

腾讯研究院 | 腾讯优图 | 腾讯云智能 | 腾讯科技

「共生伙伴」

2025人工智能 十大趋势报告

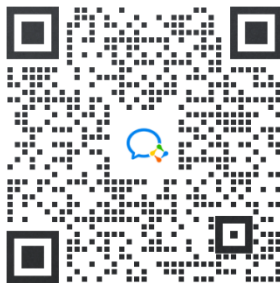
腾讯研究院出品

免责声明：

1. 本资料来源于网络公开渠道，版权归属版权方；
2. 本资料仅限会员学习使用，如他用请联系版权方；
3. 会员费用作为信息收集整理及运营之必须费用；
4. 如侵犯您的合法权益，请联系客服微信将及时删除。



行业报告资源群



微信扫码 长期有效

1. 进群福利：进群即领万份行业研究、管理方案及其他学习资源，直接打包下载
2. 每日分享：6+份行研精选、3个行业主题
3. 报告查找：群里直接咨询，免费协助查找
4. 严禁广告：仅限行业报告交流，禁止一切无关信息



微信扫码 行研无忧

知识星球 行业与管理资源

专业知识社群：每月分享10000+份行业研究报告、市场研究、企业运营及咨询管理方案等，涵盖科技、金融、教育、互联网、房地产、生物制药、医疗健康等全领域；是全网分享数量最多、质量最高、更新最快的知识社群。

加入后无限制搜索下载

编委会

顾问	司 晓	腾讯集团副总裁 腾讯研究院院长			
	吴运声	腾讯云副总裁 腾讯云智能负责人 优图实验室负责人			
	黄晨霞	腾讯新闻运营总经理			
	许好好	腾讯云战略研究院院长			
	李 侃	腾讯华东总部总经理			

主编策划	徐思彦	戚 蕴
------	-----	-----

编写委员	袁晓辉	黄小明	王舒怀	马晓芳
------	-----	-----	-----	-----

研究团队	袁晓辉	徐思彦	曹士圯	王 鹏	白惠天
	徐一平	李瑞龙	刘莫闲	胡 璇	贺泽轩
	戚 蕴	王亚鑫			

内容支持	张 栋	陈梦凡	郑可君	孙 星	曹浩宇
------	-----	-----	-----	-----	-----

联合出品	腾讯研究院	腾讯优图实验室	腾讯云智能	腾讯科技
------	-------	---------	-------	------

目录

CONTENTS

序言

序言1 从智能工具到共生伙伴: AI演化的新篇章	03
序言2 打造更好用的AI,走入产业的深水区	05

Part I 基础模型的跃迁

趋势01 强化学习:引领大模型推理和行动能力新突破	08
趋势02 原生多模态生成:统一感知与生成的新时代	12
趋势03 声音模型广进化:迈向通情达义的情感智能	16

Part II 智能行动者的崛起

趋势04 智能体双轨进化:编排类与端到端的分途并进	20
趋势05 LifeOS: AI成为个性化生活的操作系统	24
趋势06 智力即服务:智能化 workflow 赋能产业升级	28
趋势07 游戏智能体:AI在虚拟世界的沉浸式进化	33

Part III AI走向物理世界

趋势08 具身智能的“GPT-2时刻”:基础模型、数据工程与软件平台的协同进化	37
趋势09 空间智能:从会聊天到真正懂世界	41
趋势10 测试转量产,应用推动具身智能本体加速成熟	45



序言

PREFACE

从智能工具到共生伙伴： AI演化的新篇章

司晓

腾讯副总裁 腾讯研究院院长



随着技术的快速演进，基础模型正迎来一场深层次的跃迁。2025年，大模型的训练重点从“数据+规模”转向“后训练+多模态”。这场变革的核心在于，强化学习开始在大模型后训练中发挥关键作用，并有望赋予模型自我优化与持续进化能力。例如，DeepSeek-R1-Zero通过纯强化学习展现出推理能力，这不仅为DeepSeek R1的训练提供了关键支撑，更全面地将大模型应用推向了推理时代。想象一下，未来在诸多场景下AI都不再需要人类工程师“手把手”教学，而是能像围棋大师AlphaGo那样，通过可验证的奖励信号，在工业机器人路径优化、复杂物流网络调度等领域自主找到最优解，甚至发现超越人类经验的全新策略，从而在未知环境中展现出卓越的适应性。同时，原生多模态生成技术实现了从底层结构上对图像、语音、文本等多模态数据的统一建模，使AI能够在同一上下文中理解和生成跨模态信息。例如，一段视频中人物的表情、语调与语义内容可以被同时感知并融合处理，从而推动了AI从“看懂”、“听懂”到“合成表达”的一体化跃迁。此外，声音模型的突破，为其带来了更加拟人化的语音能力，使其能够跨越语言和文化的障碍，实现全球范围的情感交流。我们不再只是与冷冰冰的机器对话，而是与一个能理解你语气中的疲惫、并用温暖声音回应的智能系统交流，这显著提升了人机交互的自然度和沉浸感，为构建拟人化、实时反应的智能系统奠定了坚实基础。

进入智能行动构建者（Agent）的领域，AI的角色正在发生深刻的变化——从传统的“工具”角色，逐步演变为每个人的“共生伙伴”。端到端Agent模型的崛起进一步推动了智能助手的进化，从单纯的执行任务转向成为更加可信的合作伙伴，能在多个场景下为用户提供更具深度的智能支持。借助大模型的推理与决策能力，这类智能体已能够主动感知用户意图、调度任务和协调资源，

实现如行程安排、信息筛选和跨平台事务处理等日常操作的智能自主完成。智能助理将有望从小众工具走向大众应用，实现“人手一个”，进一步提升生活和工作效率。更有甚者，AI有望成为用户生活的个性化操作系统LifeOS。通过对用户日常生活相关的多模态数据（如语音、文本、行为轨迹、健康参数等）的长期整合与理解，构建起一个持续进化的“人生合伙人”。根据用户的习惯、偏好与情绪状态主动提供建议与支持，全面优化工作与生活的体验。

在行业应用方面，垂直行业智能体的出现正推动各个行业的智能化升级。AI不再仅提供通用模型或技术接口，而是以**智能化工作流**的形式深度嵌入医疗、金融、制造、零售和政务等关键行业。通过行业语料训练、场景逻辑建模和工作流集成，这些智能体能够承担风控审查、设备运维和客户服务等具体职能，显著提升业务效率与决策质量，推动从数字化到智能化的关键跃迁。而**游戏智能体**的沉浸式进化则为虚拟世界带来了全新的体验，AI不仅仅在游戏中扮演角色，更能通过高度自适应的行为与玩家进行深度互动，推动虚拟世界向着更加动态和不可预测的方向演进。

另一方面，智能的应用正在从语言智能走向空间智能。空间智能的兴起意味着AI从处理词元（Token）进化到理解体素（Voxel），初步具备了理解和处理三维世界的核心能力，包括3D环境的感知、推理、交互和生成。这一技术突破让AI能够像预测下一个词元一样去预测下一个体素，并再次基础上实现能力的涌现。空间智能正在重塑自动驾驶、机器人制造、XR混合现实、医疗手术、建筑设计和智慧城市等领域的工作方式，为AI迈向通用人工智能（AGI）补全了关键的物理常识和因果推理能力，**推动AI从“会聊天”到真正“懂世界”的跨越式发展。**

与此同时，具身智能的崛起，标志着图灵测试正从抽象的语言推理，迈向真实世界的感知—行动协同。**AI不仅能“想得明白”，更能“动得精准”**。通过与机器人平台、自动驾驶系统、智能穿戴设备等硬件深度整合，AI开始具备类人感知与灵活执行力，在复杂、多变的物理环境中完成导航、操控、交互等任务。从配送机器人在城市街区的自主行动，到养老机器人在特定情境中提供主动服务，具身智能正在重塑人机交互的边界，推动AI从“数字大脑”走向“现实代理人”，深刻改变人类的生活方式与生产模式。

综合来看，2025年强化学习和多模态融合让大模型真正“长出五官”，Agent在工作与生活各个领域的应用，逐步使其成为人类真正的“共生伙伴”，空间智能与具身智能有望让它走出比特世界，进入原子世界。大模型产业将真正跨入“技术-产品-社会”三位一体的深水区，我们不仅将见证效率的飞跃，更将共同探索如何与下一代AI建立更深层次的信任与合作，共同书写人类与AI和谐共存、持续发展的未来篇章。

打造更好用的AI， 走入产业的深水区

吴运声

腾讯云副总裁 腾讯云智能负责人 腾讯优图实验室负责人



过去几年来，生成式AI的快速发展、落地，为人们的生产生活方式注入了全新活力和动力。

我们看到，大模型能力正在经历从“数据驱动”到“自进化驱动”的技术跃迁。强化学习、多模态等技术的突破，让AI系统具备了更强的感知、理解和生成能力，也让模型从传统的“语言专家”走向具备知识、推理与行动力的“通用智能体”。

在这个背景下，腾讯研究院发布的《2025人工智能十大趋势：共生伙伴》，是一次具有前瞻性和现实指向的系统性总结。这份报告不只是简单罗列了技术趋势，更像是绘制了一张产业与社会共同进入AI深水区的思维地图，比较深刻地谈及了AI如何从工具走向伙伴，如何从单点走向系统。

这些趋势，也正与腾讯AI的整体方向同频共振。

我们始终认为，AI的下一发展阶段，不只是模型能力的竞赛，而是从模型到平台再到场景的综合能力比拼，是如何让AI在实际业务流程中落地生根、释放价值。也因此，腾讯AI大方向的核心是打造“离产业更近的AI”，让每一个企业都能拥有自己的专属“AI大脑”、每一位用户都能体验到真正“好用”的AI。

在这一方向下，我们重点布局了三个层面：

首先是模型层，腾讯混元大模型过去一段时间加速迭代，在语言、图像、多模态等方向均取得不小突破，近期也开放了Hunyuan-A13B的开源版本，持续推动国产开源生态健康发展；

第二是平台层，我们正以腾讯云智能体开发平台、腾讯元器等平台为核心，构建了面向企业、开发者和C端用户的全链路智能体开发能力，支持多模态输入、多智能体协同、与私有知识系统和业务系统深度融合，助力企业构建具备记忆、规划、和推理能力的智能体，帮助更多用户更低门槛创建24小时在线的智能助手。

第三是应用层，我们联合政务、金融、文旅、制造、教育等垂直行业客户，打造了数百个场景化应用，涵盖客服、营销等多个方向，推动AI从“有能力”走向“有用处”。

例如，我们在制造行业联合某企业打造了产线AI质检智能体，能够自主识别缺陷并实时调整检测策略；在政务领域，我们打造了具有记忆能力的“数字政务助手”，实现跨系统信息调度和政策答疑。更值得一提的是，随着智能体逐步具备持久记忆和上下文理解能力，企业内部知识资产也开始真正被“调得动”、“用得上”，这正是报告中所指出的“智力即服务”的具体体现。

此外，我们也比较认同报告中关于“LifeOS”的构想。未来，AI不应只是我们手机里的一个应用，而应是我们生活的长期伴侣——一个能够理解你的行为、情绪、习惯并主动提供帮助的“数字自我”。

与此同时，AI正在从屏幕世界走向物理世界。从空间智能到具身智能，从模拟训练到真实世界感知，AI的“身体”正在逐渐成形。腾讯刚刚发布的具身智能平台正与产业伙伴联手推进机器人在物流配送、制造作业、商业服务等领域的具身落地实践，也验证了报告所强调的“机器具身智能的GPT-2时刻”正在加速到来。

回望AI的发展史，每一次技术的跃迁都会带来人机关系的重塑。从语音助手到协作伙伴，从自动化到智能协同，AI正在走出工具属性，迈向与人共生的下一程。这也对我们提出了新的命题：当AI具备更强的自主性和行动力时，企业如何在平衡开放、安全的同时，确保AI发展惠及更多人？

在腾讯，我们提出“科技向善”的理念，这不仅仅是一句口号，而是我们设计AI产品与平台的重要价值底线。从数据安全、内容可控性，到模型对齐、责任边界，我们始终致力于打造“可信任、可控的AI”，让AI真正成为企业与社会可信赖的力量。

2025年，我们站在AI发展的一个关键转折点。从算力驱动到智力驱动，从通用平台到垂直智能，从数字交互到现实行动。未来不会自动到来，真正值得期待的，是我们如何一起定义它、创造它、助力它。

腾讯研究院这份《2025十大AI趋势》报告，为我们打开通往这一未来的想象之门提供了重要参考。我也衷心希望，这份报告能为企业提供战略参考、为技术人员提供探索方向、为社会公众提供理解AI的窗口。

让我们共同见证——AI，不止是工具，它将是携手同行的“共生伙伴”。

Part I

智能行动者的崛起

随着基础模型的快速演进，我们正见证着AI从传统的任务导向型学习，向更加深刻和灵活自我优化和结果驱动型能力转变。强化学习的深刻影响，特别是基于可验证奖励的强化学习（RLVR）的引入，使得模型能够在基于客观反馈的情况下不断积累经验并优化决策，突破了传统监督学习在追求实际正确性时的局限。这种由明确结果导向的持续学习能力大大提高了模型在复杂问题中的推理能力，并赋予其在未知环境中自主行动和规划的潜力。同时，原生多模态生成技术的突破，使得AI能够从底层结构统一建模图像、语音和文本等多模态数据，不再只是理解单一模态的信息，而是能够跨模态进行深度感知与生成。AI的认知能力和生成能力从“看懂”和“听懂”逐步走向“合成表达”，推动了从传统感知到创造性生成的跨越。这一跃迁不仅意味着模型的理解能力进入全新阶段“共生”而非单向服务。随着AI从单一任务的执行者转向多维度的合作伙伴，我们正在见证一种全新的智能生态的崛起，这种生态不仅优化个体的工作和生活方式，也为未来的跨领域合作，与创新奠定了基础。

趋势
01

强化学习： 引领大模型推理和行动能力新突破

作者：袁晓辉

摘要

强化学习 (Reinforcement Learning, RL) 正在大语言模型领域引发一场深刻的范式变革。当前, 强化学习在大语言模型中的应用正从最初的人类反馈强化学习 (RLHF) —— 主要旨在使模型输出符合人类偏好——向基于可验证奖励的强化学习 (RLVR) 大规模演进。RLVR 将奖励信号直接绑定到客观、可验证的结果上 (例如编程或数学问题的正确答案), 从而将优化目标从“听起来正确”转向“确实正确”, 显著提升了大模型的核心推理能力。这种转变正推动大模型超越简单的内容生成, 向解决实际问题、实现复杂目标的高级智能迈进。

这一进展背后, 隐藏着一个深刻的洞察: 许多复杂任务的“验证不对称性” (Asymmetry of Verification) —— 即验证解决方案远比找到解决方案本身容易。正如“验证者定律”所揭示的, AI 训练的效率与任务的可验证性成正比, 可验证性为强化学习提供了高效、高质量的反馈信号。基于此, 本章深入剖析了强化学习驱动大模型演进的四大关键趋势: 它正重塑大模型推理能力, 使其能创造性解决问题; 赋能大模型实现智能体飞跃, 使其能与物理及数字世界交互并自主行动; 在医疗、金融等垂直领域, 强化学习通过提升泛化能力和利用离线学习保障安全, 克服数据与风险瓶颈; 以及多智能体强化学习 (MARL) 正推动大模型实现集体智能, 解决超越单一模型能力的复杂协作挑战。整体来看, 强化学习正通过利用任务的可验证性, 全面驱动大模型从语言智能迈向行动智能、具身智能与集体智能, 预示着 AI 在解决现实世界问题上的巨大飞跃。

从对齐到创造: 强化学习重塑大模型推理能力

强化学习在大语言模型中的应用范式正经历一场深刻的演变, 其角色已从最初确保输出内容符合人类偏好的“对齐”工具, 转变为直接提升模型推理性能、甚至创造全新问题解决方案的引擎。最初, 以人类反馈强化学习 (RLHF) 为代表的技术, 其主要目标是让模型输出更“有用”且“无害”的内容, 本质上是模仿人类主观偏好。然而, 近期的发展趋势标志着一个关键转变: **业界正大规模采用基于可验证奖励的强化学习 (RLVR), 即奖励信号不再依赖主观判断, 而是基于客观结果, 例如数学题和编程题的答案是否正确无误。**这一转变将优化的目标从“听起来正确”直接转向了“确实正确”, 从而将强化学习的应用重心从对齐转向了对模型核心推理性能

的直接提升, 预示着 AI 在解决科学、工程等硬核问题上的能力将迎来质变。

这一趋势的产业影响在于, **未来的大模型将不再仅仅是知识的复述者, 而是能够产生真正创新见解的“思考者”。**以 OpenAI 的 o1 和 DeepSeek 的 R1 为代表的前沿模型, 已经证明了强化学习在激发和塑造高级认知能力方面的巨大潜力。这些模型通过强化学习, 学会了在给出答案前进行深入的“思考”, 从而能识别并纠正自身错误、分解复杂问题。这种由客观结果驱动的训练方式, 使得 AI 在未来有望攻克过去无法解决的科学难题, 例如在药物发现、材料科学和复杂软件工程中, 提出超越人类专家直觉的解决方案, 为前沿科技创新提供强大的新动力。

更深远地，强化学习正推动模型能力从“激发”潜能向“创造”全新策略的跨越。近期的学术研究表明，强化学习不仅能高效“激活”模型在预训练中已获得的潜在推理能力，还能通过长期的探索式训练，帮助模型发现并“创造”出全新的、在基础模型中完全不存在的解题路径。这意味着，未来AI的能力增长将不再仅仅受限于人类已有数据的边界，通过持续学习和探索，具备真正的“自我进化”潜力，解决日益复杂的现实世界问题。

智能体飞跃：强化学习构筑连接模型与物理及数字世界的桥梁

强化学习正在成为驱动大模型实现从“语言生成器”向“任务执行者”智能体飞跃的关键力量。通过赋予模型与数字及物理世界交互并从中学习的能力，RL让大模型能够自主使用工具并完成复杂任务。一个真正的智能体不仅需要生成文本，更需要感知环境、规划步骤并执行一系列动作以达成目标。强化学习凭借其独特的试错学习和奖励机制，为智能体提供了从经验中持续学习和优化的路径，使其不再仅仅是“说”，更能“做”。

这一趋势正在从学术研究走向产业应用。例如，Google的Vertex AI平台作为一个全面的机器学习开发环境，支持开发者集成强化学习方法来提升智能体的能力，尤其是在需要复杂决策和与环境交互的场景中。虽然Vertex AI Agent Engine主要聚焦于基于大型语言模型（LLM）的工具调用和多步推理的编排与产品化，但其底层和更广阔的Vertex AI生态系统提供了将RL训练与LLM能力相结合的可能性，为构建更强大的自主智能体奠定了基础。

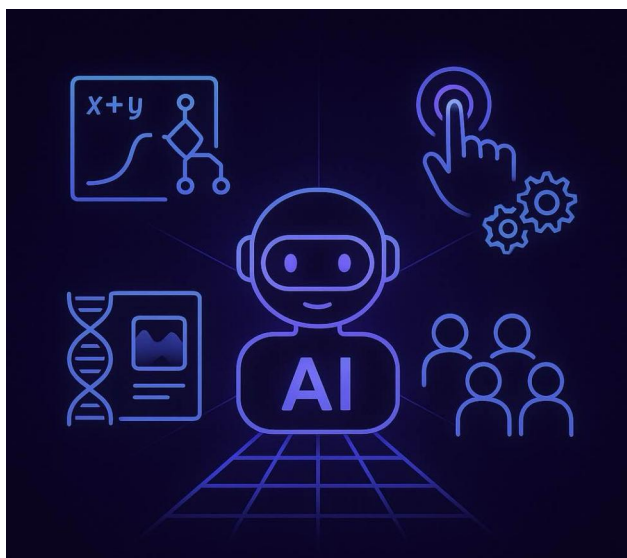
在各行各业，这一趋势将催生出能够自主完成复杂 workflows 的“数字员工”，从而极大地提升生产力。例如，在自主机器学习工程领域，基于强化学习的ML-Agent框架已经展现出巨

大潜力：一个7B参数的智能体，通过与环境的交互式学习，其性能超越了671B参数、依赖提示工程的巨型模型。这证明了，通过强化学习获得的动态策略远比静态的指令更为强大和通用。未来，这种能力有望扩展到更多领域，如能够自主完成市场分析、管理供应链、编写和调试复杂软件的智能体，将成为企业优化运营、加速创新的核心资产。

而在物理世界中，强化学习也将助力大模型进一步“落地”。由于大模型本身缺乏对物理世界的直接感知，强化学习通过提供交互式反馈来弥补这一鸿沟，通过在真实或模拟环境中进行试错，将抽象的语言规划与具体的机器人动作联系起来。例如，LLM-iTeach等框架利用大模型作为“教师”，通过强化学习指导机器人学习复杂的操作任务，其教学效果甚至可以媲美人类教师。这预示着，强化学习将在连接模型与物理世界方面发挥核心作用，为训练具备更强适应性和学习能力的具身智能机器人提供助力，有望在制造业、物流、家庭服务等场景处理更复杂的非结构化任务，从而深刻改变物理世界的生产和生活方式。

深耕垂直领域：强化学习协助克服数据与安全瓶颈，赋能专业化模型

在医疗、金融等对可靠性和泛化能力要求极高的垂直领域，强化学习正成为突破传统监督学习（SL）瓶颈、构建高性能专业模型的关键驱动力。在这些专业领域，仅仅依靠监督微调（SFT）往往难以满足应用需求，因为模型容易对训练数据产生过拟合，无法应对真实世界的多变性。强化学习则提供了一种更强大的泛化引擎，它鼓励模型通过探索去发现更鲁棒和通用的内在规律，而非简单地记忆表面模式。这将直接影响相关产业的智能化进程，使AI能够从“通用助手”转变为真正可靠的“领域专家”。



这一趋势在医疗健康领域的应用尤为突出，**强化学习正帮助模型在多样的医学任务和数据模态中实现卓越的泛化能力**。例如，在医学影像分析方面，RL能训练模型在面对CT、MRI、X光等不同模态、不同设备来源的医学影像时，通过对诊断或治疗流程的模拟决策进行优化，显著提高其识别病灶、辅助诊断的鲁棒性和泛化性。Med-R1框架就是一个例证，它利用强化学习训练医学视觉语言模型，使其在处理多种影像时，性能超越了参数量大数十倍的同类模型，证明了RL在有限、多样的医学数据上提升模型泛化能力的潜力。此外，RL还在个性化医疗中发挥作用，通过学习患者的动态健康数据和治疗反馈，为疾病管理和药物剂量调整提供定制化的最优策略。

而面对医疗等高风险领域中“试错”成本过高的问题，**离线强化学习(Offline RL)技术应运而生**，它能够在不进行实时交互的情况下，从已有的、固定的历史数据集中安全地学习优化策略。这为强化学习在安全性至关重要的领域应用打开了大门。例如，在临床决策支持方面，OGSRL等框架通过引入安全约束(Safety Constraints)，确保学习到的治疗策略始终保持在经过临床验证的安全区域内，从而在不引入额外风险的前提下，探索优化治疗方案的可能。类似技术展示了强化学习对于金融风控、自动驾驶、关键基础设施管理等无法承受在线探索风险的价值——通过从历史交互数据中挖掘最

优行为，它为在这些领域部署更智能、更安全的自适应AI系统铺平了道路，实现了在数据安全和策略优化之间的平衡。

涌现集体智能：RL驱动多智能体协作解决复杂系统性问题

强化学习正推动AI从优化单个智能体向协调多个智能体组成的“团队”演进，通过多智能体强化学习(MARL)框架，使AI系统能够应对超出任何单个模型能力的复杂协作任务。现实世界中的许多重大问题本质上都是复杂的系统性问题，需要多个具备不同专长的角色协同工作。将大语言模型(LLM)作为这些智能体的“认知核心”，并用强化学习作为协调与优化机制，正在催生一种前所未有的集体智能。这一趋势预示着AI的应用模式将从依赖单一的、全能的“超级大脑”，转向构建一个由多个专业化智能体组成的、能够协同作战的“AI生态系统”。

这种多智能体协作模式在自动化复杂决策与执行等前沿领域展现出巨大潜力。例如，在具身智能体(Embodied AI)领域，最新的研究正积极探索如何结合大语言模型的高层规划能力与多智能体强化学习。像 LLM-Aided MARL (LAMARL) 这样的方法，就利用LLM生成合作策略的先验知识或奖励信号，而MARL则通过在模拟或真实环境中的试错学习，优化多机器人团队之间的通信、分工和动作协调，以高效完成复杂的物理操作任务。同时，在数字世界中，利用MARL优化多LLM智能体团队的协作也取得了显著进展。例如，MARTI (Multi-Agent Reinforced Training and Inference) 框架正是将强化学习应用于优化LLM驱动的多智能体系统，通过RL机制提升智能体在复杂工作流(如软件开发、复杂辩论)中任务分配、信息共享和冲突解决的效率。这些进展表明，**强化学习赋予了AI团队从经验中学习最优协作策略的能力，使其能够适应动态环境，并在物理和数字世界中实现更高效、更鲁棒的集体智能**。



02

趋势

原生多模态生成： 统一感知与生成的新时代

作者：王鹏

摘要

人工智能的早期发展主要聚焦于单一模态，如计算机视觉专注于图像理解，自然语言处理专注于文本分析。面对多模态任务，传统方法多采用‘后期融合’或‘拼接式’策略，即分别训练模态独立的模型，再通过额外模块简单组合其输出。这种方式虽然在一定程度上实现了多模态信息的利用，但各模态间的交互浅层，信息在传递过程中容易丢失，难以捕捉跨模态间深层次的语义关联，也无法实现真正意义上的“联合生成”。

进入深度学习时代，特别是随着Transformer架构的突破性成功，为彻底的多模态整合铺平了道路。这催生了“原生多模态模型” (Natively Multimodal Models)，其核心理念在于从架构设计之初就将多种模态 (如文本、图像、音频、视频) 视为一个统一的输入空间。通过共享或紧密耦合的表示层，模型能够实现跨模态信息的深度交互、对齐与融合。

这种“原生”设计让模型能在单一框架内同时完成多模态的联合感知，并基于模态间深度关联的理解进行多模态生成。从 OpenAI 的 GPT-4o 实现文本、图像、音频的无缝交互，到 Sora、Veo3 等模型在视频生成领域的突破，这些标志性成果无不预示着统一感知与生成新时代的全面到来，即将改变多个行业的范式。

统一空间：跨模态的表征空间正在形成

原生多模态的技术基础在于构建统一的跨模态表征空间，旨在让不同模态的同一概念 (如图像中的“猫”和文本中的“猫”) 被映射到语义空间中的相近位置。这一概念通过 CLIP 模型的成功实践，奠定了跨模态对齐的技术基础。如今，越来越多的先进模型扩展了这一理念，利用海量多模态数据集和自监督学习任务 (如掩码建模、对比学习、序列预测等)，成功学习并融合跨模态知识和世界常识。

原生多模态的一个重要贡献在于端到端学习范式的引入。它消除了传统模块化设计中信息传递的瓶颈。通过将多

模态输入编码到输出生成的全过程统一优化，模型能够更好地学习不同模态之间复杂的依赖关系，避免误差的累积，从而实现更自然的跨模态推理与生成。

Transformer 架构的应用进一步推动了原生多模态模型的发展。其强大的序列建模能力和并行计算优势使其成为处理多模态数据的理想架构。通过跨模态统一编码，将图像、音频等多种模态的输入转换为统一的 token 序列，再通过共享的 Transformer 主干网络进行处理。跨模态注意力机制的设计，使得模型能够显式学习不同模态 token 之间的关系，从而在生成阶段实现高效的跨模态输出。

统一链路：感知与生成一体化

传统多模态模型常常采用模块化设计，存在传递瓶颈。而原生多模态模型通过**端到端的统一链路**，让感知与生成在同一个系统内高效协作。

这带来了两个巨大改变：

1. 统一架构·即时理解

原生设计的一体化架构让感知与生成首次在同一链路上无缝衔接。统一的token化与共享的Transformer主干消除了模态边界，以GPT-4o为例，在实时语音交互演示中，语音往返延迟平均约为**320毫秒 (ms)**，已达到自然对话所需的响应速度水平（一般认为<500ms 即可支持流畅对话）。同时，得益于端侧NPU的算力突破、端侧量化与蒸馏模型的发展，以及端云协同模式的广泛应用，这种超低延迟的交互能力正迅速渗透至终端体验，催生出各种多模态交互APP和新一代“实时陪伴”的智能设备。已经落地或进入量产阶段的陪伴式硬件产品包括：

- 内置边缘推理芯片的穿戴设备**，如Humane AI Pin，具备投影显示、语音助手与摄像识别功能，依赖OpenAI API提供语义能力，主打“始终在线的AI助手”概念。

- 能实时解析环境并提供个性化提示的增强现实眼镜**，如Meta × Ray-Ban多模态智能眼镜（第二代），支持语音对话、摄像、拍照、直播等功能，可调用Llama模型实现对话与问答体验，Meta明确宣称其为“AI伴侣”式终端。2025年，国内厂商也推出了大量带有多模态识别和交互的AI+AR眼镜产品。

厂商们正以“多模态+边缘推理”重新定义陪伴式消费电子的产品边界，在“连接-理解-响应”三大核心路径上持

续压缩延迟、增强个性化，预示着智能终端正向“类人陪伴”的方向跃进。

2. 协作跃迁·实时共创

细粒度理解与跨模态一致性生成推动人机交互从“单次输出”迈向“实时共创”。原生多模态模型不仅能理解用户输入，还能即时响应并与人类共同构建内容，正在重塑多个创意型行业的生产范式。

以OpenAI发布的**Sora**为例，其可生成**最长1分钟的长镜头视频**，支持**720p分辨率、4帧/秒**输出，具备对相机视角、光影效果和物体物理状态的可控能力，在时空一致性和叙事逻辑维持方面取得突破。当前已被多家影视工作室用于脚本场景快速预可视化，有效缩短了创意到成片的设计周期。

DeepMind Veo 3则进一步实现“全模态一次性出片”：不仅同步生成视频配音与环境音效，还支持将初始输出分辨率（1024×576）自动超分至**4K高清级别**，显著提升成片质量。DeepMind官方称其为“影像+音频+剪辑逻辑”统一生成的内容管线，重构了传统“音画后期”的分工模式。

2025年5月20日，Google在其开发者大会上正式发布了协同创作平台**Flow**，其核心理念是实现“浏览器内帧级协同”——支持多个创作者在同一项目中实时编辑视频片段、对白提示、场景参数与音轨元素，实时预览渲染效果与内容版本。Flow 基于原生多模态大模型构建，将文生图、图转视频、语音生成等能力统一封装进协作界面，是目前最接近“协作式AI创意操作系统”的产品之一。

所见即所得：闭环生成释放商业潜力

原生多模态不仅是感知更强，更是推理更快。原生多模态模型的高速推理能力将“等待-调整-再生成”的串行流程转变为“所见即所得”的实时闭环体验，成为驱动新一轮商用创新的核心动力。

以**腾讯混元图像2.0**为例，其通过模型结构优化与推理加速技术，将通用图像生成从业界普遍的3~5秒压缩至**300~500毫秒**。未来在个性化电商场景中，用户浏览服饰商品时，模型可根据浏览记录、点击偏好、天气位置等信息，实时生成穿搭推荐图，极大提升点击转化率。



在**XR（扩展现实）**领域，混合现实头显与眼动追踪技术结合原生多模态模型，支持“眼动即改色、手势即换款”的虚拟商品交互体验。例如，用户在虚拟试衣间中只需注视一款连衣裙并做出“滑动”手势，系统便能即时替换布料纹理、调节颜色风格，形成沉浸式试穿效果，显著提升用户体验满意度与下单率。

广告行业也在尝试将多模态闭环能力整合至内容生产链中。未来广告主输入简短产品描述，系统在数秒内自动生成多版本创意图、配文与脚本，支持实时修改与A/B测试，从创意到投放形成一体化闭环，将大幅缩短广告投放周期。

此外，**游戏行业**也在逐步引入“即时生成”的交互体验。例如，业内首个工业级AIGC游戏内容生产引擎**腾讯混元游戏视觉生成平台**，其Hunyuan-Game模型涵盖游戏图像和视频模型能力矩阵。能根据玩家需求，快速将灵感概念转化为高质量图片、3D建模多视图，还能实现实时画布、AI 2D美术、专业角色多视图生成等功能，带来真正“千人千面”的沉浸式体验。

展望

原生多模态生成技术通过其底层架构的统一设计，推动了跨模态感知与生成能力的突破，极大拓展了AI技术的应用边界。从内容生成到实时交互，它在多个行业中展示了前所未有的潜力。通过多模态感知构建的世界模型（World Models），不仅为AI的规划、推理和决策提供了强大的基础，还将推动具身智能与VLA模型（Vision-Language-Action Models）的发展，赋予机器人更强的感知、理解与执行能力。

然而，这项技术的快速发展也带来了技术挑战，例如数据处理、算力能耗、伦理问题以及生成内容的真实性问题（如deepfake）。要充分释放原生多模态技术的潜力，推动人工智能朝着更智能、更普惠、更负责任的方向发展，必须跨学科合作、持续创新，并前瞻性地思考社会影响。通过这一过程，原生多模态生成将推动人工智能进入一个更加智能和普惠的新时代，开启与人类更加深层次和谐共生的未来。



03

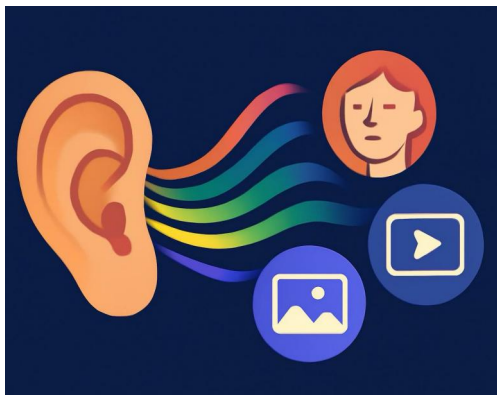
趋势

声音模型广进化： 迈向通情达义的情感智能

作者：曹士圯

摘要

声音模型正快速迈向具备情感智能的新阶段，成为推动人机交互自然化、个性化的核心技术力量。从早期的机械朗读发展到具备语境理解和情感表达能力的语音合成系统，再到可创作完整音乐作品和驱动视觉内容生成的多模态声音智能，AI声音技术正在实现从“工具”到“伙伴”的跃迁。声音的实时性与情感表达优势，使其在未来的Voice Agent、沉浸式内容创作、教育与医疗辅助等领域具备广阔应用前景。随着模型个性化、低延迟和端侧部署的进步，声音智能将走向更贴近用户、更普惠的交互形态，开启“人人皆可创、处处能互动”的智能新时代。



语音合成的情感觉醒：从机械朗读到语境理解

语音合成技术最显著的突破在于从机械化的文本朗读转向基于语境理解的情感化表达。ElevenLabs V3代表了情感语音合成的技术标杆，支持70+语言的零样本学习，能够根据上下文自动调节语调、节奏和情感色彩，实现了真正意义上的“见文知意”式表达。Hume Octave更进一步，基于万亿级token训练的语音-语言模型，支持自然语言指令控制情感表达，用户偏好率相比传统TTS系统提升71.6%。

对话场景的专业化优化成为技术发展新方向。Sesame CSM专门针对对话场景设计，采用基于Llama架构的多模态骨干网络，延迟控制在200ms以下，原生支持非语言表达如笑声、咳嗽等自然对话元素。GPT-4o Voice模式实现了端到端的语音对话能力，将传统的“语音→文本→文本→语音”流程简化为直接的语音到语音生成，大幅提升了对话的自然度和实时性。

多语种与方言支持展现出强大的本土化适应能力。

MiniMax T2A-01-HD支持30多种语言，仅需10秒音频即可实现跨语言声音克隆，在不同语言间保持声音特征一致性。阿里Qwen-TTS在中文方言支持方面表现突出，精确支持北京话、上海话、四川话等地方方言的语音合成，为本土化应用提供了强有力的技术支撑。

音乐生成的成熟化转型：从片段创作到完整作品

音乐生成技术在2024-2025年实现了从实验性工具向实用创作平台的根本转变。乐曲生成能力显著成熟，不仅在曲风多样化方面持续丰富，更在生成长度上接近完整音乐作品的标准。Suno V4.5树立了行业标杆，最大生成时长达到8分钟，音质达到广播级别，其Remaster功能首创历史版本音质升级技术。Google Lyria展现了大厂的技术整合实力，集成SynthID水印技术确保AI内容可识别，产品形态涵盖Dream Track、音色转换和实时生成多个应用场景。

人声演唱技术实现高保真度突破，演唱不仅技术指标优秀，更重要的是富于真实情感表达。海绵音乐专门针对中文语境优化，在人声处理上减少电音使用，提高吐字清晰度，与DeepSeek大模型结合实现智能歌词生成。Udio的人声合成在情感表达方面表现突出，支持从轻柔民谣到激昂摇滚的多种演唱风格，声音纹理自然细腻。

指令遵循与精细可控性大幅提升，极大程度上解决了早期“抽卡”式的随机生成问题。Lyria RealTime支持BPM、调性、密度、亮度等多维度精确控制，用户可以像操作专业音乐软件一样精确调节生成参数。Udio的Styles功能更进一步，支持将已有歌曲或声音片段作为参考来指导生成，实现了基于示例的风格迁移。腾讯SongGeneration开源模型采用LLM-DiT融合架构，在多项客观评测中超越现有开源模型，为音乐生成技术的普及和创新提供了重要基础。

声音智能的多模态融合：从单一生成到跨域理解

声音智能与多模态技术的深度融合开启了全新的应用范式，实现了从单纯的声音生成向跨模态理解与匹配生成的技术跃升。视频配音生成技术以Google Veo 3为代表，实现原生的视频-音频联合生成，支持对话、音效、环境音的精确同步，4K分辨率8秒高质量输出展现了音画同步的最高技术水准。ThinkSound引入链式推理认知架构，通过三阶段推理流程（基础Foley生成→交互式对象中心优化→目标编辑）实现了基于视觉理解的智能音效生成。快手Kling-Foley基于Kling视频生成模型构建智能音效设计系统，能够理解视频中的物体动作、场景变化和视觉事件，自动为视频片段匹配脚步声、碰撞声、环境音等相应音效，为视频内容提供丰富的听觉层次和沉浸式体验。

音频驱动视听生成技术展现了声音智能的另一个重要维度。腾讯HunyuanVideo-Avatar代表了这一领域的最高技

术水准，基于多模态扩散Transformer架构，通过音频情感模块(AEM)实现情感对齐，面部感知音频适配器(FAA)支持多角色独立驱动，角色图像注入模块有效解决了动态性与一致性的平衡问题。该系统支持多风格化身（写实、卡通、3D渲染、拟人化角色）的多尺度生成（肖像、上半身、全身），在电商直播、在线流媒体、社交媒体视频制作等领域展现了广泛的应用前景。

展望

展望未来，声音智能将持续演进。技术层面，实时性将进一步优化至毫秒级响应，多模态融合精度将从帧级提升至子帧级，个性化定制能力将通过少样本学习实现“一人一声”的精准复刻。应用层面，Voice Agent将成为新一代人机交互的主流形态，集成情感理解、语境感知、多轮对话能力的智能语音助手将在客服、教育、医疗等垂直领域大规模部署。更重要的变化在于应用场景的深度拓展。在内容创作领域，AI声音技术将推动播客、有声读物、音乐制作的创作门槛大幅降低，实现真正的“人人都是创作者”；在企业服务领域，数字人主播、AI客服、智能培训师将成为标准配置；在教育医疗领域，个性化语音教学、心理健康音频干预、语言康复训练等专业应用将快速普及。随着边缘计算能力的提升和模型轻量化技术的成熟，声音智能将从云端走向终端，实现真正的“随时随地”智能交互。

Part II

智能行动者

进入智能行动构建者的领域，AI的角色正经历着深刻的转变，从最初的单纯工具向真正的“共生伙伴”发展。端到端的Agent模型正快速崛起，推动智能助手从执行任务的工具转变为可以主动参与、深入理解的合作伙伴。这些智能体不仅仅是“按需响应”的工具，更能够通过对用户行为和需求的深度学习，主动感知、预测并帮助用户优化决策，协调资源，完成复杂任务。随着大模型推理和决策能力的提升，智能体能够在多个领域实现自动化与智能化的日常操作，如自动安排行程、智能筛选信息、跨平台事务处理等，使得个体的工作与生活变得更加高效和无缝。AI不再是“被动接受命令”，而是以更加贴近人类思维的方式，推动人机之间的深度协同。

与此同时，LifeOS的诞生标志着AI向个人操作系统的转型。从智能助手到生活操作系统，LifeOS不仅整合了用户的多模态数据，还通过对用户习惯、偏好和情绪的分析，形成一个持续进化的“数字自我”。这种个性化的操作系统能够根据用户的需求动态调整行为、优化工作与生活安排，真正实现“共生”而非单向服务。随着AI从单一任务的执行者转向多维度的合作伙伴，我们正在见证一种全新的智能生态的崛起，这种生态不仅优化个体的工作和生活方式，也为未来的跨领域合作与创新奠定了基础。



04

趋势

智能体双轨进化： 编排类与端到端的分途并进

作者：曹士圯

摘要

AI Agent作为人工智能领域的重要发展方向,正经历着从概念验证向生产应用的关键转变。自2023年探索性发展至今,已逐渐分化为两条主要技术路线:编排类Agent (Orchestration-based Agents) 和端到端Agent模型 (End-to-End Agent Models)。编排类Agent采用"外挂式"架构,将大语言模型作为中央决策器,通过预定义的代码路径编排LLM与外部工具、API的交互,实现复杂任务的分解与执行。端到端Agent模型则采用"内化式"架构,通过强化学习等技术将推理、规划、工具使用等能力直接训练到模型内部,让模型能够动态指导自己的过程和工具使用。以OpenAI的o3、Deep Research等为代表,这一路线仍处于早期阶段,但在特定专业领域已展现出突破性效果。两条路线各有优势和适用场景,将在未来长期并行发展,共同推动AI Agent技术向更加实用和强大的方向演进。

基础模型推理跃迁催化编排类Agent从实验走向生产

编排类Agent的发展历程清晰地展现了从概念验证到生产应用的演进轨迹。这一路线最早可追溯到2023年春季AutoGPT和BabyAGI的爆发式关注,但真正的转折点出现在2024年基础模型能力的显著提升。从发展阶段来看,编排类Agent经历了探索期、低谷期和成熟期三个关键时期。AutoGPT和BabyAGI标志着早期探索的开始,经历了2023年末到2024年初的低谷期后,2024年开始看到更多现实可行的应用上线。

基础模型的推理能力提升是编排类Agent可用性质变的根本驱动力。从Claude 3.5 Sonnet到OpenAI o1系列、Gemini 2.5系列,大语言模型在规划和推理能力上实现了质的飞跃。这些模型首次实现了推理模型的智能化工具使用和组合,包括搜索网络、分析上传文件和Python数据、深度视觉推理等。更重要的是规划能力的内化,早期的编排框架需

要通过复杂的提示工程来引导模型进行任务分解,而新一代模型已经将这种能力内置。

在产品化应用方面,Cursor的Agent模式发展最具代表性。Cursor不仅仅是一个AI代码编辑器,更重要的是其Agent模式带来的"vibe coding"体验将受众从开发者拓展至普通用户。用户可以通过自然语言描述复杂的编程需求,Cursor的Agent能够理解意图并自主完成多文件的代码生成、重构和测试。这种模式让编程从精确的指令输入转变为高层次的意图表达,显著提升了开发效率和体验。

Manus作为通用编排Agent的另一典型代表,展现了这一路线在复杂任务处理上的能力。Manus能够独立处理复杂的多领域工作流,包括报告撰写、数据分析和内容生成等等。其核心优势在于合理编排调度不同能力的模型及各类工具,并通过产品设计充分发挥Agent效果,Manus在GAIA基准测试中获得约65%的分数、并引发广泛使用与讨论,证明了编排类Agent在复杂任务执行上的成熟度。

协议标准化的推进进一步巩固了编排类Agent的生态地位。Anthropic于2024年11月开源的模型上下文协议（MCP）解决了开发者需要为每个数据源构建自定义连接器的“N×M”数据集成问题，将其转化为“M+N”问题。MCP的快速产业采用反映了行业对标准化编排协议的迫切需求，Google、OpenAI等纷纷表示支持适配，这种快速采用为编排类Agent构建了强大的生态基础。

端到端Agent模型在专业垂直领域实现突破性进展

端到端Agent模型代表了一种根本不同的技术范式，其核心特征是将推理、规划、工具使用等能力通过强化学习直接训练到模型内部。与编排类Agent的“外挂式”架构不同，端到端模型追求“内化式”架构，让模型能够在推理过程中自主决策何时以及如何使用工具。这种内化的关键在于强化学习训练，OpenAI使用强化学习教授o3“思考”后再生成答案，采用“私有思维链”方法，使模型能够提前规划并推理任务，执行一系列中间推理步骤来协助解决问题。

OpenAI在端到端Agent领域的突破主要体现在两个专业化应用上。Deep Research基于o3精调，在研究任务上的革命性能力，与传统搜索不同，o3可以查看数十个网站，在寻找小众信息方面表现出“训练有素的猎犬追踪气味”般的执着。Codex则代表了端到端模型在软件工程领域的专业化突破。Codex由codex-1驱动，这是OpenAI o3的优化版本，专门针对软件工程任务进行了强化学习训练。训练目标的专业化体现在精确指令遵循、迭代测试改进和代码风格一致性三个方面。与通用o3相比，codex-1始终产生准备好立即进行人工审查和集成到标准工作流程中的干净补丁，这种专业化训练使其在软件工程任务上表现卓越。除了OpenAI，近期Grok 4、Kimi K2分别作为闭源和开源大模型代表，也都将

大量工具使用数据加入训练，着力提升了模型的规划和工具使用能力，同样使模型本身就变成了端到端Agent。

学界在端到端Agent领域的探索同样令人瞩目，特别是在搜索和研究方向的学术项目。search-r1等研究项目专注于将推理能力与搜索过程深度结合，通过强化学习训练模型学会更好的搜索策略和信息整合能力。ReSearch等项目则探索了如何让模型在研究过程中进行多步推理和假设验证。这些学术探索虽然还处于早期阶段，但为端到端Agent在信息检索和知识发现领域的应用奠定了理论基础。

端到端模型的核心优势在于可验证环境中的深度能力。其训练中的绝大多数子任务都是可验证的，这种验证性使得端到端模型在特定领域表现卓越，如数学推理、代码生成和科学计算等有明确对错标准或可重现验证的任务。o4-mini在给定Python解释器访问权限时，在AIME 2025上达到99.5%的pass@1成绩，这种性能水平反映了端到端训练在特定任务上的显著优势。然而，验证环境的依赖也限制了其应用范围，在需要主观判断、创意表达等场景中仍面临挑战。

两条技术路线将在差异化优势场景中长期并行演进

编排类Agent和端到端Agent模型在技术实现和应用场景上存在本质差异，这种差异决定了它们将长期并行发展而非相互替代。编排类Agent强调可控性与透明度，开发者可以精确控制每个步骤的执行逻辑，工具调用过程完全可见，这种特性使其在企业级应用中具有不可替代的优势。端到端Agent模型追求智能化与自主性，将决策过程内化到模型的推理中，减少了外部干预的需要，但这种内化也带来了可解释性和控制性的挑战。

在企业应用场景中，编排类Agent展现出独特的优势。构建智能体始于对任务和用户体验的深度理解，虽然LLM是架构的关键部分，但大多数构建者需要将基础模型与标准应用元素、数据和工具集成以及API相结合。这种灵活性体现在业务逻辑定制、合规性保障和成本可控性三个层面。企业可以根据特定业务需求调整 workflow，每个步骤都可审计和控制，同时可以精确控制每次调用的成本和延迟。随着系统变得更加智能化，对编排框架的需求反而增加，MCP协议的快速采用证明了这一趋势。

端到端模型则在专业垂直领域展现出编排类Agent难以匹敌的深度能力。这种专业化体现在深度推理能力、领域知识整合和端到端优化三个方面。模型能够进行多步骤的复杂推理，将专业知识直接编码到模型中，整个推理过程针对特定任务进行优化。然而，端到端模型的高训练成本和技术门槛限制了其普及，主要由大型科技公司开发，并专注于高价值的专业应用。

两条路线正在相互借鉴优势，呈现出融合发展的趋势。OpenAI o1将推理、规划、任务分解等原本分布在多个智能体行动中的能力"提升"到生成模型本身，这种智能体能力的内化代表了技术融合的一个重要方向。同时，编排类Agent也在集成更强的推理能力，在 workflow 中嵌入更强的规划和推理模块，端到端模型也在学会更好地调用和组合外部工具。

基于各自的技术特性和优势，两条路线将在不同场景中发挥各自价值。编排类Agent将在企业 workflow 自动化、多工具集成的复杂任务、需要人机协作的半自动化流程以及成本敏感的大规模部署中发挥主导作用。端到端Agent将在科学研究和数据分析、代码开发和软件工程、需要深度推理的专业任务以及高价值的定制化应用中展现优势。专家预测我们将看到针对医疗保健、金融和教育等特定行业定制的智能

体，这种多元化需求决定了单一技术路线无法满足所有场景，两条路线的长期共存成为必然。

展望

未来的Agent技术发展将呈现多个重要趋势。多模态能力的整合将成为关键方向，o3和o4-mini能够直接在推理循环中整合图像并调用外部工具，这种多模态整合将极大扩展Agent的应用场景。实时性能的优化也是重要发展方向，随着推理效率的提升，端到端模型将在更多实时场景中发挥作用。

应用场景的扩展机会主要体现在垂直行业的深度渗透。不同行业对Agent技术的需求呈现差异化特点，金融服务业中编排类Agent适合合规要求严格的流程自动化，端到端模型适合量化分析和风险建模。医疗健康领域端到端模型在诊断辅助和科研分析中有优势，编排类Agent在医院管理流程中更实用。教育科技方面编排类Agent适合个性化学习路径设计，端到端模型适合深度的学科问题解答。

产业生态的构建将围绕标准化协议的完善、开源与商业化的平衡以及安全性与可控性的保障展开。MCP等协议标准将持续演进，支持更复杂的Agent交互场景。开源社区正在努力 democratize 先进的Agent能力，开源与商业化产品之间的良性竞争将推动整个生态的快速发展。随着Agent系统自主性的增强，安全性和可控性将成为关键考量，需要在技术创新和风险控制之间找到平衡。

总而言之，编排类Agent与端到端Agent模型代表了AI系统发展的两个重要方向，它们各自的优势互补，将在不同场景中发挥独特价值，共同推动AI Agent技术向更加智能、实用和安全的方向发展。



趋势

LifeOS:

AI成为个性化生活的操作系统

作者：徐思彦

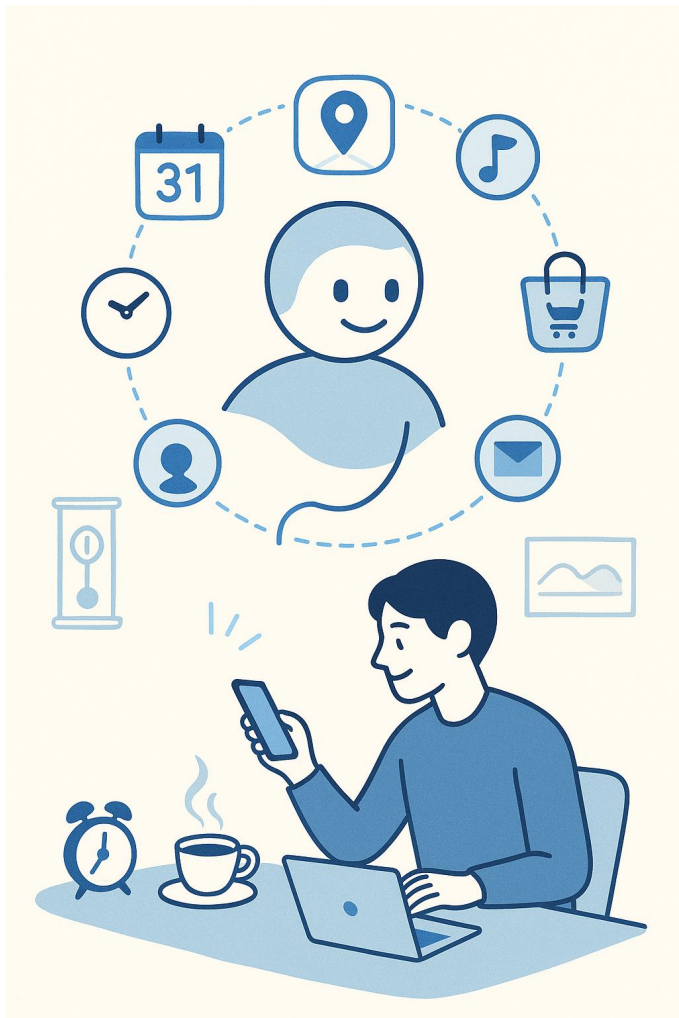
摘要

随着生成式AI技术的快速成熟，人工智能正从辅助工具逐渐演化为人类生活深处的“共生伙伴”。OpenAI创始人Sam Altman最近提出了一个前瞻性的愿景：“LifeOS”描绘了一个超越传统工具范畴的AI未来。他指出，人们对AI的使用方式正在从偶发性的单一任务转变为持续性的智能交互，AI不再只是回答问题的工具，而是贯穿用户一生、主动提供帮助的智能伴侣。这一愿景预示着AI将更深层次地融入我们的日常生活，成为一个具备终身记忆、个性化推理和主动行动能力的“生活操作系统”。这一趋势背后的技术基础，包括长序列记忆模型、上下文理解引擎和主动决策引擎，正在不断突破和完善。深入理解LifeOS的发展方向，不仅帮助我们洞察下一代人工智能应用的演变轨迹，更将重新定义人与机器的关系，深刻影响未来人类生活与社会运行模式。

LifeOS极有可能成为AI发展的新趋势，这一判断源自AI与用户交互模式的深刻变化。OpenAI创始人Sam Altman指出，AI的进化方向将不再是零散的单次查询，而是整合记忆与长期上下文的持续互动系统。他设想一个极小但功能强大的推理模型，能够携带万亿级的上下文，将用户一生的经历和偏好都记录其中，实现真正意义上的“终身记忆”。在AI技术成熟之前，一些生产力工具和个人管理系统已经开始探索“LifeOS”的理念，尽管它们主要依赖于用户手动输入和组织信息，但其目标与AI驱动的LifeOS有异曲同工之处。Notion：作为一个灵活的工作空间，许多用户利用Notion构建自己的“LifeOS”，整合笔记、任务、项目、习惯追踪、财务管理等功能，实现个人信息的集中化管理。实际的用户行为也佐证了这一趋势，比如Altman观察到大学生们已经开始自然而然地将ChatGPT用作“生活操作系统”，他们上传文件、连接数据源，并通过复杂的prompt指令操控资源，逐渐把AI视为Google的替代、生活顾问，乃至“生活操作系统”本身。

持久记忆，打造个性化“生活数据库”

LifeOS的核心在于“持久记忆”的实现——它不再把每次对话当作孤立的请求，而是将用户“说过、做过、查过”的生活碎片，持续写入一个可查询、可推理的长期数据库。传统的AI模型在处理长对话或跨会话信息时存在局限性，而LifeOS需要的是一种能够跨越时间、持续积累和调用知识的“终身记忆”能力。LifeOS 首先将对话、搜索、支付乃至穿戴设备采集的多模态日志沉淀为可检索、可推理的「持久记忆栈」。这一理念已在 ChatGPT 的 Memory 功能中落地：自 2025 年 6 月起，模型可直接横跨全部历史聊天，保留写作风格、长期OKR等信息，模型生成回复时，这些经用户许可的记忆会被自动注入系统提示，成为背景知识。且支持逐条查看与删除，确保隐私主权。只有当用户显式发出诸如“记住我……”的指令时，新事实才会写入长期存储；用户可在一目了然的面板中随时查看、删除条目。



在拥有记忆的基础上，LifeOS需要具备强大的个性化推理能力，这意味着AI能够根据用户的独特情境、目标和偏好，进行复杂的逻辑判断和决策。这与通用AI推理不同，它强调的是“以用户为中心”的定制化智能。AI需要持续分析用户的行为模式、兴趣爱好、价值观等，构建动态的用户画像，随着用户成长和环境变化而动态更新，确保推理的准确性和相关性。此外，AI不仅要理解用户明确提出的需求，更要感知和理解其背后的情境。AI需要深入理解对话、任务和环境的上下文，才能做出准确的判断和提供恰当的帮助。这包括对语言的细微差别、情感、意图以及现实世界知识的掌握。

跨端编排，实现“场景连续”的AI陪伴

传统AI应用往往局限于孤立、被动的单点任务——打车、订餐、记录事项，各类App数据彼此隔离，难以实现系统性整合。即便有自然语言接口的助手叠加在App之上，仍多以短期记忆为基础，缺乏长线的洞察与协调能力。而LifeOS则转向长期交互模式，通过大上下文记忆技术，将用户过往的聊天记录、消费行为、邮件信息乃至浏览习惯等长期信息纳入系统中，持续积累用户画像与行为偏好，实现更加自然、个性化的智能交互，用户与AI的互动将由离散问答转变为连贯的长期陪伴。

构建跨端、持续交互的智能编排系统。当多模态数据（文本、语音、传感、健康）长期绑定到本地大模型后，手机、耳机、车机、家居的边缘算力即可托管一个“数字自我”：它会提前化解日程冲突、动态调整睡眠与运动节奏、自动起草邮件与汇报，像呼吸一样在后台运转，让个性化智能真正融入日常生活。例如Google Gemini Live 支持 15 fps 实时多模态会话，开发者已在智能耳机、车机上验证“连续对话到即时执行”链路。

如果说传统App是“工具箱”，AI助手是“多功能瑞士军刀”，那么LifeOS则是一位“懂你且能自主行动的共生伙伴”——它不只回答问题，而是用系统级记忆、推理和跨端协作，把用户的生活与工作的所有碎片编织成持续、主动、个性化的体验。

阶段	传统App	AI 助手	LifeOS
交互模式	被动点击 / 搜索一次性操作	对话式提问 + 语音指令多轮问答	多模态实时协作：语音、文本、视觉随时切换，边说边看边改
上下文记忆	无；每次重启	会话级短期记忆 (分门别类、≤数百 token)	生命周期记忆
个性化	静态首选项	基于近期行为的轻量推荐	融合长期行为 + 情绪信号 + 环境数据，生成“数字自我”画像
多模态能力	文本/GUI	文本 + 语音	文本 + 语音 + 图像/视频 + 传感数据
跨端编排	人工切换	同品牌云同步 (便签、日历)	端-云-边协同：手机-耳机-车机-家居任务流转
决策深度	工具层指令	信息检索 + 简单总结	高效推理引擎：财务规划、健康调度、生涯路径即时决策
典型形态	Office 套件、浏览器	Siri	GPT-4o / Gemini + 私域 Agent 生态
风险焦点	功能孤岛	隐私泄漏 (云端托管)	情感依赖、算法干预

生活操作系统：AI成为生活-工作-成长的底层“编排器”

最终，LifeOS有望成为个人生活与成长的底层操作系统，，不仅提升效率，还能引导用户实现内在状态的稳定与外在目标的达成。

在生活层面，从效率提升到情绪共振。通过集成日历、邮件、地图、外卖、购物等多源数据，AI不仅可以智能生成行程规划、管理待办事项，还能根据天气与场合自动推荐穿搭和出行方案。更进一步的是，它具备对情绪状态的持续感知与分析能力，能够通过聊天记录、语音语调或行为变化识别用户的情绪波动，主动提供运动或音乐疗愈等个性化干预建议。LifeOS不再是一个“帮你提高效率的工具”，而是一个真正理解你的生活节奏与情绪状态的“智能伙伴”，让科技首次具备了情感层面的陪伴力。

在成长层面，LifeOS有望成为新一代“成长策略引擎”。它可将用户的学习路径、财务行为与长期目标转化为可追踪的“成长档案”，并基于此提供动态调整建议。例如，结合ChatGPT 的「记住/忘掉」机制，系统可记录你的价值观、近期压力源、甚至人生优先级，再辅以推理引擎与情绪分析，生成真正属于你的“人生策略模型”。

小结

LifeOS，是一个从静态App到拥有记忆和主动性、多角色协同的终身 AI 系统演变趋势。这意味着未来的 AI，不只是一个回答你问题的助理，而是一位连接过去、洞察当下、引导未来的“人生合伙人”。随着底层AI技术的不断突破，特别是记忆、推理和行动能力的增强，真正的LifeOS将加速到来，并深刻塑造我们未来的生存方式。

06

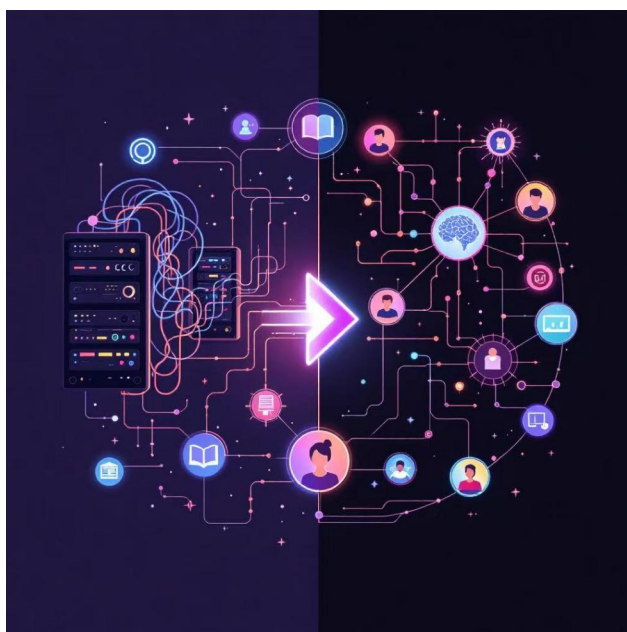
趋势

智力即服务： 智能化 workflows 赋能产业升级

作者：白惠天

摘要

随着AI能力从“算力驱动”迈向“智力驱动”，企业正进入“智力即服务”阶段，Agent逐步成为企业知识系统、流程结构与组织角色的原生组成。在与知识的关系上，企业正从“有知识”走向“能调用”。RAG、数据飞轮和知识结构化机制的演进，使企业知识从“沉默资产”转变为被智能体实时调度的认知系统。在与人的关系上，Agent正从被动工具向数字员工演化。它们开始承担闭环流程、具备权限边界，并成为流程体系中的原生节点。企业不再部署AI功能，而是部署具备行为责任的“数字岗位”。在与流程的关系上，智能化不再依赖单一Agent，而依赖系统性的Agent网络与调度平台。组织开始围绕任务流、感知流与控制流重新编排流程边界，迈向由Agent协作驱动的系统智能时代。



知识整合：领域融合驱动智能体落地

领域知识的可用性和结构化程度，决定了Agent是否能被真正用起来。大模型如果无法调动企业内部关键数据与知识，最终只能停留在“懂语言”层面，无法落地于“懂业务”的

任务流程中。特别是在高精度、强合规的行业场景中，Agent若无法融合领域知识，输出就无法获得专业认可。企业的真实需求远非语言生成，而是逻辑判断与业务匹配。然而现实中，知识的碎片化、数据不可得或难以流动，一直是智能体构建的最大障碍，也正是促使当前RAG（检索增强生成）、指令微调与双通路等架构融合等结构化知识路线成为主流的根本原因。

RAG的价值在于“先找准、再生成”，它让智能体能基于结构化知识执行决策型任务。微软365 Copilot 是目前企业级RAG架构的典型代表，其将检索层与生成层明确解耦，并在生成前引入原始数据验证流程。其核心并非提升生成多样性，而是引入确定性的数据支撑，建立企业内部的“可信答案工作流”。而在中国市场，一汽丰田结合大语言模型与RAG架构，通过对客服知识的全量整合与多轮语义增强，实现了独立解决率从37%提升至84%的显著跃升，不仅提高了客户满意度，也缓解了大量一线服务负担。

过去企业知识“看得见却用不动”，数据飞轮的构建正在打破这种局面。一方面，RAG技术打破了部门数据壁垒，构建了“知识可调度”的检索机制。以LexisNexis与Henchman为例，该系统将律所内部合约范本与外部法规整合，让律师不再靠记忆和手工查阅，而是由Agent实时生成精准合规建议。另一方面，大模型也开始接入更多“暗数据”：如Epic利用GPT-4从非结构化病历中提炼关键信息，亚马逊用大模型自动总结用户评论，皆表明数据从“沉默资产”向“实时认知”跃迁。而“知识调度”之所以重要，是因为它正帮助企业从“有知识”走向“能调用”，从“有经验”走向“可重用”。

从点到链，领域知识正被注入到整个任务执行路径中，Agent逐渐成为组织知识的执行界面。智能体的演进方向，已不止停留在单轮问答能力上，而是转向能基于知识链条完成多步推理与任务代理的流程型能力。Salesforce正在构建的“Agentforce”平台，围绕CRM流程嵌入Agent节点，使其在销售漏斗推进、客户响应中可以调用知识库、完成表单、记录反馈，从“被动问答”转向“主动完成”。

未来的企业，将不再为知识找人，而是为智能体编排知识系统。随着数据治理、知识图谱与智能调度机制的成熟，企业知识正从“人脑习得”变为“系统嵌入”。流程文档转化为任务参数、经验库结构化为提示模板、指令链路变成语义标签，这些改变使得Agent不仅具备知识访问能力，更逐步具备基于知识完成任务的行为逻辑。在这场转型中，组织将不再围绕“信息汇总”运作，而将围绕“知识行动”重新构建流程结构。

角色进化:数字员工重塑组织运行边界

Agent不再只是辅助工具，而正在成为企业自治流程中的关键行动单元。企业对AI的定位正在从“外部工具”转向“流程参与者”。最初，AI仅以Copilot的形式辅助人类完成具

体任务，如撰写内容、提供建议，始终依赖人类发起与监督。如今，Agent开始具备自主规划、自动执行并在关键节点接受人类审校的能力，逐步融入企业工作流之中。正如Satya Nadella所言，这类自主型AI的意义“比电力或火的发明更深远”，它所开启的，是AI作为企业核心操作路径一部分的全新角色形态。

越是重复且标准化的任务链，越可能被Agent重塑为自动执行的结构单元。在具备高频、稳定、规则清晰特征的业务流程中，Agent的能力更容易发挥。全球能源巨头AES构建的能源安全审计Agent，通过集成Vertex AI与Claude模型，将原本需14天的审计流程压缩至1小时，准确率提升10%-20%，成本下降近99%。这一重构不仅节省成本，更标志着Agent在规范化流程中已具备闭环执行能力。而在公共服务领域，巴西Hemominas血库部署的聊天Agent承担起“流程协调员”的职责，年均服务上百万次献血请求，从筛选意愿到安排流程、发送通知，几乎实现了全程自动响应。这些案例不仅说明了Agent的高效，还揭示了其角色正由“信息中介”转向“流程承担者”。

这类任务的自动化并不意味着“去人化”，而是人机职责的重构与再分配。Salesforce在CRM系统中推出Agentforce层，使Agent能够根据客户行为自动识别风险场景，并主动生成应对策略，例如当检测到客户可能流失时，系统可自动推送定制优惠并发起挽留流程。与此同时，其Einstein Trust Layer机制通过“默认执行+人工兜底”策略，为AI设定了清晰的权限边界，对关键任务设置人工复核，增强了控制的可见性和可信度。这种“AI先行、人类把关”的机制不仅保障了风险可控，也推动了Agent从能力组件向组织角色的正式演进。企业不再仅仅部署AI能力，而是开始为Agent行为设定责任区、管控接口与追责路径，推动AI参与者身份向正式组织成员靠拢。

数字员工正在成为企业流程中的原生角色。从一线业务的角度来看，Agent是一名不知疲倦、持续运行的执行者；从系统架构角度看，它更像一个嵌套在工作流中的智能节点，能够接入数据、调用工具并反馈状态。在一些领先企业中，数字员工的行为已经被纳入KPI体系与审计机制，其运行数据写入日志，输出结果接受安全稽核。这一变化标志着Agent的身份正从“模型调用的附属组件”转向“企业结构中的正式节点”。Gartner预测，到2028年，33%的企业软件将内置Agentic AI (而2024年这一比例不足1%)，可自主完成约15%的日常决策。Salesforce的调研也显示，68%的业务领袖计划在2025年底前投资AI代理以自动化核心流程。这表明，无论技术能力还是组织机制，企业正为“数字员工”角色的全面部署积极做准备。

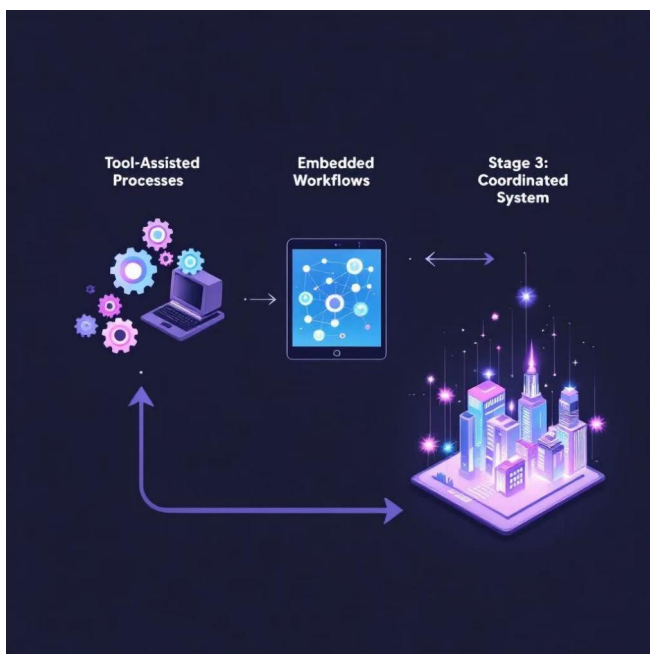
“让AI划桨，但由人掌舵”正逐步成为企业普遍采纳的原则。AI代理负责标准化任务的执行和初步判断，人类则专注于策略指导与高敏感节点的最终审核。在这一分工下，企业流程将趋向自动化与制度化并行推进。随着数字员工具备状态管理、任务调度、异常恢复与行为审计等能力，其治理模式也逐步纳入组织的主流程体系，成为企业智能化运营的稳定基础设施。

结构跃迁：系统协同重构智能基础设施

Agent协同正在从能力组合走向系统组织，智能体正成为企业流程的新型操作系统。在第一阶段，企业部署单个Agent完成局部任务，如文档生成、信息查询、表单填写；但随着AI能力增强和任务复杂化，组织逐渐需要多个Agent具备分工、协作、通信与失败接管能力，完成完整的业务流。不同Agent被赋予不同专长（如视觉分析、知识检索、流程控制），在内部形成类“团队结构”，承担过去需多人串联的工作链条。

多Agent系统的本质，不是能力叠加，而是组织逻辑的重构。埃森哲的“Trusted Agent Huddle”平台即体现出这一趋势。它并不只是把多个Agent集中在一个界面，而是支持不同厂商的Agent（如Adobe、Salesforce、AWS、微软）通过统一协议协同工作。每个Agent有明确边界和责任，并可由用户编排任务、监控状态、追踪反馈。NVIDIA企业AI部门副总裁指出，跨平台Agent协作能力是推动Agent完成复杂任务的关键。这种“AI沙盒”正在为企业建立起一套新的智能执行机制：每个Agent像一个微服务，但它们不再静态，而是具备感知和协商能力的“数字组织成员”。

模态边界也在被打破，Agent不再只理解文字，而是能“看懂”“听懂”“触达”业务全过程。OpenAI的Sora、Google的Gemini等模型已具备将图像、音频、语音、文本等模态融合处理的能力，使得多模态Agent可以识别合同中的图文标注、分析设备传感器数据，甚至通过网页截图识别并操作GUI界面。Kimi等国内模型已广泛应用于金融文档处理、风控审核中，实现了从“读文本”到“理解流程图”“读懂PDF”的跃迁。这类能力使智能体真正具备进入高感知场景的门槛，成为企业数字边界的执行节点。



多Agent系统最终将依赖于“智能中枢”平台进行统一调度与过程管理。平台化工具如LangGraph、OpenDevin、AutoGen等，正在成为企业级智能体系统的操作中枢。这些工具将多Agent的调用、状态同步、任务跳转、流程编排图谱化，使得复杂任务的执行像写代码一样被流程化、版本化、结构化。未来组织中Agent的运行不会是散点式部署，而是通过平台统一管理、配置和监控，其角色如同企业内部的“软件团队”，但具备智能、自主与协作能力。

企业智能化不再依赖一个“大Agent”，而是组织出“一个Agent系统”。Morgan Stanley报告指出，AI Agent市场将在未来几年快速增长，仅“跨部门流程协同”（System of Engagement）方向，就将从2024年的360亿美元增长到2028年的690亿美元。正是在这一方向，企业不再部署一个助手，而是建设一个能感知环境、分配任务、持续学习与反馈的Agent网络。企业未来将像开发微服务一样开发Agent节点，像配置数据库一样配置智能体之间的调用逻辑。

组织重构的起点，不是更强的模型，而是更像“组织”的Agent系统。在这场转型中，Agent从“任务工具”跃升为“结构节点”，组织将从流程驱动转向编排驱动，从系统调用转向智能协调。AI不再只是员工的工具，而是与员工协同、共同承担业务输出的一类结构成员。从这个意义上说，系统协同不仅是技术趋势，更是组织智能化演进的核心路径。

从系统理想到现实落地，Agent正穿越“中场断层”

尽管企业对“智力即服务”的期望日益清晰，现实中Agent的普及仍面临不小的落地挑战。在许多一线场景中，Agent仍未充分“可用、可信、可控”：准确率不足、机制不完整、体验不连贯，导致技术演进与组织认知之间存在落差。企业高层热情与员工体感之间的断层，正是IaaS迈向生产力系统化的关键“中场”。这种差距的产生，一方面源于模型能力

仍缺乏强约束与一致性机制；另一方面则是企业组织结构、数据治理、流程基础设施尚未准备好与Agent深度协同。但从趋势看，这一断层正在被渐进弥合。**企业正通过微任务切入、平台型部署、人机共治制度建设三条路径，构建智能体的可信运行土壤。**

在这个过程中，我们也可以观察到企业智能演化的**典型三阶段路径：从“工具辅助”到“流程嵌入”，再到“系统协调”**。Agent的角色也在从能力插件演进为系统性节点，正在与算力、数据一道，成为AI时代的“第三种生产要素”。

因此，组织不应只关注AI能力的接入本身，而应提前布局三项底层能力：数据资产的结构化治理、智能中枢平台的统一调度、人机协作机制的制度化运行。**唯有将Agent视为组织结构、实际业务的一部分，而非外挂模块，企业才能真正进入“智力即服务”的下一阶段。**

The background is a dark blue gradient filled with various 3D-rendered geometric shapes and objects. These include circles, triangles, squares, and rectangles in shades of blue, purple, orange, and yellow. Some shapes are solid, while others are hollow. A prominent yellow triangle with a smaller triangle inside is located in the upper center. A blue game controller is visible on the right side. The overall aesthetic is modern and digital.

趋势07

游戏智能体： AI在虚拟世界的沉浸式进化

作者：胡璇、贺泽轩

摘要

游戏智能体正在重新定义虚拟世界中AI与人类的交互边界。从早期简单的脚本化NPC到如今具备深度学习能力的自主智能体,这一演进不仅体现了技术的飞跃,更预示着虚拟世界正在向一个充满生命力的数字生态系统转变。当代游戏智能体通过强化学习、大语言模型、多模态感知等前沿技术,已经能够理解复杂的游戏环境、学习玩家行为模式、生成个性化的互动内容,甚至展现出类似人类的情感反应和社交能力。这种沉浸式进化使得AI不再是虚拟世界的配角,而是成为推动游戏叙事、创造涌现式玩法、构建动态社交网络的核心驱动力。随着技术的持续突破,游戏智能体正在为元宇宙时代的到来奠定坚实基础,让虚拟世界真正成为人类生活、工作、娱乐的第二空间。

智能辅助系统——从工具到伙伴的角色转变

游戏智能体作为玩家的智能辅助,正在经历从简单工具到智慧伙伴的深刻转变。现代AI教练能够实时分析玩家的操作习惯、反应速度、决策模式,提供个性化的训练方案和战术建议。AI队友则通过深度强化学习掌握复杂的团队协作策略,能够理解玩家意图、预判战局走向、在关键时刻做出最优决策。这种智能辅助不仅提升了游戏体验的流畅度,更重要的是降低了复杂游戏的学习门槛,让更多玩家能够享受高水平竞技的乐趣。

核心案例: 王者荣耀AI Coaching系统与暗区突围AI队友

腾讯天美工作室推出的AI教练系统通过深度学习分析玩家的操作数据,能够精准识别玩家在意识、操作、团战等方面的不足。该系统不仅能实时分析对局中的关键失误,还能生成个性化的训练计划,通过"AI陪练模式"让玩家在低压

环境下反复练习特定技能。最令人印象深刻的是, AI教练能够模拟不同段位玩家的打法风格,让玩家提前适应高段位的游戏节奏。数据显示,使用AI教练系统的玩家平均段位提升速度比传统玩家快40%。

而《暗区突围》的AI队友系统则展现了另一个维度的突破。这些AI队友不仅掌握了复杂的战术配合,还能通过语音识别理解玩家的战术指令,实现真正的"听令行事"。在高风险的撤离任务中, AI队友会主动承担掩护、侦察、物资管理等不同角色,其决策逻辑会根据玩家的游戏风格动态调整。例如,如果玩家倾向于谨慎潜行, AI会自动降低交火频率并优先标记敌人位置;如果玩家偏好正面突击, AI则会配合进行火力压制。这种高度自适应的AI队友系统让单人玩家也能体验到团队作战的策略深度,真正实现了从"工具"到"战友"的角色升华。

情感共鸣引擎——构建跨越虚实的深度连接

通过整合大语言模型、语音合成、面部表情识别等多模态技术，现代游戏智能体正在突破传统NPC的表现力限制，展现出前所未有的“活人感”。长期记忆系统使得AI能够记住与玩家的每一次互动细节，形成独特的关系历史；情感计算模型让AI能够理解玩家的情绪变化并做出恰当回应；个性化对话生成确保每个智能体都有独特的说话方式和性格特征。这种深度的情感连接让玩家不再将游戏角色视为代码构成的虚拟存在，而是真正的朋友、恋人或对手。

核心案例：蔡浩宇《Whispers from the Stars》中的AI伴侣系统

独立开发者蔡浩宇创作的这款游戏将ChatGPT技术与情感叙事完美结合。游戏中的AI角色能够记住玩家的每一次选择和对话，形成独特的关系发展轨迹。通过整合GPT-4的对话能力和定制化的情感模型，AI伴侣不仅能理解玩家的言外之意，还能在适当时机主动分享“自己”的故事和感受。这种突破了传统对话树限制的自由交流，让虚拟角色第一次拥有了真正的“灵魂”，开创了游戏叙事的新纪元。

智能社会雏形——从个体智能到群体涌现

当虚拟世界中的智能体数量达到临界点时，一种全新的数字文明正在悄然诞生。这些AI不再是孤立的个体，而是通过复杂的交互网络形成了具有自组织特性的智能群落。它们能够建立信任机制、发展交易体系、创造共享知识、甚至演化出独特的“文化”和“价值观”。更令人惊叹的是，这种群体智能展现出的创造力和适应性往往超越了单个智能体的能力总和，产生了真正的涌现效应。智能体之间的协作与竞争、信息传播与知识积累、资源分配与权力结构，共同构成了一个微缩但完整的数字社会生态。这不仅为研究社会学

和经济学提供了全新的实验场，更为未来人机共生的元宇宙社会提供了重要参考。

核心案例：《inZOI》的生活模拟AI生态系统

Krafton开发的《inZOI》将智能体技术推向了新的高度。游戏中的每个“Zoi”都是独立的AI个体，拥有基于大语言模型的决策系统。这些智能体能够根据自身的性格、经历和当前情绪做出复杂决策，从职业选择到社交互动都展现出高度的自主性。更重要的是，当数百个Zoi在同一个城市中生活时，他们会自发形成社区文化、经济网络和社会阶层。游戏中曾出现过Zoi们自发组织的音乐节、因经济危机导致的失业潮、甚至是基于共同兴趣形成的亚文化群体。这种完全由AI驱动的社会模拟不仅为玩家提供了永不重复的游戏体验，更展示了智能体技术在构建自演化虚拟世界方面的无限可能。

Part III

AI走向物理世界

随着AI技术的不断突破，智能应用正从虚拟空间向现实世界快速延伸，跨越了从“比特”到“原子”的重大跃迁。AI不再局限于信息处理与数据分析，它开始具备了在物理世界中感知、理解、推理和行动的能力。空间智能的兴起，使得AI从单一的词元（Token）处理，转向对三维世界的理解与生成。这一进展赋予了AI全新的能力：它能够预测和操控三维空间的动态变化，推动自动驾驶、机器人制造、XR混合现实等领域的变革。通过多模态感知与物理常识，AI正在为智慧城市、医疗手术、建筑设计等提供前所未有的技术支持，重塑我们的生活与工作方式。

与此同时，具身智能的崛起让AI开始从数字大脑走向现实代理人。与机器人平台、自动驾驶系统以及智能穿戴设备的深度整合，使得AI能够在复杂的物理环境中灵活地完成导航、操控与交互任务。例如，配送机器人在城市街区的自主行动、家庭助理机器人在具体场景下的服务反应，都是具身智能技术的实际应用。这一技术的成熟标志着AI不仅能“想得明白”，还能够“动得精准”，极大地拓展了AI在物理世界中的应用场景。通过这一系列的技术突破，AI正在完成从“虚拟计算”到“物理执行”的跨越，推动着人类社会向更加智能、高效、协作的未来迈进。



08

趋势

具身智能的“GPT-2时刻”： 基础模型、数据工程与软件平台的 协同进化

作者：徐一平

摘要

从生成式人工智能特别是GPT展现出的历程来看，具身智能领域正通过规模效应蓄积飞跃动能：2025年极有可能成为具身智能领域的“GPT-2时刻”。以VLA（视觉-语言-动作）多模态大模型突破为代表的诸多进展标志着具身智能从专用场景和单一任务向更通用、更智能、更具自主性的机器智能迈出了关键一步，其影响将进一步激活机器人在人居环境的应用潜能。具体来说：一是强大的端到端多模态基础模型正在被构建，它们将赋能机器人大小脑向更高级认知与执行能力跃升，并展现出一定泛化性；二是大规模的真实与合成数据正在以前所未有的规模被生成和利用，为模型的训练提供支撑；三是跨模态跨本体的软件平台正在统一开发流程；腾讯、英伟达等互联网公司通过构建机器人模拟和训练平台，加速技术的落地，激活万亿级的上下游生态。

一个完善的机器人系统是对人类的无限趋近（近）和仿真，因此其结构主要由机器人的大脑、小脑、本体和感知系统构成。小脑能自动或者快速地维持身体平衡、调节肌肉张力、控制步态和身体姿势、协调运动，而大脑既控制我们的运动，也进行逻辑思维、语言、识别和理解等高级认知活动。

生成式人工智能特别是以GPT为代表的大语言模型，以其惊人的涌现能力和泛化特性彻底颠覆了对AI发展速度的普遍认知。如今，一股相似的、甚至更为深刻的变革浪潮，正在具身智能（Embodied AI）领域酝酿。2025年极有可能成为具身智能领域的“GPT-2时刻”——一个能力飞跃的临界点。这一飞跃的动能，源于基础模型、数据工程和软件平台三大核心要素的协同进化与规模蓄积，它们正共同推动具身智能从特定任务和传统控制技术，迈向更通用、更智能、更具自主性的未来，其影响将深远地改变机器人和人工智能的应用前景，激活一个万亿级美元的新赛道。

基础模型向多模态进化，加速具身智能“大脑”向高阶认知跃升

当前具身智能的技术路线还在发散探索期。代表性的技

术架构有端到端、分层训练等等，并融合了包括语言、视觉、动作甚至力触觉等单一模态甚至多模态的能力。如同大语言模型赋予AI理解和生成文本的能力，具身智能领域的基础模型，特别是视觉-语言-动作（Vision-Language-Action，VLA）大模型，正在成为机器人实现高阶认知与执行能力的核心。它们是具身智能的“大脑”，让机器人不再是简单的执行工具，而是能够理解意图、感知环境、积累经验并自主决策的智能伙伴。

传统机器人系统通常采用模块化设计，即感知、规划、控制等环节各自独立，通过预设规则或特定算法进行连接。这种模式在特定、结构化任务中表现良好，但在面对复杂、动态、非结构化环境时，其泛化能力和鲁棒性严重受限。每一个新任务或新环境都需要大量的重新编程和调试，效率低下。

VLA大模型的出现，标志着具身智能迈向端到端学习的新范式。这些模型能够直接将原始的多模态传感器输入（如摄像头图像、深度信息、触觉数据，甚至语音指令）映射到机器人的低级或高级动作输出。VLA模型将视觉、语言和动作紧密耦合的能力，通过大规模数据训练，学习到视觉、语言和动作之间的复杂关联。

VLA模型具有如下特征：一是多模态理解与指令遵循。VLA模型能够解析自然语言指令，结合视觉信息理解场景，并将其转化为具体的机器人操作序列。二是高阶认知与情境感知。VLA模型能够超越简单的物体识别，理解场景的语义和功能。这种情境感知能力使得机器人能够做出更智能的决策。三是泛化性与零样本/少样本学习。基础模型通过在海量多样化数据上进行预训练，VLA模型能够学习到通用的世界模型和操作技能。这意味着，当面对一个全新的任务或未曾见过的物体时，机器人无需从头学习，而是可以利用其已有的知识进行推理和泛化，甚至通过少量示范或自然语言指令就能完成任务。

从基础模型的发展趋势来看，机器人的大小脑正在协同进化，从感知到执行逐步实现端到端统一。VLA大模型正在模糊传统机器人架构中“感知”与“执行”的界限。它们不再是独立处理视觉信息或语言指令的模块，而是将这些信息直接融入到动作生成过程中。这种“大小脑”协同进化的模式，使得机器人能够更流畅、更自然地响应环境变化和人类指令。

真实与合成数据双路并进，驱动具身智能“燃料”生产向规模化迈进

大模型的成功离不开海量高质量数据的支撑。对于具身智能而言，数据不仅包括视觉图像、语言文本，更重要的是与机器人动作、环境交互紧密结合的具身数据。大规模的真实与合成数据正在以前所未有的规模被生成和利用，为具身智能模型的训练提供源源不断的“燃料”。

真实数据成本高昂但不可或缺。真实世界数据具有无与伦比的真实性和多样性，是训练具身智能模型不可或缺的组成部分。然而，收集高质量的真实具身数据面临巨大挑战：采集和标注成本高昂、安全性与伦理、数据稀疏性与长尾等问题。尽管面临挑战，但通过众包、远程操作、以及在受控环境中进行大规模数据收集（如谷歌的Open-X Embodied Datasets），研究者们正在努力克服这些困难，为具身智能模型提供宝贵的真实世界经验。

合成数据是一种容易规模化、多样性与可控性的新范式。为了弥补真实世界数据的不足，合成数据（Synthetic Data）正成为具身智能数据工程的关键组成部分。通过高度逼真的模拟器生成合成数据并带动了上游动作捕捉，下游数

据工程等生态发展。

具身智能的数据工程并非简单地选择真实数据或合成数据，而是二者的有机结合，形成一个“数据飞轮”，如此可以逐步闭合逐步闭合sim-to-real-to-sim gap。例如，NVIDIA的Isaac Sim和Omniverse平台，以及腾讯等公司构建的机器人模拟和训练平台，正是这一数据飞轮的核心。它们提供了高保真度的物理仿真、丰富的资产库和灵活的API，使得开发者能够高效地生成合成数据、测试机器人策略，并加速从模拟到真实世界的迁移（Sim-to-Real）。这种大规模数据生成和利用的能力，是具身智能模型能力飞跃的根本保障。

专用软件平台实现应用，推动具身智能统一开发与生态融合

具身智能的迈向通用离不开开源开放的软件平台支撑。这些平台不仅为基础模型的开发和部署提供基础设施，更重要的是，它们正在实现跨模态、跨本体的统一开发流程，极大地降低了机器人开发的门槛，加速了技术的落地。

过去，机器人开发往往是高度碎片化的，不同的机器人硬件、不同的传感器、不同的任务都需要定制化的软件栈。这导致开发效率低下，知识复用性差。如今，具身智能的软件平台正致力于构建统一的开发框架和工具链。

这个平台往往具备如下特征：一是跨模态集成。平台提供统一的接口和库，方便开发者集成和处理来自不同模态的数据，如视觉、听觉、触觉、本体感知等。例如，ROS（Robot Operating System）作为机器人领域的“操作系统”，虽然不是专门为大模型设计，但其模块化和通信机制为多模态数据的集成提供了基础。二是可以支持模型训练与敏捷部署。平台集成了主流的机器学习框架，并提供针对具身智能优化的训练工具和部署方案。开发者可以在同一平台上完成从数据预处理、模型训练、到模型部署和推理的全流程。三是无缝衔接仿真与真实世界。高度逼真的仿真环境是软件平台的核心组成部分。这些平台不仅提供物理引擎和渲染能力，更重要的是，它们与真实机器人硬件和传感器接口保持高度一致，使得在仿真环境中开发的算法和模型能够无缝迁移到真实机器人上。NVIDIA的Isaac Sim和Omniverse平台在这方面表现突出，它们提供了从机器人设计、仿真、训练到部署的端到端解决方案。

具身智能软件平台的一个重要目标是实现“一次训练，多处部署”，即一个训练好的基础模型能够适应不同形态的机器人本体。例如，谷歌的Gemini Robotics On-Device模型，虽然在训练时使用了特定的炒虾机器人ALOHA，但实验表明其模型能够扩展用于双臂和人形机器人。同样，腾讯用于“浇花”的大脑模型，也支持了腾讯自家的Max机器狗和宇树的机器狗。这种跨本体的泛化能力，极大地提升了模型的复用价值，降低了机器人开发的边际成本。

为了加速具身智能的普及和应用，领先的科技公司正致力于构建开放的软件平台和生态系统。如同GPT等大模型正在变成服务生态的开源开放平台，具身智能的基础模型和软件平台也将走向开放。腾讯、英伟达等互联网公司通过构建机器人模拟和训练平台，不仅加速了自身技术的落地，更重要的是，它们正在激活一个庞大的万亿级上下游生态。这个生态将包括硬件制造商、传感器供应商、算法开发者、系统集成商、应用服务商等，共同推动具身智能的商业化进程。Robotics X 实验室正在搭建具身智能开放生态，定位为开源开放平台，助力机器人企业，从数据共建、算法能力和仿真环境支持，协助机器人企业进行更好的商业场景落地，全方面建立具身智能的开放生态平台，推动机器人行业的发展。

展望：具身智能“GPT-2时刻”临近、开启万亿市场新纪元

具身智能的“GPT-2时刻”，它是基础模型、数据工程和软件平台长期积累和协同进化的必然结果。强大的端到端多模态基础模型赋予机器人高阶认知与泛化能力；大规模真实与合成数据为模型训练提供充足燃料；统一的软件平台则为开发、部署和生态构建提供了坚实基础。

资金正是为这种认知世界的高阶潜力买单。2024年最受VC青睐的10家机器人企业中，明确把“机器人脑大模型”列为核心研发对象的就有6家。当前市场共识已经从“比谁的关节多、负载大”转向“比谁的大脑更会认知”。大脑认知尺度的不断升维，将让机器人实现跨模态、跨任务、跨本体的能力跃迁，从而让AI和机器人真正实现泛化和通用，点燃巨大的商业想象空间。

当前，具身智能的应用仍然是分散的，但趋势已经非常显著：借助具身智能大模型，机器人正快速聚合自然语言理解、多模态感知、运动规划、任务记忆等多种能力，为机器人植入了可迁移的“通用大脑”。这一技术突破正在推动“模型定义场景”的模式在柔性制造、商业零售和家庭服务三大领域同时应用落地。

基于人工智能技术迎来从符号推理到具身认知的范式跃迁，通用人工智能 (AGI) 的曙光初现，具身智能 (Embodied AI) 作为打通数字与物理世界的“神经中枢”，正重新定义机器人的能力边界——从被动执行指令的机械体，进化为能主动感知环境、理解意图、动态决策的“智能生命体”。

在柔性制造领域，人形机器人凭借多模态感知大模型和模仿学习能力，能够完成摘取、移动、组装、质检等一系列精细操作，甚至能像特斯拉Optimus那样，在充分感知周围物体和环境的基础上，快速响应并执行工程师的口头指令，展现出前所未有的适应性与生产效率。在商业零售场景，机器人正重塑客户体验，通过自然语言对话深入挖掘顾客需求与偏好，结合实时库存与权益给出精准商品推荐；它们还能对价格、配置、技术参数乃至购车政策进行专业深入问答，并依据顾客意图导航至对应货架或展车完成现场演示与扫码下单，实现从获客到成交的完整闭环服务。而在家庭服务中，机器人则能提供任务理解、精准家政与情感陪伴等多重价值。例如，人形或移动操作机器人只需听到“帮我收拾房间”这一句话，即可调用大模型的推理与规划能力，将水果、餐具、垃圾等不同类别物品自动识别、分类并放置到指定区域；当主人询问“面前这幅画表达的是什么”时，机器人又能通过工具调用（本地知识库或网络检索）快速获取背景信息，给出符合场景的解释，真正实现“能干活也能聊天”的智能陪伴。

展望未来，机器人模型的能力将如同GPT等大模型，演变为通用平台，释放出驱动万亿级市场的磅礴动力。从工业制造到物流仓储，从医疗护理到家庭服务，从探索未知到辅助科研，具身智能将以前所未有的深度和广度融入人类社会，开启一个由智能机器人驱动的全新时代。我们正见证着一个新物种的诞生，一个能够理解、感知、学习和行动的智能实体，它将深刻改变我们的生产方式、生活方式，乃至整个社会的面貌。

趋势
09

空间智能： 从会聊天到真正懂世界

作者：李瑞龙

摘要

随着AI技术的不断突破,智能的应用正在从语言智能走向空间智能。空间智能的兴起意味着AI从处理词元(Token)进化到理解体素(Voxel),具备了理解 and 处理三维世界的核心能力,包括3D环境的感知、推理、交互和生成。这一技术突破让AI能够像预测下一句文本一样去预测三维空间并涌现。空间智能正在重塑自动驾驶、机器人制造、XR混合现实、医疗手术、建筑设计、智慧城市等领域的工作方式,为AI走向通用人工智能(AGI)提供关键的物理常识和因果推理能力,推动AI从“会聊天”到真正“懂世界”的跨越式发展。

技术架构变革,从二维信息处理向三维空间理解跃迁

李飞飞指出,当今大语言模型(LLM)的基本单位是文本词元,而“世界模型”(即对3D世界的生成式模型)的基本单元则是像素或体素。这意味着AI正从处理平面信息迈向处理立体信息的新阶段。她创建的新公司World Labs的使命,正是攻克这一3D生成式基础模型难题——即让AI像预测下一句文本一样去预测三维空间。通过从图像或文本生成可交互的3D场景,模型将不再仅输出二维平面上的像素,而是输出构成三维环境的体素。World Labs已经展示了“一张图生成一个3D世界”的原型模型,实现了从单幅图像或一句话生成完全模拟且可交互的三维场景。目前在大型语言模型竞争白热化的阶段,国内外不少企业已经开始提前部署这项下一代的核心技术,除李飞飞的World Labs,国内如群核科技等均已开展提前布局。

随着深度学习与传感技术的发展,一系列支撑空间智能的关键技术正逐步成熟,包括:

三维感知:赋予机器获取3D信息的“眼睛”。例如立体相机、LiDAR激光雷达、深度传感器等,使AI能够获取场景的深度和几何结构。Google研究人员近年开发的算法可将多张照片重建出3D场景(如神经辐射场NeRF技术),标志着机器3D视觉取得突破。3D信息捕捉技术,正在为空间智能积攒越来越多高清数据源,借鉴大型语言模型的经验路径,这是空间智能的scaling law的重要驱动因素。

空间表示与场景生成:有效表达和生成三维环境的“脑”。这包括点云、体素网格、mesh网格以及神经场等3D表示学习,以及结合物理法则的世界模型构建。值得注意的是,当前部分视频生成模型也可以生成360°视频环境并号称世界模型,但事实上存在本质的区别。预测整个3D场景比逐帧生成2D图像有诸多优势:场景具有持久性(视角改变时世界保持一致),且允许实时交互控制(生成场景后,用户可自由移动视角观察细节)。未来发展方向之一是需要找到更加高效的空间表示方式。

模态融合：将视觉、语言、运动等多源信息整合的"中枢神经"。比如，机器人是高度多模态的系统，需要将触觉、视觉、听觉和语言等感知整合才能完成复杂任务，6月初，Figure AI展示了其人形机器人Helix在真实的物流仓库中的类似人类处理包裹、扫码的微妙操作，这种复杂的行为并非通过硬编码实现，而是通过海量的数据中自主学习得到。当前的研究趋势是让大型模型（不再满足于或局限于语言、词元）同时处理文本、图像和3D数据等，实现跨模态理解能力。基础模型正在从纯语言向语言-视觉-空间多模态的拓展。

整体来说，空间智能基本技术栈正在日趋完善，以上3D感知（深度相机、LiDAR等）、空间表达与生成（点云/体素/NeRF等表示，世界模型构建）、模拟仿真训练（虚拟环境中的交互学习）、跨模态大模型（同时理解文本、图像、3D的基础模型）等，上述技术的综合，正在使AI具备"从词元到体素"的预测生成能力，真正理解、创造三维世界。

智能进化驱动，空间智能成为AGI发展的关键拼图

Meta首席AI科学家、图灵奖得主杨立昆(Yann LeCun)一个著名的论断："当前AI系统虽然在语言处理上表现出色，但在理解和交互物理世界方面仍远远落后于一只家猫"。这也是他认为当前大型语言模型无法实现AGI的一个重要原因。空间智能在AGI中有着十分关键的作用：

赋予AI"看见即理解"的常识推理能力。空间智能被李飞飞视为通往人工通用智能(AGI)的关键拼图。在人类和动物智能中，视觉与运动进化出感知-洞察-行动的良性循环：看见转化为理解，理解指导行动。AGI不仅要会聊天（语言智能），更要"懂世界"。空间智能让AI具备物理常识和因果推理：知道水杯若被推到桌边就会掉落，明白支撑与平衡、遮挡与碰撞等基本常理，这样的常识性理解正是当前大模型普遍缺乏的。当下最前沿的AI公司正竞相赋予模型这种常识推理和物

理直觉，以弥补纯语言模型在逻辑一致性上的不足，但表现仍有较大的提升空间。包括杨立昆在内的众多专家认为，这种基于对现实世界的理解而形成的"理性"是AGI的必备条件。

支撑具身智能与行动能力。AGI被期望不仅能思考，还能像人一样行动。若希望AI超越当前能力，我们需要的不只是会看会说的AI，还要会行动的AI，也就是Embodied AI（具身智能）。AI以机器人等形态嵌入物理世界，与环境交互。空间智能为具身智能提供了"眼睛"和"大脑"支持：一方面，3D感知和定位使智能体知道自身在空间的位置和周围物体的形状、距离（例如自动驾驶汽车通过多传感器融合"看"周围环境）；另一方面，空间推理使智能体能够规划路径、躲避障碍、操纵物体，实现从感知到行动的闭环。有了空间智能的AI，才能驱动机械臂精准抓取，令服务机器人灵活导航，真正参与到人类的日常活动中。可以说，空间智能是当前国内外发展人形机器人、具身智能的关键卡点，一旦突破，机器人才可能真正的走入我们的日常，而不是只在工厂的流水线上，通过硬编码执行机械式的操作。

支撑AGI实现其因果推理、反思的正循环。当前的大型语言模型在缺少物理语境时容易产生荒谬结论，而融入空间智能的AI由于理解真实世界的物理规律，将更擅长因果推理和常识判断。大型模型迭代的过程中所需要的核心反馈，不管是RLHF人类反馈强化学习，还是RL大规模纯强化学习，正确有效的"反馈"是稀缺和珍贵的。而当具有世界模型的AI可以通过视觉观测推断事件可能的后果（如杯子掉落会摔碎），从而采取合理行动预防。这种"会思考的视觉"对于机器人自主决策、自动驾驶规避风险、智能体规划复杂任务至关重要。空间智能让AI像婴儿那样通过与环境互动逐步学习因果关系，在模拟环境中尝试推倒积木、打开门锁等，失败和成功的经验积累为模型提供对物理世界机制的理解，从而为AGI奠定了与现实对接的认知基石、反馈链路。

产业应用爆发机会，多行业共性需求驱动技术迭代

虽然关联AGI，但空间智能并不是一项遥远的技术，其作为AI理解和重构现实的能力，在多个行业都有共性需求，驱动着新一轮技术变革，蕴含重大发展机遇，比如：

自动驾驶与无人机。智能交通工具是空间智能的重要用武之地。自动驾驶汽车和无人机都必须实时感知3D道路/空域环境，识别行人、障碍物和动态变化，并据此规划安全路径。它们依赖激光雷达、摄像头、雷达等多传感器融合构建高精度3D地图和环境模型，然后利用因果推理预测其他车辆和行人的运动轨迹，从而决策驾驶/飞行行为。共性需求在于高精度的空间感知和定位（厘米级定位、毫秒级反应）、可靠的环境建模与预测（包括夜间和恶劣天气条件），以及冗余安全机制。驱动力主要来自交通安全和效率提升：空间智能有望大幅降低交通事故率，实现车辆群体协同和拥堵缓解。当前，理想汽车、元戎智行等提出的VLA模型（Vision-Language-Action）等，正在重塑未来交通运输方式。

机器人，含工业制造。无论是工厂里的机械臂还是服务型机器人，都需要准确感知三维环境和自身位置，以安全高效地完成操作。共性需求在于精准的环境感知与避障、灵巧的操作规划以及人机协作安全。这在服务型机器人应用中尤其关键，因为服务环境的多样化与不确定，如养老机器人等，机器人需要通过3D视觉+力觉识别零件位置和姿态，再辅以空间推理规划装配路径，避免与人或其他机器碰撞。驱动力主要来自于机器人的商业化，空间智能有望让机器人胜任更复杂精细的工作（提高服务效率和质量），并通过环境理解提升安全性（防止事故）。腾讯Robotics X实验室发布的居家服务机器人“小五”，正在通过空间智能的提升，推进养老场景中的实际应用。

未来终端，XR混合现实。手机已经不是AI的最佳载体，AI眼镜、AR/VR/MR头显等，正在抢夺未来入口与生态位。空间

智能将帮助用户更加自然地与现实世界和数字世界进行交互，比如Meta的Oculus Quest系列、苹果的Vision Pro头显设备已初步具备一定的空间计算能力，让佩戴用户可以连接现实世界与数字内容，创造独特的交互方式；包括两年前火热的元宇宙概念，因为模型资产成本过高，生态内容缺乏而折戟，而通过3D场景的自动化重建技术提升，以及随着空间智能逐步理解、创造三维世界的能力提升，可以预见，元宇宙必然会卷土重来，成为人类现实生活空间的一个重要延伸。在这个新的赛道，得益于智能终端的产业链优势、国人的眼镜使用习惯，以及3D资产（包括游戏资产）积累等等，中国具备反超的可能性。

医疗、建筑与智慧城市等。诸多传统行业正出现融合空间智能的创新应用。在医疗行业中，通过增强现实（AR），外科医生术中可叠加患者体内三维模型，获得导引从而提高手术精准度，减少创伤；在城市管理中，通过城市的摄像头、传感器，AI可以实时感知城市的人流车流分布和空间状态，从而进行智能调度和管理。比如：城市交通摄像头结合计算机视觉，能监测道路车辆密度，自动调整红绿灯配时优化通行。通过空间智能或取代传统数字孪生技术，低成本的建立城市3D模型，规划者可以模拟新建筑或道路对周边的影响，进行更科学的设计决策。类比元宇宙的发展，过往由于成本高昂、协同难度大的数字孪生行业应用，生成式AI赋能的数字孪生与仿真正焕发新机。



10

趋势

测试转量产， 应用推动具身智能本体加速成熟

作者：刘莫闲

摘要

2025年3月发布的《2025年国务院政府工作报告》，首次将智能机器人定位为“新一代智能终端和智能制造装备”，其纳入“人工智能+”行动计划。同时《工作报告》还明确提出“培育具身智能等未来产业”，这标志着具身智能已上升至国家战略高度。随着技术突破和应用场景的不断拓展，具身智能本体 - 机器人正从实验室走向产业化，迎来从测试到量产的关键转折点。本专题深入分析具身智能本体发展由“测试”向“量产”的重大转变，其硬件配置、能力提升和产业影响三个维度的最新发展趋势，并探讨其如何在应用驱动下加速走向成熟。

具身智能“大脑”的快速进化，补齐了具身智能机器人应用的一块关键的“拼图”，具身智能的生产力得到明显提升，本体机器人向规模化量产和应用快速迈进。

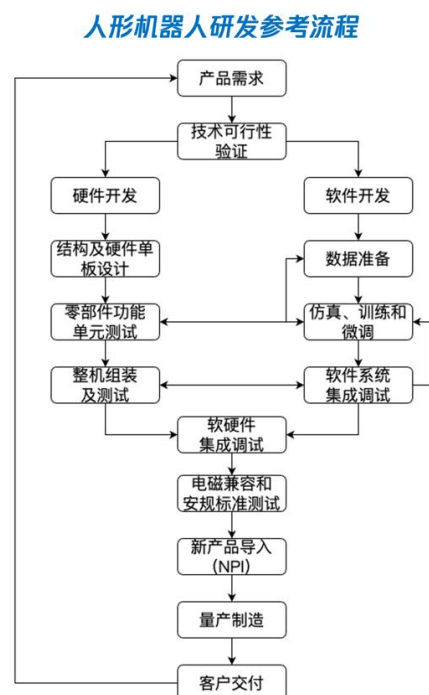
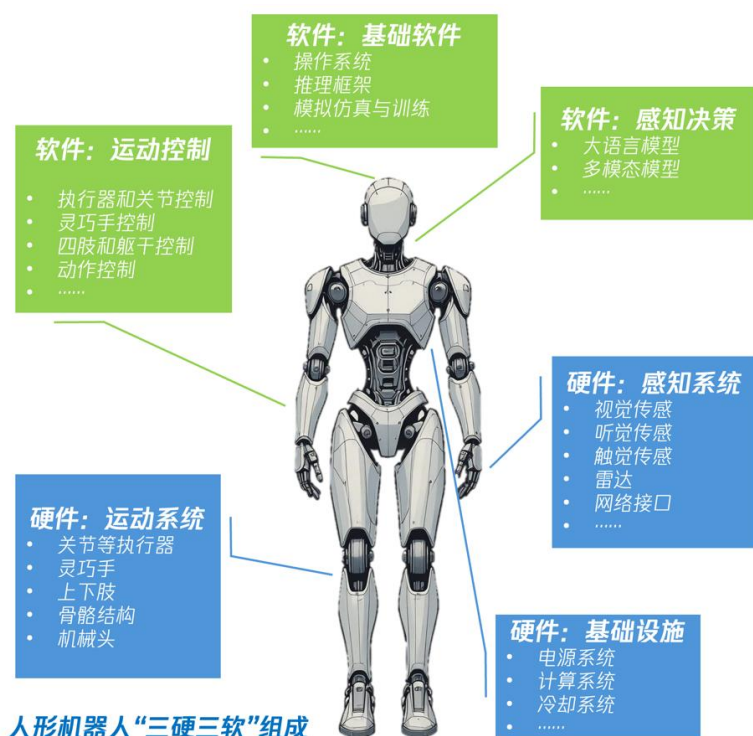
2025年，各大具身智能机器人厂商开始加大机器人在工业、物流、仓储、零售等领域的试点力度和深度，并基于试点的反馈持续迭代和优化具身智能本体的硬件配置、控制策略、以及运行效率和稳定性。产业界开始迈出从“测试”走向“量产”的重要一步。根据公开信息报道，Tesla Optimus、1X Neo、Agility Digit、优必选、傅里叶GR-1和众擎机器人等具身智能厂商均有1000台左右的量产计划。随着具身智能本体量产的开始，其配置会更具体稳定，能力更贴近实用，并会逐步开始在高危险、低技能、高重复的应用场景提供有力的劳动力补充。

配置：“运动、感知和基础”三系统配置基本定型，大规模量产和应用在即

具身智能本体的整体架构设计全面服务于大规模量产。呈现出三方面特点：人形形态更受关注。市场主要形态包括人形机器人、多足机器人和轮式底盘机器人三大类。其中，人形机器人因其能够更好地体现人工智能技术优势和满足人机交互期望，正成为市场关注焦点。Figure 02、1X Neo等代表性产品都在具身智能大模型等技术应用的推动下快速迭代。

集成度显著提升。本体各部件的集成度和模块化程度持续提高，这一趋势体现在两个方面：一是关键元器件的集成化程度不断增强，如最新的转动关节组件已实现电机、减速器、控制器、编码器等部件的深度集成，有效保障了质量和性能的一致性；二是外部结构设计更加简洁轻巧，灵巧手和上下肢的优化设计显著提升了运动敏捷性和交互友好程度。**模块化程度深化。**模块化设计理念的深入应用，大幅降低了制造、组装、定制和维护等环节的复杂度，为大规模生产制造、标准化运输组装以及集中化维护服务奠定了基础，这对于具身智能产业化具有重要意义。

运动系统作为具身智能本体的“骨骼和肌肉”，技术和配置趋于收敛。运动系统的发展水平直接决定了具身智能本体在物理世界中的行动能力和生产效率。其基本构成包括关节执行器、灵巧手、上下肢组件以及骨骼结构等核心部件。优秀的运动系统在运动控制算法和系统的精确控制下，能够高效、平顺、协调地完成各类复杂动作和人机交互任务。**下肢配置基本收敛为双足和轮腿两种主流方案。**双足配置如Tesla Optimus、宇树G1提供了更高的灵活性和表现力，适用于复杂地形和精细操作场景；轮腿配置如腾讯“小五”护理机器人，则更适合需要平稳可靠移动的应用场景。**上肢配置方案已收敛为“双臂+末端操作”，**且末端操作技术日臻完善。根据应用需求的不同，末端操作主要分为2-3指灵巧手和5指多自由度灵巧手两类。当前灵巧手技术方案正处于快速迭代期，



驱动方案普遍采用空心杯电机或无刷有齿槽电机。典型的多自由度传动方案采用三级传动架构：一级传动采用行星齿轮箱实现减速和扭矩放大；二级传动采用高精度丝杠（如行星滚柱丝杠）实现旋转-直线运动转换；三级传动采用腱绳方案，充分利用其远距离传动优势和布置灵活性。**运动系统未来的迭代趋势更加明确。**在减速器、丝杠、电机等执行器方面，高扭矩、准直驱和优异散热性能成为重要发展方向，以满足制造业、物流仓储等应用场景对负重能力和高频重复作业的要求。灵巧手方面，尽管存在不同技术路线，但整体趋势是向“够用而不贵”的方向发展，通过成本优化提升实用性。

感知系统中“听看力触网”五感成标配，感知技术基于应用需要持续演化。感知系统作为具身智能本体的“五感和神经网络”，通过视觉、听觉、触觉等多模态传感方式，实时感知周围环境的空间关系、物体位置和物理特征，为机器人实现自然人机交互、自主决策和任务执行提供核心数据输入。感知系统的技术水平直接影响具身智能本体的智能化程度和应用效果。其中，**视觉方案基本定型。**当前视觉方案主要收敛为两种技术路线：一是双目摄像头方案，代表产品包括Tesla Optimus Gen2、Figure 02和1X Neo等；二是深度摄像头方案（采用3D结构光和ToF传感器），代表产品包括宇树G1、

Sanctuary Phoenix、众擎SE01、优必选 Walker S1等。同时，针对特定应用场景，视觉方案会适配额外的传感器组件，如后置摄像头、激光雷达、红外热成像传感器、毫米波雷达等，以实现更全面的环境感知。作为与运动系统想强化的力反馈技术，**力矩传感器配置区域集成化。**力反馈方案已基本收敛为集成化解决方案，主要包括两种配置：一是将一维力矩传感器集成在线性执行器和旋转执行器内部；二是将三维/六维力矩传感器集成在灵巧手末端，实现对功率和负载的实时精确感知。**触觉感知技术持续创新演进。**触觉传感技术尚未完全标准化，但基于压阻和电容原理的电子皮肤技术已成为主流方案。腾讯RoboticsX实验室和香港城市大学联合研发的电触觉传感器代表了该领域的技术前沿，实现了高达76点/厘米的空间分辨率和4kHz的快速刷新率。**其他感知组件配置稳定。**麦克风阵列、扬声器、WiFi和蓝牙模块已基本成为具身智能本体的标准配置，为语音交互和网络连接提供基础支撑。**感知系统面向场景定制化趋势明显。**感知系统正朝着场景化定制优化方向发展，Figure 02 在物流分拣应用中的视觉系统升级就是典型案例。该产品原本采用单目摄像头配置，但为了在物流分拣场景中更好地识别形状各异的包裹，进一步采用了上下排布的双目摄像头设计，显著提升了多角度、多尺度视觉感知能力。

基础设施系统延用自动驾驶新能源车配置，并逐渐形成专有方案。作为具身智能本体的“内脏”，为硬件运行提供动力支撑，为AI模型运行提供算力保障。**电源系统因普遍延用电动汽车行业的成熟锂电池技术而相对成熟**，单块电池组可支持机器人连续工作2-6小时。主电源多采用躯干集成或背负式设计，快拆结构便于连续作业场景中的快速更换，如优必选天工行者、宇树G1等。未来，具身智能本体的电源系统技术发展趋势包括：全固态电池技术、锂硫和高镍三元锂等新材料应用、智能电池管理系统、快充和无线充电技术等。这些技术的应用将实现更高能量密度、更大功率输出、更优散热性能、更高安全性以及更强的极端环境适应能力。**计算系统逐渐从类服务器配置升级为具身智能专有方案。**具身智能本体底层执行器控制采用MCU内部集成方案，负责执行部件的精确控制。中央计算单元的传统方案采用的Intel CPU + Nvidia GPU配置。而随着AI专用计算需求的增强，中央计算单元逐渐采用SoC方案，进一步集成了专门用于神经网络计算的张量处理单元（NPU），形成CPU+GPU+NPU的三合一配置。Tesla Optimus采用的FSD方案、Figure 02和波士顿动力Atlas采用的Nvidia Jetson方案等都体现了这一趋势。同时，高通、AMD、地平线、寒武纪、华为等厂商也在积极布局具身智能专用计算芯片。未来计算系统将呈现两大发展趋势：计算能力方面，将更好地适配端侧部署的VLA等多模态AI模型，提供更强大、更实时的推理能力；在系统设计方面，Chiplet、ASIC等芯片级集成技术将逐步应用，实现更高计算带宽、更低能耗和成本优化。此外，神经形态计算等新一代低功耗计算范式也将逐步进入实际应用阶段。

能力：应用推动机器人协调和协作能力更强，综合素质进入“小学生”水平

人形机器人手、眼、脑更协调，实现从感知到认知的升级。遥控技术普及和仿真训练环境的完善，使得视觉-语言-行动（VLA）多模态模型开始发挥其技术价值，并进一步推动具身智能实现从“能看到听到”向“能看懂、听懂和做到”的重要跃升。应用试点中，具身智能机器人在工业装配、物流分拣、仓储搬运、家庭服务等领域展现出更高的手眼脑协调性，**任务执行的有效吞吐量大幅提升**。例如，Figure 02在物流分拣场景中的表现充分展现了这种协调能力，机器人能够准确识别不同形状、大小的包裹，理解分拣指令，并精确执

行分拣动作。Figure02的整体协调能力在执行速度提高50%后依然可以保持，并保持执行的高准确度。Agility机器人在仓储搬运场景中同样表现出色，能够在复杂的仓储环境中自主导航，识别货物并完成搬运任务。1X Neo Gamma在家庭环境中使用洗衣机洗衣服的演示，更是展现了具身智能在日常生活场景中的应用潜力。

人形机器人上下半身协调，复杂动作执行更加稳定流畅。过去一年中，人形机器人在实际应用场景中的大量测试和训练，使得运动控制算法得到更快的优化和完善，整体运动协调性显著提升。机器人上下半身已能够针对统一任务进行同步协调执行，而非传统的先后顺序完成模式。这种协调能力使得机器人能够流畅完成行走、坐下、蹲下、爬楼梯等基本动作，为复杂任务执行奠定了基础。进一步在实际应用场景中，**机器人已能够协调完成家务劳动、仓储搬运、工业装配等复杂任务**。Figure 02在宝马汽车装配线上装配汽车零件，展现了其在制造领域的应用潜力。Agility机器人在仓储货架上存放物品的稳定表现，使其在物流仓储领域加快规模化应用。1X Neo Gamma能使用吸尘器清洁地面、从洗衣机里收捡衣服，使其在家庭服务领域的应用前景广阔。此外，马拉松比赛、拳击赛、足球赛等**运动竞技类应用已成为人形机器人验证和改善运动协调性、敏捷性的重要训练场所**。这些极限场景的训练不仅提升了机器人的运动能力，也促进了感知系统与运动系统的深度融合和高效配合。

多本体协作展现集群智能雏形。基于统一控制指令和多模态推理大模型的发展，多机器人协作能力正在快速提升，集群智能应用开始显现。统一执行方面，多个机器人已能够基于统一控制指令同步完成相同动作，如集体舞蹈、队列行走等。协作执行方面，多模态推理大模型已能够控制2台甚至更多台具身智能本体进行智能协作。**当前多机协作已呈现出任务分工、空间协调、时序配合等多种模式，为不同应用场景提供了灵活的解决方案**。Figure公司的Helix VLA系统能够控制两台Figure 02协同完成超市物品的分类和收纳任务，展现了多机协作在零售、养老、家庭等领域的应用前景。两台傅里叶GR01机器人基于Nvidia GROOT N1人形机器人基础模型协同完成工业产线零件分拣的案例，也是多机协作在制造业应用的重要探索。

影响：多形态、多协作的具身智能机器人将成为未来劳动力市场的有力补充

多形态、多协作将成为具身智能发展的重要方向。在自动驾驶和人形机器人研发及应用的推动下，具身智能算力、模型、数据集和工具集正在快速丰富完善并具备通用泛化能力。这为机器狗、轮式底盘机器人、无人机、工业和农业机械等各种形态机器人的智能化升级提供了技术基础。而**各类具身智能机器人应用场景也将持续拓展**。正在从单一场景向多场景、跨场景扩展。工业制造、物流运输、农业生产、家庭服务、医疗护理等领域都将迎来具身智能机器人的广泛应用。Weave Robotics 推出了 Isaac 具身智能轮式底盘机器人，可以完成多项家居任务，并计划在今年向用户做第一轮交付。进一步，**多种形态的具身智能之间的协作模式也将不断创新，集群智能将更加多元化**。在统一的具身智能模型和平台协调控制下，不同数量、不同类型的机器人正在应用场景内部以及跨场景之间展现出更多协作可能。工业制造与物流仓储的协同、农业生产与运输配送的联动等跨领域协作模式正在涌现，具身智能生产力将获得进一步提升。

以人形机器人为代表的具身智能将成为未来劳动力市场的有力补充。当下国际就业市场需求巨大。根据Agility Robotics的调研数据，2023年美国在仓储、物流和制造行业存在超过100万个物料搬运岗位空缺，且这一数字仍在持续增长。**虽然全球就业结构正经历深度调整，但未来就业岗位总体需求仍呈现增长趋势**。世界经济论坛2025年1月发布的《就业未来报告2025》显示，综合考虑人工智能数字化转型、全球经济波动、气候变化、人口结构变化和地缘政治等因素，2025-2030年期间全球劳动力市场22%的就业岗位将面临更新，预计新增14%约1.7亿个岗位的同时消失8%的岗位，最终实现7%约7800万岗位的净增长。结合具身智能进入大规模生产和应用的趋势，可以预见在农业、物流、建筑、制造、护理、教育辅助等工业、农业和服务业一线，包括人形机器人、机械狗、无人机、自动驾驶车辆等**各类具身智能机器人将成为劳动力市场的重要补充**。预计具身智能机器人将**主要承担重复性、低技能要求和高危险性的工作岗位**，与人类形成优势互补的协作关系，共同推动生产效率的提升和工作环境的改善。

具身智能评测体系逐步建立，“机器人绩效”将成为未来行业应用投入产出考量。目前具身智能的发展仍然处于早期快速发展阶段，相关评测正在研发和技术验证层面逐步探索和建立，例如谷歌Deepmind Robotics的评测数据集、复旦大学主导的VLABench基准测试，等等。而具身智能机器人在各行业的规模化应用日益临近，建立标准化、科学化的评测体系已成为产业发展的迫切需要。产业评测体系需要涵盖技术性能、应用效果、经济效益、安全可靠等多个维度，目前国内外相关企业也已经开始探索。Figure公司在物流领域逐渐探索建立适应“包裹分拣”等具体应用场景的评估指标“有效吞吐量”，用来衡量单位时间内有效处理包裹在总包裹数量的占比，并基于该指标探索Figure02+Helix VLA方案的峰值输出功率和持续调优。美团也针对家居和零售空间服务场景，提出RoboCAS评测基准。可以预见，具身智能评测体系将随着工业、服务等领域的应用而逐步健全和完善，并逐步演变成“机器人绩效”评估的重要参考，为企业的具身智能投产和运营的决策提供量化依据。

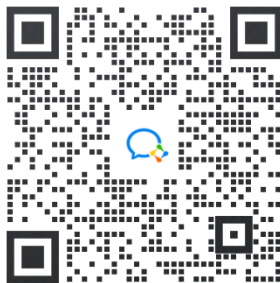
具备基本空间感知能力、能够使用多模态大模型思考、并基于人形等形态利用灵巧手完成低技能重复性工作的具身智能机器人，正站在从技术验证向规模化产业应用的重要拐点。未来几年，更多的具身智能将走进应用场景，与个人、产业和社会产生更多的互动，为经济社会发展注入新的动能。政府、企业和学术机构应加强协同合作，共同推动具身智能产业的健康快速发展，创造向善价值，为构建智能化社会贡献力量。

免责声明：

1. 本资料来源于网络公开渠道，版权归属版权方；
2. 本资料仅限会员学习使用，如他用请联系版权方；
3. 会员费用作为信息收集整理及运营之必须费用；
4. 如侵犯您的合法权益，请联系客服微信将及时删除。



行业报告资源群



微信扫码 长期有效

1. 进群福利：进群即领万份行业研究、管理方案及其他学习资源，直接打包下载
2. 每日分享：6+份行研精选、3个行业主题
3. 报告查找：群里直接咨询，免费协助查找
4. 严禁广告：仅限行业报告交流，禁止一切无关信息



微信扫码 行研无忧

知识星球 行业与管理资源

专业知识社群：每月分享10000+份行业研究报告、市场研究、企业运营及咨询管理方案等，涵盖科技、金融、教育、互联网、房地产、生物制药、医疗健康等全领域；是全网分享数量最多、质量最高、更新最快的知识社群。

加入后无限制搜索下载

