链条持续传递、叠加、放大，形成推理精度的累积衰减效应。同时，长序列下注意力权重稀释、信息信噪比下降，叠加长期记忆调度、跨层级任务规划能力的缺失，导致模型难以稳定胜任长周期、高复杂度、高精准度的持续性任务，无法满足行业应用中的确定性、高可靠、可复现的输出要求。

自主演进闭环缺失，智能迭代依赖人工赋能且伴随安全风险

现有大模型的能力边界由离线预训练、人工微调、指令对齐等人工干预流程决定，模型不具备自主感知任务缺陷、自主采集高质量数据、自主迭代优化参数、自我修正能力的闭环进化机制。模型迭代高度依赖人工数据筛选、标注、调参与方案设计，智能化迭代效率低、进化成本高、场景适配滞后。与此同时，当前自迭代学习、自主强化学习等前沿自进化方案尚未建立成熟的价值约束体系，若完全开放模型自主演进能力，将面临人类价值对齐失效、能力不可控、行为不可预测等安全治理难题，形成“无自主进化则能力固化，有自主进化则安全失控”的结构性矛盾。

多模态表征体系割裂，缺乏统一原生世界表征

现阶段绝大多数多模态大模型存在模态融合不充分、跨模态推理不一致、感知与生成链路割裂等问题，无法形成全局统一、逻辑自洽、可跨模态迁移的完整世界表征体系，制约了通用智能对复杂现实场景的整体感知与综合推理能力。

六者相互关联，任何单一维度的突破都不足以实现 AGI，唯有这些挑战协同推进，AGI 才可能从概念走向现实。

实现 AGI 的可能研究方向

人类智慧的形成，根植于语言交互、环境任务实践与科学创新探索三大核心过程。对标这一演化规律，AGI 的发展可依托语言智能、具身智能、科学智能三大核心方向协同推进，三者对应通用智能的底层核心维度，相辅相成、缺一不可。其中语言智能搭建思维骨架，承载逻辑推理与知识传承；具身智能筑牢物理认知根基，塑造空间感知与原生因果直觉；科学智能驱动认知层级跃迁，赋予智能体主动探索规律、开展科学发现与自主迭代理论体系的能力。RSI 作为科学智能的关键，使 AI 将每次发现沉淀为可复用资产，自主提出新假说并递归优化模型，形成自我加速闭环，使 AI 从静态训练迈向持续进化，是实现通用智能的必须能力。三者共同构筑了 AGI 从理论构想落地、迈向通用自主智能的核心发展路径。

方向一：语言智能——以语言为核心，依托缩放定律（Scaling Law）持续突破模型智能上限，通过架构迭代、记忆体系升级、多模态

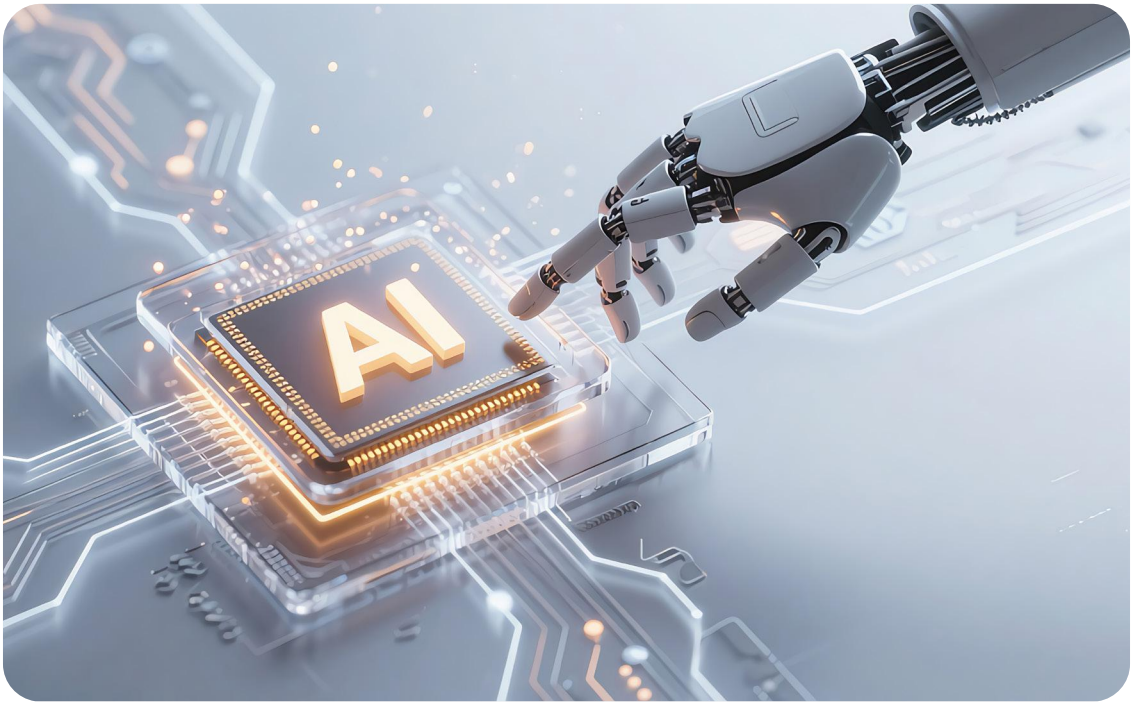
统一与因果推理优化，构筑稳定、可解释、可迁移的通用思维与推理底座。

方向二：具身智能——聚焦 3D 物理世界的感知、理解与动态交互，推动世界模型、空间智能、物理因果建模、端到端具身控制与长程自主规划等研究，让智能体能够深度适配真实环境、执行复杂实时任务，补齐 AI 物理认知与场景泛化短板。

方向三：科学智能——以类人高阶智慧为目标，开展跨时空、跨尺度世界规律的理解与挖掘。依托递归自主演进智能（RSI）机制，构建主动假设、实验验证、规律提炼、自我迭代的闭环系统，推动 AI 从被动数据拟合走向主动科学探索与认知升级。

当前主流大模型仍局限于概率统计拟合框架，本质是基于海量数据压缩形成的检索式知识库，输出效果高度依赖训练数据质量。这类概率黑盒模型不具备人类“知其然、知其所以然”的结构化逻辑认知，也是模型幻觉、推理脆弱、无法真正泛化的根源。因此，模型结构范式改变，融合符号主义思想，通过抽象符号表征、显式逻辑推演与因果结构化建模，弥补纯统计深度学习的认知缺陷也可能是走向 AGI 的关键方向。

几个研究方向并非独立演进、而是深度耦合、相互赋能：语言推理能力可在具身交互与科学探索场景中持续校验修正，提升推理真实性与稳定性；具身真实场景数据持续反哺语言表征与世界模型建模精度，消解纯文本学习的认知偏差；科学智能的递归自主迭代机制，为全域系统提供持续自我进化的核心动力；符号逻辑则为全链路推理、因果建模与可信决策提供规则支撑。几条路径最终深度融合，迭代构建统一自洽、可持续进化的通用世界模型，推动 AGI 从理论愿景，落地为可工程化、可迭代、具备自主认知能力的通用智能系统。



更多免费行业报告 并购家 www.ipoipo.net

算力集群规模百倍提升，智能体任务成本千倍下降

面向 2035 年，推动 AGI 与物理 AI 从前沿技术探索深度落地、赋能千行百业，行业将迎来十年维度的终极考验：持续推动智能体全任务综合成本跨越式下降，实现成本千倍级优化，让智能算力像水电一样普惠易得、随取随用。这一产业愿景，亟需算力系统整体运行效率实现质的跃升，彻底打破复杂智能业务的落地瓶颈与成本桎梏，全面降低通用人工智能的应用门槛，为 AI 规模化、产业化普及筑牢核心基础。

我们正面临严峻的供需端“剪刀差”：模型 Scaling Law 与智能体推理思考链（Test-time Scaling）驱动算力需求超万倍增长，但单芯片性能的发展受物理极限约束，不可能等比例增长。这一需求与供给的巨大的鸿沟，倒逼算力集群规模必须再提升百倍，迈向百万 / 千万卡时代。

然而，简单的“规模堆砌”会导致数据搬运开销急剧增加、有效算力利用率（MFU）大幅下滑。因此，行业演进的核心矛盾已从单纯追求“单芯片大算力”转向追求“超节点算力系统 Token 的高转换效率”。打破规模扩张下的效率瓶颈，并重构以“超节点”为核心的计算架构，将是未来算力基础设施发展的重要方向。

“超节点”的发展面临三重技术瓶颈

将算力集群规模推向百倍提升、同时实现千倍成本降低的过程，本质上是一场对抗物理规律与系统熵增的较量。未来挑战主要集中在超节点与其互连架构的三重“技术墙”：

物理极限与规模蔓延下的“搬运墙”

在传统冯 • 诺依曼架构下，计算与存储分离。随着集群膨胀至百万卡级别，简单的 Scale-out 拓展导致网络拓扑复杂度呈指数级上升。数据的跨芯片、跨机柜搬运所消耗的算力与功耗，已远超计算本身，海量 Token 若在



网络中遭遇严重拥堵，将会导致有效算力大幅衰减。

高动态 AI 负载与静态确定性架构的“错位墙”

Agent 智能体任务一般包含长上下文、MoE 动态路由与多轮工具调用，计算负载呈现极高的动态性与突发性。然而，现有超节点与集群主要基于静态拓扑、确定性路由设计，面对高动态流量极易出现尾部时延激增与计算单元阵发性闲置，从而造成巨大的算力浪费。

超大规模下的“可用性墙”

当算力集群组网规模提升百倍达到百万卡级别时，硬件故障将从“偶发”变为“常态”，传统集群因故障中断和等待恢复带来的开销将吞噬大部分有效生产力。

破局之道：未来可能的研究方向

未来只有摆脱“单纯依赖芯片微缩”和“堆砌服务器”的路径依赖，在应用 - 模型 - 架构 - 互联 - 芯片 - 能源的多维空间中寻求最优解，才能全面重构超节点形态与组网逻辑，可能有以下几个解题方向，值得业界共同探索：



方向一：重构计算范式，探索构建“机柜 / 晶圆级超节点”

推进晶圆级集成（Wafer-Scale System）与板级封装，将成百上千计算核心与大容量内存紧耦合在机柜内部，并研究存算一体（PIM）等非冯 • 诺依曼架构，通过消除单机物理壁垒，大幅缩短数据搬运距离，实现提高有效算力的目的。

方向二：突破 Scale-Up 极限，打造“光电融合”的低时延网络

全面引入共封装光学（CPO）与光交换（OCS），用光子替代电子完成超节点内及超节点间的高速数据交换，降低 70% 以上的互联功耗并实现亚微秒级时延。

方向三：模型 - 架构 - 互联深度协同，打造随需而变，动态自愈的“超节点操作系统”

依托模型、架构与互联机制的全方位深度协同，打造具备动态适配、自主自愈能力的超节点操作系统，实现集群计算资源与网络拓扑的微秒级动态调配、按需微粒度切片调度，精准匹配智能任务的实时算力需求。同时引入硬件遥测感知与 AI 态势预测的智能化自愈机制，搭建多层次容错体系与快速微重构能力。针对单计算核心、单链路突发故障场景，构建亚微秒级故障旁路与无感切换能力，推动算力集群从传统“故障中断、被动恢复”模式，升级为带伤稳态运行、动态自愈迭代的全新范式，显著提升大规模异构算力集群的整体可用性、稳定性与容灾能力。

同时，如果未来几年量子计算等多元新计算范式得到突破性的进展，那么量 - 超 - 智协同计算也可能是一个重要的方向，发挥超算、智算、量子计算的各自优势，形成新的计算范式，突破经典计算的物理上限。

更多免费行业报告 并购家 www.ipoipo.net

构建因果准确、溯源可信的新型存储与记忆系统

当 Agentic AI 从对话工具逐渐走向具备自主执行能力的智能体，一个关键的命题随之浮现：支撑 Agent 推理的数据存储与记忆机制，是否已经为这场变革做好了准备？

浅层化知识导致推理逻辑失真：当前，大语言模型的概率生成本质，天然存在“幻觉”风险。现有系统多采用被动语义提取与线性知识图谱模式，缺乏对多模态异构数据的立体化认知，难以将碎片信息转化为高阶逻辑知识，极易出现语义扭曲、逻辑断层等问题，无法满足高合规业务对确定性的严苛要求。

静态记忆机制导致场景泛化难：现有记忆系统依赖统一向量检索或固定提示词，缺少随时间推移自主演化的能力，当智能体面对新场景或长周期任务时，适应力就可能受限。

存算网架构低效：由于企业数据分散、格式不一，且在传输前缺乏本地预处理，导致算力中心面临存储和能耗的巨大压力，在传统存网算分层架构下，数据供给效率不可控、跨层搬移开销高、异构调度低效，制约 AI 应用的实时响应与规模化运行。

安全信任机制缺失：现有安全体系多为“边界筑墙”的外挂补丁，未将信任机制融入底层架构。Agent 推理全生命周期缺乏原生免疫能力，易受数据污染和恶意攻击，且缺乏内生防篡改机制，导致企业在开放共享数据时面临巨大的合规风险与信任危机。

这些问题激发我们深入思考，构建因果准确、溯源可信的新型存储与记忆系统时，需要重点研究如下几大方向：

构建基于因果逻辑的推理引擎，摆脱单纯的数据拟合模式。通过主动探索未知场景，建立事物间的深层关联而非表面关联；依托高维知识图谱，整合多维度信息以验证结论的合理性。最终实现从“概率猜测”到“逻辑确证”的认知升级，确保回答有据可依。

建立模拟人类联想与遗忘的智能记忆体系，打破传统数据库的僵化检索。通过情景记忆快速复用过往交互经验，实现对新任务的零样本或少样本适应；依托语义泛化能力，将特定场景知识迁移至相似领域。在降低反复训练成本的同时，保持记忆系统的动态更新与高效复用。

建立数据与计算深度融合的新型处理范式，消除传统架构中的数据搬运瓶颈。通过在存储层直接执行计算任务，实现“数据不动计算动”，大幅降低延迟与功耗；依托以总线为中心的层次化内存架构，优化高频数据的就近访问效率，发挥 DRAM、NAND、HDD、磁带、光盘等不同存储介质的各自优势，在提升响应速度的同时，显著降低整体算力运营成本。

建立内生于系统架构的安全防御体系，摒弃事后补救的传统防护思路。通过代码级、数据级的原生隔离与验证，确保核心逻辑与数据在运行中不被篡改或泄露；依托持续的安全机制，实时识别并阻断潜在威胁。在保障业务高效运行的同时，守住不可逾越的安全底线。

因此，未来智能体系统需要建立在“近数据计算 + 高速互联 + 多层存储”的全新范式之上，构建因果准确、溯源可信的新型存储与记忆系统，通过全对等互联打破传统通信瓶颈，让数据与算力的协同不再受制于搬运效率。这不仅是存储介质的一次迭代升级，更是行业从“数据管理”走向“认知智能”的深刻转变。唯有为数据底座注入原生的“认知”能力，智能体才能真正突破算力与数据的双重瓶颈，完成从简单的任务执行者，到准确、可信的决策伙伴的跃迁。

AI 全面融入通信系统，构建体验为中心的智能联接底座

随着人工智能技术的迅猛发展，其在各行各业的应用日益广泛，通信网络作为现代社会信息传递的基础设施，也正迎来与 AI 深度融合的新时代。

未来十年，在保证网络稳定性和安全性的前提下，AI 如何与通信网络结合，推动网络能力持续升级成为关键命题。结合通信网络架构特征，AI 与通信融合的主要研究方向将包含三个层面：



AI 与通信融合的主要研究方向

网元智能，实现网络能力持续升级

网元作为网络的基础构建单元，承担着数据转发、信号处理和资源管理等核心功能。将智能引入网元，能够实现网络能力的持续升级。例如，通过智能算法驱动的高效资源管理，可以动态调整网络设备资源分配、频谱利用率和能效；利用智能算法优化的低扰信号收发技术，可以增强信号质量和抗干扰能力；无线网络的实时功效寻优则通过智能算法不断调整网络参数，实现性能与能耗的最佳平衡。智能与网元结合，不仅提升了网元设备的智能化水平，也可为整个网络的高效运行奠定基础。

网络智能，推动网络自动化走向单域自智

网络管理是保障通信服务质量的关键环节。传统网络管理依赖人工配置和经验规则，未来十年，随着网络从联人到联智，覆盖从地面到空天地一体之后，人工模式将难以应对日益复杂的网络环境和多样化的业务需求。

引入智能后，网络管理将实现从单域自治向单域自智的转变。通过构建网络智能体，运营商不仅能够实现网络零退服、用户零投诉、业务零等待的目标，还可以用于实时监测和预测网络状态，甚至在异常发生前主动干预，从

更多免费行业报告 并购家 www.ipoipo.net

而极大提升网络的自愈能力和服务连续性。这种智能化管理模式不仅降低了运维成本，也提升了用户体验。

截止到 2026 年 8 月，全球已有 97 家产业伙伴签署《AN 宣言》，自智网络（AN）已形成产业共识。我们预计：2030 年将实现整体 L4 级别的能力跨越，2035 年将达成整体 L4+/L5 级别自智网络的构建与应用。

为了实现 L5 级别的自智能力，网络需要具备高精度数据实时感知和网络全域、多模态、异构、多层数据的本体建模能力，通过构筑虚拟的数字孪生网络，以开放接口的方式提供给智能体动态调用、并利用强化学习等技术增强网络知识，可有效提升复杂任务的意图理解、目标分解、工作流编排等自主决策能力。通过大模型与知识图谱融合，并增强大模型可靠性，可提高知识图谱未知问题推理能力，破局已知到未知问题的高效可解释性。未来，网络智能体可以动态生成决策闭环思维链，并基于网络数字孪生底座仿真验证，实现接近 100% 的准确性。

业务智能，赋能业务焕新提升业务体验

通信网络的最终目的是支撑丰富多样的业务应用，并为业务提供良好的体验保障。

一方面通信系统可引入业务智能平台（AISF），赋能主营业务，焕发新活力。AISF 通过深度学习和大数据分析，可以精准地理解用户需求和业务特征，并支持沉浸式通话、Token 体验保障、智能体通信等创新业务场景。例如，沉浸式通话依赖于低延迟、高带宽和智能信号处理，借助智能算法能够动态优化网络资源分配，从而保障通话质量；未来 Token 体验保障可通过智能身份认证和安全策略，来提升用户安全感；而智能体通信则可以支持人、物、智能体之间的多样化交互，从而推动万物互联的实现。

另一方面基础网络也可以结合不同业务的特点进行方案创新，提升业务体验。如针对入算网络，在保持高性能数据转发的同时引入数据面推理，可以实现业务识别、体验保障、威胁检测等，提高业务性能、降低业务时延；针对算内通信，通过将通信开销隐藏于计算时间中，突破“通信墙”瓶颈，使集群有效算力接近线性扩展；针对算间互联，构建波长级、空间级全光交换网络，实现算力、存储资源的无阻塞光层直达，让动态算力调度的时延降低一个数量级。

AI 与通信网络的融合是未来通信技术发展的必然趋势。面对 AI“幻觉”带来的挑战，必须坚持高可靠、高可用的网络原则，稳步推进 AI 技术在网元设备、网络管理和业务层面的应用创新。面向 2035 年，一张融合了 AI 的通信网络，将不仅满足传统的人与人通信需求，更能支持物联网设备和智能体的多样化联接需求，从而实现网络智能化的全面升级。



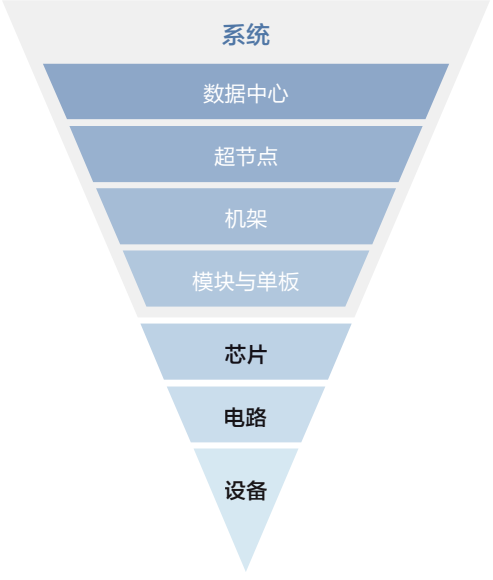
深耕韬定律下芯片设计方法和工具，支撑时延与能效协同优化

韬定律提出以“时间微缩”替代“几何微缩”，通过多层级优化逼近百倍性能提升。但韬定律不止是理论框架，要真正走向工程实践，还需要建立与之匹配的设计方法和工具体系，而这首先需要突破以下两大核心约束：

逻辑折叠设计方法学缺失：逻辑折叠是韬定律的核心机制——通过纵向多层堆叠重构物理路径，在同制程下实现功耗下降 41%、面积下降 37.5%。“折叠”彻底改变了芯片物理拓扑，互连从水平走向垂直，信号路径从二维变三维，寄生参数分布模型完全重写。传统 EDA 工具链为平面展开设计而建，面对三维拓扑既无法准确建模垂直互连寄生效应，也无法在折叠约束下做自动化物理综合。这不是打补丁能解决的，需要从底层算法开始，面向折叠重构整套设计方法学。

多层级能效挑战：芯片功耗持续增长，热阻和热传导控制成为关键挑战。从器件的毫瓦到系统的吉瓦，横跨 12 个数量级；从皮秒级开关到秒级热平衡，时间维度跨越多个尺度。器件层的寄生电容优化影响电路层开关速度，电路层折叠方案改变芯片层功耗密度分布，芯片层功耗布局决定系统层热管理策略。真正的能效优化必须是跨层级的系统权衡——在功耗 - 时间两个维度中做联合优化。

$$\tau = f\left(\tau_{tr}, \tau_{circuit}, \tau_{chip}, \tau_{sys}\right)$$



层级	目标函数	时间尺度
器件	$\tau_{device} = (R_{gate}+R_{net}) * (C_{gate}+C_{load})$	ps
电路	$\tau_{circuit} = f\{\sum(\tau_{transistor} \,\&\, \tau_{interconnect}), DTCO\}$	ns
芯片	$\tau_{chip} = f\{\sum\tau_{cir_{XPU}}, \sum\tau_{cir_{MEM}}, \sum\tau_{cir_{IO}}, N_{parall}, \eta_{eff}, M_P\}$	μs
系统	$\tau_{sys} = f\{\sum\tau_{chip}, \sum\tau_{comm}, N_{parall}, \eta_{eff}, M_P, RAS, STCO\}$	毫秒 ~ 秒

τ - 缩放：普遍且可持续的法则

更多免费行业报告 并购家 www.ipoipo.net

这两大约束分别对应韬定律落地的设计与能效两大关键环节，也是配套技术体系需要重点突破的方向，并进一步形成以下两个核心研究方向：

方向一：逻辑折叠配套设计方法和工具

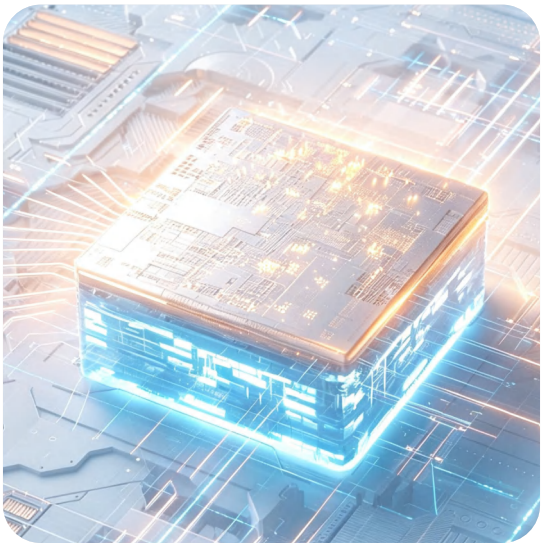
攻克垂直互连的物理约束。逻辑折叠将多层晶圆堆叠，垂直互连是层间唯一通道，也是寄生效应最集中的区域。垂直互连的寄生参数直接决定信号完整性和传输延迟，必须精确建模。更复杂的是 KOZ（禁止区域约束）——每层晶圆堆叠在相邻层产生禁止布线区域，随工艺偏差动态变化，需在布局阶段自动规避。晶圆间工艺偏差的相互耦合——阈值电压、驱动电流、互连 RC 参数——使传统假设参数均匀分布的模型不再适用。

应对晶圆间偏差的系统性放大。折叠后，不同晶圆间的偏差在垂直方向叠加——阈值电压偏移影响开关阈值，驱动电流偏差影响时序裕量，互连 RC 偏差影响信号完整性。解决这些问题，需要开发面向折叠的统计型 EDA 工具：在晶圆间偏差分布下做蒙特卡洛仿真，在 KOZ 约束下做自动布局，在垂直互连寄生模型下做时序闭环。这不是单点工具升级，而是从设计意图到物理实现的完整工具链重构。

方向二：多层级系统能耗优化技术

如果说逻辑折叠工具链解决“怎么设计”，那么多层级能耗优化解决“怎么运行更省电”。以下是可进一步探索的优化方向：

通过内存语义互连架构，消除协议栈开销。传统互连基于网络协议栈，数据逐层封装，每层增加头部和校验开销。在计算集群内部，协议开销已成能效隐形杀手。内存语义互连让远端内存像本地内存一样被直接访问，消除中间



协议层的封装、解析、拷贝开销——从“发消息”变为“读写内存”，互连延迟从微秒级压缩到百纳秒级，同时大幅降低每比特能耗。

使用近封装光学技术，达到每比特能耗降低数个数量级。电互连在板级和机架级已逼近极限，铜线 RC 延迟和高频损耗使每比特能耗难以继续下降。近封装光学将光收发器移至芯片封装附近。从电互连的皮焦耳级每比特向光互连的亚皮焦耳级演进，能耗降低数个数量级。当互连能耗在总功耗中占比从 30% 降到个位数，释放的电力可全部用于有效计算。

结合背面供电与存内 / 近存计算，实现从供电到计算的全面能效优化。背面供电将供电回路移至背面，消除正面供电网络损耗和散热压力。存内计算和近存计算从范式上消除数据搬移——让计算发生在数据所在处，化解“搬数据比算数据还耗电”的困境。

从方法到工具，从工具到产业，两个方向并非独立——逻辑折叠带来的更高集成度放大了功耗密度挑战，而多层级能耗优化又为更激进的折叠方案提供热管理空间。两者相互驱动，共同构成韬定律从理论走向工程的完整配套体系。

从驾驭环境到自主演进，Agent OS 构建智能体原生技术底座

AI Agent 正从概念验证走向规模化落地，但一个更根本的问题也随之出现：随着 Agent 数量增多、任务越来越复杂、消耗的 Token 越来越多，Agent 到底只是在被动适应现有的技术环境，还是有能力主动驾驭、甚至改变这个环境，这可能正是判断 Agent 能不能从“效率工具”升级为“系统级智能服务平台”的关键区别，Agent OS 将会在其中发挥关键作用。

业界对 Agent OS 尚未形成统一认知，这本身或许并不是一件坏事——它意味着这一底座的形态仍有巨大的探索空间。我们倾向于认为，Agent OS 不应被简单理解为传统操作系统上新增的一个 Agent 入口，而更可能是一种以 Agent 为核心，重新组织模型、工具、数据与业务系统的系统级底座。我们认为未来的框架：

Agent OS =（协同力 × 执行力 × 记忆力 × 连接力）^ 演进力

协同力、执行力、记忆力与连接力作为能力基座，共同界定了 Agent OS 在某一时刻“能做什么”；而演进力作为指数项，则试图回答一个更开放的问题——系统能否通过持续的反馈、学习与迭代，让既有能力不断放大，决定了前面的能力基座能产生多少复利。可信治理贯穿始终，也需要伴随这一演进过程动态调整，而非一次性划定边界。



更多免费行业报告 并购家 www.ipoipo.net

那么，走向“可持续演进的 Agent OS”，中间隔着哪些尚未解决的挑战？

- 多个 Agent 与人类专家共同参与任务时，角色边界、任务交接与结果验收该如何界定，才能避免重复执行与目标偏移？

- Agent 任务具有长上下文、多轮推理、频繁工具调用的负载特征，与传统单次推理的调度方式完全不同，如何面向智能体任务提升资源利用率、降低单位任务成本？

- 记忆能否成为系统级能力，而非上下文窗口的补充？或许是 Agent 能否“自主演进”而非“重复劳动”的分水岭，而这离不开权限、隐私与审计的平衡。

- 多 Agent 系统仍依赖预定义工作流，Agent 能否自主发现伙伴、动态组队并可信协作？

- Agent 系统的自我改进，能否摆脱对人工更新模型、Prompt 与工具的依赖，让成功与失败的经验自动转化为新能力？



Agent OS 框架设想

面对上述挑战，我们尝试提出五个可能的探索方向：

协同工程：从模型的 Scale Up 走向 Agent Scale Out

协同工程（Coordination Engineering）构建起大规模异构智能体的动态协商分工与分布式状态一致性保障；长链路任务下协作逻辑漂移与执行断点自愈；角色策略与团队协作范式的标准化及跨平台可移植复用；多 Agent 调度和多模型路由。最终构建从协作编排、资产共享、持续迭代到风险管控的完整工程体系，弥补现有单智能体框架在组织级自主协同上的能力短板。

Agent 原生运行时：重构任务执行效率

面向 Agent 负载重新设计“通算、智算、网络、存储”一体化资源调度与运行时体系。根据任务依赖、上下文长度、模型类型，以及 CPU、NPU、DPU 等异构资源状态进行统一编排，优化推理、工具调用与沙箱执行之间的协同链路，系统性降低任务时延与单位任务成本。

多层记忆范式：让经验成为可持续积累的资产

构建工作记忆、情景记忆、语义记忆、程序性记忆四层体系，具备主动编码、压缩、巩固、遗忘、更新与溯源能力，支持时序抽象与因果关联，能区分事实、偏好、失败经验与可复用策略。完善的记忆管理与持续沉淀机制，使任务经验从临时上下文升级为可继承、可复用、可审计的系统资产。

生态互联：统一交互、可信协作的协作规则

通过统一的交互、协作与可信机制，解决不同 Agent 之间“如何理解彼此、如何合作、如何建立信任”的根本问题。使跨平台、跨组织的 Agent 能够自主发现能力、动态建立协作关系，逐步从人工编排走向任务驱动的自主组织，形成类似“智能体市场经济”的开放生态。

递归自演进：让 Agent OS 具备持续成长能力

递归自我改进（RSI）不是让 AI 单纯把任务做得更好，而是让 AI 自主优化“改进自身的机制”，形成复利式正反馈闭环——每一代更强的系统，会变得更擅长制造下一代更强的系统。这是 Agent OS 区别于传统软件的本质特征：传统软件上线后趋于稳态，Agent OS 在部署后仍然持续自我递归成长。

最终，Agent OS 围绕智能任务，将重新构建协同、执行、记忆、连接与自主演进这五项原生能力。当这些能力逐步打通，Agent 将从一次性完成任务的智能应用，演进为能够长期协作、积累经验并持续成长的系统级智能生产力，成为 AI 进入千行百业的关键基础设施。

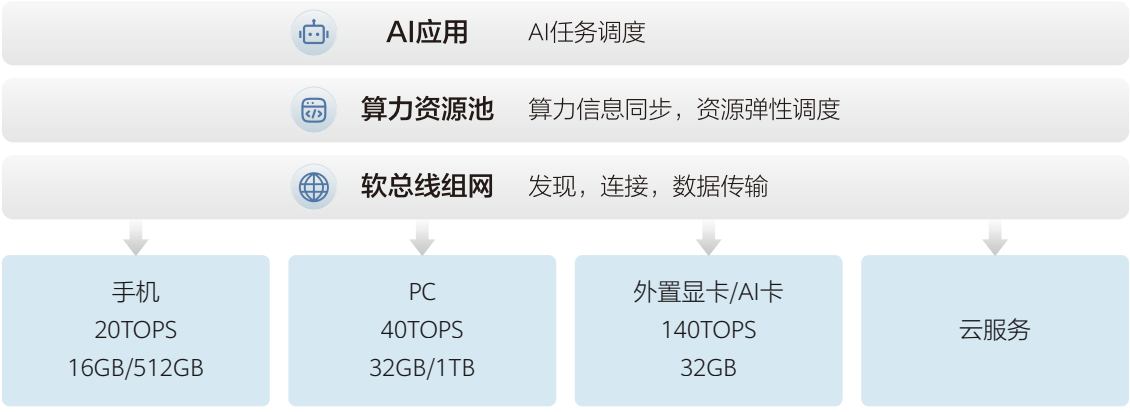


更多免费行业报告 并购家 www.ipoipo.net

定义无感适配、全域随行的终端智能新形态

在智能终端的演进过程中，物理硬件正逐步弱化其算力承载功能，轻量化为专注感知、交互与呈现的 Agent 交互载体。传统终端的人机交互以用户主动下达指令、设备被动执行为特征，本质上仍是工具化载体。随着智能服务向无感适配、人机共生演进，终端将升级为持续在线、深度协同的个人数字分身：它不再局限于被动响应任务，而是能够全程理解用户思维、校准服务意图，与用户协同完成复杂目标的拆解与任务规划，甚至在无人监管场景下，也能以高度贴合用户偏好的方式自主持续运行，最终形成无感适配、全域随行的新型人机服务形态。

要实现这一构想，几个关键挑战是不可绕开的。

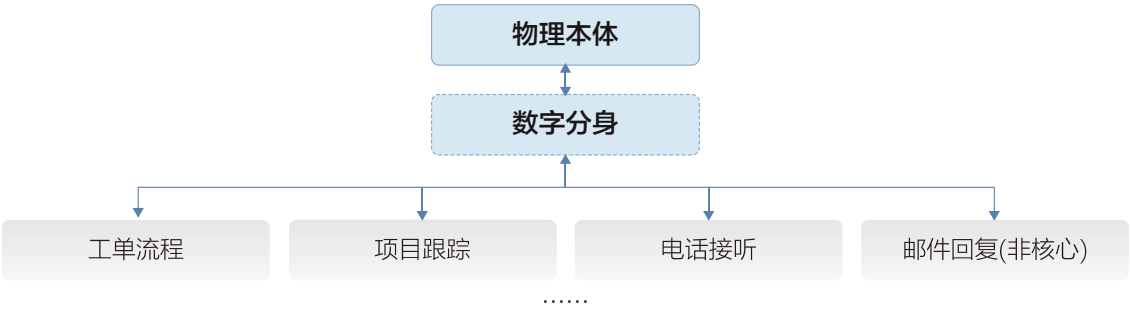


跨端异构算力融合的一体化计算架构

挑战一：跨端异构算力融合

传统智能终端的服务体验上限受限于单体设备的硬件算力配置。伴随终端设备持续向轻量化、便携化形态迭代，单设备难以承载高负载、高功耗的 AI 计算任务；而完全依赖云端算力的模式，又存在网络时延不可控、实时性不足、敏感数据上存带来的隐私泄露风险等突出问题。因此，算力范式亟需向终端、边缘、云端全域资源动态池化的方向升级，构建即插即用、按需调度、弹性扩展的泛在算力服务体系。

跨端异构算力的深度融合，面临技术与产业的双重瓶颈。技术层面，各类终端底层指令集、硬件架构差异显著，难以实现全域统一的算力抽象与标准化调度；网络节点动态拓扑更迭，无法保障计算任务的连续性与稳定性，同时无线信道的实时波动、带宽抖动，极易造成 AI 推理、实时控制等关键场景卡顿、失效，严重影响智能服务的可靠性。产业层面，头部厂商普遍以生态闭环、市场竞争为核心诉求，普遍采用私有通信与算力协议，缺乏开放通用算力接口的内生动力，导致跨品牌、跨设备的闲置算力资源无法实现近场共享与高效流转，制约全域泛在算力体系的落地普及。



人机共生的新型交互示意

挑战二：人机交互范式重构

从指令式交互迈向人机共生协同，范式重构面临三重瓶颈。

其一，用户个性化偏好、思维逻辑与决策习惯多为难以量化、文档化的隐性知识，现有静态记忆体系无法实现长效、精准的个性化建模。

其二，面对模糊化、跨域、未明确的高阶需求，系统缺乏人机协同研判、动态澄清意图的认知交互框架，难以将模糊诉求转化为可落地、可执行的精准任务方案。

其三，大模型生成的不确定性特征，在无人监督的自治运行场景中，极易引发决策偏差与不可逆的系统失效，制约无感共生交互的安全落地。

针对上述挑战，以下几个技术方向值得共同探索：

方向一：探索物理解耦、可信安全的弹性融合计算架构。

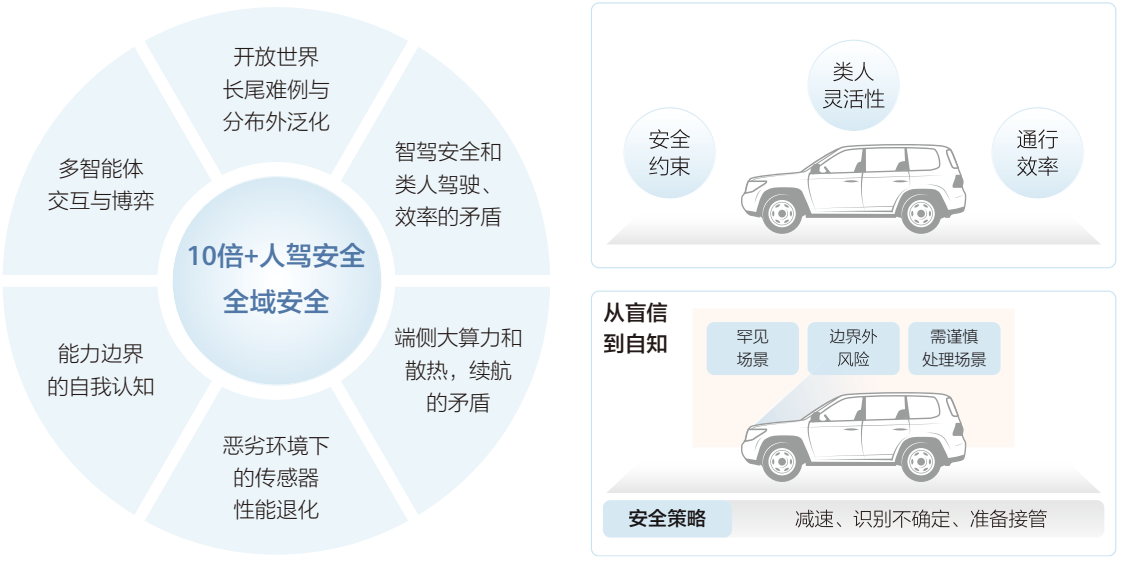
针对跨端异构算力碎片化、调度不稳定、数据不安全等问题，需构建物理解耦、逻辑统一的弹性融合计算架构：通过对跨厂商、跨架构、跨场景的异构终端进行标准化抽象封装，将分散的物理计算节点虚拟化为逻辑一致、可统一调度的全域算力资源池；同时配套隐私安全与可信验证机制，保障算力迁移与任务调度过程中的数据隔离、隐私保护与防篡改。该架构可动态适配节点上下线、网络抖动、设备异构迭代等复杂变化，实现智能任务的无感迁移与连续运行，从而保障全场景智能服务的稳定性与体验一致性。

方向二：构建可认知共生、可自治演化的全天候个性化数字分身。

面向传统指令式交互能力单一、意图理解层次浅、自治能力不足的短板，打造具备用户思维克隆、人机认知共生、持续演化能力的全天候数字分身体系。意图认知层面，突破显性指令局限，持续挖掘、建模并迭代用户的隐性偏好与决策习惯，形成个性化思维模式；协同交互层面，建立人机双向认知、动态校准的共生框架，支持 AI 与用户共同研判模糊目标，协同完成澄清、拆解与路径规划；安全自治层面，构建边界可控、行为可约束、过程可追溯的自治机制，确保数字分身在无人在环场景下稳定、高可靠地执行复杂任务，实现自主性与安全可控性的统一。

更多免费行业报告 并购家 www.ipoipo.net

实现智能驾驶的自我学习与极致安全



智驾的关键挑战

自动驾驶基于端到端网络架构已跨越 L2+ 级的技术鸿沟，正进入产业爆发期。行业开始追求更高级别的自动驾驶，目标是在大部分场景下不需要人类接管，并达到 10 倍于人类驾驶的安全水平，实现全域全场景极致安全。但这一目标能否真正落地，仍然面临一系列尚未解决的难题。

开放世界的长尾难例与泛化问题。真实道路是非结构化的开放场景，杂物、临时施工、极端交通行为等 corner case 几乎无法通过数据堆砌实现全覆盖。模仿学习在面对分布外场景时，容易产生感知虚检漏检和决策失误，全域安全从何而来？

多智能体交互博弈的不确定性。各类交通参与者的行为充满博弈与不确定——抢行、对峙、试探等复杂交互频发，现有系统在应对这些场景时可能出现决策偏差，安全和通行效率能否兼得？

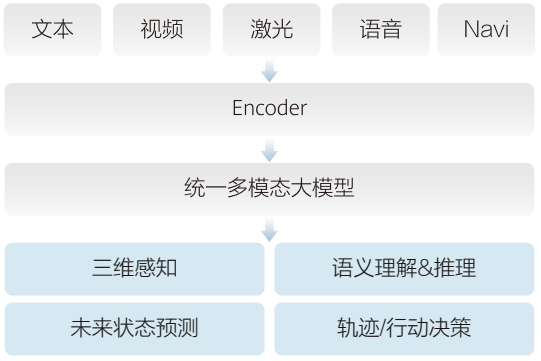
多重目标的内在矛盾。极致安全与类人驾乘体验、高效通行之间存在制衡：保守策略可能降低通行效率，激进策略又会压缩安全冗余。与此同时，端侧大算力带来的散热和功耗压力，也在影响车辆续航。

系统能力边界的认知缺失。系统是否具备“知道自己不知道”的能力？在雨雪雾、强光等恶劣环境下，传感器性能不可避免会衰减，传感器降级后的故障安全兜底机制是否足够可靠？

面对上述挑战，我们认为，智驾的安全与进化或许不能仅靠堆砌数据和算力来解决，而需要在模型架构、学习机制、感知体系和硬件协同等多个层面同步突破。可能的攻坚方向包括：

智驾原生基础模型

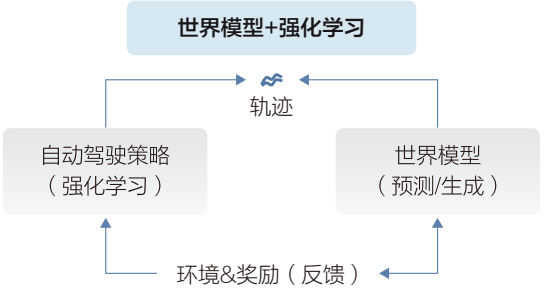
面向真实物理交互场景构建智驾专属基础模型，突破传统感知的认知局限，精准识别复杂交通交互状态，输出厘米级位置与速度信息。通过融合视觉、激光、导航等多模态数据，一体化实现语义理解、几何定位、物理推理与行为预测，依托大规模视频预训练、反事实增强训练植入因果常识，赋予模型强泛化、可外推的场景推理能力。



智驾原生基础模型

交互式世界模型

作为智驾想象力引擎，兼具数据生成、策略评估、实时预测功能，具备时空一致、物理可信、超实时推演、可控可交互四大特性，可保障长时序推演无偏差、场景生成符合物理规律，以超高速推演支撑海量算法评测，并可按需生成极端难例、模拟交通参与者动态交互行为。



交互式世界模型

安全可控的自进化机制

构建主动学习、持续迭代、经验沉淀的自主进化体系，主动聚焦高风险、边界场景定向优化，在吸纳新场景能力的同时留存成熟驾驶经验。通过长期记忆模块沉淀道路特征、交互经验与风险案例，同时设置刚性约束机制，保障模型迭代全程可控，严守安全底线。



安全可控的自进化机制

全自动化数据闭环

自动挖掘路测数据中的失效场景、隐性风险与长尾短板，依托世界模型扩充难例场景库，搭配标准化自动评测体系，精准量化模型性能缺陷，形成问题发现、数据扩充、模型训练、效果验证的全链路自动化迭代闭环，摆脱人工迭代依赖。

恶劣环境高鲁棒感知体系

依托 RGBD、涉水传感等新型硬件补齐极端场景感知短板，通过多模态算法融合实现传感器故障识别与自适应切换，搭配系统降级、安全驻车兜底策略。结合除霜、疏水、超声去水等硬件防护技术，大幅提升全天候工况下的感知稳定性与车辆可用率。

高效推理与软硬芯协同

通过稀疏注意力、MoE 架构降低模型推理开销，依托算子优化、图算融合、编译精简推理链路、降低时延。同时依托芯片协同设计，实现算力分片并行推理，平衡大模型算力需求与车载功耗、散热和续航。

更多免费行业报告 并购家 www.ipoipo.net



攻克供电与散热难题，支撑 AIDC 走向更大算力规模

大模型与 AI 智能体应用快速普及，推动训练与推理算力需求指数级攀升，传统中小型分散式数据中心已无法适配 AI 规模化落地，GW 级超大规模 AIDC 集约化建设成为产业发展必然趋势。面向 2035 年，超大 AIDC 将面临能源、基建、组网、芯片调度及智能算力适配等系统性技术难题，其中供电保障与散热控温是制约大规模 AIDC 落地的核心难题。

更大规模 AIDC 供电难题与突破路径

大规模 AIDC 迈向多园区集群、总容量突破 GW+ 后，供电挑战从单园区问题升级为区域级能源基础设施与算力负载强耦合的系统性难题，主要体现在四方面：

微小单级变换效率劣化会在多级变换与海量馈线上逐级叠加，单个百分点的效率下降直接对应数十 MW 的损耗。

高压直流短路 di/dt 极高、直流电弧无自然过零点特性，极易引发拉弧放电与跨园区故障扩散。

海量电力电子器件耦合放大扰动，引发同步 / 次同步谐振，传统小信号分析法难以适用于数十万节点规模。

算力、变流器、储能、电网分属独立控制域，跨域感知、亚微秒级同步、异构通信协议，亟需构建多目标联合决策模型实现供电全局最优。

应对上述供电难题，我们认为需要优先突破的方向是：

方向一：宽禁带化合物直变供电，搭建超低损耗新架构

摒弃传统多级降压架构，采用高压大容量宽禁带化合物器件，优化 AC/DC 直变拓扑，缩减工频变压层级，根本性抑制损耗累积，解决更大规模 AIDC 的 MW 级损耗问题。

方向二：高压直流器件，破解高压直流拉弧与短路扩散难题

针对高压直流拉弧、高 di/dt 短路及故障跨区扩散问题，集成电弧感知、限流隔离、混合



断路器硬件，实现故障快速预警与微秒级隔离，规避集群供电安全风险。

更大规模 AIDC 散热难题与突破路径

AIDC 迈向更大规模，传统液冷能力趋近上限，散热技术将迎来迭代升级，主要体现在如下三个方面：

随着 AIDC 向更大规模演进，单机柜功率密度持续飙升，传统液冷方案（如冷板式）的换热效率与热通量承载已接近物理极限，亟需散热底层架构的颠覆性升级。

先进制程芯片在峰值负载下易形成毫秒级、微米尺度的极端热点，传统散热手段无法精准响应，导致芯片性能受热节流制约。

当集群规模扩展至十万卡乃至百万卡级别，超节点功耗和发热总量急剧增长，不同层级（芯片、板卡、机柜、机房）的热特性差异显著，单一散热技术无法覆盖全链路需求。

应对上述散热难题，我们认为需要优先突破的方向是：

方向一：全液冷与芯片级直接接触散热

推行机柜高效冷板散热技术，将芯片产生的高密热流快速带走，探索浸没式冷却，实现极致无死角冷却与节能；同时发展芯片嵌入式微流道散热，使冷却介质直接接触硅晶圆，消除界面热阻，解决单芯片超高热流密度难题。

方向二：零水耗相变制冷与热电一体化余热回收

结合 AI 与数字孪生技术，毫秒级动态调度算力任务，消除局部热点并优化冷源效率；采用闭式干冷塔配合地热 / 深层海水及 CO_2 相变循环，实现零水耗冷却；探索热化学制冷等前沿技术，协同提升冷却与余热利用效能。

方向三：探索液氮低温散热前沿技术，持续突破超高算力集群的散热技术瓶颈

液氮低温散热利用其 -196℃极低沸点与高潜热特性，可获得远超传统液冷的热通量移除能力，是破解百万卡级超大规模集群散热困局的颠覆性路径。为此，需研发高效相变换热器、低热阻超低温冷板及智能液氮流量调控系统，构建稳定可控的液氮供冷链路。

更多免费行业报告 并购家 www.ipoipo.net

面向自主 Agent 的安全与隐私保护，为智能世界保驾护航

面向 2035 年，自主 Agent 将由辅助工具演进为具备感知、规划、推理和执行能力的数字行动者，持续参与生产、生活与治理。其规模化应用，有赖于一个模型可信、架构可控、隐私可保的安全底座。但 Agent 的底层机理与传统软件截然不同——它拥有开放的行为空间与不确定的演化路径，这使得既有安全范式可能不再适用，而新的防护体系尚未成熟。

从“传统安全”走向“面向自主 Agent 的内生安全”，存在一些关键的挑战：

模型是否真正可信？认知投毒可能通过全网多模态数据流植入系统性偏见，伪对齐与奖励黑客可能让模型在训练中表现完美却在部署时暴露破坏性，而持续学习带来的“泛化漂移”也可能使安全边界悄然偏移。

Agent 的行为是否始终可控？长周期、多目标的自动化执行中，环境反馈的随机性与模型涌现效应可能导致长程决策链失控，权限需求随任务意图动态变化，传统静态规则难以防止“合法越权”，多轮交互的不确定演化也使系统状态趋于黑盒化。

人的隐私边界是否依然清晰？当 Agent 跨场景调用工具、关联多模态数据时，传统脱敏与匿名化可能难以抵御高阶逆向工程，无处不在的感知使个体趋向透明，而“知情同意”在 AI 长程自主决策面前或许流于形式。



自主 Agent 安全与隐私保护的核心挑战

面对上述命题，我们倾向于认为，安全与隐私保护应当从“事后拦截”前移至模型生产的初始阶段，并贯穿 Agent 运行与演进全过程，可能的探索方向包括：

可证明的模型内生安全：将安全约束与伦理边界作为硬性数学约束项注入优化器目标函数，从数学上论证模型在不同优化路径下的安全性；同时探索神经元激活表征的实时侦测，以及解耦表征学习，使模型吸收知识而不继承潜在的毒化逻辑。

意图驱动的动态权限与端到端守护：根据任务意图、因果影响和环境状态按需授予权限，并在任务结束后及时收回；建立长周期任务的全流程监视与策略校验，防止目标漂移、提示注入和恶意劫持。

持久化状态的完整性与可追溯性：对上下文、长期记忆和向量数据实施抗污染、防篡改与可信校验，确保 AI 决策轨迹、执行上下文与系统状态改变可追溯且不可篡改。

隐私保护的新范式探索：构建人机双向隐私心智模型，让用户理解 Agent 的数据使用边界，也让 Agent 识别隐私敏感情境；结合零知识证明与同态加密，构建去中心化的“可信记忆空间”；并将“人类自主权不可侵犯”作为最高伦理红线，转化为不可违背的数学惩罚项，从底座机理上确保 AI 在任何优化路径下都不触及侵犯人类主体自主权的风险区域。

最终，自主 Agent 的安全与隐私或许不能仅靠外围防御来解决，而需要在模型训练、行为执行、记忆管理与隐私保护等多个层面同步构建内生机制。当这些能力逐步打通，Agent 才能在深度参与生产、生活与治理的同时，守住可信、可控、可保隐私的底线。但这些路径能否真正落地、如何在效率与安全之间取得平衡，仍有赖于持续的探索与验证。



更多免费行业报告 并购家 www.ipoipo.net





行动

面向 Agentic AI 时代
构建坚实的智能基础设施
重塑用户体验和生产模式

更多免费行业报告 并购家 www.ipoipo.net



Agentic AI 正在推动数字世界从“以应用为中心”走向“以智能体为中心”。Token 作为智能体感知、推理、决策与交互的重要载体，其规模化生成、实时流动和高效应用，使计算、云和网络等基础设施面临远高于传统应用的要求。面向 Agentic AI 时代，需要围绕 Token 构建“生产—运载—应用—防护”的全链路智能基础设施体系。

在 Token 生产侧，算力需要支撑更长的推理链条、更复杂的模型协同和更高频的工具调用，实现智能能力的高效生成；在 Token 运载侧，通信网络需要具备低时延、高并发、高可靠的能力，保障 Token 在云、边、端之间高效流动，以及智能体与人、设备和物理世界的实时交互；在 Token 应用侧，云基础设施需要承载智能体的开发、运行、协同与状态管理，使模型、数据和工具围绕任务高效组织；同时，通过原生安全的能力，实现 Token 全生命周期的可信防护。

最终推动算力、数据、模型、智能体、云和网络深度融合，从底层重构智能生产与应用体系，全面释放 AI 驱动的智能生产力，推动基础设施能力向确定性用户体验和智能、高效的企业生产模式加速转化。

计算系统与软件: 超节点架构创新与开放的软件生态，实现 Token 的高效生产

超大参数模型、超长上下文和极致推理时延等新型负载，对算力底座提出了更高的要求。超节点通过内存统一编址、超低时延和超大带宽，可实现计算、网络和存储资源的系统级协同，成为下一代 AI 基础设施的重要演进方向。

算力底座之上，基础软件是连接芯片能力与 AI 应用的纽带。软件成熟度，既决定生态创新的速度，也决定系统性能优化的效率。面向智能体与新编程范式，需要一套分层解耦、开放协同的软件体系，推动芯片、模型、框架、工具链与应用深度融合，加速技术创新与产业价值落地。



开创超节点架构，满足算力持续增长及负载多样化的需求

人工智能技术快速发展，算力基础设施的应用场景不断延伸，涵盖 MoE 大模型训练、大规模专家并行推理、慢思考长链推理，并深度赋能生成式推荐与自主智能体等。训练侧，集群规模越大，通信耗时占比越高。以某万卡集群 175B 稠密模型为例，单步计算约 10 秒，通信耗时却达 10 至 90 秒。传统集群采用以太网互联，跨节点带宽仅为节点内的 1/15 左右，已成为训练效率的核心瓶颈。推理侧，目前主流的大模型提供 10ms 级的 TPOT 时延，且在追求更低的时延，以提供更好的用户体验。传统以太网络不仅带宽低，通信时延比总线技术高出 10 倍以上，无法满足低时延推理需求。

面对这些趋势，算力竞争的基本单元也在发生深刻变化——从单颗芯片、单台服务器，逐步走向整机柜、超节点乃至数据中心级计算系统。当模型参数达到十万亿、集群规模达到

更多免费行业报告 并购家 www.ipoipo.net

十万卡，系统瓶颈从“计算”走向“计算 + 通信”。未来算力的竞争，并不只是单颗芯片峰值性能的比拼，更是将大量计算、存储与网络资源高效组织起来的系统能力之争。“并行”与“批处理”是提升大模型计算性能与效率的两大支柱：更大的并行规模，是支持更大模型、实现更低时延、缓解单卡内存压力的基础，也为更大的批处理规模创造条件；而更大的批处理规模，直接决定了 GPU/NPU 等算力单元的利用率。

超节点通过架构创新释放系统潜力，以系统级扩展弥补单点能力的不足，实现极致并行规模，同时攻克“超大参数规模、超长序列、超低时延”等多重挑战。

围绕超节点定义和特征，业界已形成共识。在 2026 年发布的《超节点定义与实践白皮书》[1]，已给出明确定义：超节点是一种在物理上由多个计算节点通过高效的互联协议紧密连接组成，具备跨物理节点的统一内存编址能力，逻辑上具备“一台计算机”特征的计算系统。

具备统一内存编址、超低时延、超大带宽的超节点成为 AI 基础设施建设新范式

基于跨物理节点统一内存编址，超节点有三大核心技术特征：“内存语义”和“统一内存编址”、超低时延、超大带宽。这些特征为 AI 大模型等“单体紧耦合”软件构建了统一地址规划与数据布局的计算环境，使所有处理单元高效协同，为大规模、高性能计算场景提供近乎线性的算力扩展能力。

技术特征一：基于“内存语义”和“统一内存编址”实现跨物理节点的内存访问

现代处理器（CPU、NPU、GPU）均遵循冯 • 诺依曼计算架构，其指令集主要分为四类：数据存取、数据计算、程序控制及系统管理。其中，数据存取指令承载“内存语义”操作，核心指令为 Load 和 Store，由处理器 Core 的硬件单元直接发起对指定地址的读写，是指令集与微架构的固有组成部分。在编译器配合下，软件程序的数据访问模式被映射为这些内存语义指令。

超节点通过系统总线拓展内存语义操作的可达域，使得任一处理器 Core 都能直接对超节点内的内存地址执行 Load/Store 操作。这一扩展伴随统一内存地址空间的建立，从而在编程模型和执行模型上获得与单系统一致的语义。

在这一结构下，跨处理器 Core 的数据同步在机制层面与芯片内部一致，均表现为内存语义访存。这种统一性使系统能充分利用处理器微架构中的预取、乱序执行、多硬件线程等机制，由硬件、编译器和算法协同完成访存调度与同步隐藏，无需在软件层引入额外的格式编码或协议栈。

[1] 2026 年鹏程实验室（PCL）与全球计算联盟（GCC）联合发布《超节点定义与实践白皮书》

对于不支持内存语义的节点间互联，数据交换必须依赖软件对信息进行 TLV、KV 等格式封装，再通过通信协议栈收发。这条路径无法融入处理器的微架构优化，有效带宽和延迟均受限于软件开销，也无法利用硬件自动预取和乱序执行来掩盖通信延迟。

统一内存编址是实现跨节点内存语义访问的前提——只有将不同节点的内存被纳入统一编址，Load/Store 等内存语义操作才能够访问远端内存地址。因此，统一内存编址与内存语义访问构成超节点的核心架构特征。缺失这些特征，即便互联链路提供高带宽和低时延，系统在并行计算中的性能与时延收益仍将显著受限，且这一制约随单芯片性能提升而加剧——处理器 Core 执行越快，软件通信路径的相对开销越突出。

技术特征二：超低时延

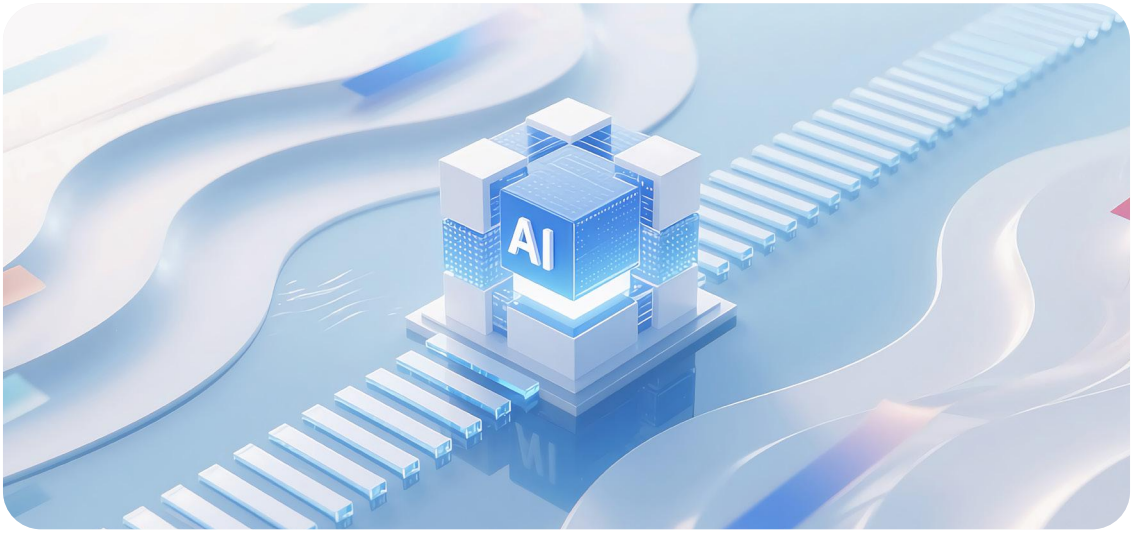
MoE 架构已成为大模型发展的共识，且大模型普遍采用同步并行的计算模式，节点间通信同步时延如果不能被计算时长掩盖，计算单元就会阻塞等待，直接拉低算力效率。同时，大模型一次完整的前向计算要经过几十层的神经网络，每层多次节点间数据同步，每次同步包含多次的数据发送、接收确认和同步控制。微小的总线 RTT 时延增加，都会被放大成百上千倍，最终影响大模型的计算效率。

因此，在具备“统一内存编址”和“内存语义访问”核心架构后，超低时延总线也是超节点的关键特征。通过超节点的微秒级 RTT 通信时延，可大幅提升大模型训推效率。

技术特征三：超大带宽

AI 大模型的激活值是成千上万个隐藏维的高维向量，数据量被放大成千上万倍；加之数十至上百层神经网络，一次完整的前向计算的整体数据负载可增加数十万乃至上百万倍。这给多节点并行计算带来了巨大的同步数据开销。

因此，超节点总线必须具备 TB/s 级高带宽，并能随算力增长持续提升，才能承载海量的通信流量，确保超节点在运行大模型时的性能优势。否则，跨节点通信开销将严重抵消并行加速收益，实际性能提升大幅受限。



更多免费行业报告 并购家 www.ipoipo.net

超节点已成为万亿大模型训练和推理的必备

超节点通过高速互联将多台物理服务器聚合为一个逻辑上的“超大单一节点”，构建统一的算力与内存池，大幅压低跨节点通信时延，显著提升模型算力利用率。它既为万亿级大模型的预训练与后训练提供了高效稳定的算力保障，也为超长上下文推理、多智能体并发等低时延推理场景奠定底座。

大模型训练场景：模型训练时，激活值和优化器状态消耗海量显存，对传统服务器挑战巨大。通过超节点的统一内存编址，长序列（如128K/256K）训练时的全局 Batch Size 可大幅提升，无需频繁进行跨节点的 Checkpoint 保存和参数重分布，算力利用率显著提高。同时，超节点规模越大，MFU 提升越明显。

训练加速场景：大模型训练的难点之一是实现 ALL2ALL 通信的计算通信掩盖。张量并行（TP）和 ALL2ALL 对通信带宽和延迟极度敏感，传统服务器 RoCE 带宽不足，ALL2ALL 通信无法被掩盖，导致严重的通信气泡和长尾延迟。以昇腾超节点为例，超大带宽使 ALL2ALL 性能相比 RoCE 提升 16 倍以上，实现 ALL2ALL 通信掩盖，彻底打破“通信墙”。

强化学习（RL）后训练场景：RLHF/GRPO 后训练，推理生成与训练更新通常分离。传统框架采用僵化的“批处理”同步等待，且变长序列需大量填充，算力浪费严重。TransferQueue 将数据流改造为细粒度的“流处理”。结合超节点架构，推理生成的 Token 流无需经过外部以太网，直接通过超节点内部的统一内存编址和超大带宽，以纳秒级延迟“流式”注入训练更新的 TransferQueue 中，彻底打破训推间的同步阻塞与网络拥塞。

低时延推理场景：Decode 阶段具备典型的“访存密集型”的特征，瓶颈在于读取权重和 KV Cache 的速度。超节点 TB 级的超大带宽配合百 TB 级的统一内存编址，允许单节点加载千亿参数模型的全部权重，并驻留超大规模的并发 KV Cache。Decode 请求在节点内直接命中内存，避免了传统分布式推理中因 KV Cache 跨节点迁移带来的巨大网络延迟，实现推理性能的倍级提升。

大 EP 推理场景：大模型推理采用 Prefill 和 Decode 分离部署，P/D 分离架构下的资源利用率达到极致。相较于 P/D 同节点部署，吞吐性能提升 1.4 倍。集合通信（小包场景）对时延极度敏感，超节点超大带宽、超低时延的特征天然亲和小包场景，进一步消除了“Prefill 卡等 Decode 卡”的算力闲置。

面对 MoE 模型对通信、显存的严苛诉求，超节点不再简单叠加算力芯片，而是以系统级协同提升集群算力利用率。训练上保障大规模并行的效率与稳定性，推理上支撑复杂业务的超大并发与超低时延需求，为大模型规模化产业落地筑牢算力根基。



构建开放易用的计算基础软件生态，加速算力释放与应用落地

计算生态的建设是一项涵盖芯片架构、软件栈与工程能力的系统性工作。当前，全球算力规模正朝着 2035 年 10 万倍的增长目标迈进。在硬件持续迭代的背景下，无论是超节点架构对极致规模的探索，还是芯片、互联与存储等技术的协同演进，都在不断拓宽算力边界；而软件栈的完整性与易用性，正日益成为释放硬件效能的关键变量。AI 算力的规模化部署与应用场景的持续分化，对开发效率与适配灵活性提出了前所未有的要求。

构建开放、易用的计算基础软件生态，正是应对这一挑战的关键所在。唯有通过标准化的接口、完备的编程范式与开源协作，才能降低迁移成本、提升开发效率、加速创新迭代，真正打通从芯片算力到业务价值的最后一公里。

开源开放的计算基础软件加速 AI 创新

计算产业的演进长期围绕算力与模型两大核心要素展开，算力决定了规模上限，模型决定了智能上限。但硬件算力的快速增长与软件释放效率之间的落差日益凸显，“木桶效应”正成为制约算力转化为生产力的关键瓶颈。

在传统封闭生态下，大量基于特定厂商编程模型编写的科学计算与 AI 代码面临着“生态异构”的困境，不同芯片的计算单元架构各异、编程模型彼此割裂、算子库重复建设。硬件纸面性能峰值与用户实际可获取的有效算力之间存在显著鸿沟，大量算力在数据搬运、调度等待和低效算子中空转。与此同时，业务场景从预训练集群追求峰值吞吐，向推理场景中对内存带宽与低时延的苛刻要求分化，统一僵化的闭源软件栈难以应对碎片化的真实世界需求。

更多免费行业报告 并购家 www.ipoipo.net

由此，开发者的诉求日益清晰。一方面，开源模型能力持续提升，逐步逼近头部闭源模型的 SOTA（State-of-the-Art，当前最优水平），叠加其成本优势，已成为行业主流选择。模型的快速迭代要求软件栈能够随业务演进灵活调整，传统闭源体系以季度为发布周期，无法匹配“日更、周更”的开发节奏。另一方面，面对性能瓶颈或精度异常，开发者需要深入底层硬件细节进行定位与调优，而非被动等待厂商响应。开源正是回应这些诉求的必然选择——它赋予开发者代码的可读性、可修改性与可追溯性，将创新的主动权真正交给开发者。

计算基础软件生态，正是连接芯片算力与业务应用的“神经系统”。向下使能硬件，向上赋能应用，其优劣直接决定了纸面算力有多少能转化为真实生产力。唯有构建开放、易用的计算基础软件生态，让接口标准化、能力透明化，才能打通从芯片算力到实际业务价值的最后一公里。

开源开放的计算基础软件架构全景



开源已从早期的技术理想演变为产业共识，成为计算基础软件的主流形态。当前，从上层训练推理框架、中间 AI 框架、基础算子库到底层编程语言、运行时与驱动，开源项目已全面覆盖计算基础软件架构的各个层级。

各层级开源生态的具体特征体现在以下方面：

训练推理框架层，开源社区主导了技术路线与迭代节奏。Megatron、vLLM、SGLang 等项目通过开放合作，聚合全球开发者的算法与优化经验，持续拓展训练与推理场景的性能与功能边界，形成自加速的创新循环。开发者可直接基于主流开源框架高效完成模型训练、推理部署及二次开发。

AI 框架层，开源生态确立了事实标准并主导了技术演进方向。PyTorch 等 AI 框架通过开源社区汇聚全球贡献，持续迭代分布式训练与

并行计算加速等核心能力。开发者基于开源 AI 框架进行模型开发与训练，可充分利用社区生态资源，实现跨平台迁移与灵活部署。

算子库与加速库层，开源社区预置了丰富的核心算子实现，开发者无需从零编码，可直接调用预置 API 完成计算逻辑，聚焦于业务专属算子的定制与优化。开源算子库的持续积累，为上层训练推理框架提供了高性能的计算基座。

编程语言与编译器层，开源生态降低了异构硬件的开发门槛。Triton 等开源编程语言通过高层抽象屏蔽底层硬件差异，消弭异构硬件的编程鸿沟，使开发者能够以接近 Python 的语法编写高性能算子，由编译器自动完成硬件映射与优化，大幅提升算子开发效率。

基础使能层，开源驱动与运行时提供了设备管理、内存管理及任务编排等核心能力，使开发者能够深入理解硬件调度逻辑，进行底层性能调优与问题定位。

以上各层的开源开放，共同推动了产业协作模式的深刻转变，开发者从被动的使用者转变为主动的贡献者，生态从封闭的单点创新走向多方协同。

以昇腾为例，CANN（Compute Architecture for Neural Networks）作为异构计算架构的参考实践，通过分层解耦与全面开源开放，使能上层框架与底层硬件高效协同。开源内容覆盖算子库、通信库、领域加速库、图引擎等加速库，算子编程与毕昇编译器抽象能力，运行时与驱动等资源管理组件。此外，Mind 系列开源全场景 AI 框架昇思 MindSpore、训练加速库 MindSpeed、推理引擎 MindIE 及全流程开发工具链 MindStudio 等，覆盖从模型开发、训练到推理部署的完整流程。二者协同，共同构成从硬件使能到应用开发的开源基础软件栈，为 AI 产业的持续发展提供坚实的创新底座。

计算基础软件生态的三大核心演进趋势

完备的基础软件栈架构是起点，真正释放算力潜能的关键在于让开发者好用、易用。实现“开放易用”，不仅是降低开发门槛的手段，更是让算力从专业工具走向普惠基础设施的核心支撑。围绕这一目标，计算基础软件生态正经历从底层编程范式到上层开发模式的深刻重构，具体表现为三大趋势：

趋势一：软件栈分层解耦与开源开放，共建开源技术生态

开源已是计算基础软件的既定方向，但演进并未止步。当前产业正从“是否开源”迈向“如何更好地开源”的深水区，通过分层解耦与社区共治，使开源生态从代码的简单开放走向可持续的创新协作体系。

在这一阶段，开源的价值不再局限于代码透明，而在于构建开放并长期稳定的软件底座，降低适配成本，让开发者得以聚焦于业务本身。同时，开源社区的共治模式使得技术路线的选择不再由单一厂商主导，而是由多方参与者共同决策，从根本上避免了产业碎片化。

更深层次的变革在于创新模式的转变。开源不仅赋予 AI 厂商深度定制与性能调优的能力，也使高校及科研机构能够基于开源代码开展底层技术创新。这些贡献通过社区回馈生态，

更多免费行业报告 并购家 www.ipoipo.net

形成“使用—贡献—再创新”的正向循环。开放的本质，是汇聚群体智慧、释放生态合力，以协作替代封闭，加速创新迭代。

趋势二：硬件使能走向 Tile 级抽象，编程门槛大幅降低

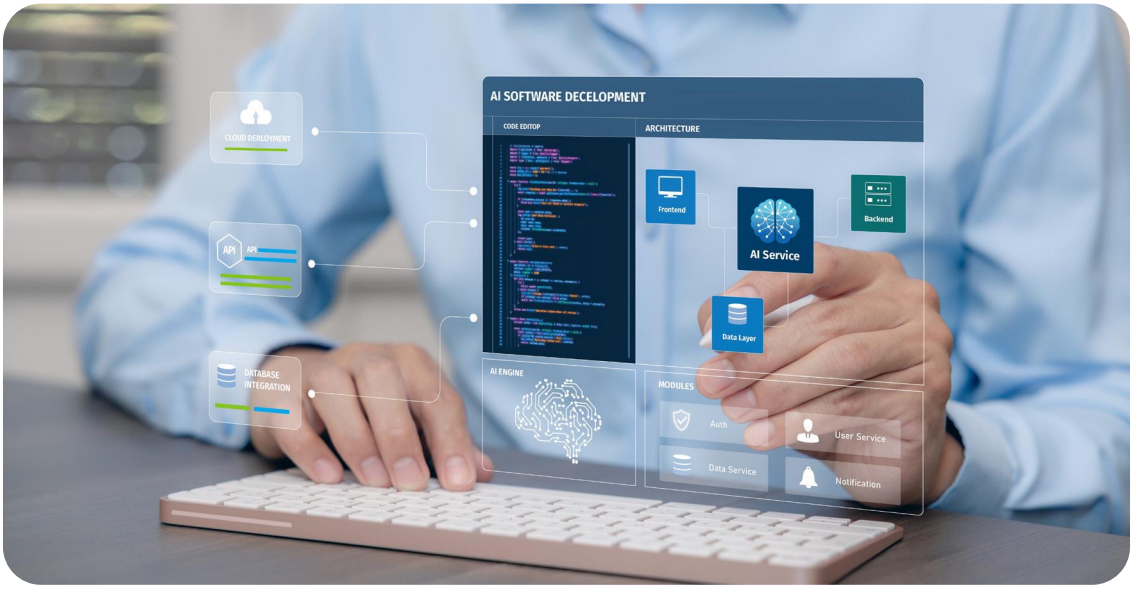
传统算子开发需开发者深度掌握硬件架构，手动管理片上内存、寄存器分配及同步流水调度，高度依赖人工经验。这一模式虽可产出高性能算子，却抬高了开发门槛，制约了开发效率与人才供给。

当前，基于 Tile（块）的编程范式正成为行业主流。以 Triton、TileLang 为代表的新一代编程语言，通过 Python 前端语法封装底层硬件差异，使开发者能够以类数学公式的形式描述分块计算逻辑，由编译器自动完成循环展开、内存调度与异步流水线优化。开发者的关注焦点从“如何调度硬件”转向“如何设计计算”，在保持性能竞争力的同时显著降低开发门槛。Tile 级抽象已在 DeepSeek-V4 等主流模型中验证了其可行性，有效打破了编程模型与特定硬件的强绑定。

趋势三：Agent 赋能算子开发，从人工编码到人机协同

算法创新是推动人工智能持续演进的核心驱动力。新算法持续涌现，每一篇论文的创新都对应着算子层的实现与适配——将数学公式转化为硬件可执行的代码，使理论创新真正落地于实际计算平台。算子开发由此成为一项长期且持续的工作。

这一模式高度依赖人工编码，曾是 AI 落地流程中最耗时的人力瓶颈。如今，Agent 辅助代码生成正成为提升开发效率的重要方向。已有计算平台支持基于 Agent Skills 的算子自动生成能力，通过可复用的工作流覆盖算子开发的全流程，形成自动化迭代闭环。开发者仅需下达自然语言指令，Agent 即可调用 Skills 能力，



自动完成代码生成、编译与调优。安装方式上支持插件化集成，可适配 Open Code、Claude Code、TRAE 等本地已有工具，实现快速部署。

与此同时，一个自然的问题随之产生：Agent 能力日益增强，是否还有必要进行手写算子开发？答案是肯定的。Agent 虽能不间断地调用工具和工作流，但其底层仍依赖已有的算子库代码、API 文档及调试调优的经验沉淀——这些知识资产均由开发者创造和积累。现阶段，Agent 尚不具备自主发明全新调优策略的能力，其生成的代码质量与性能优化水平，本质上仍受限于既有经验库的丰富程度。因此，持续拓展算子生态库、覆盖更多行业场景，不仅服务于当前业务需求，也为 Agent 提供更充足的学习素材，形成“生态越丰富、Agent 越智能”的正向循环。人与 Agent 协同，仍是当前及未来相当长时期内算子开发的基本范式。

开源共建的昇腾计算基础软件生态

昇腾坚持开源开放战略，全面开源 CANN 驱动、运行时、算子编程、算子库、通信库及工具等全量代码，使能客户、伙伴及开发者自主创新，充分释放昇腾硬件性能。在开源社区适配层面，兼容 Megatron、vLLM、SGLang、PyTorch、Triton 等 90 余个业界主流开源项目，匹配开发者使用习惯，有效降低迁移与开发成本。在应用使能层面，构筑昇腾亲和的 Mind 系列应用使能套件并全量开源，为业界提供从模型训练到推理部署的全流程参考实践。同时，昇腾持续将 Mind 系列中经过验证的昇腾亲和特性回馈至上游开源社区，推动三方开源框架原生支持昇腾硬件。

生态规模持续壮大，产业合力加速形成。基于开源开放的计算基础软件生态，昇腾已携手超过 3000 家行业合作伙伴，累计孵化超过 7000 个行业解决方案，广泛覆盖运营商、制造、金融、交通等关键领域。软件能力的开箱即用与持续迭代，显著提升了新模型与新架构的适配效率——DeepSeek、Kimi、GLM 等多款主流模型均在发布后快速上线，为千行百业的智能化转型提供了坚实基座。

社区影响力持续深化，逐步走向原生与共治。昇腾深度融入全球开源体系，与主流开源项目建立紧密合作：Triton Ascend 成为 Triton 官方社区在中国首个支持的加速后端，TorchNPU 已正式上线 PyTorch 官网，成为首个获得 PyTorch 官方支持的中国芯片硬件，亦是首个拥有官方分支仓下载支持的加速后端。从适配到原生、从跟随到共治，昇腾计算基础软件生态正逐步嵌入全球 AI 软件栈的主流技术生态，为开发者提供更顺畅、更原生的使用体验。

开源开放的创新模式，赋予了昇腾生态持续进化的活力。组件化的架构设计使开发者能够贴合自身业务需求，对特定模块进行自主深度调优，不再受限于封闭系统的一体化发布节奏。开放完备的编程体系有效降低了高性能算子的开发门槛，典型融合算子的开发周期可缩短至两周以内；在 Agent 辅助开发的加持下，算法验证更可达到天级完成，显著加速了从算法创新到业务落地的转化进程。从硬件使能到应用开发，从开源社区融入到产业生态共建，昇腾正逐步构建起开放活跃、具备长期竞争力的 AI 计算生态。

更多免费行业报告 并购家 www.ipoipo.net

从存数据到存知识 / 记忆，释放数据价值

数据的规模与质量决定了 AI 智能的高度。千行百业在数智化转型中，无论是基础大模型的训练与微调，还是特定场景的推理应用，都依赖大规模高质量数据的持续供给。随着 AI 从“会聊天”走向“能办事”，深入企业核心生产场景，企业还必须构建具备长期演进能力的知识库与记忆库，让 AI 不仅“知道”，更能“记住”与“思考”。数据存储将从“存数据”加速向“存知识 / 记忆”迈进，全面释放数据价值。



建设 AI 数据湖，存好、管好数据，实现高质量数据汇聚与供给

企业想让 AI 落地，让 AI 算力真正释放潜能，首先需要跨越“数据量大、数据管理难、数据检索慢”的鸿沟。以某车企为例，为实现 2030 年“取消方向盘”的完全自动驾驶，需汇聚传感器、摄像头等多源数据，总量超 1000PB，存储成本高昂；同时，这些数据分散在不同数据中心，难以统一整合与集中调度；更关键的是，高阶训练高度依赖暴雨、突发障碍物等极端场景，而从海量数据中精准定位并快速提取这类关键样本，无异于大海捞针。

为破局上述挑战，建设 AI 数据湖已成为行业共识。作为集“存储、管理、处理”于一体的新型数据集约化平台，AI 数据湖深度融入 AI 全生命周期，具备三大核心特征：EB 级大容量、全局可视可管、超高速检索。在实践中，AI 数据湖正为企业带来三大收益：

释放存储成本压力，打造极致 TCO：通过高密度硬件布局与极致的数据重删压缩算法，AI 数据湖可显著降低单位 TB 存储的功耗与机

架占用。在应对 EB 级数据爆发式增长时，实现总体拥有成本（TCO）降低 30% 以上，让海量数据“存得起、用得好”。

全局数据可视可管，提升数据利用效率：基于统一元数据视图，数据科学家与工程人员无需再耗费 70% 的时间在“找数据、洗数据”等繁琐环节。全域数据资产一目了然，结合精细化权限控制与全链路合规审计，让数据流转效率提升数倍，激活沉睡资产。

千亿数据秒级检索，加速 AI 模型创新：通过将海量文件检索时间从“分钟级”缩短至“秒级”，彻底压垮数据供给的时间瓶颈。无论是自动驾驶场景下的难例挖掘（Hard Example Mining），还是大模型预训练中的海量语料筛选，均可实现“随需即取”。这不仅是效率的提升，更是创新速度的质变——帮助企业实现更快的试错、更敏捷的调优，最终将 AI 算力潜能充分释放为业务增长的强劲动力。

AI 数据湖正在从“打破数据孤岛，实现海量数据的高效存储与管理”，走向“打通语义鸿沟，实现多模态数据的智能理解与高效检索”；未来将迈向知识湖——基于企业级本体与知识图谱，将碎片化数据提炼为可理解、可联想的上下文网络，驱动实时智能决策。通过与智能体生态无缝对接，知识湖能够实时感知业务状态、推演决策路径，完成从“被动响应查询”到“主动协同决策”的跨越，真正让 AI 成为企业运行的指挥中枢。

建设知识与记忆平台，让检索更精准、让推理更高效、让模型越用越聪明

在 AI 推理过程中普遍面临三大痛点：模型容易产生幻觉、长序列推理响应缓慢、多轮交互缺少持续记忆，导致大模型能力难以转化为企业的业务价值。为解决上述挑战，业界领先厂商开始尝试两类解决办法，一是通过部署上下文记忆存储，解决推理过程中 KV Cache（键值缓存）爆炸增长，导致 HBM、DRAM 空间不够，推理效率低的问题；二是建设 AI 数据平台帮助企业的数据治理智能化，将传统数据转化为 AI 可识别的知识和记忆，解决算力很强，但数据拖后腿的问题。

上下文记忆存储：提供 PB 级 KV Cache 存储能力，正成为超节点集群的标配

多轮推理会产生海量 KV Cache 数据，传统架构仅依靠 HBM 显存、DRAM 内存承载，容量极其有限，极易出现缓存溢出，导致推理中断、反复重算等问题。新的上下文记忆存储，专门补齐显存与 DRAM 在推理场景下内存容量不够，导致推理缓慢的问题，核心解决大模型 KV Cache 扩容难题。作为 HBM 与 DRAM 能力的外延层级，上下文记忆存储必须满足超高带宽、极低时延的严苛需求。因此上下文记忆存储配套 SSD 需要支持 GPU/NPU 直通，实现数据直接传输，消除主机 CPU 中转造成的性能损耗。与此同时，上下文记忆存储持续进行 KV 块写入、淘汰、覆写循环，叠加随机 IO 引发较高写放大效应，对闪存耐久能力提出高要求，生产稳态部署下 SSD 需要满足更高 DWPD（Drive Writes Per Day，每日全盘写入次数）。在互联网、大模型厂商、运营商等建设超节点集群的实践中，上下文记忆存储正在成为超节点集群标配，为客户带来两大价值：

算力效能最大化与 TCO 优化：打破 KV

更多免费行业报告 并购家 www.ipoipo.net

Cache“困在单机”的宿命，实现跨节点 / 跨实例的缓存共享与免重复 Prefill 计算，大幅减少 GPU/NPU 空转等待，降低推理综合算力成本。

极致长文本体验与集群弹性：实现海量历史上下文的秒级调取，显著降低多轮 Agent 推理时延；同时实现计算与缓存解耦，保证集群在动态扩缩容时缓存命中率不下降，赋予业务极高的敏捷度。

AI 数据平台：知识库、记忆库、KV Cache 库“3 合一”，提升推理效率与体验

AI 正加速从通用对话向金融、医疗、制造等行业生产环境纵深落地，在 AI 编程、高精知识检索、智能查询、商业数据洞察等典型应用中，通过部署 AI 数据平台可有效提升企业 AI 推理体验与效率。

知识库提供高精度多模知识，让检索更精准

在企业级 AI 应用中，检索的精准度直接决定了业务系统的可用性。以智能查询为例，客户日常需要处理海量信息，涵盖产品手册、行业白皮书、政策文件等。传统关键词检索只能匹配文档标题或摘要，无法穿透图片和表格提取有效信息，这导致检索结果往往“答非所问”。

领先厂商构建的知识库能力，通过多模无损解析、Token 级编码等方式，将非结构化的文本、图片、视频等资源转化成海量细粒度的知识单元，结合多维度语义检索与比对，能够实现对用户意图的精准理解，为客户提供“有据可依、拒绝幻觉”的高精准检索查询体验。

记忆库提供上下文记忆管理，让模型越用越聪明

大模型在处理复杂商业逻辑时，往往面临“转瞬即忘”的困境。以商业数据洞察为例，企业每月需处理销售数据、促销效果反馈、产品库存周转率等大量业务数据。这就要求模型的推理不能每次都从零开始，而是需要记住过去半年甚至一年的营销成败经验，并能够在新策略生成时主动调用这些历史记忆进行对比与校准。

领先厂商构建的记忆库能力，通过上下文状态保持、上下文凝练等技术，可精准萃取历史数据与经验，并将其沉淀成可召回的记忆。模型记得更多，推理就更精准，进而实现“越用越聪明”的持续进化，让商业洞察从“事后复盘”真正走向“智能决策”。

PB 级 KV Cache 容纳海量历史数据，让推理更高效

长文本、长序列、多用户多并发的推理已成为典型负载特征。随着输入序列增长，KV Cache 数据量显著上升，其容量通常随上下文长度近似线性增长，并同时受模型层数、批处理规模、KV 头数量和数据精度等因素影响。多项变量叠加后，长上下文与高并发场景下的缓存总量将快速膨胀，对存储容量、访问带宽和调度效率提出更高要求。如果基础设施无法支撑大规模的缓存存储，系统响应就会变得迟缓，导致推理体验恶化。

通过领先厂商构建的 PB 级 KV Cache 扩展能力，系统能够完整保留大规模、多轮次的缓存数据。在新请求到达时，系统可以直接复用历史缓存，大幅减少推理过程中的重复计算，显著降低推理时延，提升推理吞吐与用户体验，为长序列、复杂逻辑的智能体推理提供关键性能支撑。

建 AI 时代全场景数据保护能力，守护企业核心数据资产

AI 智能体作为企业新型数字员工，在大幅释放生产潜能的同时，也带来了全新数据安全风险，比如 Agent 自主执行高危删除、越权访问、敏感数据泄露，甚至沦为勒索病毒与恶意攻击者的潜在载体与联动工具。

面对 AI 时代愈发复杂的风险态势，企业亟需构建面向 AI 全场景的数据安全纵深防御体系，打造全方位的“免疫力”：

防滥用：防范 AI 智能体被滥用，关键在于行为基线建模与实时阻断。通过建立每个智能体的“正常行为画像”——包括其访问的数据对象、执行的操作类型、调用的 API 序列等，系统可实时检测偏离基线的异常行为。依托 AI 推理引擎，能在毫秒级内判断操作风险等级，对高危操作（如批量删除、权限批量变更、敏感数据批量导出）实现 99% 以上的识别率，并在分钟内主动拦截，同时触发告警与人工复核流程，将风险消灭在萌芽状态。

防投毒：构建隔离的 AI 安全沙箱环境，在 Agent 应用与模型上线前实施全流程安全评估，AI 智能体在上线前，其配置、权限声明、关联知识库及第三方插件均需经过安全“体检”。沙箱区全功能安全评估机制可在隔离环境中模拟运行 Agent 的所有预设操作链，自动检测是否存在越权倾向、指令注入风险或未知外部依赖。只有通过全部安全评估项的 Agent 方可获得上线授权，有效防止“带病入职”的数字员工进入生产环境。

防篡改：底层 WORM（Write-Once Read-Many，一次写入多次读取）与安全快照筑牢信任基石，在存储层深度集成一次写入多次读取 WORM 技术与底层安全快照机制。对于核心训练数据集、模型参数、业务日志及合规文档，需启用 WORM 特性，确保数据在保留期内任何人均无法修改或删除。同时，结合安全快照技术，可按分钟级粒度记录数据状态，即便发生逻辑错误或恶意篡改，也可一键恢复到最近的可信时间点。两层机制叠加，实现数据的“不可变”刚性防护。

防勒索：打造“主存 - 备份 - 隔离区（Air-Gap）”的多层立体防勒索架构。勒索攻击已从“广撒网”演进为“精准狙击”，AI 时代的防勒索必须构筑分层纵深。在主存层，部署实时行为检测引擎，识别加密行为特征并即时中断 I/O；在备份层，采用不可变存储与多版本备份，确保即使主存沦陷也有干净的恢复副本；在隔离层，建立物理隔离的安全恢复环境，定期执行自动化恢复演练。通过三层联动，实现 99.99% 以上的勒索攻击检测率，确保企业在遭遇极端勒索攻击时，仍能实现数据的一致性高效可信恢复。

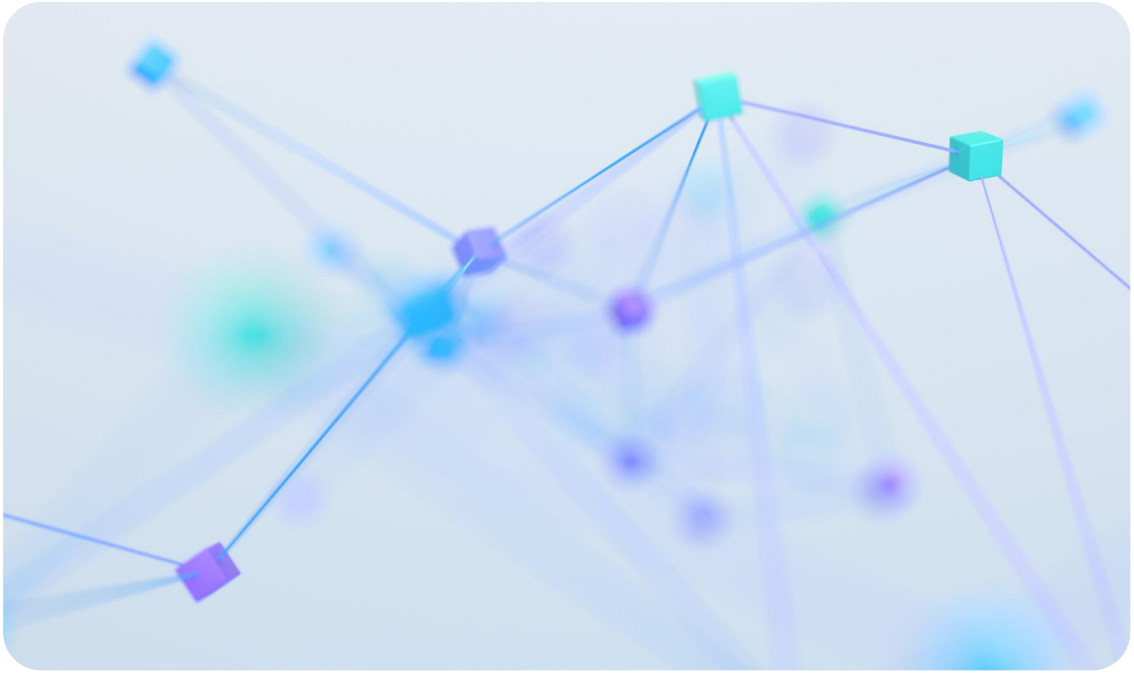
AI 时代的数据安全不仅是合规底线，更是 AI 业务可持续创新的前提。通过构建防滥用、防投毒、防篡改与防勒索一体化的全场景纵深防护体系，企业将能彻底破除安全后顾之忧，推动 AI 智能体在安全可控的轨道上释放无尽潜能。



更多免费行业报告 并购家 www.ipoipo.net

通信系统与网络：构筑以体验为中心的智能联接底座，实现 Token 的高效运载

随着智能体从简单的信息交互走向自主感知、实时决策、跨设备协同和任务执行，智能体将更加深入地进入个人、家庭、汽车、机器人以及千行百业，对网络连接广度、密度、质量都提出更高的要求。网络不仅是信息的传输通道，更是直接参与到 Token 生产、分发到使用的全流程之中。未来的网络不仅需要“连接得上”，更需要“连接得快、连接得稳、连接得智能”，并能够根据智能体业务的实时需求，在云、边、端之间动态调配连接、算力和资源。5G-A、全光网络以及确定性网络等能力的持续发展，将进一步降低智能体业务的连接成本和交互时延，提升实时智能、沉浸式交互、智能驾驶、机器人和工业智能等场景的体验。



从“连接人”到“连接智能体”，构筑智能世界的全域移动联接底座

移动通信网络正步入以智能体互联为核心的新阶段。联接对象从百亿级的人互联，扩展到人、物、智的多元协同共存，联接规模将百倍增长，同时承载全球 90 亿人的通信需求与 9000 亿智能体的联接需求，网络能力面临根本性升级。

智能体是这一轮演进的核心变量。不同等级的智能体[1]对联接的诉求差异明显：L1/L2 工具级智能体约占 70%，需要持续在线、网络稳定；L3 辅助级智能体约占 25%，需要高可靠性与隐私保护；L4 协同级智能体约占 5%，需要支持海量数据上传、多路高并发及百毫秒级协同。

面向这些变化，移动通信网络需要从“连接人”扩展到“连接智能体”，商业范式也从同质化服务走向差异化分层。其核心演进方向是承担人、物、智的多样化联接，支撑智能生产、分发和使用的全流程。具体推进路径是三个方向：以网强智、以智赋网、平滑演进。

以网强智：构筑分层协同的全域联接能力

随着智能体深度融入消费生活、行业生产与全域泛在场景，对移动网络的承载容量、覆盖广度与体验确定性提出全方位升级要求。无线网络需要以频谱资源拓展与高效利用为主线，从大带宽中频筑基、全频段灵活协同、空天地立体组网三个维度稳步推进联接能力的系统性进阶。三类技术路径分别匹配容量承载、体验适配、全域覆盖三类核心诉求，共同构筑分层协同的全域联接能力体系，为多元智能体业务落地提供坚实网络支撑。

大带宽中频构筑核心容量层，承载海量智能体应用

智能体规模化部署与多场景深度渗透，对移动网络容量与并发能力形成刚性挑战。热点区域智能体密度预计将超过 1000 万个 / 平方公里，远超人口承载密度；智能体间点对多点自主交互、7×24 小时持续运行的特征，推动流量规模指数级增长。网络需提供稳定的 Tbps/平方公里量级的吞吐能力，以保障大量智能体“背景式”持续运行。

引入 200MHz~400MHz 超大带宽，是满足海量 Agent 业务诉求的核心技术路径。更大的速率动态区间，可承载智能体瞬时爆发的海量数据交互需求；充足的时频域资源支撑多用户并行调度，保障高并发场景下的业务体验一致性；充裕频谱资源可部署多链路冗余备份，满足工业级高可靠场景的业务要求。

U6GHz 已在 WRC-23 被确立为下一代移动通信的核心频段，具备覆盖对齐 3.5GHz、频谱资源充裕的优势，是构筑中频大带宽体系的核心载体。针对 U6GHz 频段高、穿透损耗大、覆盖受限的特性，重用现有站点引入 U6GHz 以规避加站成本，成为新一代网络部署的核心要求。依托超大规模天线阵列与 MIMO 精准波束赋形技术，在不增大天线物理安装尺寸的基础上集成更多天线单元，结合超精度波束赋形与扫描技术，可实现 10dB 覆盖增益，支撑 U6GHz 与 2.6GHz/3.5GHz 共站点部署，并带来小区容量多倍提升。当前 U6GHz 已具备芯片与终端生态基础，预计将在全球部分区域落地基于 5G-A 的先行商用验证。

[1]《人工智能终端智能化分级》（GB/Z 177—2026）

更多免费行业报告 并购家 www.ipoipo.net

全频段灵活协同，匹配多样化业务体验需求

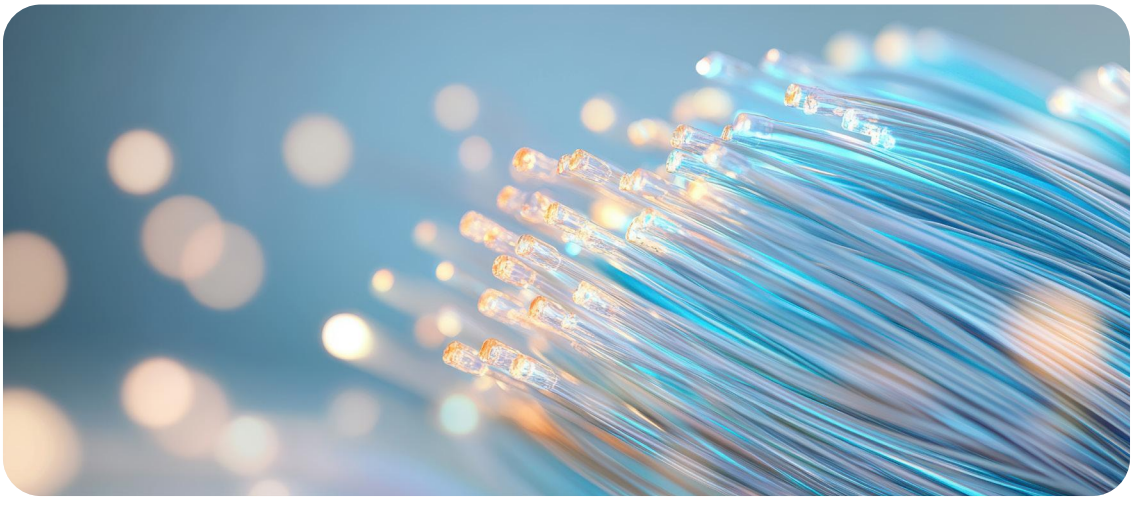
当前用户 MBB 需求体验已经开始分化，AI Agent 普及对多维体验差异化进一步拉大，包括大上行、时延确定性等都有不同的诉求，多维能力成为关键。业务的话务模型将从以下行为主的内容获取，转变为上下行均衡的智能协作。多模态实时感知、上下文同步、协同训练与推理将驱使上行流量激增，也对上行速率、时延提出极高要求。当前 AI 最大上下文已扩展至 10M Token 以上，单次交互上行流量将突破 45MB。同时为保障云端大模型对环境的精准识别，网络必须大幅提升上行带宽能力。

应对多样化业务体验诉求，需充分发挥不同频段在覆盖、容量层面的差异化优势，实现全频段资源高效协同。部署 SUL 补充上行与上下行解耦方案，将 TDD 中频大带宽与 FDD 低频广覆盖能力结合，在不损耗下行传输资源的情况下，大幅提升上行容量与边缘速率。采用灵活时隙配比与帧结构动态调整技术，可将专用频段上行资源占比提升至 40%~60%，上行容量提升 30% 以上。通过 FSA 灵活频谱接入技术实现时隙级动态选频，有效提升现有上行频谱利用率，全时段激活多频段上行资源潜力。

空天地立体组网，补齐全域覆盖场景短板

全球仍有超 3 亿人未被移动网络覆盖，在 AI 飞速发展的今天，数字鸿沟可能加剧，数字包容更需提速。同时智能体终端深度融入行业流程，进入人难以到达的环境，人、物、智活动范围不断向空、海全域扩张，地面网络覆盖存在天然盲区，全域泛在联接的诉求持续提升。

联接需求走向全域全场景覆盖，需要构筑空天地立体通信网，通过高速 NTN 与地面网络按需互补，实现 100% 地理覆盖。二者可通过差异化部署模式实现高效协同，当前主要形成两类典型落地路径。两网独立漫游模式下，地面网络与卫星网络各自独立运营，依托手机直连能力覆盖海洋、山区、沙漠等广域覆盖白区，用户通过漫游机制实现跨区域联接接续。地面主导中继模式下，由地面网络统一管控卫星链路资源，依托专用宽带终端承载低空、远海等场景的定制化业务，卫星资源作为地面网络的延伸能力池统一调度。两类模式可按需适配不同场景诉求，有效补充地面网络覆盖短板，填补全域联接服务的空白。



以智赋网：内生智能驱动网络效能全面跃升

随着联接规模持续扩容与业务形态不断丰富，移动通信网络正向资源高效利用、运行低碳节能、运营智能自治的方向持续演进。AI 能力融入无线网络的资源调度、能效管理、运营维护全流程，通过体系化的智能升级驱动网络效能全面跃升，支撑构建高效、低碳、自治的下一代网络运行体系。

联接数百倍增长，全局智能调度最大化谱效能效

智能时代，联接规模的爆发式增长与交互方式的变革，推动了业务复杂度的激增。同时，更丰富的频谱、更大的带宽、百倍增长的联接数，也给下一代移动通信的精细化管理和干扰抑制带来了全新挑战。伴随用户数、频谱带宽、天线数等十倍增长，求解维度百倍增长，传统单域串行调度模式难以满足资源实时调度需求，无法实现资源最优配置。

面向多用户全场景一致性体验目标，无线网络需要构建全局资源智能调度与收发体系。推动资源调度架构从传统单域串行向时频空功全域并行计算演进，实现全域资源动态最优分配，系统性提升多用户并发场景下的资源利用效率。面向实时交互类智能业务，基于智能干扰抑制算法优化抗干扰接收能力，有效降低用户间互扰，持续提升上行链路传输性能。智能算法将全面提升频谱效率、覆盖能力与用户体验，适配百倍联接规模下的精细化资源管理需求。

智能业务峰均比增大，精细化原生绿色管理

海量智能体 7×24 小时持续运行，业务流量动态波动加剧，峰均比持续拉大，驱动网络绿色节能能力向更精细化方向升级。



践行 Native Green 原生绿色理念，移动通信网络需要从空口设计、架构优化、产品设计三个维度系统性构建精细化节能能力。在空口设计层面，采用信道稀疏化与自适应设计，缩减移动性与连接建立相关信号的传输周期，并将超精简运行模式向空域与频域延伸，实现时域、频域、功率域、空间域的资源灵活配置与按需供给。在架构优化层面，通过深度并行化设计解耦组件间依赖关系，实现网络设备毫秒级快速唤醒，支撑网络全天候动态节能运行；同时引入唤醒信号机制，在保障终端快速响应上下行数据传输的前提下，使终端多数部件可长期处于低功耗睡眠状态，同步实现终端侧能效提升。在产品设计层面，引入业务智能预测算法与新型功放架构，基于业务负载动态调整功放工作点，保障系统在不同负载场景下均可运行于最优效率区间，从硬件底层降低系统整体功耗。

运营复杂度激增，网络驾驶智能体提质增效

网络频段、带宽、站点规模持续扩张，叠加智能体业务的差异化、动态化特征，网络运

更多免费行业报告 并购家 www.ipoipo.net

营复杂度呈指数级上升。传统运维模式依赖人工处置与固定自动化规则，故障发现滞后、处置周期长，难以适配海量网络单元的精细化运营需求，运维效率成为产业发展瓶颈。

无线网络正在构建意图驱动的单域网络自治，实现从运维提效到网络提质。基于超过 10 万篇案例库和 1 万个以上专业 API/Skills，业界唯一的无线网络智能体（RANSpirit）通过数字孪生和无线专业模型，达到 95%+ 意图理解准确度，实现覆盖、速率和节能等多目标的分钟级感知、预测及最优求解，先于人自主发现并解决问题。目前网络运维自动化已服务全球 20+ 网络，覆盖超 100 万小区，赋能 2000+ 工程师；平均故障恢复时长降低了 35%，运维效率倍增。面向下一代通信网络演进，无线网络引入生成式和决策式大模型、无线智能体自演进工程框架，最终达成新业务即时开通零等待、业务 SLA 100% 满足零投诉、网络永远在线零退服的目标。



价值跃迁：以智能体验牵引网络价值持续升级

面向智能世界的发展需求，无线网络的价值衡量逻辑在传统比特成本最优的前提下，进一步向 Token 交付效率与任务价值最优延伸。智能体分级差异化的业务诉求，也将驱动产业探索 Byte 与 Token 融合的增量价值空间。与此同时，网络代际升级需要充分兼顾产业现存价值资产，避免激进架构切换带来的产业成本与落地风险，需要在技术演进与商业价值方面同步探索可持续的发展路径。

面向智能业务需求的持续演进，移动通信网络的能力体系与架构形态需要长期迭代升级，适配内生智能、全域覆盖、差异化 SLA 等多元方向。需要全产业联合探索可延续的统一演进框架和跨制式频谱共享技术，以现有网络架构为基础，采用模块化、分阶段的演进思路，支撑新能力按节奏有序导入，充分复用现有网络资产，实现移动通信网络的平滑演进，兼顾长期技术演进目标与短期产业落地节奏。在此演进框架下，频谱资源的统筹规划与产业生态协同是网络能力稳步升级的核心支撑。全球产业需协同推进中频频谱生态构建，围绕 U6GHz 打造协同一致的高质量产业链，在已完成频谱标识与发放的国家加速生态培育，降低产业落地

成本，推动全球共享 U6GHz 产业红利；同时积极参与 ITU 框架下新频段的共存研究，推动 WRC-27 完成部分国家的频段标识，持续补充中频大带宽资源储备。此外，产业各方需合力推动各国规范 U6GHz 的授权使用，避免宝贵频谱资源的低效与碎片化利用，为网络长期平滑演进提供可持续的资源保障。

基于持续演进的网络能力，移动通信的价值覆盖范围与兑现模式不断拓展，深度融入智能的生产、分发与使用全流程，形成多维度的增量价值空间。面向消费领域，差异化 SLA 能力可支撑智能体验分级、沉浸式通信保障等多元服务形态，满足不同场景下的确定性体验需求，提升用户数字生活品质；面向行业领域，专网即服务、端到端动态切片、网络能力开放 API 等可承诺、可计量的网络服务，可为工业、交通、能源、园区等各类场景提供确定性联接支撑，赋能行业数字化智能化升级；面向全域泛在联接场景，星地融合的网络架构可支撑远域专网、低空智联网等新型服务形态，填补远海、偏远区域及低空领域的覆盖空白，拓展移动通信的服务边界与产业价值。

创新业务：下一代 AI 原生架构核心网，使能智能体新业务

随着智能体持续向全终端、全业务、全场景深度渗透，AI 智能手机加速规模普及，各类 AI 创新终端迎来爆发式增长，具身智能开启规模化商用。移动网络的联接对象从人和物延伸至海量智能体，人、物、智能体对网络的需求存在显著差异。如图，人联的主要需求是以视频为主的下行流量，要求长时间在线，部分应用要求 SLA 保障；物联的主要需求是以传感器数据为主的上行流量，周期性在线，要求低功耗；智能体联接的主要需求是实时多模态推理任务，7x24 小时突发，要求大上行和任务级保障。一张网智能的满足百倍联接和差异需求，成为移动网络必须具备的核心能力。

	人	物	智能体
接入	长时间在线	周期性在线	7x24小时突发
流量	视频	传感器数据	实时多模态
流向	下行	上行	上行
SLA	应用级	连接级	任务级
功耗	中	低	高

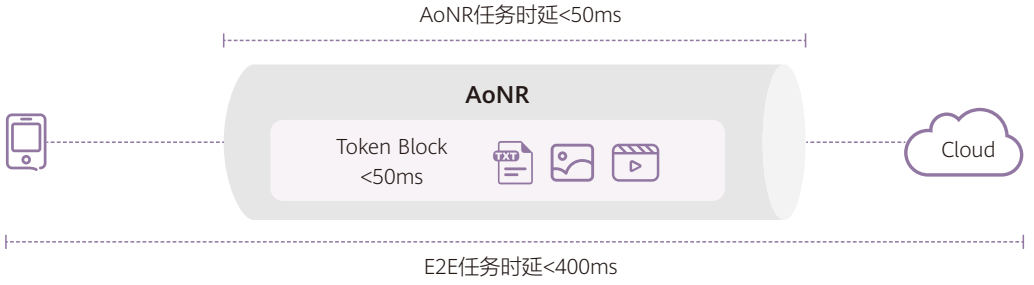
人、物、智能体对网络的需求

智能体业务成为移动网络的典型场景，需要智能的任务保障

随着 AI 终端的规模化普及，智能体业务逐渐成为移动网络服务的典型场景，智能体的目的是完成特定任务，例如端云协同完成 AI 推理任务，与其他智能体通信完成协作任务等。受终端算力差异与任务诉求多样化影响，对移动网络提出三类全新的智能任务体验保障诉求。

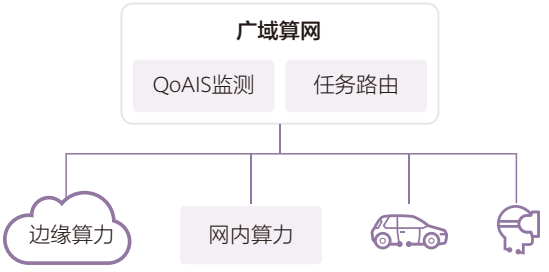
更多免费行业报告 并购家 www.ipoipo.net

AoNR 保障任务时延：AI 手机、AI 穿戴设备等随身 AI 终端的算力受功耗、体积、成本制约，端云协同完成 AI 任务是主流模式，E2E 任务时延即“用户提交任务到收到首 Token”是智能体业务体验的关键，AI 交互类任务占比从 8% 提升到 20%，E2E 任务时延小于 400ms 才能达到优秀体验。如图，移动网络为 AI 任务提供专用通道 AoNR（Agent over New Radio），保障一次 AI 任务的完整 Token 在 50ms 内完成传输，从而支撑 E2E 任务时延小于 400ms。



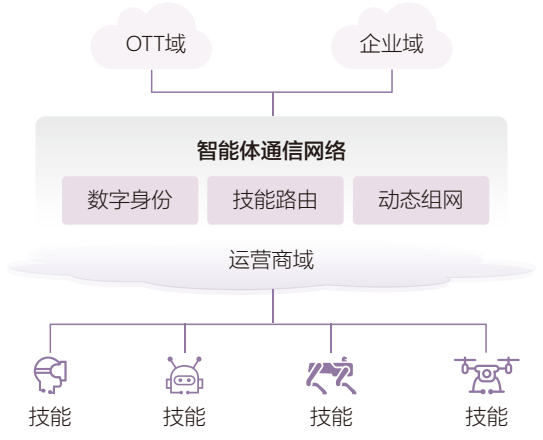
AoNR

广域算网提升算力效率：一方面 AI 手机、AI 穿戴设备的的算力不足，一方面部分具身智能、网络设备的算力闲置，如静止充电状态的自动驾驶汽车、潮汐空闲的网络设备。如图，广域算网通过移动网络将各类终端的算力、网内算力、边缘算力连接在一起，构建基于任务的算力调度能力，推出标准化算力服务，按需为算力紧缺的终端与云端业务提供弹性 AI 推理服务，高效盘活全域算力。



广域算网

智能体通信网络：单智能体受限于模型能力、上下文容量、角色冲突、串行瓶颈等，难以解决大量分布式、多目标、动态变化的复杂问题，跨域、跨生态的多智能体协作是 AI 进阶发展的必然选择。如图，基于移动网络构建智能体通信网络，通过跨域可信数字身份、基于技能的智能体发现和路由、基于任务的动态组网三大能力，使能跨域、跨生态的智能体高效协作。



智能体通信网络

基于意图接口的端网协同成为移动网络核心能力，核心网是实现载体

移动网络联接对象从人物延伸到智能体，

要求移动网络能够按照任务意图智能调整接入策略，如动态调整 DRX 参数等。智能体业务成为移动网络的典型场景，要求移动网络支持 AoNR、广域算网、智能体通信网络等智能的任务体验保障，驱动移动网络的服务模式从无差别流量转发到基于任务意图的任务时延保障、算力效率提升、协作安全保障。

无论是智能接入策略还是智能任务体验保障，均需要移动网络能够快速获取终端的任务意图，进而智能决策并匹配最佳的算网资源。

当前移动网络是信令驱动的固定流程模式，无法快速获取终端的任务意图，也无法即时理解任意图并智能执行操作。移动网络需要构建基于意图接口的端网协同机制，快速获取终端的任务意图，打造适配智能体业务的智能决策能力。

核心网向下联接终端，向上使能业务，是移动网络的接入控制中心，业务使能中心，数据汇聚中心，是移动网络承载端网意图接口、实现面向任务的智能决策能力的核心载体。

围绕 AI 原生定义下一代核心网架构成为产业共识，意图接口 +DAP 是关键特征

2025 年 6 月，3GPP 立项下一代移动网络架构研究课题，围绕 24 项关键课题推进方案设计与架构选型。截至 2026 年 6 月，AI 架构、意图 NAS（Non-Access Stratum）、数据框架、计算服务、自适应 QoS、智能体通信网络等方向成为共识度最高的课题。产业界对下一代核心网架构的关键方向已经达成共识，采用 AI 原生架构，用 AI 全面赋能核心网接口、功能与流程。如图，AI 原生架构可以归纳为三个关键特征。



下一代 AI 原生核心网

接口：信令到意图，支持传递终端任务需求（体验 / 组网 / 算力等），简便使用网络服务。

功能：两平面到三平面，在控制面和用户面的基础上，新增数据 & 智能面 DAP（Data & AI Plane），理解终端任务意图，实现终端任务诉求与网络资源的动态适配，实现智能的接入策略、任务体验保障。

流程：AI 叠加到 AI 内生，DAP 通过 ABI（Agent based interface）协议与控制面、用户面协同，AI 决策深度嵌入到控制面 / 用户面基础流程，提升智能执行能力

更多免费行业报告 并购家 www.ipoipo.net

从“Bit 管道”到“Token 运载网”，构筑智能世界的全场景算力互联底座

在这场由 AI 带来的深刻产业变革中，固定网络正在发生根本性的转变：从传统时代单纯的内容传递者、比特传输管道，升级为支撑全产业智能化发展、承载分布式智能体持续运行的 Token 运载网。这要求固定网络加速升级——打通 Token 生产、Token 流转、Token 应用全链路，实现端到端高效承载、可信安全与低碳绿色协同发展。

高效高可靠的绿色数据中心网络，支撑 Token 的高效生产

推理业务快速增长推动 Token 消耗量爆发式提升，算力竞争的焦点逐步由“算力规模”转向“Token 生产效率”。在算力供需剪刀差持续扩大的背景下，通过底层网络架构革新释放算力效能，已成为破局的关键路径。

当前智算数据中心建设中，全栈方案和开放以太方案凭借各自独特优势成为两条主流路线。端到端全栈方案带来高算效及一体化开发运维，具有效率优势，例如英伟达的 NVLink+IB 全栈、华为端到端 UB 灵衢方案；以太网标准开放、产业链成熟、综合成本更低，具有更好的可扩展性，业界主流厂家均基于以太的 RoCE v2 技术推出了智算数据中心网络方案，如华为推出的 NSLB 网络级负载均衡技术，通过算网协同统一算路，解决多任务碎片化场景全局均衡的问题。

但随着推理业务快速增长，传统面向训练设计的网络架构正面临新的挑战。传统智算网络采用 400GE/800GE 参数面、100GE/200GE 存储面独立组网模式，带宽资源静态分割，无法实现跨平面动态调度。在线推理场景下，随着会话变长、并发提升，大量 KVCache 需要贮存在存储设备，催生更高的 KVCache 访问带宽需求，使传统面向训练规划的网络带宽配比出



现结构性失衡。因此，业界亟需更经济、高效的智算网络演进方案，支撑激增的推理业务。

与此同时，智算集群规模扩大也对网络可靠性和能效提出更高要求。大规模智算集群依靠多节点协同完成任务，任一链路、设备的单点故障，都可能引发任务中断、服务异常，造成算力资源的损耗并影响 AI 服务的连续性；另一方面，随着集群规模持续扩张，数据中心能耗总量快速增长，网络设备作为数据中心基础设施的重要组成部分，其自身功耗及对整体 PUE 的影响不容忽视，绿色低碳已成为智算基础设施建设的重要约束。

综合以上需求，智算数据中心亟需重构底层互联体系，打造兼具高效、高可靠和绿色等特征的新一代网络底座，构筑 Token 规模化供给的坚实联接根基。

高效：存算网融合架构和 UB 灵衢全栈，提升 Token 生产效率

KVCache 读取效率直接决定 Token 生成速度与业务体验，针对 KVCache 带宽瓶颈问题，业界主要通过存储面与参数面融合打通带宽资源壁垒。此时，网络需同时承载 KVCache 写入、读取等多类数据流，为保障 KVCache 读取效率，网络设备需要识别不同类型数据流，并提供智能调度的能力，优化推理业务体验。实测显示，在架构融合基础上，星河 AI 数据中心网络，依托精准的流量识别与调度策略，将 Token 生产效率提升 20% 左右。

同时，UB 灵衢全栈一体化方案通过协议栈的深度统一实现端到端协同优化。网络侧支持全语义逐包级智能调度，精准识别并适配 AI 训练、推理、存储等多样化业务流量特征，实现 98% 的网络吞吐效率。计算侧免专用 RDMA 网卡部署，配合交换机的低时延转发架构，全栈时延较传统以太网方案降低 40%。

高可靠：三级高可靠架构提升算力可用性，保障算力长稳释放

智算数据中心网络可以通过设备级、链路级、模块级三层高可靠架构，保障智算业务的稳定运行。

在设备级，传统交换机重启时长普遍超过 120 秒，一旦发生故障将造成训推任务回滚、推理请求超时。转控分离架构与极速转发保持技术，在控制面异常期间保持交换芯片持续转发；设备升级、重启时，业务转发中断时长小于 5 秒，保障智算集群训推任务不中断。

在链路级，大规模智算集群使用海量光链路，极易出现端面脏污、接头松动等难以提前识别的隐性问题。短距光回损定位技术能分钟级排查光纤松动、脏污等常见问题，事前消除故障，大幅缩短故障处置周期，为Token稳定生产筑牢链路根基。UB 灵衢全栈方案中，通过协议层面融合，提供通道抗损和链路层重传，实现误码闪断 0 丢包，并支持跨链路重传，保障单点链路故障 0 丢包。

在模块级，光模块作为智算集群互联的核心元器件，端口故障多数起源于光模块性能衰减与可靠性不足。通过规模化部署高稳定性商用光模块，结合光模块实时状态感知能力，降低端口侧随机故障概率，使光模块故障引发的中断次数降低 90% 以上。

绿色：光电混合架构打破算力互联“功耗墙”，Token 能效比倍增

光电混合架构是绿色低碳的智算中心主要建设方向，其中液冷交换机、NPO 光电交换机和 OCS 创新架构，是网络层面破解算力互联“功耗墙”的关键路径。

液冷交换机具备高效散热能力，能够支

更多免费行业报告 并购家 www.ipoipo.net

撑更高端口密度部署，显著降低散热系统能耗，助力智算中心优化 PUE 指标。当前，51.2T/100T 盒式液冷交换机已成为业界主流，高密的 400GE/800GE 为构建全栈绿色液冷的智算集群提供丰富选择。

NPO 光电交换机将光引擎贴近交换芯片，大幅降低高功耗 DSP 信号补偿电路，端口功耗相比传统方案下降 20% 以上。当前 NPO 光电交换机已经处于试点验证向规模商用过渡阶段，华为等厂家的多款 51.2T NPO 交换机已完成实验室与机房灰度测试，预计 2026 年底至 2027 年将开始规模商用。

超大模型公司也开始探索在 DCN 中引入光电路交换（OCS）。光电路交换（OCS）无需光电转换与数据包解析，可实现接近物理极限的传输性能，其具备四大特征：一是极低功耗单端口功耗小于 1W，整机功耗远低于同规模电交换机，并且随着模块速率升高 OCS 功耗不变，功耗可以减小到交换机的百倍分之一；二是亚纳秒级转发时延，OCS 内数据的单跳时延低至 1-10 纳秒，进一步提升大模型的训练和推理效率；三是支持平滑演进，OCS 对信号速率与协议无感，从 400G 到 6.4T 以及更高速率都可以处理，升级无需更换交换硬件，保护投资成本，而且可以支持集群规模从 32K 到 64K 甚至到 512K 及以上按需平滑扩容；四是高可靠，OCS 无数字转发芯片，且无光模块，硬件故障大大降低，MTBF 改善 25% 以上。因此数据中心 DCN 网络引入 OCS 后形成“光电协同”混合架构对于提升整体性能和降低功耗有显著的益处。

超宽无损安全的毫秒级跨域算力互联网络，支撑 Token 的高效运载

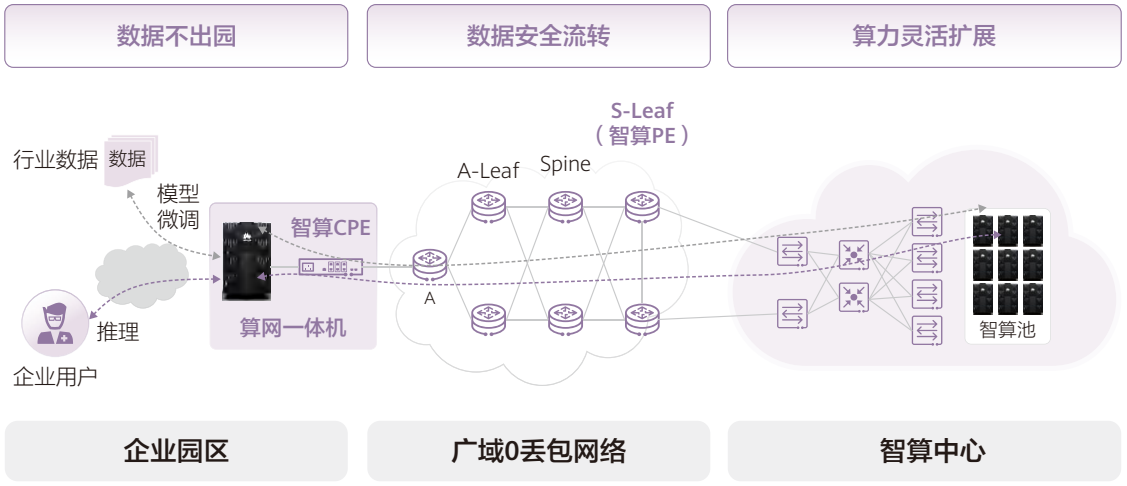
行业预测，2027 年分布式训推业务量将全面超过集中式训推，催生大规模跨地域算力互联与 Token 跨域运载需求。在跨区域模型训练、远程推理调度、多智能体跨节点协同等场景下，海量参数、特征信息与流式 Token 需要在网络中传输，传统广域网存在拥塞丢包、有效传输效率不足等短板，一旦报文丢失将引发重传开销激增，拉长训练周期、降低 Token 产出速率。因此，Token 跨域运载网络必须具备无损传输能力，消除跨域协同带来的算效损失。与此同时，训练样本、业务交互 Token 是企业核心资产，超过 50% 的企业提出数据不出园区的刚性诉求。算力互联网络需构建端到端的安全体系，保障 Token 跨域交互的安全合规底线。以“Agent 自动检索网络素材并下载到本地”为例，华为 iLab 测试显示网络连接时延从 5ms 增加 20ms 时，任务完成时间增至 1.6 倍，Token 消耗费用增加 16%；再叠加丢包率 0.01% 的情况，任务完成时间将增至 5 倍，Token 消耗费用共增加 70%。

针对以上普遍诉求，Token 跨域运载网络需要同步构筑无损传输与全域安全底座，支撑高效、安全的分布式算力协同与 Token 供给。

基于精准流控和自适应重路由的广域无损网络，Token 跨域算效可达 95% 以上

AI WAN 智能 IP 广域网是当前业界应对智算业务对网络提出新需求的主要路线。其中，广域 RDMA 技术是跨域算力高效互联的核心。传统 RDMA 原生适配数据中心短距组网，在百公里、千公里级广域复杂网络中应用存在明显短板。广域网络时延波动、微突发拥塞引发的微量丢包，造成 AI 业务算力收敛变慢、算效大幅损耗。对此，业界主要围绕租户级精准流控

（SPFC）与自适应流重路由（AFR）等开展针对性创新。SPFC 以租户、智算任务为粒度，实现精准反压与拥塞隔离，规避广域拥塞扩散风险；AFR 通过 AI 流感知能力识别智算大象流，实现全网动态调度，替代传统静态转发模式。二者协同构建广域无损传输体系，有效解决长距 RDMA 丢包难题，跨域算效可达 95% 以上。



千公里无损算力协同广域网

基于超宽联接、超低时延、超高可靠的毫秒算力网，提升 Token 跨域运载效率

分布式智算集群的跨节点协同推理、多机分布式训练及跨域 Agent 交互，既依赖 DCA（Data Center Access 数据中心接入网络）解决同城内 Token 调度，也需 DCI（Data Center Interconnection 数据中心互连）承载跨区海量 Token 与参数同步。因此，构建 DCA 与 DCI 的传送协同至关重要，其本质解决 Token 在 1ms/5ms/20ms 等多层数据中心形成的时延圈下“传得快、传得稳、传得安全”，避免算力硬件因网络短板形成算力空转浪费。

以 AI DC 为中心，构筑超宽联接、超低时延、超高可靠的毫秒级算力网络已成为建网趋势。首先，部署 400G/800G/Beyond 1T 到骨干网、100G 至城域网，构建超宽网络底座；其次，基于 OXC 技术实现城市内 1ms、省内 5ms、国家枢纽间 20ms 的确定性低时延，保障用户用算体验；再次，将 WSON 从骨干网延伸至城域网，提升网络可靠性，实现业务抗多次断纤，支持算力网络多点故障下的多轮次 50ms 重路由保护恢复及波长资源动态调度。

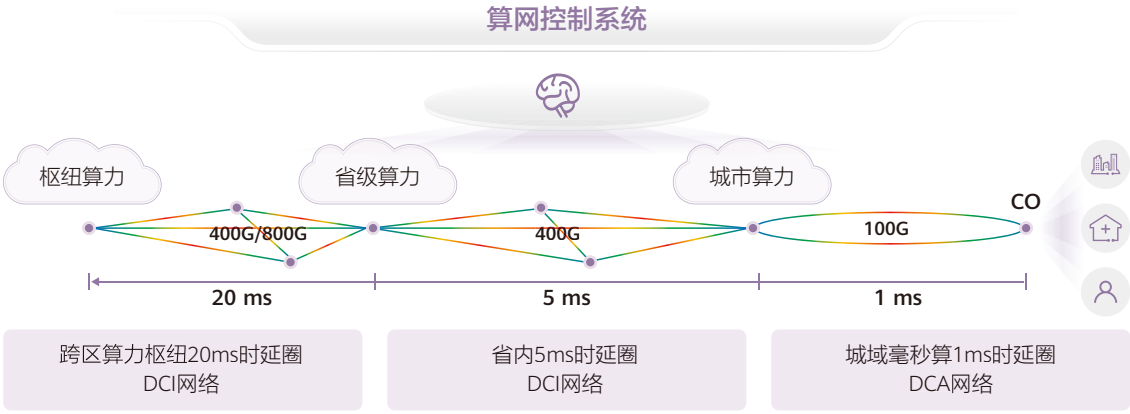
三层主动安全保障体系，使能 Token 可信流转

面向跨域算力协同，可从设备、链路、网络搭建主动安全保障体系，确保 Token 在广域网络内可信、安全流转。

在设备层，广域路由器内置基于 AI 的安全检测引擎，实现异常流量智能识别与威胁闭环拦截，筑牢网络基础设施第一道安全防线；基于端到端波分设备硬件形态的加密信号与密钥合波共纤传输，为数据高效安全传输打下物理基础。

在链路层，构建量子密钥分发（QKD）与

更多免费行业报告 并购家 www.ipoipo.net



以 AI DC 为中心全光目标网

后量子密码（PQC）融合的新一代安全通信体系，实现链路物理级加密防护，信号传输做到通密一体。同时，当前业界主推的在网络设备中内生 QKD 的路线，不仅可以抵御未来量子破解风险，同时可以降低 60% 以上的量子加密网络建设成本。

在网络层，基于 APN6（Application-aware IPv6 Networking）的数据围栏技术，依托 IPv6 扩展头携带 APN 业务标识，让网络精准识别不同敏感度数据流。通过在网络转发层设置流量边界，依托 APN ID 实施精细化管控，划定数据合法流转范围，实现敏感数据不出域、不越界。

构建泛在超宽、多维感知和全域安全的家庭和园区网络，以确定性体验赋能 Token 品质应用

随着 AI 与 Agent 加速融入家庭生活、企业生产和园区治理，数字空间正从“连接人与信息”迈向“连接人与智能、人与空间、人与服务”。AI 摄像头、数字人、云游戏、智能康养等多模态应用不断涌现，带来大规模上行、高并发、突发性流量和实时交互需求；AI 推理、Agent 协同也对带宽、时延、可靠性和安全提出更高要求。面向未来，家庭和园区网络需要从“尽力而为”的连接底座，向泛在超宽、多维感知和全域安全演进，以确定性体验支撑 AI 高品质应用。

泛在超宽：大上行超千兆 / 万兆光网，从“尽力而为”走向确定性体验

AI 业务普及正在重塑网络流量模型，推动家庭和企业网络由下行主导向上下行并重、适应突发高并发演进。网络不仅需要更大的双向带宽，还需要从“有带宽”进一步走向“体验可保障、SLA 可承诺”。通过 Wi-Fi 6 升级至 Wi-Fi 7、10G PON 升级至 50G PON，驱动光纤宽带向超千兆、万兆持续升级，构建泛在超宽能力，保障 AI 业务确定性入云入算；并通过 Wi-Fi 调优、网络切片等技术和体验管理平台等，保障、管理并优化用户体验，做到体验可承诺；同时在终端和 OLT 内置算力，实现端、边、云、算力协同，来承载家庭智能体和各类智能应用；最后通过升级运维能力，实现网络问题自动检测，自动修复，主动上报，并主动识别质差用户，使能高效运维。

智慧家庭：以家宽 AI Agent 为核心，基于“业 - 云 - 智 - 网 - 端”五层架构构建主动智能生活空间

随着端侧大模型、多模态交互和智能终端持续发展，智能体将成为智慧家庭新的统一智能入口，推动家庭服务从“人找服务、设备执行”向“理解意图、主动服务”演进。

智慧家庭将以家宽智能体为核心，基于“业 - 云 - 智 - 网 - 端”五层架构构建主动智能生活空间。业层通过从传统带宽、组网、IPTV 拓展至视频、游戏、康养等多元智能应用；云层基于统一的云 OS 平台，实现内容在大中小屏之间无缝流转，为用户提供跨场景、跨终端的统一体验；智层通过打造家庭 AI 智能体，为用户提供统一的入口；网层大带宽是极致体验的基础，从千兆迈向超千兆和万兆；端层推动家庭网关向家庭智能中枢升级，实现各类智能设备的高效互联与协同，融合设备接入、本地 AI 算力和智能控制能力，成为家庭智能交互与感知执行的重要入口。

在此基础上，融合光感知、Wi-Fi 通感和空间智能等技术，实现对家庭人、物、空间和环境的多维感知，推动服务由被动响应向主动服务升级；通过端边云协同和分布式智能，实现算力与网络资源动态调度，推动家宽终端向“家庭算力调度枢纽”演进，支撑低时延、高可靠的智能交互。同时，智能流量调度主动识别并调度 AI 流量，确保高带宽业务确定性体验。

智慧园区：构筑泛在万兆、多维感知和全域安全的智慧园区网络

AI 应用在园区蓬勃发展，驱动园区从传统物理空间跃迁为数字智能空间。园区网络不仅承担连接功能，更将成为园区智能感知、业务协同和安全治理的重要基础设施。

在连接方面，以 Wi-Fi 7、50G PON 等技术构建泛在万兆网络，融合 Wi-Fi、星闪、蓝牙等多种物联网技术，实现园区终端泛在接入和统一管理。

在感知方面，通过光感知、Wi-Fi+X 通感一体以及空间能力三大技术协同，形成覆盖人、机、物和环境的多维感知能力，为后续智能决策与服务提供坚实数据支撑，形成智能园区高阶智能的基础。

在安全方面，以“数字 + 物理”双域融合为核心，构建覆盖连接、资产、空间与隐私的立体化安全防护体系，打造全域感知、全域防护的智慧园区网络。通过无线空口保护、MACsec 叠加 PQC 算法等技术增强连接安全；通过哑终端识别、异常流量秒级检测与主动阻断有效防止网络入侵扩散，提升资产安全；并面向会议室、办公空间等敏感场景通过 CSI 技术（信道状态信息）和 Wi-Fi 感知能力实现空间安全状态感知、摄像头精准检出等，提升隐私风险识别与防护能力。

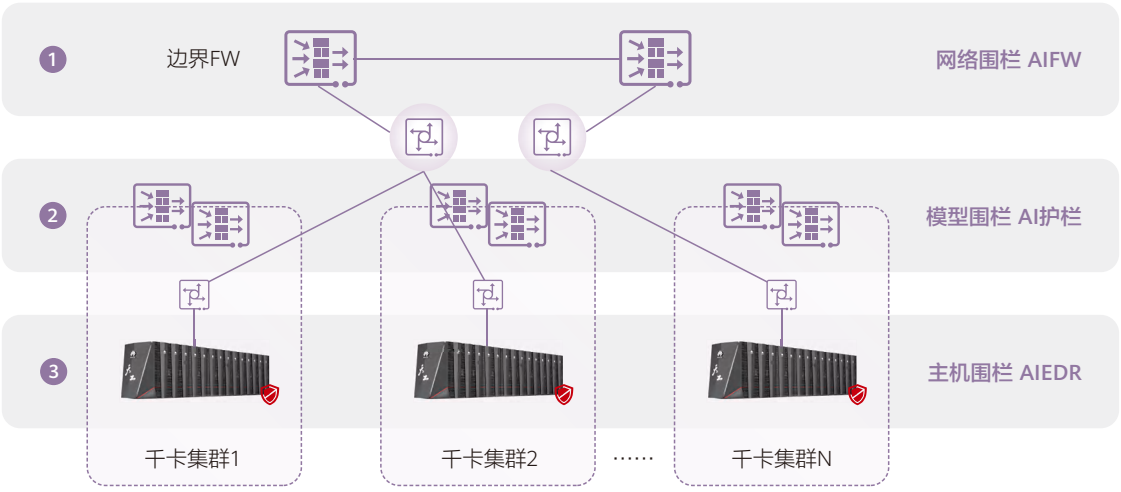


更多免费行业报告 并购家 www.ipoipo.net

以 AI 守护 AI， 支撑 Token 的智能防护

AI 业务规模化落地同步催生全新安全风险。一方面，攻击者依托大模型可以快速生成攻击脚本、自动化漏洞探测工具等，大幅降低攻击门槛和成本。IBM《2026 X-Force 威胁情报指数》监测显示，AI 驱动的漏洞攻击同比增长 44%，攻击周期由数天压缩至分钟级。另一方面，大模型与 AI 智能体自身安全隐患突出，多项联合安全研究显示，已投产商用的 AI Agent 中 91% 存在工具链攻击漏洞，94% 面临记忆投毒、提示词注入等攻击风险。

为保障 Token 的安全、高效应用，我们建议在网络、Agent、主机三个层面构筑纵深防御体系。



三层安全围栏，护航 Token 安全

第一层为网络围栏，设备本地通过内置威胁识别模型，恶意流量检出率可达 95% 以上，同时，通过内置专用 AI 算力实现本地研判，毫秒级恶意流量阻断。

第二层为 Agent 围栏，通过构筑提示词注入检测模型，风险识别准确率达到 95%，并依托异构算力实现检测性能优化，将检测时延由传统 500ms 压缩至 200ms，在有效抵御指令篡改、智能体越权攻击的同时，规避业务交互卡顿，保障终端用户体验。

第三层为主机围栏，部署于大模型算力节点，持续监测异常恶意进程。传统主机安全组件平均占用服务器 5% 左右 CPU 资源，挤占训推算力；新一代轻量化防护方案目标将 CPU 资源占用压缩到 1% 以内，兼顾算力环境安全与 Token 生产效率，最大限度减少防护机制带来的算力损耗。

三层围栏协同联动，分层拦截流量入侵、智能体攻击与主机侧恶意行为，形成闭环防御能力，平衡安全管控与智算效能。

AI 与网络深度融合，驱动超宽带网络迈向 L4 高阶自治，提升 Token 使用体验

随着人工智能加速融入各行各业，智能体应用迎来爆发式增长。面对千行百业、千家万户对算力接入稳定可靠、高效便捷的诉求，自智网络旨在提供 Zero-X（零等待、零故障、零接触）的极致用户体验和 Self-X（自服务、自发放、自保障）的网络运维能力，为 Token 使用体验提供确定性的网络保障。

从单场景自治到多场景协同自治，以确定性网络保障赋能 Token 使用体验

良好的 Token 使用体验高度依赖网络的确定性传输能力——无论实时交互还是离线批处理，网络波动带来的丢包、拥塞或中断都会直接拉低服务效率与可用性，网络运维亟需大幅提升效率，并实现运维模式从“被动响应”向“主动预防”转变。高阶自智网络秉承“分层智能、开放协同”理念，构建网络运维智能体，最大化实现网络故障自愈、性能自优与安全高稳，从而保障 Token 使用体验。

在业务发放场景中，Token 服务的差异化特征要求网络具备按需、即时分配资源的能力。高阶自智网络基于业务意图自主完成最优路径推荐与开通，实现敏捷入算。

在故障管理场景中，Token 服务对连接稳定性有着极高要求——一次网络链路故障便可能导致大规模推理任务中断，造成算力与 Token 的双重浪费。高阶自智网络通过构建涵盖故障预测、识别、定位及恢复的闭环自治能力，既有效规避故障引发的 Token 浪费，又显著降低人工运维成本。

同样，在网络性能优化、能效优化等场景中，通过场景化自治能力构建，为算力应用服务提供稳定可靠、绿色低碳的网络传输保障。同时，依托网元内生安全能力，高阶自智网络可构建网络传输链路安全防御体系，为 Token 使用营造安全、可信的环境。

此外，随着高阶自智网络从单场景自治向多场景协同自治演进，确定性网络保障能力将得到进一步夯实。多场景协同自治基于统一的业务意图，自动对故障管理、网络优化、体验保障等多场景任务进行统筹规划与冲突处理，最终实现整网最优，从而最大程度保障用户体验。

依托 TMF Forum 标准和产业 OpenAN 开源项目共建共享，加速高阶自智网络规模应用

TM Forum 联合产业伙伴与标准组织持续完善高阶自智网络标准体系，围绕架构、场景、成效指标与解决方案包等形成系列成果，为运营商推进高阶自智网络部署提供标准依据与方法论指导。

在此基础上，为加快高阶自智网络规模化部署，破解多智能体互联标准与语义不统一、跨层跨域集成难度高、公共组件复用率低等痛点，产业界发起 OpenAN 开源项目，依托 Linux 基金会网络子基金会（Linux Foundation Networking，简称 LFN）开展建设与运营。项目汇聚运营商、设备厂商、OSS 厂商等多方产业力量共建共享，围绕智能体管理组件、智能体框架和自智网络本体（AN Ontology）开展开源协同：提供统一的注册中心、编排中心与 A2A-T SDK，实现跨域跨厂商智能体间高效协作；开放智能体框架源码，加速智能体开发；对齐通信语义，抑制模型幻觉，实现确定性推理。

依托这套产业开源协同机制，可有效加快自智网络标准、技术与实践的衔接，推动高阶自智网络从单点验证迈向规模化应用。

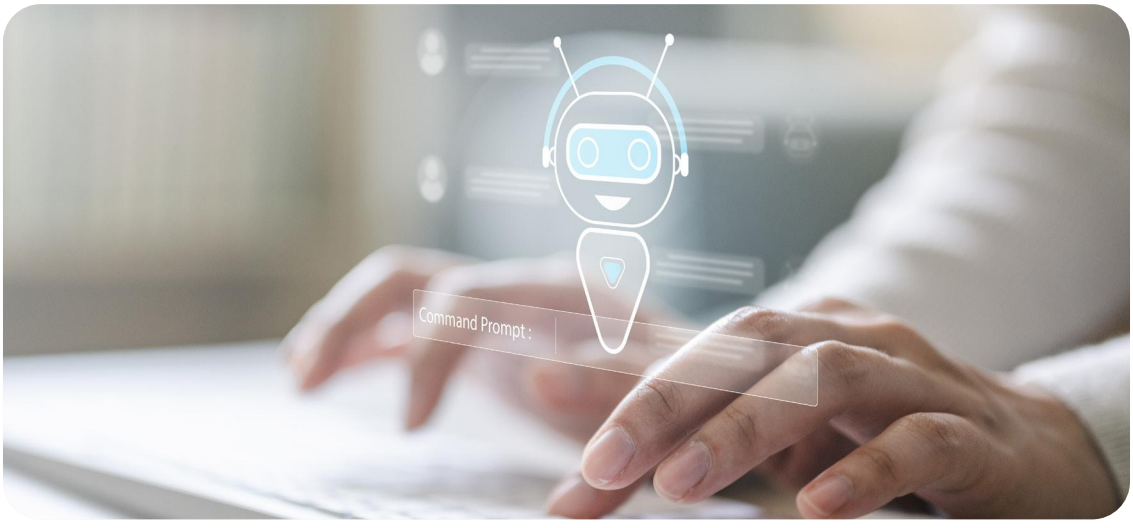
更多免费行业报告 并购家 www.ipoipo.net

云基础设施与软件：打造 Agentic 原生的云基础设施与应用平台，加速 AI 在行业深化落地

Agentic AI 时代，云基础设施与软件正在加速重构，成为企业智能化的核心底座。智能体的规模化商用，推动软件形态持续演化，也对底层算力架构、平台能力提出全新要求。既要适配企业智能体的业务应用变革，也要打造大规模、低时延、高效推理和通算、智算一体化调度的云上 AI 原生基础设施，同时构筑面向智能体时代的平台服务，以此破解智能应用落地过程中算力低效、任务不可靠、规模化难、安全治理不足等痛点，加速 AI 在企业深度落地。

企业智能体重构应用软件，向自主智能的“数字员工”演进

企业智能体正在从实验室走向生产应用，其核心能力在于自主完成感知、决策与执行的闭环，能够处理文字、语音等多种信息，并针对不同业务场景进行泛化与自进化。如今，它们早已跳出问答和检索的范畴，进化为能够自主拆解目标、灵活调度工具、并在人机协同（HITL）框架下保障执行质量的“数字员工”。从发展趋势看，企业智能体的兴起与软件形态的演变紧密相关：流程稳定、规则明确的业务将沉淀为固化的智能办公应用，而灵活多变的敏捷业务则催生临时生成、用完即弃的动态软件服务。两类软件形态的并行演进，也对应着企业对智能体在“开发态”与“运行态”的两类核心需求。开发态的“造软件”以代码智能体为代表，深入软件生产环节，覆盖编写、测试、运维等全生命周期；运行态的“用软件”以办公智能体为代表，覆盖日常办公、财务、人力、法务等泛办公场景，将重复性认知劳动自动化。这两条路径既是当前技术最成熟、落地最广的实践，也指明了从辅助执行到自主决策、从局部提效到全局优化的演进方向。



打造能自主交付软件的代码智能体，重塑企业软件生产方式

代码智能体已经能够理解项目级代码库，完成跨文件修改、功能开发、测试执行、缺陷修复和代码提交，正从“辅助开发者编写代码”走向“在组织约束下承担软件交付任务”。随着代码生成速度大幅提升，软件交付的主要瓶颈正在从编码环节转移到规划、评审、测试、部署和治理等上下游环节。传统软件开发生命周期以人工执行、串行交接和阶段性审批为基础，已经难以匹配代码智能体的执行速度。因此，未来十年不仅需要持续提升模型能力，更需要同步重构智能体协同、Skill 沉淀、安全治理和 Agentic DevOps 体系，推动代码智能体从“自主编写代码”迈向“自主交付软件”，将个人提效转化为组织级研发效能提升。

强化基础模型的编码与多模态能力，夯实自主交付的智能底座

编码能力是基础大模型理解和驱动数字世界的通用核心能力。多模态大模型将同时理解需求、代码、文档、界面、测试结果和运行反馈，这需要基于真实项目的完整过程数据进行强化学习，模型完成任务后，还要在可执行环境中检验结果，将成功经验和失败教训用于后续训练。通过这种“执行—验证—学习”的循环，模型将具备处理复杂软件工程任务的能力，减少无法运行或只完成一半的代码，成为代码智能体可靠的能力底座。

建设代码智能体自适应协同能力，让复杂任务高效执行

复杂软件交付跨越需求理解、方案设计、开发测试和上线运行等多个环节，单个智能体很难同时处理所有问题，固定工作流程也难以应对任务中的变化。研发智能体自适应协同，就是先把人的需求转化为目标、范围、约束和验收标准清晰的任务，再根据实际需要动态选择合适的模型、智能体和工具，决定哪些工作并行、哪些工作依次完成，并随执行结果及时调整。当任务中断或失败时，系统能够从最近的有效节点继续，而不必全部重来。由此，多个智能体可以像一支按需组建的研发团队一样协同工作，让复杂任务完成得更快、更完整，也更容易控制。

构建可验证、可版本化的 Skill 自进化体系，将组织经验转化为可执行能力

项目的代码、文档、历史修改和运行记录中沉淀着大量经验，但这些经验往往分散在个人、代码库和 DevOps 工具之中，难以被后续任务直接复用。研发知识经验自进化，就是让智能体在开发工作中持续记录有效做法、失败原因和人工纠正，并将其整理为可查找的研发知识和可重复使用的研发技能。新经验只有经过测试和验证后，才会进入研发能力库，避免错误知识被不断放大。这样，研发经验便不再随着任务结束或人员流动而消失，而是从个人经验转化为组织资产，使代码智能体在持续使用中越来越熟悉业务、项目和工作方式。

护好研发安全硬边界，保障代码智能体可管可控

代码智能体能够自主修改代码和调用系统，也可能因判断错误或外部信息误导而带来风险，因此安全不能只依赖模型“自觉”，而必须由系统设置不可突破的硬边界。每个智能体都应拥有独立身份和完成任务所需的最小权限，开发、测试和生产环境彼此隔离，文件、网络、账号和密钥受到统一管控。生成代码、运行测试等低风险操作可以在边界内自主完成，而版本发布、删除数据和改变权限等重要或高风险

更多免费行业报告 并购家 www.ipoipo.net

操作必须经过严格门禁或人工审批。所有关键操作和验证结果都会形成可溯源的证据，使代码智能体既能高效自主执行，又不会越过组织设定的安全底线。

打造智能体亲和的 DevOps 生产线，使能多种智能体高效协同

成熟 DevOps 平台已经能够通过 CLI、API、Webhook 和 MCP 等方式被代码智能体稳定调用。下一步需要从面向人工串行操作的传统流程，转变为面向人机协同和智能体持续执行的 AI 原生流程。AI 原生的 DevOps 生产线，需要让各类工具提供明确的能力说明、统一的输入输出、清楚的错误原因，以及重复执行和失败回退能力，让智能体能够理解、调用并判断结果。同时，系统将需求、代码变更、测试结果、发布过程和线上表现连接起来，使智能体可以持续处理问题分流、失败分析、版本升级和测试补充等日常工作，再由确定性的质量门禁工具把守最终上线关口。这样既能减少重复性人工操作，也能让每一次软件交付看得见、查得清、控得住。

未来，软件交付将形成从理解意图、组织协作和执行任务，到验证结果、积累经验和控制风险的持续闭环，让代码智能体真正具备自主、可靠且可控的软件交付能力。

使能办公智能体自主操作和执行，从员工个体提效到企业全局效率提升

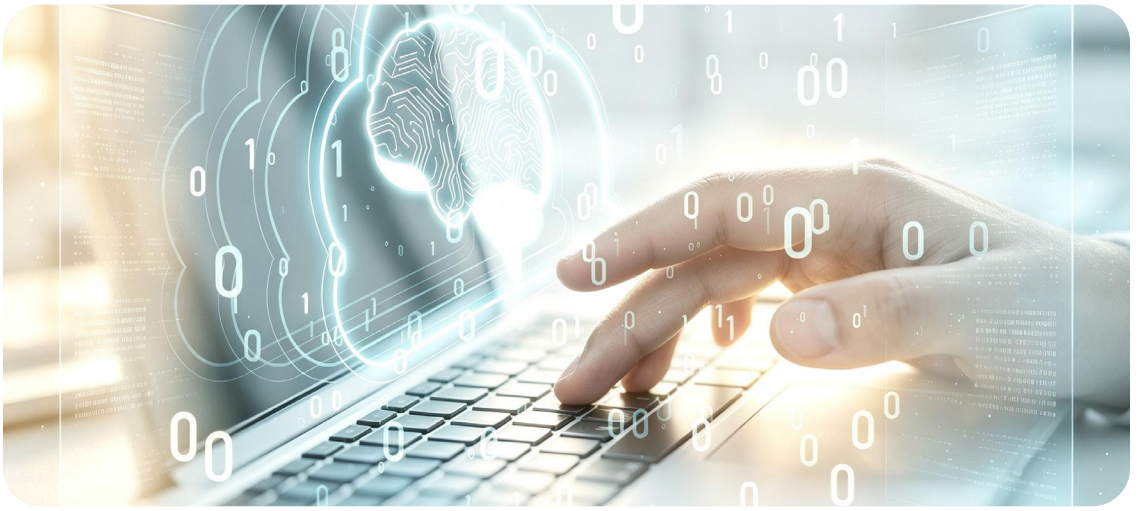
2026 年，办公智能体正从对话交互转向多步骤操作执行，企业已将其接入日常办公系统。未来在自主任务执行、分布式群体智能协同、环境感知交互模式及全链路安全治理等方向的持续演进下，办公智能体将从辅助工具演变为承担实质性操作的自主执行体。

构建全链路自主执行框架，支撑办公任务的长程运行与持续自主调度

在办公场景中，大量任务具有执行周期长、链路复杂、需持续追踪等特点。随着任务时长和交互轮次增加，智能体容易出现上下文连贯性衰减；在工具调用过程中，也可能因参数幻觉、权限越界或执行环境污染导致调用失败。同时，部分办公任务还要求具备 7×24 小时后台自主运行与跨时段持续追踪能力，对智能体的长时记忆、任务调度和稳定执行能力提出了更高要求。针对这类需求，需构建全链路自主执行框架：任务执行前先锁定目标，将最终目标分解为可独立验证的子步骤并建立状态机；每完成一步即写入状态存储，后续步骤从存储读取前置信息，避免依赖上下文窗口；执行中持续校验与目标对齐，异常时支持回滚。工具调用时，请求经 Schema 强校验、最小权限校验和环境隔离层后执行，成功则组合哈希化避免后续重复调用，失败则捕获异常并写入失败模式库。该框架使智能体持续追踪任务目标，跨时段恢复执行状态，并在异常后快速恢复，支撑小时到天级、7×24 小时运行的长程任务。

建立分布式群体智能协同能力，让办公智能体动态适应复杂流程与多端场景

办公场景中复杂任务（如新品发布会需协调物料、会场、人员、法务等）涉及多角色协同，单智能体难以覆盖全部流程；员工在多端移动办公，需跨端同步。这需要构建分布式异构智能体的群体协同机制：每个智能体提供唯一标识和能力描述，注册中心动态发现并生成



拓扑，按需路由；系统提取任务的并行度、依赖关系、时延预算等特征进行建模，根据运行时反馈（通过率、分歧等）自适应调整多智能体的协同拓扑与分发策略；云端统筹核心调度和并发执行（沙箱隔离），本地操作下沉至 PC 或移动端，任务状态统一存于云端，支持多端同步查看与续接。该机制使系统动态适应复杂流程，并在多端保持状态一致。

支持环境感知交互模式，使开放办公任务在共创中沉淀并可靠执行

泛办公场景下的开放任务（如组织策划多地客户拜访，需协调时间、会议室、材料，并应对突发变化）直接自动化容易放大目标误解、上下文缺失，还需人的判断与修正。为避免人的智慧、灵感、经验被排除在外，办公智能体需支持环境感知的交互模式：持续在线，融入办公现场，通过多模态理解、应用插件或服务扩展等方式，感知屏幕、文档、会议，以及浏览器、办公软件等应用环境，理解当前任务位置与对象。主动判断介入时机，提供辅助支持或理解人员隐含上下文的指令。经过多次人机协作、修正和验证后，沉淀为可复用的流程；一旦满足可信条件，任务便切换至主动模式，由智能体在授权边界内后台执行同类任务。该模式实现前台共创与后台执行结合，使开放场景下的复杂办公任务更易沉淀和可靠执行。

打造全链路安全治理框架，保障办公智能体安全合规与成本可控

办公智能体在企业中规模化部署面临操作归属难追溯、单点故障扩散与审计记录不完整及成本管控（如 Tokens 消耗与计算资源配额治理）等问题。这需要建立“事前 - 事中 - 事后”安全框架：事前为智能体分配独立虚拟身份，权限授权最小化且与任务周期绑定；事中监测运行状态，失败率过高或触发高敏感操作时自动挂起并通知人工复核；事后完整记录操作日志，写入中心化的不可修改的审计系统。业务治理层面支持企业集约型成本治理与动态配额调控，覆盖 Tokens、高成本服务 API 调用、及计算资源。该体系确保每笔操作可追溯、越权可拦截、成本可度量，满足企业内控与合规。

未来，随着上述方向持续迭代，办公智能体将具备处理更长时间跨度和更高复杂度任务的能力，逐步成为企业泛办公场景中可依赖的数字员工。

更多免费行业报告 并购家 www.ipoipo.net

建设通智一体 AI 原生云基础设施，结合大规模资源池化实现算力高效利用与灵活组合 / 切分

随着云基础设施与软件从“云原生（Cloud-Native）”向“AI 原生（AI-Native）”进化，围绕算力与内存资源的灵活组合和切分，下一代云基础设施呈现三个主要技术特征：

技术特征一：以内存为中心的 AI 原生基础架构

传统基于 Cloud-Native 的云调度系统，构建在虚拟化软件之上，CPU、内存固定配比，网络拓扑复杂，弹性效率低下。迈向 AI 原生（AI-Native）的云调度系统，必须借助内存总线相关技术打破服务器的物理边界，将分散的显存与内存重构为全局共享的内存池。调度系统将不再管理离散的服务器节点，而是基于模型、缓存、上下文等新数据形态分配和管理流动的内存资源，实现从“CPU 围着数据转”到“数据在内存总线上自由流淌”，确保万亿参数大模型的推理与训练不再受限于单机物理瓶颈。

技术特征二：通智融合的高效资源编排调度系统

云资源的调度与编排是云基础设施效率的核心承载，不同场景下业务对资源的需求差异显著，而数据中心不同资源池中各类资源的形态与配置亦存在显著差异，资源效率随业务需求演进与硬件代际更迭持续变化，通算场景追求高密度部署与 QoS 均衡，智算场景强调拓扑感知与资源发放效率，Agent 场景则要求极低调度时延与超高并发吞吐，三者的系统设计目标显著不同甚至相互矛盾——高密度部署依赖细粒度切分与强隔离，而极致吞吐则需要轻量



化调度与快速路径分发，基于单一架构无法同时满足全场景对资源效率的差异化需求。未来，下一代资源调度与编排系统正在朝着多模态感知、分级调度与 AI 驱动优化的方向演进，通过构建统一的资源抽象层屏蔽底层异构差异，以分层架构实现控制面与数据面解耦，并引入强化学习等智能算法实现全局资源效率的动态寻优，在保障各场景关键指标的前提下实现全场景资源效率最优，进而将云基础设施的整体资源利用率提升数倍。

技术特征三：分时复用的智算算力切分机制

智算场景算力销售通常是把单卡或多卡独立分配给单一租户，但随着单卡算力和显存逐代增强，对于轻量任务（如推理、开发测试）或大小模型混合部署的场景将存在大量的资源闲置和浪费，基于传统的硬件静态切分技术或 API 拦截分发技术，无法同时满足场景对算力灵活分时切分和安全隔离的需求。未来，下一代云智算的算力切分技术正在结合云基础设施平台以及推理训练框架，朝着既能保障云上多租户与多业务间的性能独立与安全、又能实现毫秒级的算力调度的方向发展，进而数倍提升云智算的算力资源利用率。

构建面向智能体的平台服务，为规模化智能体提供可靠、安全、可持续优化的技术底座

随着智能体数量的爆炸式增长，智能体平台将从开发、编排和运行支撑为主的“辅助型工具链”，全面演进为面向规模化智能体的平台服务。下一代智能体平台不仅需要提供高效的任务执行与资源供给能力，更需要围绕持续优化、长期记忆、安全治理和智能调度构建系统级技术底座，实现智能体从“可运行”向“可自优化、可持续记忆、可安全治理、可全局调度”演进，为海量智能体提供可靠、安全、高效且可持续演进的运行环境。基于此，下一代智能体平台将呈现四个主要技术特征：

构建全栈自优化系统，持续提升智能体任务执行的准确率、稳定性和完成质量

智能体的优化，传统流程高度依赖人工介入，从异常案例收集、根因分析、到配置修改与重新部署。然而随着企业智能体数量每增加一个数量级，人工介入成本呈指数级上升。新一代智能体优化系统将沿着记忆自优化、Skill 自优化、模型自优化三个方向演进。

记忆自优化按热度分层，辅以重要性评分与时间衰减曲线，让数据按“价值密度”流动；同时，将成功路径压缩成可复用的“经验单元”入库，新旧记忆语义冲突时触发重要性再评估。

Skill 自优化通过遥测或真实线上链路回流分析，理解调用图、构造参数化模板，将组合操作流程自动合成，沉淀为新的“原子技能”。

模型自优化将智能体真实运行轨迹作为训练语料，以用户反馈、工具调用结果、任务完成度构成多通道奖励信号，模型在沙箱中持续更新，形成在线强化学习与基于 AI 反馈的强化学习闭环。

上述优化均依赖智能体的真实运行轨迹，

更多免费行业报告 并购家 www.ipoipo.net

通过持续采集与迭代形成数据飞轮，使智能体在准确率、稳定性与任务完成质量上持续提升，同时将优化响应周期从天级缩短至小时级。

打造多模态融合记忆系统，强化智能体处理复杂任务时的记忆表现

记忆系统是智能体具备持续能力的关键。当前主流方案以基于语义相似度匹配的向量检索为主，在多跳推理与关系型查询场景下存在明显局限。下一代记忆系统需向融合向量、标量、全文与图结构的多元架构升级，提升访问性能并节省 Tokens 消耗。通过对“人”与“组织”等实体及其关系的建模，支持跨团队跨 Agent 的群体记忆；并对每条记忆标记明确时间区间，以支持精确回溯等记忆事务管理。最终使智能体在复杂业务场景中的记忆检索精度与响应速度均得到显著优化。

建立统一治理框架，保障智能体稳定安全运行，提升业务成效

面向未来，智能体治理需从安全、运维、运营三个层次构建统一框架。

基础安全层聚焦于全链路风险防控，融合意图护栏、执行隔离与安全边界三重机制。首先，在入口和工具调用全过程实时解析意图，拦截恶意指令并过滤敏感数据；其次，将任务隔离细化至子意图级，每个工具（Skill）在独立沙箱中运行；最后，对每一次调用和协同均实施最小权限与最小网络开放约束。

运维层提供全链路因果可观测能力，将推理、决策与工具调用转为可查询的知识图谱，支持业务指标的快速提取、异常识别及根因自动回溯。

运营层在安全与运维能力的基础上，扩展



成本优化与反馈跟踪等运营场景，使平台在稳定运行、安全合规的前提下持续改善业务效果与资源效率。

构筑面向任务 SLO（服务等级目标）的 Agent 通智调度引擎，支撑智能体整体效率达到全局最优

不同类型任务对 SLO 的诉求差异显著。交互式任务对响应延迟极度敏感，而定时批量任务则更关注成本消耗。面向未来，平台引入面向任务 SLO 的通智调度引擎，根据任务类型与历史执行轨迹，动态调度智算与通算资源：在模型选择上，考虑 KVCache 配置前提下，自动匹配合适的强弱模型；在执行环境上，基于任务意图识别预判所需资源，综合运用就绪缓冲池分配、环境状态快照恢复与按需加载等策略，提前完成子智能体与工具的依赖准备与加载，消除冷启动等待；在工具调用上，按必要性自动选用付费或免费工具，避免非必要开支。通过上述多维度的协同调度，该引擎能够在满足任务 SLO 要求的基础上，使整体执行效率与资源利用率达到全局最优状态，为大规模智能体应用提供高效、经济、可靠的运行支撑。

智能的应用和落地：重塑个人用户体验和企业生产模式，实现 Token 的高效应用

当 Token 能够被高效生产、灵活运载，并依托云与智能体基础设施实现规模化调用，智能的价值将进一步从底层能力向应用端释放。对于个人用户，智能将突破传统的人机交互方式，以更加自然、主动和无感的方式融入生活与工作；对于企业，智能则将深入业务流程和生产经营的各个环节，推动人与智能体、智能体与智能体之间形成新的协作模式，重构企业的生产与运营模式。

人和智能体共生，终端智能全面走向全场景智能

智能终端将成为智能革命走向现实世界的核心载体，人与智能体也将从简单的人机协作走向深度共生。未来，人机交互将从“点击 App、执行指令”转向“表达意图、智能体办事”，智能体将通过算力、智力和终端的技能树，自主拆解任务，实现人的意图达成。同时，以人为中心的全场景设备，将在人、车、办公、家等空间下形成不同的能力集合，并朝着空间智能体的方向演进，变成人的 AI 生活助手或 AI 机器管家，提供主动智能服务。

个人移动空间：摄像头将成为空间智能体的“眼睛”，麦克风是“耳朵”，扬声器将成为空间智能体的“嘴巴”，智能手表等穿戴设备则让空间智能体能够感知人的身体状态，端云协同的大脑将这些分布式的“具身”以及每个终端的“技能树”粘合起来。它拥有你的记忆，了解你的身体，感知你身处何处，并理解你的意图，成为你形影不离的 AI 生活助手，为你提供便捷的主动智能服务。

车空间：车载摄像头、毫米波雷达、UWB 传感器、麦克风阵列、扬声器组成强大的智能汽车感知系统，结合手机、手表、眼镜将进一步形成更全面的感知能力，车空间全场景设备构成的空间智能体，将为人带来前所未有的智能体验。

办公空间：电脑、平板、手机、手表、眼镜、伴随机器人等全场景设备无缝组成另一个新的群体智能集合，不仅能处理日程提醒、会议帮记、文档编辑、AI Coding 等，成为你的生产力工具，更是你搜索、洞察、思考、创作的智慧助手，也是 24 小时能够帮你完成工作任务的“下属”。



更多免费行业报告 并购家 www.ipoipo.net

家空间：一个由全屋智能设备、手机、电脑以及具身智能机器人等终端共同构成的智能生态系统，它的核心是 AI 机器管家。这位管家的“感官”无处不在：摄像头看清家中人员和宠物动向；麦克风捕捉儿童的啼哭、长辈的呼唤、宠物的叫声；家空间所有智能设备，也全部是它的“手脚”。它不仅能认识家人和宠物，基于意图和视觉感知实现家居控制，更能帮你干活，例如识别到地面脏污自动指挥扫地机器人清扫，识别到下雨自动关闭窗户，识别到人离开家后自动关闭空调灯光。这个 AI 机器管家拥有你全家人的长时记忆，同时也装载在具身智能机器人的脑子里，可以走出家门，陪你逛街、散步、办事。

围绕上述智能空间，终端将在脑子聪明、感知发达、技能丰富、隐私与安全四个方向全面革新升级，从智能终端跃迁为 AI 泛在的全场景智能。

脑子聪明：终端 AI 算力架构走向端云协同 + 端侧自治，端侧 AI 算力将爆发式增长

端侧自治是终端解决 Token 成本可承担、任务闭环时延可控和隐私与安全可管理的必然选择。智能终端本地算力，将轻量化、常规化的交互与任务处理全流程在端侧自主完成，实现任务自主闭环，摆脱对云端的过度依赖，让交互时延、任务响应速度全程可控，大幅提升智能体验的流畅度与稳定性。而针对超长文本解析、深度复杂逻辑推理、图文音视频多模态融合处理、跨设备联动协同等终端难以独立承载的高复杂度、高算力消耗任务，将需要采用端云算力协同一体化运行的架构来完成。本地终端完成数据筛选、格式规整、需求预判等前置预处理工作，精简无效数据、锁定核心任务需求；再依托云端超强算力和更大参数模型进行高速运算、全局资源调度与复杂逻辑推演；最后回流至终端完成结果优化、本地化适配及



落地执行等后置处理。这套端云算力协同机制，既兼顾了复杂任务的处理精度与效率，又能精准压缩云端无效算力损耗与 Token 调用开销，实现算力资源最优配置、成本与体验的双向平衡。

模型入端将成为终端的标配：瞄准细分场景需求的轻量化、高适配的小模型，将是入端模型发展的关键方向。针对影音娱乐、智慧出行、智慧办公、运动健康、智能家居等场景做精准优化，剔除冗余参数、压缩模型体积，在适配终端有限算力与功耗的同时，保障场景专属任务的运行精度与响应速度，让终端无需频繁依赖云端调用，即可独立完成高频、专属、轻量化智能任务，真正实现终端智能能力的本地化、常态化、高效化落地。

端侧算力将爆发式增长：终端侧 NPU、TPU、LPU 及各类独立 AI 算力模块将迎来算力爆发式增长，解决模型入端的算力基础。存算一体架构、异构计算技术凭借高效的算力调度、数据处理优势，将成为现阶段端侧 AI 性能突破的核心路径。在实际落地应用中，端侧 AI 算力的部署将以精度、时延、功耗、成本四大核心维度为核心评判标准，结合不同终端的硬件规格、使用场景、功耗限制和商业化定位，制定差异化、精细化的算力部署与适配策略。其中，智能手机、PC、智能汽车、家空间四大终端场景，凭借庞大的用户基数、高频的智能交互需求、充足的硬件升级空间，依旧是当前及未来端侧 AI 算力落地迭代、规模化爆发的核心主阵地，持续承载各类轻量化、实时性 AI 智能任务，支撑全场景端云协同群体智能体系的落地运行。

长时空记忆加持的持续学习是脑子聪明的关键：AI 智能体将不仅能够识别用户身份，还将构建起覆盖超长上下文对话、用户长期行为习惯、全场景空间感知的长时空记忆体系，彻底突破当前 AI 短时记忆、单次交互、无场景积淀的局限，将提升决策的准确性和及时性。智能体精准识别个人移动空间、车空间、办公空间、家空间等不同物理空间的环境特征、设备布局与场景属性，形成物理世界的长时空记忆数据库。依托长效且全面的记忆积淀，智能体能够深度理解用户潜在需求、适配不同空间下不同场景的使用逻辑，实现更精准、更贴合用户习惯的自主决策，有效提升智能服务的响应及时性、适配准确性与个性化程度。

感知发达：多模态融合感知系统增强，是智能体理解人的意图的基础，驱动终端硬件能力持续增强

传统单一、碎片化的感知方式已无法适配自然人机交互需求，随着视觉、听觉、身体感知等多维度感知技术的深度融合与统一建模，终端硬件的环境感知、用户感知、场景感知能力全面跃升，为智能体读懂人、读懂场景、主动服务提供了扎实的数据与技术支撑。

听得清：融合听觉、视觉的多模态感知算法与智能化建模推理能力，同一空间下的全场景设备构建连续融合的感知能力，让空间智能体拥有“耳朵”。系统可精准区分人声与环境背景噪音，完成人声分离、声源定位、身份识别、场景环境建模，既能清晰听懂用户的语音指令，也能精准判断当前发言人及所处环境状态，有效规避杂音干扰，告别“聋子”。

看得见：通过视觉感知、雷达探测、毫米波传感等多类感知硬件的协同融合，搭建统一

更多免费行业报告 并购家 www.ipoipo.net

特征空间建模体系，搭配多模态智能推理能力，同一空间下全场景设备构成的空间智能体将拥有全方位、高精度的“眼睛”，可实现对物理空间、环境物体、设备状态、用户动作的全方位识别与立体感知，做到看得清、看得全、看得准。

情绪感知：穿戴设备感知人的身体体征数据，融合视觉、听觉等多维度感知数据，通过多模态融合推理与情绪建模算法，让全场景设备终端在空间下具备高阶情感感知能力。系统可实时捕捉用户的面部表情、肢体体态、语音情绪等细微变化，精准判别用户的喜怒哀乐、疲惫、放松、烦躁等情绪状态，打破传统 AI 无情感、无温度的交互短板，实现人性化、情感化、有温度的智能交互，让服务更贴合用户当下心境与状态。

意图理解：基于上述构建的多模态感知系统，围绕人的智能终端在空间下拥有了“耳朵”、“眼睛”和“情绪感知”，叠加长时空记忆能力的深度赋能，智能体实现感知能力的跨越式升级。系统不再局限于简单的识别人、识别动作，而是能够深度读懂用户行为逻辑、生活习惯、活动规律与场景诉求，精准理解用户的潜在意图与真实需求，实现从“被动应答”向“主动服务”的智能跃迁。

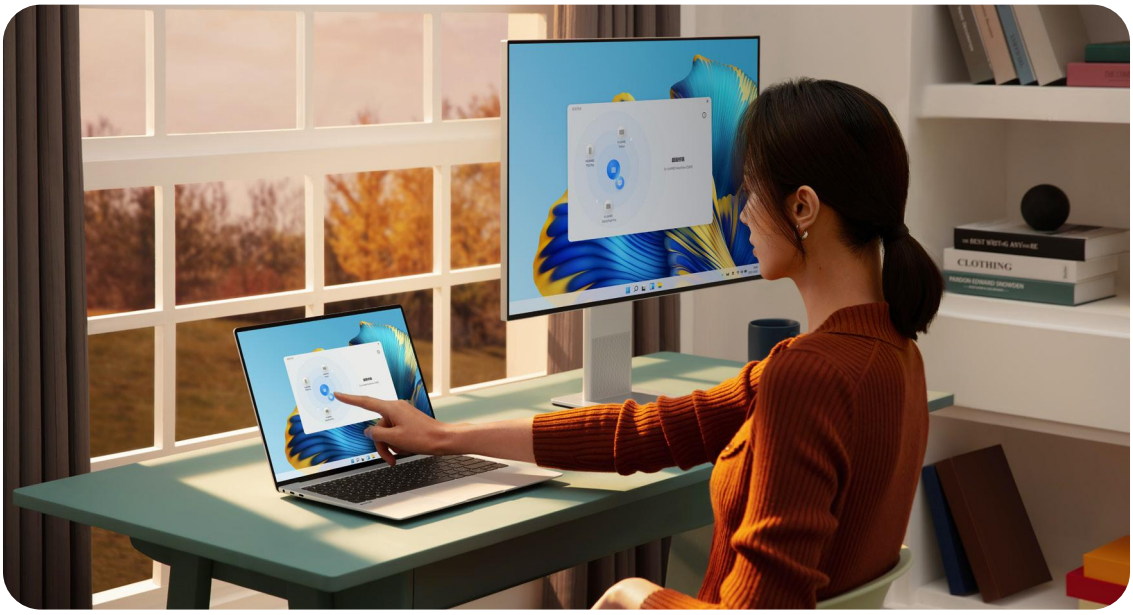
技能丰富：Agent OS 和生态协同，将提升长时主动策划任务的协作能力，提高 AI 执行能力的上限

要实现终端智能体强大的“技能树”，操作系统必须全面革新为 Agent OS，这是突破智能体技能上限的基石；同时平台要搭建友好的南北向 AI 生态体系，以简单、高效、低成本的接入通道，将北向应用、南向设备的原子化功能和技能接入到 Agent OS，供智能体无缝调用。这两点直接决定了智能体能否开展长时主动策划任务，也决定了 AI 执行能力的天花板。

革新终端 Agent OS：终端操作系统将完成从传统 OS Agent 辅助工具形态，向原生 Agent OS 智能操作系统进化，底层架构、运行机制、交互逻辑将全面重构。在传统点击式、指令式的被动交互范式的基础上，以用户意图为核心的自然交互、主动服务范式将成为主流，实现人和智能体共生。在全新的系统架构下，AI 智能体不再是依附于系统的应用插件，而是成为系统数据调度、资源调配、服务调用的核心主体，提升长时主动策划任务的协作能力，有效提升智能体做事的“技能树”。

南北向生态协同：软硬件生态将完成全方位 AI 适配与 Agent 化改造，构建双向贯通、全域协同的智能生态体系。面向用户的北向应用生态，将通过 AI 友好架构实现从交互入口到核心功能的全面 Agent 化重构，原生适配 AI 意图交互与自主任务执行，提升全链路 AI 亲和性。面向硬件的南向设备生态，将各类终端的硬件能力拆解为标准化、可调用的原子化功能，统一对接操作系统底层架构并完成 AI 适配调用。依托双向生态打通，智能体可自由跨设备、跨应用、跨空间调取数据、资源与服务能力，打破设备壁垒、应用孤岛，实现全域资源高效联动，提供场景化服务。

多设备互联互通：同一空间下的全场景设备，为智能体和人提供技能组合和群体智能，万物互联的操作系统催生强大的设备群体智能



体系，组合成空间智能体。以 HarmonyOS、OpenHarmony OS 为代表的万物互联操作系统，具备天然的分布式协同能力，无论终端设备算力大小、形态差异，均可实现自动组网、自动感知、无缝互通、能力共享、协同运算。这种原生协同特性，让碎片化的终端能力聚合为群体智能，大幅拓展智能体的服务边界与能力维度，随着终端生态超大规模发展，从联接到智能的优势将持续放大，成为全场景智能的核心支撑。

AI 走进物理世界的步伐永不停歇。手机、家居、车载、运动健康各类智能终端，将依托全场景设备融合能力与群体智能优势，加速向空间智能体方向迭代进化，成为用户日常工作生活中不可或缺的 AI 生活助手和家庭的 AI 机器管家；具身智能机器人将整合算力、智力、感知系统、灵巧手等能力，完成软硬件一体化升级，发展成为新的场景化 AI 终端，满足更多潜在的人类需求，成长为未来新的万亿产业，带领人类迈向一个全新的智能时代。

隐私与安全：需要面向智能体访问 OS 设计隐私和安全规范，全方位保障用户数据与设备使用安全

伴随 Agent OS 的革新升级，设备操作系统的服务对象已经从传统单一的人这个使用者，转变为人 + 智能体双运行主体，传统只围绕人的行为搭建的数据风控、权限管控、隐私防护与系统安全规范，已经无法适配智能体自主调度、跨技能联动、自动执行用户意图的全新运行模式，整套安全框架、监管逻辑、运行约束机制需要全方位重构。

端侧优先：隐私与数据安全是人工智能规模化落地、常态化使用的绝对底线，整套智能体系严格遵循端侧优先的安全原则，最大限度

更多免费行业报告 并购家 www.ipoipo.net

保障用户数据安全。用户各类行为数据、交互数据默认留存本地、不出设备，从源头杜绝隐私安全风险。针对身份信息录入、健康监测、金融交易等高度敏感场景，系统强制执行本地运算、本地存储、本地闭环的运行机制，全程不向云端传输原始敏感数据，彻底规避云端数据泄露、抓取、滥用的安全隐患，筑牢全场景智能服务的安全根基。

数据可控：终端所有本地留存的用户行为数据、场景数据、个人隐私数据均采用全链路加密存储机制，依托端侧安全芯片与加密算法实现数据闭环防护。针对用户个人数据实行分层分级精细化管控机制，根据数据敏感等级、使用场景、服务需求划分不同权限等级。在用户授权环节，严格遵循最小权限原则，仅开放当前任务必需的最小颗粒度数据权限，杜绝过度授权、超范围数据调用、隐性数据采集等行为，让每一次数据使用都精准可控、有据可依。

交互透明：Agent OS 需要始终坚守“人拥有绝对数据控制权”的核心准则，重构 AI 数据使用的设计范式，对智能体的数据调用、权限申请、资源使用行为进行明确且刚性的约束。在智能体和机器交互的过程中，数据采集范围、数据使用用途、数据流转链路、权限开启状态全部要实现可视化、透明化展示。通过直观易懂的交互设计，让用户清晰知晓智能体正在调用哪些数据、用于何种服务、权限有效期时长，保障用户对个人数据的知情权、使用权、处置权。

权限可控：基于智能体丰富的设备技能与跨场景服务能力，系统对所有终端技能、智能服务实行分层分级权限管理体系。针对涉及用户隐私、用户协议、资金安全、人身安全等高敏感行为，强制要求经过用户主动交互确认后方可执行，杜绝智能体自主触发高危操作。同时，将主动鉴权、协同鉴权作为端侧隐私安全设计的核心关键，通过设备间协同校验、用户主动授权确认、系统安全机制多重校验的方式，构建多维度、立体化的安全鉴权体系，全方位保障用户数据与设备使用安全。



AI 驱动汽车迈向“移动具身智能体”，安全与体验双重跃升

汽车产业正经历从电动化向高阶智能化的深层演进，技术迭代速度显著加快，辅助驾驶加速迈向自动驾驶，“AI+ 汽车”大幅提升安全和用户体验，产业生态与商业模式正在改变，新的纪元已经到来。

自动驾驶规模商用时代正在来临，AI 驱动汽车从交通工具迈向移动具身智能体

高阶智能化在感知决策、多模交互、协同控制等方面正在重新定义汽车的功能边界与产业价值。基于先进算法模型，辅助驾驶已具备深度感知与复杂决策的自我进化能力，实现人驾更安全、辅助驾驶更舒心、泊车更省心的体验闭环；多模态大模型的规模化应用正在重塑人机交互范式，推动系统从被动响应迈向主动服务，座舱演变为万物互联的移动智能空间，更懂用户、更有温度；集中计算与区域接入的整车架构日益普及，软硬解耦、跨域协同，软件定义汽车加速落地，软件订阅的商业模式已成为行业趋势。

随着 AI 能力不断成熟，自动驾驶从乘用车向干线物流、支线配送、环卫、矿山、港口等全品类载运工具延伸；应用场景从限定范围向更复杂的城市道路甚至非结构化道路深度拓展，载人、载物的无人化进程持续加速。

商业化落地也正在从单点示范走向规模应用。在载人领域，Robotaxi 已从示范运营走入规模化商用的快车道；载物领域，无人配送、无人重卡等因低使用成本已实现商业闭环。载人、载物场景的多元，最终将构建起海陆空一体的三维立体交通新生态。

但从“智能汽车”真正迈向“移动具身智能体”，仍需要解决一系列底层能力与产业化难题，核心集中在四个方面：

挑战一：模型规模膨胀与算力 / 功耗之间的冲突。端到端大模型与多模态交互的应用，使参数量、数据量与训练成本持续攀升，车端与云端算力成本、能耗压力激增。如何在安全、成本与功耗约束下实现大模型高效部署与实时推理，是汽车智能化迭代的首要瓶颈。

挑战二：长尾场景难以遍历。极端天气、突发路障、行人异样行为等长尾场景无法被数据覆盖穷尽，决策失误的风险长期存在，这仍是从辅助驾驶走向全无人化的重大难题。

挑战三：网络安全与隐私保护难度急剧增大。智能汽车已成为移动终端，车载传感器与云端数据频繁交互，控车功能日益复杂，攻击面持续扩大。从数据泄露到恶意控车，网络安全与隐私保护必须在汽车全生命周期加以重视并解决。

挑战四: 趋严的法规使合规成本持续上升。自动驾驶准入、数据跨境传输等规则不断收紧，导致车企在产品智能化与国际化进程中，拉长了车型上市周期，技术适配与合规成本增加。

更多免费行业报告 并购家 www.ipoipo.net

“AI+ 汽车”助力高阶智能化快速落地，让出行更安全、体验更卓越

智能汽车迈向移动具身智能体，必须具备阶段性商业落地能力，在自动驾驶、智能座舱、智能车控、安全体系与生态之间形成协同闭环，给用户带来更安全更高效的出行体验。

自动驾驶：技术范式加速向世界模型演进，推动自动驾驶商业闭环

世界模型赋予 AI 对物理世界的深度理解与空间场景的想象力，使自动驾驶系统能像人类司机一样，通过观察现实道路环境，自主归纳并理解物体运动规律、天气变化和复杂的交通博弈，更准确预测潜在风险，提前采取减速、避让等防御性措施。

车端采集的海量数据持续喂养世界模型，随着高质量数据积累、模型泛化能力提升以及数据闭环，系统对复杂场景的处理能力将持续增强。当系统在特定运行设计域内达到可验证、可监管的安全水平，Robotaxi、无人物流等场景才能扩大运营范围，实现可持续的商业闭环。

智能座舱：AI Agent 驱动智能座舱进化为主动服务的第三生活空间

AI Agent 快速融入座舱，使人车交互从单模式控制走向集语音、手势、视线、情感于一体的多模交互，多模态大模型不仅能理解用户言谈举止，还能精准捕捉用户情感，实现对舱内环境和用户状态的深度感知，完成从被动响应到主动预判的服务跃迁。

基于 AI Agent 自主思考与推理，结合用户的驾驶习惯、生物体征及交通环境等信息，通过接入第三方服务，智能座舱能主动为用户提供贴心的出行解决方案。智能座舱还具备自进化能力，不仅越来越懂用户，更能通过云端模型的持续迭代，实现功能的自我生长，最终使智能汽车变成一个有温度、与用户共同成长的专属移动生活空间。

智能车控：构建整车级端到端控制架构，满足自动驾驶的高冗余要求

智能车控正加速向高带宽、低时延的中央计算 + 区域接入的架构演进，以满足高阶自动驾驶毫秒级响应和极致安全需求。新架构打通自动驾驶域与数字底盘域的一体化控制链路，实现从智能感知到执行机构的协同冗余控制。当主控制链路失效，系统将毫秒级切换至备控制链路，确保汽车单点故障依然能靠边停车或人为接管，甚至驶向预期目的地，构建全栈冗余的安全底座。

线控部件逐步上车，驱 / 制 / 转 / 悬控制算法上移，AI 赋能智能底盘协控算法与故障预测，实现底盘异构安全备份，持续开发新的操控特性，适应更复杂行驶环境；提升云端与车端的智能化协同能力，给用户带来极致的安全与驾乘体验。

安全体系：六维覆盖，迈向全场景主动安全新时代

告别被动防御，聚焦主动安全。车企正全面构筑五大安全屏障：确保计算与通信架构的功能安全、解决认知局限的预期功能安全、保障人机可靠协作的交互安全、防范大模型幻觉与漏洞的 AI 安全，以及抵御恶意攻击的网络安全。在此体系支撑下，智能部件将深度赋能整车，实现全时速、全方向、全目标、全天候、全场景与全时域的六维核心安全目标。安全体系的落地，为高阶自动驾驶构筑起不可逾越的安全底座。



以安全为底座：从事中安全走向全时域安全

产业生态：智能汽车的全生命周期价值创造

智能汽车已成为数字世界的移动智能终端。头部车厂正通过整合资源，构建包括停车、充电、保险、维保、二手车交易等的闭环生态，为用户提供“用车无忧”的全旅程体验，推动车厂从汽车制造商向智能出行服务提供商转型，实现软件服务、生态运营、数据增值的全生命周期价值创造。

高阶智能的终局是移动具身智能体，行业必需在技术、安全、法规上同步突破

高阶智能的终局不是单点能力的叠加，而是打造可规模化运行的移动具身智能体，深度理解环境、可靠执行任务，并在合规约束下持续进化。为此，必须在三条战线上同步突破：

技术栈：推动模型能力与工程效率同步突破

大模型是汽车智能化的核心引擎，仅追求参数规模，对车端算力与存储都不可持续。汽车对功耗、成本和实时响应要求严苛，这就需要全行业在模型轻量化、分布式训练、车端推理效率、数据合成与仿真等工程层面同步突破。只有通过极致的工程能力，将大模型压缩并高效部署在车端，才能以可控成本支撑起汽车智能化能力的持续迭代。

安全栈：安全是商业化落地的先决条件

从辅助驾驶向自动驾驶的跨越，责任主体将改变，系统对安全的要求已上升到不允许发生任何致命事故的高度。在自动驾驶的方案策划与定义阶段，安全就必须作为顶层设计的核心。为了构筑这道不可逾越的安全防线，行业及车企必须共同推动汽车安全的系统级融合设计，端到端筑牢智能汽车安全防线。

法规栈：合规将成为全球化企业的核心竞争力

全球数据跨境、隐私保护、自动驾驶监管政策的收紧，要求车厂在智能化方案立项阶段，就将本地法律法规的底线纳入整车架构设计，同时应携手行业伙伴，积极参与目标市场法律法规与标准的制定，才能从根本上化解准入风险，实现企业国际化发展战略。

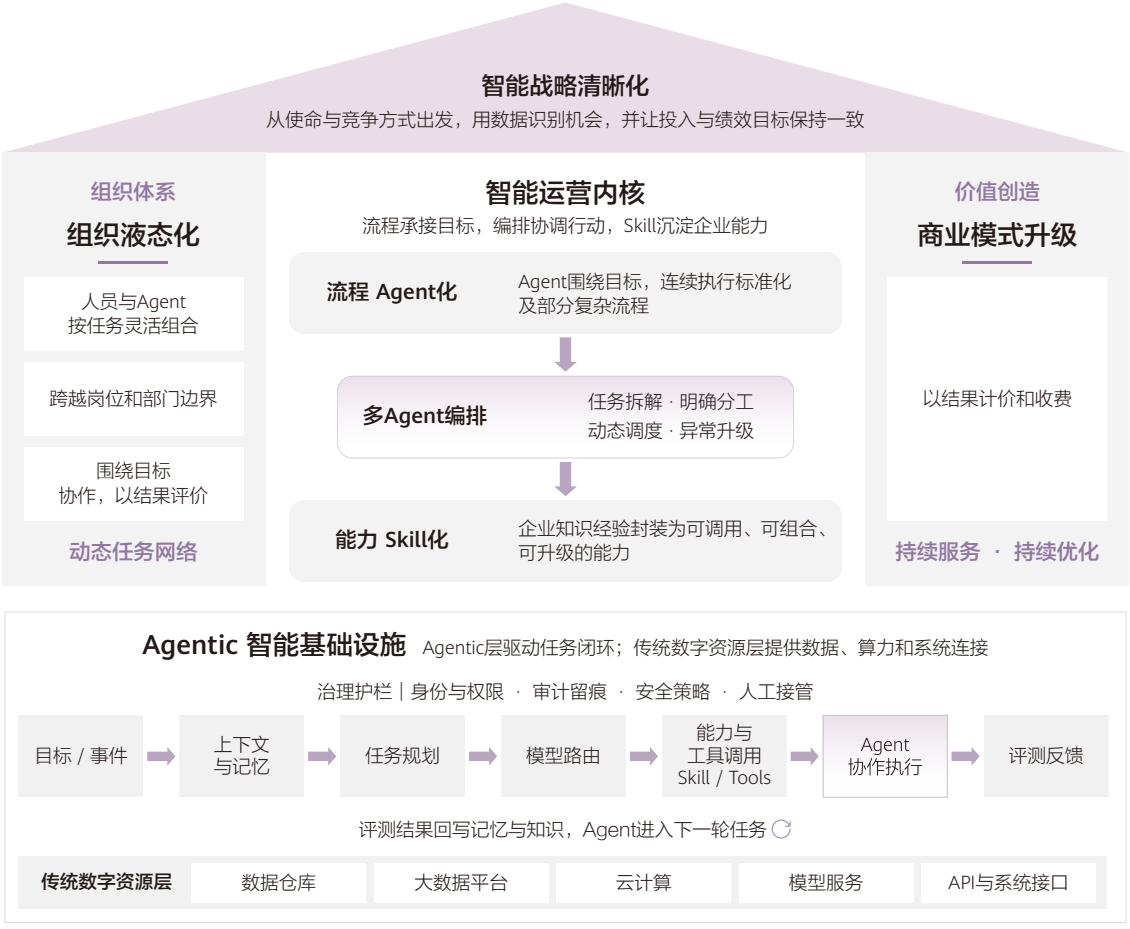
更多免费行业报告 并购家 www.ipoipo.net

Agentic 企业应用：从单点智能走向原生智能，实现企业级智能规模化落地

随着 Agent 加速进入企业核心业务流程，企业智能化的竞争焦点也由单一模型能力和应用数量，逐步转向能否将 Agent 与企业数据、业务系统、组织机制和专业能力深度融合。而这一趋势也将进一步推动企业自身架构发生系统性重构，催生出全新的组织形态——“智能型企业”。

智能型企业：战略、组织、流程与能力全面重构，实现企业智能原生化

智能型企业的“智能”，内生于企业的战略、组织、流程、能力、商业模式和基础设施，实现全要素生产率跃升，推动企业敏捷适应环境变化。



智能型企业的架构图

流程 Agent 化：从“人驱动流程”转变为“人设定目标，Agent 围绕目标端到端自主执行”

智能型企业的业务流程不再是一条由人逐节点推动的固定 SOP，而是转变为“人设定任务目标，Agent 自主拆解并执行，过程中持续感知并动态调整”的智能执行系统。首先，由人明确业务目标、约束条件和执行边界；随后，主 Agent 围绕目标自主制定执行计划，将复杂任务拆解为多个子任务，并分配给不同专业 Agent 并行处理。各 Agent 在执行过程中可连续调用 CRM、ERP、数据库、邮件、代码工具等业务系统，完成信息获取、分析判断与操作执行，并将结果持续汇总至统一目标之下。当外部条件变化、执行结果偏离预期或出现异常时，Agent 能够识别问题、重新规划路径、调整任务优先级并重新调度资源；只有超出授权边界或涉及重大风险的异常事项，才升级至人工处理。由此，人的角色逐步从“流程节点执行者”上移为目标设定者、规则制定者、异常决策者和最终责任承担者，而 Agent 则成为围绕目标持续规划、执行和优化的主要流程主体。

能力 Skill 化：将经验沉淀为可调用、可组合、可评估和持续升级的 Skill，使企业能力能够像软件模块一样复用

智能型企业将原本依附于个人、岗位和部门的专业知识与方法论，逐步沉淀为可调用、可组合、可评估并持续升级的“Skill”，从而实现企业能力的模块化和资产化。Skill 需具备三项能力：可调用，可被不同 Agent 按需调用；可组合，多个原子 Skill 可编排形成复杂任务链；可评估，围绕任务完成率、准确性、稳定性、时效、成本、工具调用成功率、人工介入率及合规性等指标持续评测。最终通过业务反馈和评测结果不断迭代，实现企业能力的模块化、资产化和持续升级。

系统智能化：数据、模型、算力和业务系统具备自主感知、动态调度、故障恢复和资源优化能力，为 Agent 规模化运行提供智能企业底座

系统智能化是智能型企业规模化运行 Agent 的底座。智能型企业持续加强算力基础设施投入，形成可弹性扩展、动态调度和故障恢复的计算能力；同时，将垂直行业长期积累的业务数据、知识和场景数据沉淀为核心智能资产，持续提升模型与 Agent 的专业化水平。对于金融、医疗、制造等高敏感行业，通过本地化或私有化部署，将模型、数据与核心业务系统部署在可控环境中，并结合权限管理、数据隔离、审计追踪等机制保障隐私与安全。最终，数据、模型、算力和业务系统形成统一、可调度、可信赖的智能基础设施，为 Agent 规模化运行提供支撑。

组织液态化：人与 Agent 突破固定岗位和部门边界，围绕任务动态组合、按目标协作

智能型企业的组织形态从“固定岗位、固定部门、固定汇报关系”转向围绕任务与目标动态重组的“液态组织”：在企业组织架构上，随着 Agent 承接大量信息传递、任务协调与标准化管理工作，传统层级和部门边界被弱化，组织架构更加扁平；在具体项目上，企业围绕具体项目和业务目标，灵活组合业务人员与不同专业 Agent，并根据任务进展、资源约束和异常情况动态调整分工。

智能型企业不仅要具备新四化，同时还需要明确的智能化战略与与之匹配的商业模式升级：

智能型企业将智能化作为企业级战略能力，并以此重新定义企业的竞争逻辑与资源配置方式。当智能化被嵌入企业战略制定、业务运营

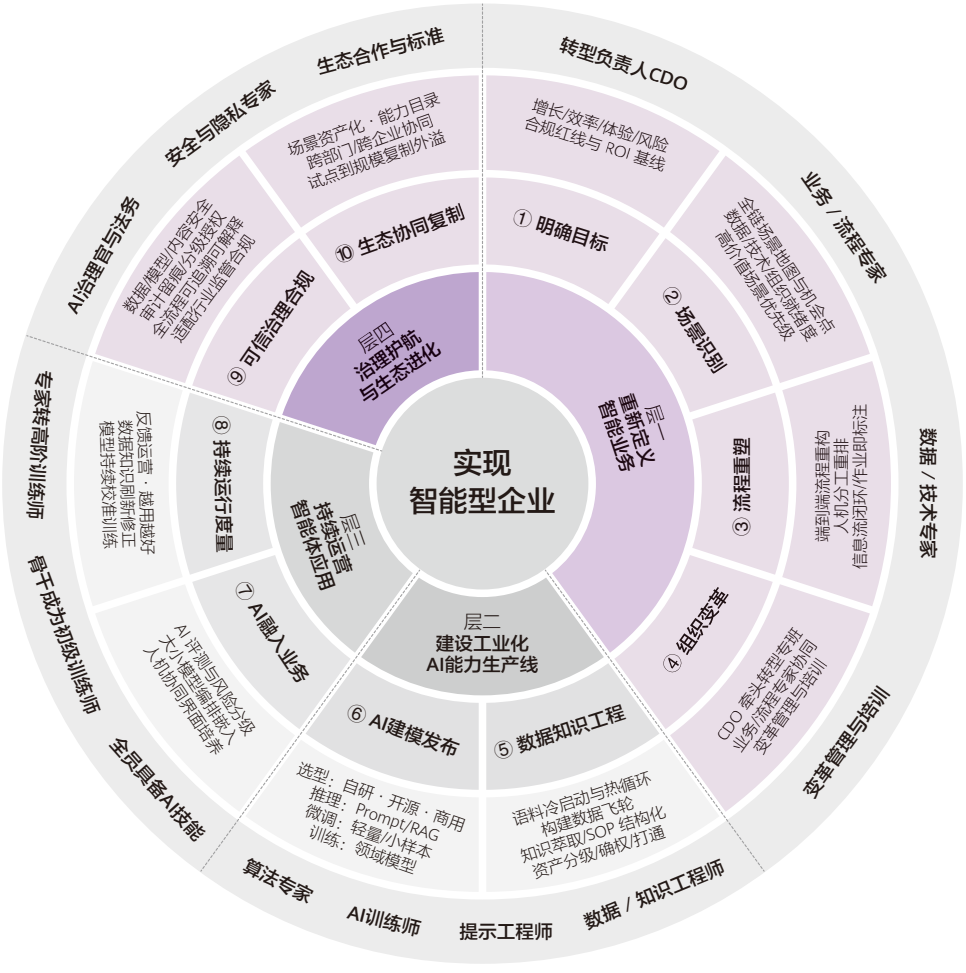
更多免费行业报告 并购家 www.ipoipo.net

与组织管理的全过程后，AI 才会从局部效率工具演化为企业持续获取信息优势、决策优势和执行优势的基础能力，最终形成难以被简单复制的长期智能竞争壁垒。

在 Agentic AI 时代，企业交付对象逐步演变为由“产品 / 软件底座 +Agent 能力 + 数据闭环”共同构成的持续智能服务：企业的商业模式也从“卖产品、卖 License”转向“以结果计价和收费”，客户购买的不再只是某一时点的功能，而是一套能够伴随业务环境变化持续学习、持续执行和持续优化的智能系统。

智能化转型路径：以“四层十步法”重构企业能力体系，加速迈向智能型企业

企业智能化不是一次技术升级，而是一项贯穿业务、流程、组织、数据、技术与治理的系统工程。基于对各行业企业智能化转型实践的观察，我们将智能化落地的方法论归纳为“四层十步”：以四个层次界定转型的推进纵深，以十个关键步骤界定可执行、可检验的落地动作。



智能化转型“四阶十步法”

第一层：重新定义智能业务，形成有优先级、有边界、有责任人的智能业务蓝图

重新定义智能业务，本质上是回答“AI 优先改变什么业务、创造什么价值，以及业务应该因此变成什么样”，最终形成一份有优先级、有边界、有责任人的智能化转型蓝图。



重新定义智能业务

在智能化转型蓝图中，流程重塑是其中核心的价值杠杆，起到了关键性作用。据沙利文研究显示，重新设计工作流与 AI 价值释放具有明显相关性：重新设计工作流的组织实现企业价值增长的比例约为 37.2%，而未改变工作流的组织仅为 5.4%。真正决定智能化成效的并非“有没有 AI”，而是企业是否围绕 AI 重新定义了工作流。

某全球性智能电力管理企业与一家技术服务商合作，重构了供应链短缺管理。该电力企业拥有 300 多家工厂、超过 10 万笔日常销售订单、3200 多万种零部件以及 70 多个 ERP 环境，传统模式下，供应链人员需要跨系统寻找短缺信息、判断影响，再依靠个人经验决定优先级和解决方式。技术服务商并没有简单给供应链团队增加一个“AI 助手”，而是重新定义了整个短缺管理流程：首先统一不同 ERP 中的供应链数据，形成供应链本体；其次让系统自动识别物料短缺，并按照其潜在收入影响进行优先级排序；进一步结合历史处置经验，由智能化平台自动提出替代物料或解决建议，最终将人的工作重点从“搜集数据、逐项排查”转变为“判断关键例外、确认高价值决策和处理复杂异常”。这一工作流帮助该电力企业实现了约 25% 的生产率提升。

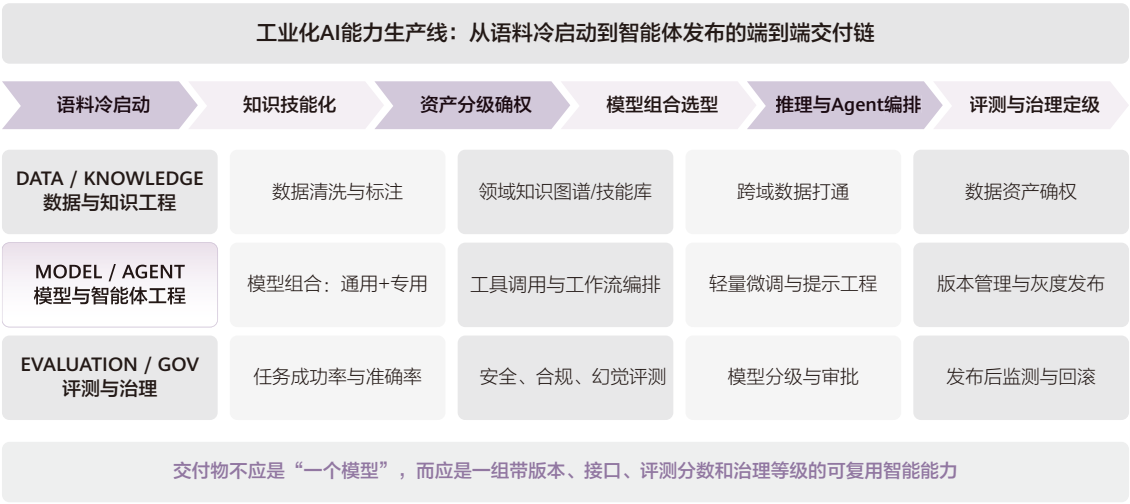
第二层：构建工业化 AI 能力生产线，推动智能能力复制升级

随着智能化从场景试点走向规模化落地，企业竞争的关键开始从“能否做出一个 AI 应用”，转向“能否持续、稳定地生产 AI 能力”。面对跨场景复制、持续迭代与统一治理的需求，传统以项目制开发为主的模式难以支

更多免费行业报告 并购家 www.ipoipo.net

撑大规模智能化应用。为此，企业需要构建工业化 AI 能力生产线，将数据、知识、模型与工具能力持续加工为可复用、可评测、可治理、可直接调用的智能能力，为后续智能应用规模化部署提供统一能力底座。

首先，企业需要夯实数据与知识工程底座，进行语料冷启动，将散落在文档、数据库、业务系统和专家经验中的隐性知识进行清洗、分类、标注和结构化，搜集用户反馈和数据回流，形成持续迭代的数据飞轮。同时，推动领域知识“技能化”，将专家方法、业务规则与标准作业流程转化为 AI 能够检索、理解和执行的知识与工具。在此基础上，企业需要根据不同任务在大模型、小模型、检索增强和专业模型之间进行组合选型，通过 RAG、工具调用与 Agent 编排形成复合智能能力，并利用提示词优化、轻量微调和持续评测不断提升模型表现；与此同时建立从准确性、稳定性、安全性到合规性的分级评测与治理机制。其最终目标不是“训练一个模型”，而是建立一条从数据、知识、模型到 Agent 的工业化 AI 能力生产线，让同一套智能能力能够被不同业务场景反复调用和持续升级。



构建工业化 AI 能力生产线

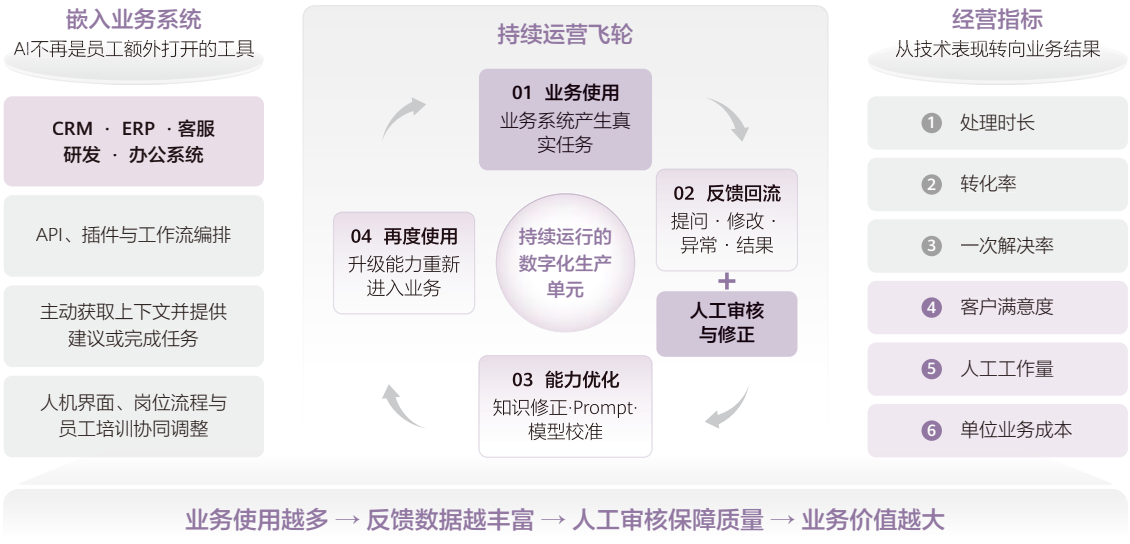
以某全球领先的金融机构把数十年积累的专业知识转化为可规模调用的“专家能力”为例。在财富管理业务中，该金融机构与某 AI 技术服务商合作构建机构专属的 AI 助手，要求模型基于该公司内部经过控制的内容生成答案，将研究报告、投资观点、业务流程和专业知识转化为顾问可以自然语言调用的企业知识能力。目前，该 AI 助手已被超过 98% 的财富顾问团队采用。通过这个过程，公司完成了“从知识资产到 AI 能力资产”的转换：过去分散在报告、制度和专家头脑中的知识，需要员工自己搜索和理解；经过知识工程、模型封装和评测治理之后，这些知识被转化为可检索、可推理、可直接嵌入业务流程的标准化智能能力。企业建设工业化 AI 能力生产线的核心，是让数据、知识、模型和 Agent 可以持续生产、持续评测、持续迭代，并被前台业务按需调用。

第三层：建立 AI 应用的持续运营机制，释放企业级 Agent 的长期价值

随着 AI 能力逐步标准化并具备规模复制条

件，企业智能化的竞争重点开始从“能力建设”转向“持续运营”。仅有模型和 Agent 并不足以形成长期价值，真正决定 AI 能否转化为生产力的，是其能否深度嵌入业务流程，并在真实使用中不断获得反馈、修正和升级。由此，AI 应用也需要从一次性部署模式转向持续运营模式，使模型与 Agent 在业务环境中不断学习和演进，逐步成为能够稳定承担任务、持续优化的企业级 Agent。

因此，企业一方面需要解决“嵌入”问题，将模型与 Agent 能力通过 API、插件和工作流编排等方式嵌入 CRM、ERP、客服、研发及办公系统，使 AI 不再成为员工额外打开的独立工具，而是在既有业务流程中主动获取上下文、提供建议乃至完成任务；同时通过人机协同界面、岗位流程调整与员工培训降低使用门槛，让 AI 真正进入员工日常工作习惯。另一方面，企业需要建立 AI 应用的持续运营机制，将真实业务中的用户提问、人工修改、异常案例和任务结果重新回流至数据、知识库与模型体系，通过知识修正、Prompt 优化、模型校准和能力扩展形成“使用—反馈—优化—再使用”的数据飞轮。衡量 AI 价值的标准也应从准确率、Token 成本等单纯技术指标，进一步转向处理时长、转化率、一次解决率、客户满意度、人工工作量和单位业务成本等经营指标，最终形成“业务使用越多—反馈数据越丰富—人工审核保障质量—业务价值越大”的正向循环。



建立 AI 应用的持续运营机制

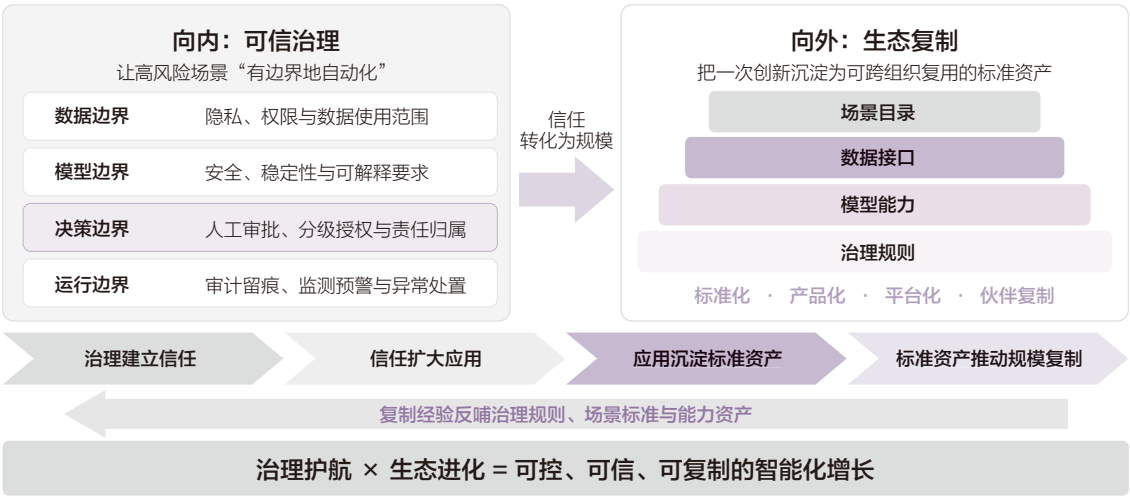
以某全球大型商业银行把 AI 虚拟助手从一个聊天机器人运营成持续进化的企业级生产 Agent 为例。这家银行通过把同一套 AI 能力嵌入消费者移动银行、财富管理、企业银行以及内部员工服务等多个业务系统，使 AI 直接进入客户和员工原有的工作与服务流程。在人的监督之下，将 AI 能力嵌入业务系统获取真实使用，并通过人机协同产生高质量反馈，推动知识和模型持续修正，再以通话时长、服务量、人工负荷等业务指标检验价值，最后实现将 AI 落地经验继续复制到更多业务场景。项目自启动以来已有超过 1.8 万名客服人员使用该 AI 助手，其生成式 AI 增强能力可在约 3 秒内提供上下文指导，并使单次客户通话平均缩短接近 1 分钟。

更多免费行业报告 并购家 www.ipoipo.net

第四层：构建可信治理体系，规模化应用智能能力，辐射产业上下游生态

当 AI 能力真正进入核心业务并开始规模化运行后，企业智能化面临的关键问题将进一步从“如何用好 AI”转向“如何在可控风险下放大 AI 价值”。随着 Agent 承担的任务越来越复杂、调用的数据与系统越来越广，模型幻觉、数据安全、权限越界、责任归属等风险也随之放大；与此同时，经过验证的智能能力如果仍局限于单一部门或单一企业，也难以充分释放其规模价值。

因此，企业需要在“控制风险”与“放大价值”之间建立新的平衡机制：向内构建可信治理体系，通过清晰的责任、权限与风险边界，让 AI 能够被放心地规模化应用；向外推动智能能力生态化复制，将经过验证的数据、模型、场景和运营方法沉淀为可复用资产，向产业链上下游企业复制。

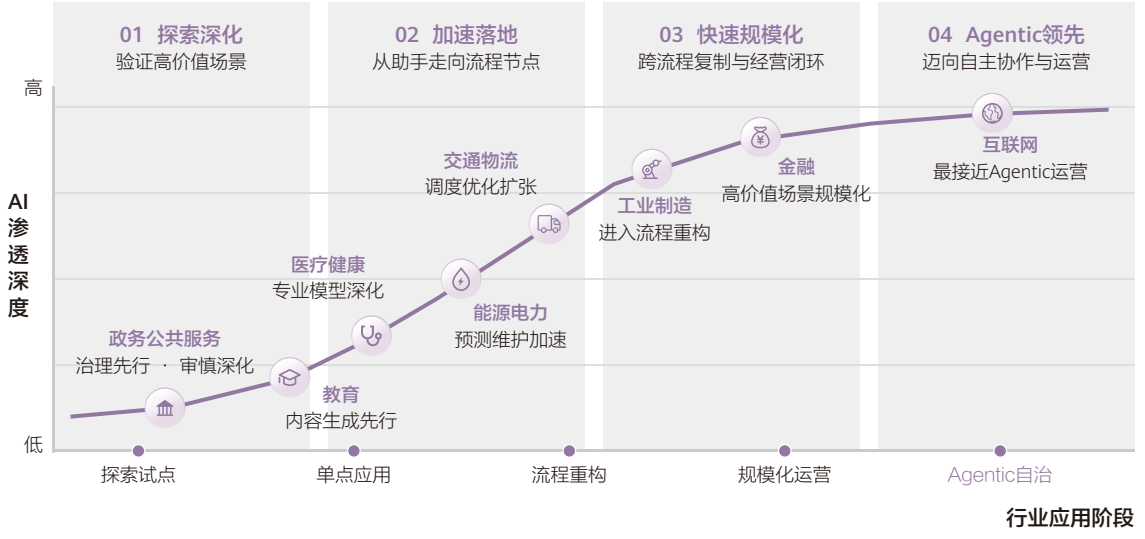


构建可信治理体系

以某全球医疗诊断平台为例，其以“可信治理 + 平台生态”推动医疗 AI 从单点算法走向产业级复制。 医疗是典型的高监管、高风险 AI 应用场景，患者隐私、算法偏差、临床安全和可解释性都直接决定 AI 能否进入真实诊疗流程。为此，该平台并没有采用传统的“集中汇集所有医疗数据”方式，而是建设联邦式数据网络，合作医疗机构的数据经过清洗和去标识化后仍由原机构保持控制，数据无需离开各自的安全边界，合作方可以在隐私保护架构下开展跨机构分析和 AI 开发。其治理逻辑并非简单禁止数据流动，而是把“数据所有权、访问权和使用权”拆分，使数据在不失控的前提下产生跨组织价值。

AI 规模化并不是“治理”和“创新”的二选一。恰恰相反，只有把数据权限、隐私、安全、算法验证和责任机制标准化，企业才能把一次性的项目经验封装成可复用的可信资产，并向更大的组织和产业生态复制。 因此，智能化转型第四层真正要建设的，不只是一个“AI 治理委员会”，而是一套“可信规则 + 标准资产 + 平台生态”：对内让业务敢用、能管、可追责；对外让成熟 AI 场景能够被更多组织快速复制，实现从单点智能化向生态智能化跃迁。

智能应用实践：沿高价值场景持续渗透，拓展智能落地边界



智能应用场景渗透

*注：本图用于展示不同产业可能的采用顺序和演进方向，并非基于完整统计数据形成的市场渗透率预测；受篇幅和研究范围限制，未系统展开农业、林业、牧业和渔业等第一产业场景。

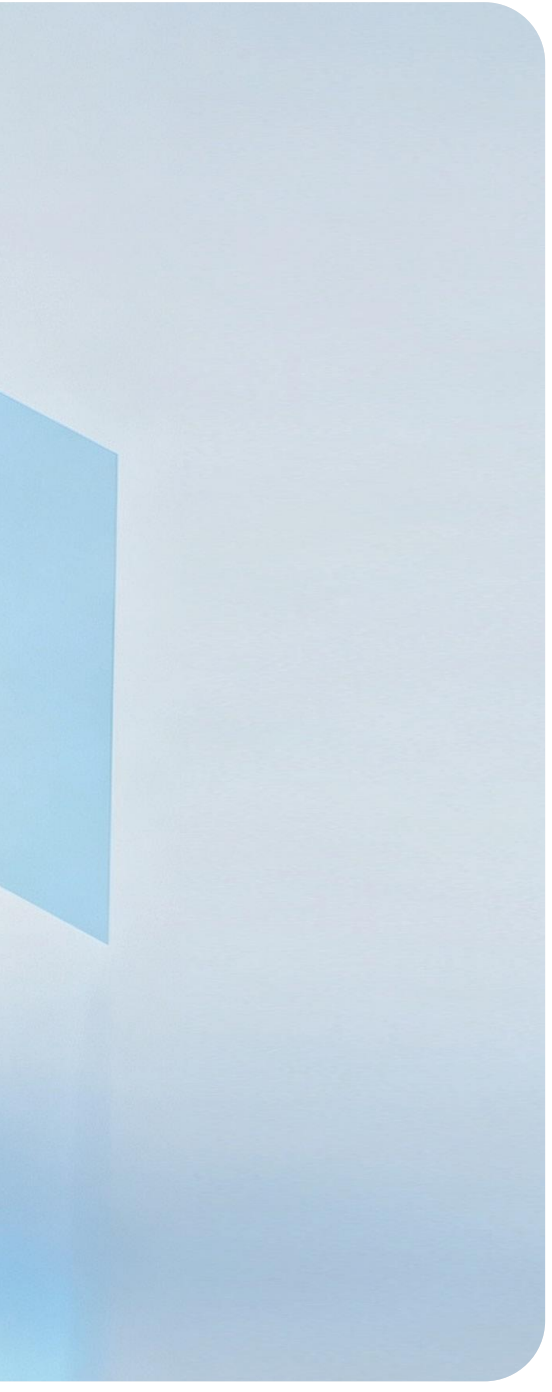
当前行业智能化并不是同步进入 Agent 时代，而是呈现明显的梯次发展。互联网平台已经开始从“员工使用 AI”走向“AI 直接参与运营”；金融和工业制造正从单点应用进入核心流程重构；医疗、教育、能源和交通物流已经形成一批高价值场景，但仍以人机协同为主；政务公共服务则受责任认定、数据安全和公共治理要求影响，整体处于审慎探索阶段。

从各行业的实际落地情况看，当前 AI 应用规模化推广面临的最大难题，并不只是模型能力不足，而是 AI 交付结果仍然“不够可信”，导致业务人员“不敢用、不能放权给 AI”。尤其是在医疗诊断、金融风控、工业控制、能源调度等高风险场景中，即使 AI 能够显著提高分析和执行效率，只要结果存在幻觉、逻辑不可解释、数据来源不可追溯、执行过程不可审计等问题，企业就很难让 AI 直接进入关键决策和执行环节。因此，当前大量 AI 应用实际上停留在“生成建议—人工复核—人工决策”的辅助模式，而尚未真正进入“AI 自主判断—自主执行”的 Agent 化阶段。未来 AI 价值释放的关键，将从单纯提升模型能力，进一步转向建立覆盖数据、模型、流程和执行结果的可信体系，使 AI 的每一次判断都能够被验证、被追溯、被约束，并在明确的责任边界内运行。

因此，未来智能应用的演进并非简单增加 AI 应用数量，而是沿着“更核心的业务、更复杂的任务、更高的自主程度”持续深化。随着模型能力、企业级 Agent 能力与可信治理体系逐步成熟，企业将能够持续扩大 Agent 的任务范围与决策权限，使 AI 从辅助人工作的智能工具逐步演变为可被托付任务的生产助手。智能应用也将由外围场景向核心流程不断渗透，最终推动 Agent 跨越“能用”到“敢用”、再到“规模化应用”的关键门槛，持续拓展智能落地的价值边界。

更多免费行业报告 并购家 www.ipoipo.net





共识

构建人与 AI 和谐共生的产业生态倡议

更多免费行业报告 并购家 www.ipoipo.net

面向 2035，人工智能带来的变化将远不止于一轮技术升级。随着 AI 从数字世界走向物理世界，从单点工具进入生产、科研、组织和社会运行的核心环节，智能正在逐步成为新的基础能力，并由此改变生产方式、创新模式和人与技术之间的关系。

面向未来，我们提出五点倡议。

首先，坚持长期主义，让智能发展建立在可持续的基础之上

AI 的价值最终要在真实世界中得到检验。面向制造、能源、交通、农业、城市等实体经济场景，应持续提升算力、能源和资源使用效率，让智能技术真正转化为生产效率和社会效率的提升。同时，也要把资源约束、环境影响和长期社会价值纳入技术发展的考量，使今天的技术进步不以透支未来为代价。

其次，坚持以人为本，让技术进步服务于人的发展

AI 可以承担越来越多的计算、分析、决策和执行任务，但技术发展的目的不是让人退出，而是让人从大量重复性工作中释放出来，把更多精力投入创造、判断、协作以及对真正重要问题的探索。与此同时，需要加快培养兼具领域知识与 AI 能力的复合型人才，降低智能技术的使用门槛，让更多个人和组织能够参与并受益于智能化进程。

第三，坚持科技向善，让能力的增长始终伴随责任的增长

当 AI 开始进入医疗、金融、交通、工业控制等关键领域，并拥有更强的自主决策和执行能力，安全、隐私、可信和责任边界就不再只是技术之外的约束，而应成为智能系统本身的一部分。需要在创新与治理之间建立动态平衡，通过透明、可追溯、可验证的机制，让技术能力始终运行在明确的规则和责任边界之内。

第四，坚持开放合作，让智能世界建立在连接而非孤岛之上

AI 的发展越来越依赖算力、网络、数据、模型、能源、终端以及行业知识的协同，任何单一企业、单一行业都难以独立完成智能世界的建设。产业各方需要进一步开放接口、共建标准、共享能力，在保障安全和权益的前提下推动技术、产业与生态协同，让创新成果能够跨企业、跨行业、跨区域流动和复制，使智能化真正成为广泛共享的发展机遇。

第五，坚持系统思维，让技术创新真正转化为产业进步

AI 的价值不会因为模型能力提升而自然产生。只有当数据、算力、网络、模型与业务流程、组织机制、人才体系和治理方式协同演进，智能能力才能真正进入生产系统并持续创造价值。因此，面向智能世界的建设不能停留在单点技术突破，而要从基础设施、产业应用到组织治理进行系统布局，让技术创新最终沉淀为长期、稳定、可持续的生产力。

从今天走向 2035，我们需要回答的不只是“AI 还能做什么”，更是“我们希望用 AI 创造什么”。

智能世界 2035 不是一个等待抵达的终点，而是由今天每一次技术选择、产业实践和生态协作共同塑造的结果。

智能世界的未来，不只取决于我们能够创造多强的智能，更取决于我们选择用智能创造怎样的世界。

华为技术有限公司
深圳龙岗区坂田华为基地
电话：+86 755 28780808
邮编：518129
www.huawei.com

商标声明

HUAWEI，HUAWEI，是华为技术有限公司商标或者注册商标，在本手册中以及本手册描述的产品中，出现的其它商标，产品名称，服务名称以及公司名称，由其各自的所有人拥有。

免责声明

本文档可能含有预测信息，包括但不限于有关未来的财务、运营、产品系列、新技术等信息。由于实践中存在很多不确定因素，可能导致实际结果与预测信息有很大的差别。因此，本文档信息仅供参考，不构成任何要约或承诺，华为不对您在本文档基础上做出的任何行为承担责任。华为可能不经通知修改上述信息，恕不另行通知。

版权所有© 华为技术有限公司 2026。保留一切权利。
非经华为技术有限公司书面同意，任何单位和个人不得擅自摘抄、复制本手册内容的部分或全部，并不得以任何形式传播。

更多免费行业报告 并购家 www.ipoipo.net