腾讯研究院 | 腾讯云智能 | 腾讯优图

协同进化

2026人工智能十大趋势报告

腾讯研究院出品

人工智能正在从单点突破走向系统性进化。模型在使用中持续学习，环境在交互中被重塑，商业在落地中寻找真实价值，组织在协作中摸索新形态——四条线索彼此咬合，构成一幅协同进化的全景。

编委会

顾　　问　司　晓 | 腾讯副总裁 腾讯研究院院长

袁晓辉 | 腾讯研究院副院长 资深专家

研究团队　李瑞龙　陈东明　茹炳晟　刘莫闲　余　一

曹士圯　翁金塔　白惠天　王　强　吴朋阳

王　鹏　黄　浩

联合出品　腾讯研究院　腾讯云智能　腾讯优图

目录

CONTENTS

Part I

进化

学习、感知和发现

趋势

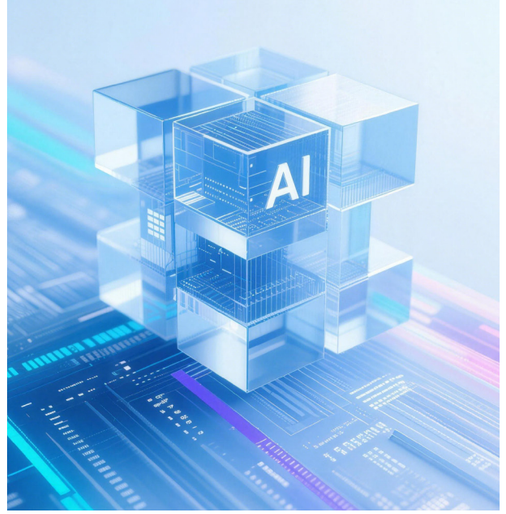
01

在线进化：进化不止于训练，模型部署后的再学习

作者：李瑞龙　黄浩

过去三年，基础模型进步的默认逻辑是“Scaling”：更多数据、更大参数、更多GPU。2026年大模型能力仍在持续提升，甚至没有人明确看到天花板。但随着时间拉长，有一个问题开始浮现：模型智商提高之后，要真正为人类所用，模型必须学会coding和数学之外的更多事。正如谷歌DeepMind的科学家Hassabis所说，Scaling时代远没有结束，但仅靠Scaling已经不够了，AGI的实现还需要一两项重大创新才能真正突破。

新东西究竟是什么？远期或许在于架构的创新，继续刷新智能前沿。但在2026年，我们认为是三条同时发生的学习曲线。这也许是在下一代颠覆性架构来临之前，进一步最大化基础模型应用的工作。模型的进化不再只发生在训练阶段，部署之后还在学，使用过程中还在长，研究本身也在被AI加速。



强化学习走出代码和数学，向更多可验证领域扩散

过去两年训练范式最大的变化发生在反馈方式上。以前是RLHF，让人类标注员判断哪个回答更好；现在是RLVR，让程序直接验证答案对不对。OpenAI的o系列和DeepSeek-R1先后证明，纯靠程序反馈做强化学习，推理能力就能自己涌现，不需要监督微调。优势在于绕开人工标注，自主产生高质量训练信号，模型在不断试错中变强。代码和数学最先被突破，因为验证信号最容易获得。问题是这件事并不自然发生在所有领域。代码幂律曲线已开始递减，材料科学、超导体发现等方向远未饱和。RLVR正从接近天花板的领域，向这些空白地带扩散。

为什么之前扩不出去？Anthropic在一项研究中提出了相关的思考。设想一下，让AI在生物数据基础设施中导航，就像在汽车发明之前设计的意大利山城里开车，街道虽美，但现代车辆根本开不进去。代码领域最先被突破，一方面是因为答案容易验证，另一方面整个软件基础设施天生为程序化访问而建。但生物学、材料科学、临床医学，大量关键数据被困在手动网页流程和分散数据库里，模型触达不到。障碍正在被清除。一个病毒学领域的实验（VirBench，120个真实查询）显示：只要加一个确定性检索工具层，所有模型准确率都超过90%，最好达99.7%。换句话说，基础设施质量对结果的影响甚至大于模型能力本身，铺好数据访问的路，模型自然跑得通。这件事的产业含义在当下值得重视：如果基建质量的权重高于模型能力，那么科学AI的话语权就会部分从模型公司，转移到那些握有领域数据通路的机构。谁控制数据进出的闸门，谁就掌握了RLVR下一波扩散的入场券。模型不再是唯一的护城河。

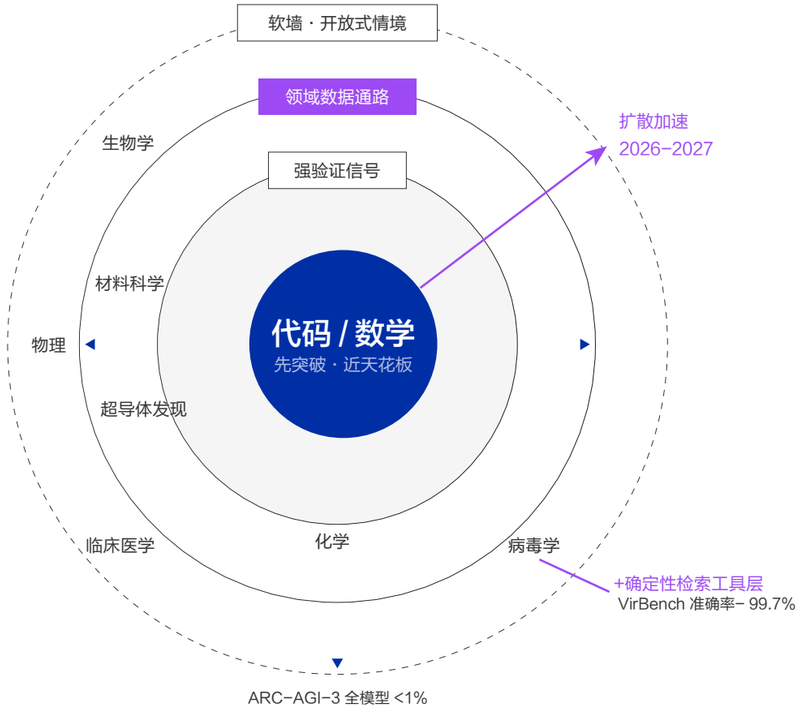


图1–1 · RLVR 同心扩散，从代码数学向科学前沿加速

更多专业领域的天花板仍待攻破。一方面是各专业领域的数据基建还停留在上个阶段，真正值钱的领域数据还锁在国家实验室、学术机构和半导体工厂里。验证本身没问题，基础设施还停在马车时代。谁先铺好程序化访问的公路，谁就先拿到RLVR的燃料。另一方面，某些领域的适用性还有待探索。ARC–AGI–2基准上，前沿模型实现了四个月内从53%冲到了98%；但ARC–AGI–3一经推出（交互式、不给说明书），所有模型又被打回不到1%，人类照样能拿到100%。在规则明确、可验证的领域里，推理能力正在快速逼近甚至超越人类；但面对开放式、需求模糊的真实情境，差距依然很大。面对ARC–AGI–3时模型效果的崩塌不是更多训练就能补上的，它标记的可能是RLVR这条路本身的边界，可验证奖励只能适用于有标准答案的任务，而开放式情境恰恰是答案需要被定义、而非被验证的地方。这意味着这条路在2026–2027还会撞上另一道软墙：可验证领域逐个被攻克，但越过去仍然需要验证一种“验证不出来的能力”的新范式。这道墙在哪一年撞上，是判断Scaling之后下一次架构创新何时到来的最好观测点。

模型越用越懂场景，最强不再等于最大

前沿模型会越来越聪明，但那不一定是大部分人用得上的模型。处理日常工作事务的用户规模，可能是前沿模型用户的十倍，比如我们生活中看到各种work工具，就比code工具，规模要高一个数量级。对这些人来说，大模型需要在日常应用中持续变强。

在生成式AI高速发展这几年，行业长期有个隐含假设，认为最强的模型一定是最大的模型，而2026年的事实正在动摇这一判断。腾讯首席AI科学家姚顺雨主持的CL–bench实验提出了一个行业不太愿意面对的事实：当前的前沿大模型，其实并不擅长从上下文学东西。实验设计大致如下，尝试全部采用虚构内容（编造星际法律、假SDK之类）防止模型背答案，来让模型执行。结果发现，哪怕信息明明摆在上下文里，最强模型的任务解决率也就17%出头。模型倾向于用自己预训练时记住的老办法，无视眼前的新规则。这就像一个固执的老员工，宁可按老流程走，也不看你刚发的新文档。基于此，团队提出了一个务实的判断：大模型最终需要服务于人类，需要应用于各类生活、工作的上下文场景。2024年的主旋律是Scaling，2025年是Reasoning，2026年是Context Learning。而更远的战场应该是让模型学到的东西能够持久化，是Memory Consolidation（记忆巩固），否则清空窗口就全忘了，何谈持续进步？模型将从一次成型的产品，变成一个会随你长期共同进化的个体。

技术路径上，推理时变身和记忆巩固两条线正在同时推进。前者让模型权重不再静态：比如，NVIDIA的TTT–E2E在推理时把上下文压缩进权重，腾讯混元的HY–WU（无相）则训练参数生成器直接按条件生成个性化权重，都在探索让模型根据场景实时换脑。后者让学到的东西能跨会话留存：腾讯混元的Hy–Memory为Agent设计了六层记忆架构和演化链，在LongMemEval基准上得分85.2远超同类，记忆条数仅为同类1/3但信息密度高出3–4倍。两条线加上CL–bench对上下文学习能力的度量，构成一个完整的在线进化技术栈：度量、适应与持久化，AI在全人类的使用与反馈下变强。

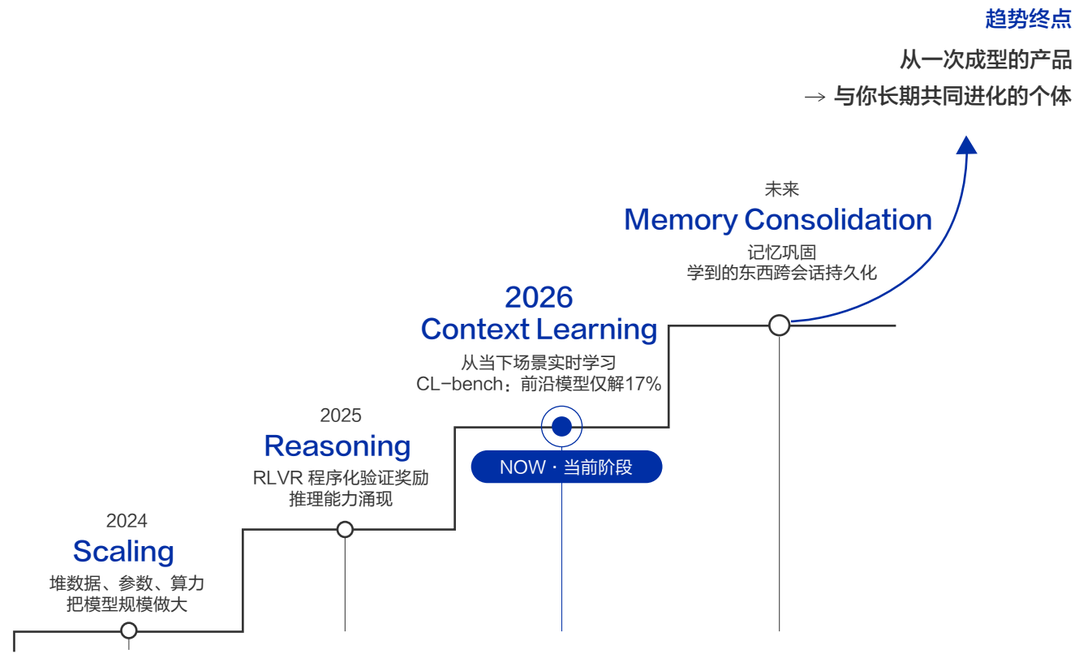


图1–2 · 在线进化的四级递进

再具体一点，产业侧将会如何推进落地？首先需要明确一下，这里的实用性模型指腾讯混元HY3这类云端模型，跟手机上跑的端侧小模型两回事。HY3采用MoE架构，295B总参、21B激活、256K上下文、开源。设计思路是比谁用得好，不是比谁参数更大：在成本相对可控的参数规模下，通过MoE架构，在自建CL–bench专测上下文学习能力，让推理效率提升40%，成功率99.99%以上，可稳定跑完495步Agent工作流。CL–bench这类能力基准的出现代表评测标准的迁移，评测标准比模型发布往往更早预示范式转向。在这个过程中，参数规模让位于场景中持续进化的能力。利用Context Learning的结构性优势，让AI从场景中实时学习，就不需要把所有知识预训练进去，而是用更小的激活参数覆盖更大的能力范围。对普通用户、work用户来说，AI正在从极其聪明但不认识你的陌生人，变成越用越懂你的伙伴。在千行百业与个人消费者里，AI比拼的焦点已经从谁的模型更大转向谁先让模型真正读懂此时此刻的用户。

AI加速AI研究，自我进化的引擎开始换挡

一些数据其实可以让我们大致的感受到AI参与自身研发已经走到了什么程度。Anthropic今年6月披露：合并到生产环境的代码超过80%由Claude编写；训练代码优化能力一年内从3倍跃升到52倍；在一次端到端研究实验中，Claude agents用800小时恢复了97%的性能差距，同一任务两个人类研究员干了一周才恢复23%。

OpenAI在更早的时间，提出过一条明确的时间线：2026年9月前达到“实习生级研究员”、2028年3月前造出能独立做完科研项目的AI。其首席科学家Pachocki今年4月澄清了核心瓶颈：智力水平已经够了，卡在持续自主性上。AI能解决高级问题，但仅限短暂爆发；区分实习生和独立自主研究员的关键变量是能连续可靠工作多长时间。

Google DeepMind没有承诺时间线，但其实也已经在实打实在用AI加速研究。比如，AlphaEvolve以大模型为变异引擎，对算法源代码做大规模变异、评估和筛选，在远超人工穷举能力的设计空间中自主找到更优解。在生产环境跑了一年，恢复Google全球0.7%计算资源，FlashAttention加速32.5%，超越困扰数学家300年的11维亲吻数问题。AlphaEvolve的哲学略有不同，它绕开人类研究者的工作方式，直接在人类不可能穷举的空间里暴力搜索，只要能自动验证结果，这条路就走得通。

在AI加速AI研究这个方向上，三家在一件事上趋于一致，承认即现阶段仍是渐进过程。但是，窗口可能比预想的更短。Anthropic的判断最具代表性：递归自我改进尚未到来，也并非不可避免，但它可能比大多数机构准备好的时间更早到来。为什么还没到？核心瓶颈在持续自主性，“研究实习生”和“自主研究员”之间的真正鸿沟在于能连续可靠工作多长时间。目前AI可靠完成的任务时长每4个月翻倍，2024年是4分钟任务，2026年是16小时，但距离数周级别的持续自主研究还有距离。人类当前的比较优势是研究品味：选什么问题重要、信什么结果、什么时候该换方向。同时，递归自我改进也存在诸多问题，例如对齐成本随迭代轮次快速上升，导致中断等等。但是，逐渐加速的趋势是确定的。在可验证的执行层，写代码、跑实验、优化算法，AI已在系统性超越人类工程师。定义什么问题值得解决、判断什么时候该换方向，目前还无法被验证器程序化，因此尚未被AI通过搜索学会。但边界已经在松动。

回到开头，2026年是AI落地应用的元年，仅靠Scaling不能满足所有需求，更重要的是学习方式的多样性。强化学习拓展能力边界、从使用场景中实时学习、AI加速自身研究。模型的未来需要在反馈中迭代进化，从训练一次定型，走向与场景共生进化。我们可以设置三个锚点，12个月后回来验证：第一，RL驱动的推理是否在3个以上非代码领域产出规模商业产品；第二，实用性模型是否在多数企业场景成为默认选择；第三，OpenAI和Anthropic递归自我改进比多数机构准备好的更早到来的这些预测兑现了没有。总的来说，基础模型正在从训练一次、固定部署走向在线进化。算力和参数之外，真正稀缺的是为进化设定方向的能力。

趋势 02

多模态认知：从渲染器到创作者，多模态AI开始理解世界

作者：陈东明　茹炳晟

2026年，多模态AI的角色正在发生位移。过去两年，行业主线是看起来“更逼真”，现在变化已不止于画质。生成系统开始表现出理解意图、规划步骤、自主迭代的能力，生成行为本身从一次性输出变成了多步决策过程。

我们围绕“从渲染器到创作者”这条主线，梳理出四条相互缠绕的趋势，它们共同指向一个更大的命题：多模态AI开始理解世界。

生成系统认知升维：从被动渲染到具备意图的创作Agent

过去，图像和视频生成更像是条件反射：用户输入提示词，模型直接出图。能力高下只看画面精细度和风格稳定性。2026年的核心变化是，模型在生成之前先做理解和规划，先判断用户想表达什么，再决定画面该如何组织，而非把提示词机械地翻译成像素。

当前的Agent化智能存在明显的结构分层：理解和规划主要由语言模型承担，视觉模型在很大程度上仍是执行器。腾讯混元团队的实验尖锐地揭示了这一点：所有模型在隐喻迁移任务中只会做字面替换，无法抓住类比关系。这说明，对表面语义，单模型已然够用；对深层逻辑，仍需系统级协作。真正的挑战在于，让视觉模型自身也具备叙事逻辑理解能力，到那时才会发生下一次范式质变。

但Agent化有一个前提：可控性。2025年AI视频可用率仅约20%，是典型的“抽卡”体验。2026年上半年，字节Seedance 2.0将15秒视频可用率拉升至90%左右，越过进入生产流程的临界点。你无法导演一个不可控的系统。先让每一步可靠，多步协作才有基础。

图2–1 · 认知升维，从渲染到 Agent

再往上一层是可编辑性。可用率突破阈值后，创作者的需求从“能不能用”升级为“能不能改”。可编辑性意味着系统能根据新指令对已生成视频做语义级局部修改：更换背景、调整角色动作、改变光影氛围，同时保持主体和时空连续性。模型必须同时理解“生成什么”和“保留什么”，本质上是对视觉内容的选择性重构。有了这一层，创作者与Agent的关系从“一键出片”的单次交易，转入“生成–反馈–修改”的迭代协作。

可控、可编辑、Agent化，三层递进，将生成AI从执行层推向认知层。

原生多模态架构：缝合模态裂缝，催生视觉推理涌现

GPT Image 2放弃了 DALL · E 的命名，图像生成从“扩散模型外挂语言模型”的拼接模式转向原生多模态架构。视频领域走得更远：可灵Kling 3.0已将智能分镜做成原生功能，用户写的是镜头列表而非提示词。交互模式从“描述画面”变成了“编排叙事”。

上一节暴露的“语言模型调度、视觉模型执行”分层缺陷，正被下一代架构修正。图像、视频、音频正成为统一模型内部的“母语”。Google将Veo视频生成并入Gemini Omni全模态模型，GPT Image 2也在架构上完成统一。这些动作指向同一个判断：多模态AI正在走完从“多模型协作”到“单模型统一感知与推理”的最后一步。

拼接模式有先天缺陷。语言模型负责调度，视觉模型负责执行，意图在模态转换中被简化、扭曲。腾讯混元团队的实验证明了这一点：即使是GPT Image，在处理隐喻迁移时也只做字面替换，无法抽象出关系逻辑。隐喻理解需要视觉和语言在同一个表示空间里同时被加工，拼接架构做不到这一点。

原生多模态架构试图消除这道裂缝。在统一架构内，图像和文本被编码进共享的潜空间，模型在Token级别进行跨模态注意力交互。一个画面中的构图、颜色、物体间的关系，可以被模型像阅读文字一样“阅读”和“思考”。由此涌现的能力超出生成质量本身：模型开始表现出早期的视觉推理能力。它能根据一张图表推断背后的社会趋势，从一张建筑照片中读出文化隐喻，在没有文本提示的情况下主动补全一段视频缺失的逻辑环节。

世界模型迈向行动：从仿真物理到预演后果

能生成逼真视频，不等于理解物体为什么这样运动；能生成3D场景，不等于知道人在其中碰撞、操作时会发生什么。世界模型要回答的问题是“这个世界如何运转”，而不是“下一帧看起来像什么”。

李飞飞团队2026年的分类框架把世界模型拆为渲染器、模拟器和规划器三个层次，这帮助我们看清产品所处的阶段。World Labs 的Marble是渲染器，腾讯混元HY-World 2.0走模拟器方向，NVIDIA Cosmos 3和DeepMind Genie 3则在逼近规划器。但严谨实验表明，纯视频生成模型在训练分布内可以完美模拟物理运动，一旦遇到未见过的条件组合，误差会高出一个数量级。利用大规模的数据可以逼近分布内的模仿，却无法外推到未知的物理条件。

这让非生成式路线重新获得关注。Meta的V-JEPA 2-AC不重建像素，在抽象表示空间做预测，仅用62小时无标注机器人视频就实现零样本控制，运行速度比生成式路线快约15倍。LeCun长期坚持的判断在这里得到验证：理解世界不一定需要重建世界。两条路线正围绕“具身世界模型”展开竞赛，核心问题已从“下一帧像什么”转向“行动之后的下一个状态是什么”。

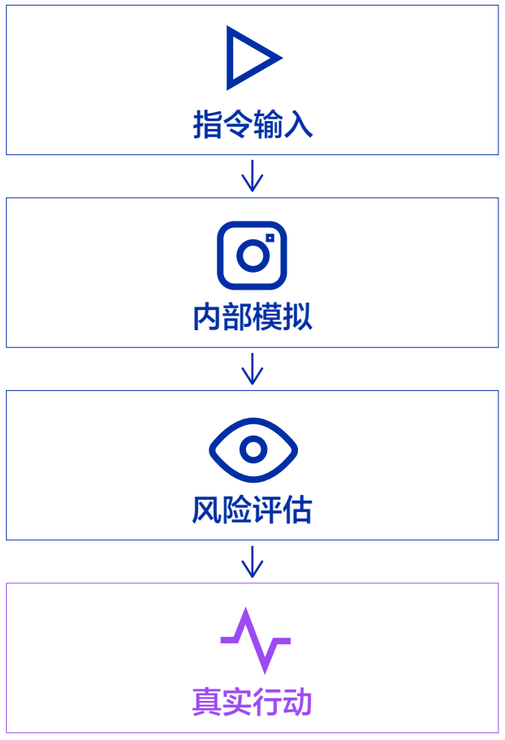


图2–2 · 想象引擎，世界模型驱动的物理预演闭环

对机器人和自动驾驶而言，物理世界的试错成本极高。世界模型的价值在于充当“想象引擎”：接收动作指令后，先在内部快速模拟未来可能的状态序列，评估风险，再选择最优动作。这要求它理解重力、碰撞等刚性物理，也要掌握拧瓶盖时摩擦力变化、端起咖啡时液面晃动幅度等软物理常识。一旦世界模型能回答“要达成这个目标，最好采取哪一系列动作”，多模态大模型就从内容创作者变成了物理世界中的决策者。

多模态记忆：从一次性生成到拥有“视觉经验”的AI伴侣

可控性解决了“这一次生成是否可用”的稳定性问题，但更深层的缺口随之暴露：每一次生成都是孤立的。用户今天生成的产品海报、家居设计方案，明天AI就全然忘记；反复调整过的色调偏好、构图习惯，一点都留不下来，每次交互都要从头“教育”模型。

多模态AI的下一道分水岭是记忆。它需要多层次的“视觉经验”：情境性记忆记住用户的家居环境、办公空间乃至孩子的成长过程，保持时空一致性；偏好性记忆从反复的编辑、选择和抛弃中，学习对光影、色彩、叙事节奏的偏好；情感性记忆把过去共同创作的设计、修复的老照片，攒成可引用的“共同历史”。当AI说出“像我们去年夏天那种风格”，那是真实的索引，不是比喻。

技术上，这需要多模态模型与检索增强生成、参数高效微调以及端侧存储的配合。一些产品已尝试在设备端构建个人视觉知识图谱，让隐私敏感的视觉记忆不出本地。由此可能诞生一种参数不大、但认知里充满“你”的个人多模态模型。

但视觉记忆远比文本记忆更亲密。当模型开始记住你的家、你的面孔和日常习惯，隐私问题会变得尖锐。如何让AI拥有丰富的个人视觉史，又不沦为监控工具？视觉层面的“遗忘权”和价值对齐机制必须同步构建。

还有一个维度：当多模态AI开始创造世界，安全对齐就从文本逻辑延伸到了视觉伦理。渲染器只是工具，责任在使用者；但当Agent自主规划画面、编排叙事，它本身就参与了意义的创造。2026年，多模态安全仍滞后于能力发展，而滞后的代价正在急剧上升。让AI能够创作，就必须让它同时承担相应责任。

图2–3 · 经验与良知，记忆分层与价值对齐

四条趋势环环相扣：可控性让生成结果稳定，Agent化让系统学会规划和迭代，原生多模态架构缝合感知与推理的裂缝，世界模型将理解推向行动预演，记忆和伦理则为AI注入连续性和边界感。它们共同将多模态大模型推向一个问题面前：它是更逼真地反映世界，还是开始参与世界?

2026年的多模态AI值得被单独观察。图像、视频、3D正在共同回答一个问题：AI能不能通过多模态经验建立对世界的理解。判断标准已经变了。以前看模型画得有多像，现在看它是否知道自己在画什么、为什么这样画，以及画面背后那个物理的、情感的、伦理的世界如何运转。

趋势

03

科学智能：三位一体，AI for Science的攻坚之战

作者：刘莫闲

2026年5月，谷歌连续发布Gemini for ScienceAI工具集和Co-Scientist科研智能体，将文献洞察、假设生成和计算发现纳入统一科研工作流，并与其专用模型和工具构建起一套完整的AI for Science技术栈。同月，Isomorphic Labs以 21 亿美元的 B 轮融资将累计外部资金推至约 27 亿美元，首个AI设计药物已进入临床试验阶段。在国内，上海人工智能实验室开源了万亿参数科学多模态大模型Intern-S1-Pro，中科院磐石·科学基础大模型覆盖八大学科，服务科研创新全流程，晶泰科技自主实验室方案持续服务海内外客户。

科学智能（AI for Science）的产业成熟度在快速提高，并进入一段技术体系构建、产业原始积累、基础设施重筑三浪叠加的攻坚时期。技术基础已呈现出体系化运转的雏形，“基础模型+科研智能体+自主实验室”三位一体的AI驱动科研范式从探索期进入了成形期。AI在生物医药、材料发现、地理气象和数学四个典型场景中的应用也开始阶段性产出和价值验证。业界对既懂AI、又懂科研的π型人才加大了发掘和争夺力度，也在努力构建可信的AI科研体系，为科学研究的未来可持续护航。

基础大模型成为科研操作系统，从单点工具走向系统平台

科研基础大模型正在成为科学研究的操作系统。 科学智能在技术和范式层面最明显的变化，是多家头部公司均以基础大模型为核心，将科研AI从单点工具升级为系统性平台。除了谷歌的技术体系以外，Anthropic在该领域的进展也值得关注和学习。最近，Anthropic最新模型Mythos 5在药物设计、分子生物学、基因组学等领域展示出更强的科研统筹能力。Anthropic把自身模型和Claude Cowork等产品逐步适配科研场景的需要，也通过收购及合作等方式延伸科研技术生态。Anthropic以约4亿美元收购 AI生物公司Coefficient Bio，探索药物研发自主全流程执行，并持续展开与生物医药行业核心数据平台、药企、医学研究机构、以及医疗服务网络等行业生态合作，一个围绕Claude模型贯穿从文献查阅、实验方案设计、药物研发到临床治疗全流程的科研创新网络正在逐步显现。国内基于科研基础模型展开的行业体系构建也有不错进展。上海AI实验室基于其最新的书生大模型Intern-S1-Pro发布的“AGI4S珠穆朗玛计划”联合上下游打造科研生态，中科院以磐石·科学大模型技术栈支撑中科院50多家单位上百个科研场景。

专用科学大模型在多个学科形成了纵深矩阵。通用AI模型保障科学智能的广度与下限，而专用科研模型则不断探索细分科研场景创新的深度和上限。在生命科学领域，谷歌DeepMind Alpha系列模型覆盖了从序列到功能、生成、调控的蛋白质研究各环节。OpenAI基于GPT-5.5面向生命科学研究打造GPT Rosalind系列模型和工具。亚马逊也公布了其用于抗体药物研发的垂直模型Bio Discovery。在材料科学领域，北京科学智能研究院联合北大和深势科技推出的DPA4新一代高效能原子模型架构，微软推出新一代材料大模型MatterSim-MT模拟复杂多属性的材料现象。在物理和地球科学领域，上海AI实验室发布的书生科研大模型、中科院磐石大模型均在物理和地球科学领域相关的专业任务基准测试中名列前茅，谷歌AlphaEarth也探索基于卫星遥感和图像数据的地球科学研究。

科研智能体参与探索人机协作新边界。各大头部AI厂商、机构和初创公司纷纷在科研智能体领域发布重要更新和实验探索：谷歌有科研智能体团队Co-scientist和数学智能体Aletheia，Anthropic有基于Claude面向生物、化学的智能体，FutureHouse有Robin生物医学科研智能体，等等。同时，这些探索也在不断塑造着人与 AI 协作边界和生态系统。从人机协作模式看，科研人员与智能体的协作方式探索正在三个层面逐步推进。

第一层，让智能体更加主动参与，包括部分决策。谷歌Co-Scientist“思想竞赛”机制使AI在多假设间自我辩论和收敛，角色从执行转到构思和策划。第二层，让智能体的工作内容从实验验证向假设推演延伸。谷歌DeepMind的ERA自主设计、生成并测试科学软件，AlphaEvolve通过进化计算自动发现更优算法，AI开始帮科学家产生想法而不仅限于验证想法。第三层，人机分工的边界进一步重新划定。“人类定义问题+评估结果，AI生成假说+编排流程+分析数据”的模式正在逐步成为行业新共识。从生态构建路径角度看，大家构建科学智能生态的方式比较多元，却也充分结合自身特点：Anthropic Claude目前主要是模型的平台化，Google DeepMind则在构建模型、Agent和工具一体化体系，国内上海AI实验室、中科院等探索开源开放和生态共建。

自主实验室逼近规模化前夜，从能做走向规模化运行

自驱动实验室（SDL）逐步进化成可自适应的循环科研系统。 相对于高通量自动化实验室系统，自驱动实验室的最大区别在于：SDL 会把实验设计的决策权部分交给AI。基于“设计－执行－表征－学习－再设计”的实验工作回路，SDL可以基于实验结果自适应的选择下一步实验，而不完全按照人类预设的脚本来执行。参考过去生命科学、材料和化学等领域头部实验室和专家在过去几年发表的论文，一个典型的AI驱动SDL工作回路通常依赖三大系统中五层工作的相互配合（如下图）。

图3-1 · 自主实验室的流程闭环

政策资本的多方推动正在加速各国SDL规模化建设。美国和英国在2025年后的相关策略最为明确。美国NSF、DOE、CHIPS等相关部门和组织均面向自主实验室启动建设项目和加速计划。 英国《AI for Science Strategy》将自主实验室写入国家动作。欧盟则通过Horizon Europe中的专项以及RAISE项目支持自主实验室自动化科学发现。我国同样在“人工智能+”等国家级政策中明确提出“加快探索人工智能驱动的新型科研范式”，并通过重点实验室升级、重大设施智能化改造和地方AI+计划的方式加大相关投入。此外、日本、韩国、新加坡等亚洲国家也都陆续推出支持自主实验室规模化建设的政策、目标和试点计划。

SDL商业化初见规模，更多学科开始SDL应用探索和升级。 一方面，我国公司领跑全球自主化实验室商业化探索。晶泰科技可编排数百个工作站和自动化实验站，围绕AI模型预测、机器人实验执行、数据反馈和Multi-Agent调度构筑研发飞轮，支持7×24小时研发，每月产生超5万条反应产率数据，30万过程数据。并在国际市场与亚洲领先药企韩国JW Pharmaceutical达成合作，提供药物发现自动化工作站和基于人工智能的化学反应优化平台。晶泰科技已于2025年首次实现年度盈利，成为行业新里程碑。英伟达风投领投的Lila Sciences累计融资已达5.5亿美元、并已在美国大波士顿地区签署23.55万平方英尺的实验室租赁合同建造AI科学工厂。另一方面，除了药物研发以外，更多的学科领域开始探索SDL升级之路。在无机材料领域，Berkeley Lab的A-Lab展示了以最少人工干预、按天推进无机固相合成的能力。在化学合成领域，英国利物浦大学的移动机器人化学家率先证明了机器人可在标准化人用实验室中跨设备自主操作。在电子和能源领域，美国阿贡实验室的Polybot SDL将“少至无人工干预”的材料合成、样品转运、表征和分析纳入统一平台。

应用落地多赛道铺开，从单点验证走向并行攻坚

技术基础和实验能力逐步就位，科学智能在应用层面的机遇和挑战也变得更清晰。应用落地、人才储备和可信基础设施三个维度，共同构成了当前攻坚阶段的核心议题。

四大应用领域持续积累AI4S价值验证。生物医药的潜在价值最大，第一批AI制药进入上市前关键阶段。自Alphafold问世以来AI在生物医药领域便进入快速发展期，其中新药研发的潜在社会价值和商业价值巨大。据行业机构估计截至2026年初，全球已有约170个AI参与发现、设计或优化的药物项目进入临床开发，其中多个项目进入Ⅲ期开发项目或试验。英矽智能已有13款候选药物获批IND，其中4款由公司自主推进临床开发，6款通过与中国或海外合作伙伴联合开发或授权的方式持续推进。**材料科学是AI4S的下一个前沿，商业化探索已开始。**美国伯克利A-Lab实现每天超2种新材料的发现速度。国内鼎犀智创正构建RhinoWise材料创新平台，Periodic Labs、CuspAI等众多AI4S初创公司均以材料科学为核心赛道。材料科学的优势在于实验反馈周期相对较短，有望更快的进入商业化和市场化验证。**气象是AI4S渗透最深的赛道，AI模型开始参与中期天气预报。**该领域天然具备大数据、明确的物理规律和数十年建立的标准化评估框架，是大参数模型应用最早的科学领域。相比传统数值计算模型，AI模型带来的算力节省也很可观，甚至超过90%。但由于气象预测对全球各国安全运行影响巨大，以及AI仍存在的各类不成熟问题，全球的气象预测系统仍在较为谨慎的对比和引入AI模型预测，目前以0-15天的中期天气预报为主要应用试点。**数学领域出现了从解题到发现的拐点信号。**OpenAI GPT-5被用于探索埃尔德什难题，谷歌基于AlphaEvolve推进数学和理论计算机科学研究。多个大模型已在奥林匹克数学竞赛中不断刷新金牌成绩，但在全面解决现有数学难题、独立定义新问题等方面，AI仍存在较大差距。

AI4S跨领域人才需求明显增大，全球开始发掘和培养原生π型人才。AI4S业内一场面向“π型人才”的全球争夺正在展开。仅靠现有科学家学学用AI、抑或仅靠现有AI工程师转行做科研，均难以解决AI4S原生创新问题，其最大瓶颈在于能将AI与学科知识更天然且深度的结合。世界需要更多像DeepMind创始人Demis Hassabis这样的原生“π型人才”，而面向这类人才的全球发掘与争夺已经展开。Anthropic近期举措最多，AI for Science Program提供免费API额度，STEM Fellows Program以三个月全额资助将专家引入旧金山总部协作，Anthropic Fellows Program更是面向欧美各地招募研究员。谷歌成立专门的AI for Science基金，并于2026年投入3000万美元发起“AI4S影响力挑战赛”，面向全球公开征集项目，加速科学发现。谷歌还与Wellcome Sanger Institute、FutureHouse等创新科研机构合作建立研究员资助计划，发掘AI原生科学家。OpenAI Residency以六个月跨界培养面向数学、物理、神经科学等背景的定量分析人才。国内清华大学、北京大学、上海交通大学等顶尖高校也在积极布局AI+科学的交叉学科培养。

加速建设AI for Science可信基础设施势在必行。技术越是光鲜，可信度基础越值得重视。AI生成的虚假参考文献正爆炸式增长。《柳叶刀》近期一项研究显示，基于250万篇生物医学论文和1.26亿条引用的系统分析，每万篇论文中检出的虚构参考文献数量，已从2023年的约4.0条升至2026年初的56.9条，增幅超过12倍。《Nature》针对2025年学术出版物开展的抽样分析估计，全球可能有超过11万份学术出版物至少包含一条无效引用，其中部分呈现出疑似AI生成或“幻觉引用”的特征。《Science》期刊群总编辑日前也发表《Resisting AI slop》社论。由此可见，业界自律已经无法遏制AI虚假信息对科研文献的污染，业内组织应对正在从自愿披露转向强制审查，相应的人工审查机制、模型算法、鉴别系统亟待加快建设。

技术拓展上限，产业筑牢中坚，可信坚守底线。2026年的AI for Science，正在爬坡量变，破局攻坚。

Part II
落地

从可用到好用的基础设施

趋势04

驾驭工程：给野马套上挽具，竞争前沿移到模型外部

作者：余一

2026年开年有一个信号，一个多月里，Anthropic、Cursor接连发文讲同一件事：怎么给模型搭一套环境，让它能自主完成任务。Anthropic那篇叫《Harness design for long-running applications》，Cursor那篇叫《Towards self-driving codebases》。这个正在被命名的方向，就是Harness Engineering。

Harness这个词原意是马身上的挽具。模型越来越强，释放出来的能力像一匹野马，核心问题随之转变：如何驾驭它、让它为你的目标服务、给你确定性和掌控感。国内把Harness Engineering译成“驾驭工程”，取的正是这层含义。

MODEL

中枢怎么思考

预训练边际收益递减
推理时计算成为新发力点

HARNESS

在什么规则下跑

记忆、工具、规划、回复、评估
模型外的运行容器

三个可独立发力的增长维度

模型智力	上下文	Harness
更强推理	更干净的工作记忆	更可靠的运行框架

图4-1 · 前沿从模型内部移到模型外部，Harness决定智能能否可靠释放

Harness阶段显性化，从外部搭建走向被模型内化

Harness先是一层运行容器。模型本身负责推理，Harness负责给它安排记忆、上下文、工具、规划和多智能体协同。模型决定中枢怎么思考，Harness决定模型能看到什么、能用什么工具、遇到失败怎么办。这个词出现以前，做这件事的人已经做了很久：搭记忆、调上下文、接工具、处理失败。后来大家才发现，这些工作都在处理同一层问题。先有工作，再有名词。Harness是被实践逼出来的，一群人各自摸索，最后被同一个词收拢。放到更长的时间线上看，它的位置很清楚。22到24年是Prompt Engineering，靠把指令说到位；25年Andrej Karpathy把这件事重新命名为Context Engineering，强调精心设计填进上下文窗口的内容，Anthropic同年九月把上下文工程定调为提示词工程的自然演进；到26年，重心又往外挪了一层，变成Harness Engineering，开始同时处理模型看到什么、能用什么工具、遇到失败怎么恢复、多个agent之间怎么协同。这条线的方向很清楚：从模型内部（怎么问），到模型的工作记忆（给它看什么），再到模型外的整个运行框架（让它在什么规则下跑）。每往外走一层，关注点都会从单次回答质量，转向整个系统能否可靠地完成任务。

原因在模型侧的边际收益变化。单纯把预训练做得更大，回报在递减，数据墙也在逼近。与此同时，推理时计算（inference-time compute）成了新的发力点。微软研究院2025年的工作发现，模型参数固定时，靠多次采样、树搜索、自我验证这些推理阶段的手段，复杂任务的表现就能往上抬一截。让模型在运行时多做几轮推理和自我验证，收益开始超过单纯扩大模型参数。编排这些推理步骤的，正是Harness。可以把Harness和模型智力、上下文拆开看。上下文做好，结果会更稳；模型更聪明，结果会更好；模型和上下文都不变，只换一套更好的Harness，结果也会变化。差距会落到那些会搭运行环境的人身上：同一个模型，别人只能问答，他能让它连续做事。

这个独立维度有一个终局：模型会逐步把Harness内化。今天需要人在外部搭建的记忆管理、错误恢复、工具选择、多步规划，模型自身正在学会做这些事。推理时计算本身就是内化的第一步，模型开始用自己的推理链替代外部脚本的调度。再下一步，模型会学会自己管理上下文窗口、自己判断什么时候该停下来验证、自己在多轮任务中维持一致性。每一代模型发布，都在把上一代需要外部搭建的能力吃进去。

Harness现在处在一个很短的窗口里：它刚刚被看见、被命名、被工程化，终局却是消失进模型能力里。就像马鞍，骑马的人不用管它内部怎么和马贴合，只知道骑起来更稳。今天行业还在大谈特谈Harness，说明它还没有被内化完。这个显性化窗口里，散落的实践经验正在被编目成可传授的设计模式。

ADPS（Agent Design Patterns Society）用上图两个轴收拢了28个命名模式，配套Manning出版了Designing AI Agents作为选型手册。这代表行业正在把Harness往工程学科的方向推。一旦这些模式被内置进平台和模型本身，使用者就不用再自己拼装。Harness从显性走向无感，最终被模型吃掉。

这个过程会比多数人预期的快。现在是普通人理解和亲手搭建它的好窗口，越早把它当回事的人，越能吃到这一波由“会搭环境”带来的个体差距。等模型把这些能力全部内化，窗口就关了。

工作重心从指挥AI做事转向设计AI怎么工作

前沿外移，最直接改变的是人的工作方式。过去程序员追求过程确定性，每一步都要可控可复现。用agent协作时，重点开始转向结果确定性：把预期结果定义清楚，用一个能存状态的Harness接管流程，让AI自己发现问题、自己修正，直到达到验收标准。人的精力从逐步监控中释放出来，转向定义终点和设计环境。

这个转变绕不开一个老问题：AI会出错，凭什么敢把流程交给它？许多人用AI的状态，类似招了一个极聪明但刚入职的新人，能力强，但不了解组织的禁区。更有效的做法，是把任务边界、工具权限、上下文和交接方式设计清楚，让他先在一套环境里工作。

一旦把注意力从“每一步怎么指挥”挪到“环境怎么设计”，输入格式、底子复用和平台入口都会变成工作方式的一部分。

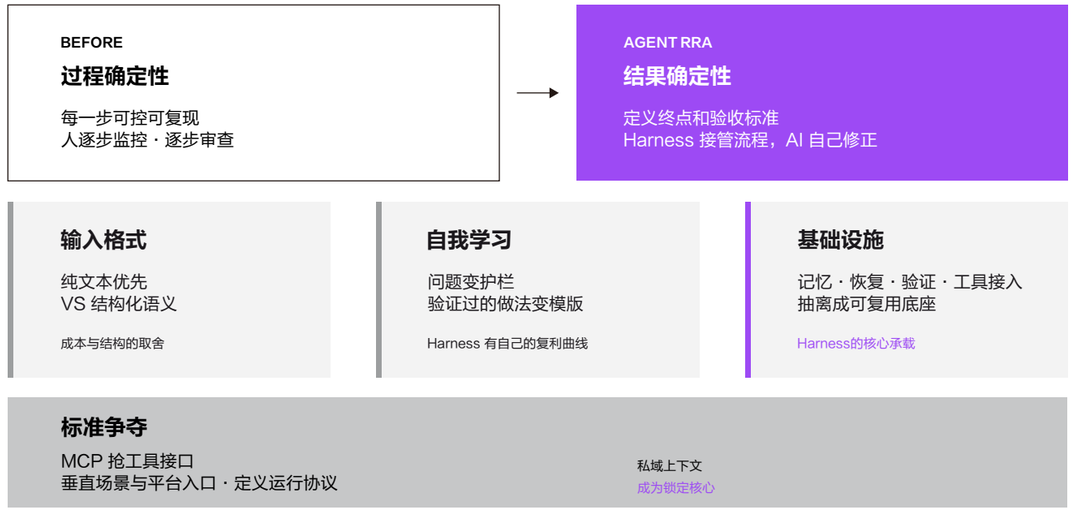


图4–2 · 从指挥AI做事，到设计AI怎么工作

输入格式是实践者中间最活跃的争论之一。一派主张纯文本优先：HTML的标签和样式消耗token、分散模型注意力，Markdown或纯文本结构更清晰、成本更低，上下文窗口是稀缺资源，应该把每一个token都留给有效信息。另一派认为结构化格式（包括HTML）在复杂任务中提供的语义线索反而帮助模型理解，省token不等于省对了地方。争论指向的是上下文窗口的使用策略：什么时候优先压缩成本，什么时候保留结构。Local First、纯文本优先的倡导者把前者推到了极致，但共识尚未形成。

第二个分歧，是要不要把Harness当成基础设施。今天大多数人还是一个项目临时搭一套，项目结束就废弃。但Harness中能积累复利的部分，包括记忆存储、错误恢复、工具接入、权限边界，完全可以抽出来、跨项目复用。它正在变成AI时代的一类基础设施，像数据库，像云。谁先把自己这层infra做厚，后面每搭一个新东西的边际成本就越低。

开发者只是第一批被卷进Harness的人。广义的 Harness，比如一个Agentic Loop、一个Claude Code、一个命令行工具，对每个人都息息相关，会下沉成普通人的基础能力。但无论哪种定义，Harness都不能搭一次就结束。不少人把规则、记忆、模板堆在一起，以为就有了系统，但堆叠不等于工程。一套可用的Harness，要能被复用、能被交接、能在不同任务里稳定运行。从一个文件收纳系统，到一个能运行的架构，中间隔着一次跃迁。多数人卡在第一步，把整理当成了工程。

当Harness从个人行为变成行业行为，标准争夺就不可避免。2025年MCP已经在抢“模型怎么连工具”的标准位。接下来竞争会落在两类地方：一类是金融、医疗、教育、代码、内容生产这些垂直场景，另一类是IDE、云平台、办公套件、企业知识库这些平台级入口。它们会各自定义运行接口，规定领域知识怎么接入、上下文怎么隔离、工具权限怎么授权、多agent怎么移交状态。谁定义接口，谁就把第三方工具、私域数据和用户工作流吸进自己的生态。ADPS的开源参考实现试图给出另一条路：展示模式如何在成本、延迟和治理约束下组合，不被单一平台锁定。

这一层很可能重演浏览器、IDE当年的格局：少数几个平台级Harness会像今天的IDE一样普及成基础设施，垂直Harness则会在各行业里长成默认工作台。伴随标准争夺，私域上下文会成为锁定的核心。模型在趋同，大家用的基础模型差不多，差距落在喂进环境里的东西。随着个人把越来越多判断、记忆、工作流存进某一家的Harness，迁移成本会高到形成事实锁定，数据和上下文能不能带走，会从技术洁癖变成真实的使用者问题。

评估前置为控制层，从静态脚手架走向自学习系统

Agent越自主，结果确定性的难点越少停留在模型智力，越多落在系统能不能知道什么叫做得好、什么叫偏离目标、偏离后怎么纠正。评估系统在这里要前置为Harness的控制层：它定义任务的验收标准，记录agent的执行轨迹，判断每一步是否偏航，把错误转成可追溯的证据，再把这些证据喂回下一次运行。缺少评估时，Harness只是给模型接了更多工具；评估接进来以后，它才开始像一个能自证质量、能积累经验的工程系统。

图4–3 · 评估，Harness的控制层

好的Harness应该能在每次运行中自动把出过的问题变成下次的护栏、把验证过的做法存成可复用的模板。它需要有自己的复利曲线：运行越多，越了解边界在哪、什么时候该停下来请示、怎样避免重复犯错。只会被动接收指令的Harness是静态脚手架，能自我积累的Harness才具备工程系统的特征。这个自我积累的入口，就是评估。

垂直场景里更是这样。医疗、金融、教育、法律、代码这些领域，行业标准和用户偏好不能只写进提示词，还要变成一套稳定的评估系统，让agent在具体场景里被持续校准。ADPS的模式选择方法论里有一条隐含判断：选型本身就是一种评估，你选什么模式，说明你先定义了什么叫“对”。当设计模式可编目，选型变成有据可查的工程决策，评估就会前置到架构选择阶段。

再往前一步，评估数据会成为Harness回流学习的入口：今天它先用来做自检、回放、回滚和人工审计，下一步会用来改写提示词、调整工具路由、攒下训练样本，甚至接进强化学习或在线更新流程。这也是第一章讲的“模型内化 Harness”的具体路径。评估数据就是模型学习的素材，Harness运行得越久，积累的评估数据越多，模型从中学到的也越多，直到它不再需要外部框架来替它做这些判断。

错误恢复越好，人对agent的警觉会越低。当一个系统总能自己接住错误、自己改、自己跑完，使用者会越来越少去看它到底干了什么。单次错误确实变少了，代价是系统性偏移变得更难被发现。评估系统的价值，最终体现在它能不能把这些安静积累的偏移变成可见的信号。

持续学习的现实路径大概率会从这里长出来。模型不会靠一次更大的预训练就学会一切，它会靠一层越来越好的环境，在真实任务里一点点变强。而搭建和维护这层环境的，是每一个使用者。驾驭工程被放在腾讯研究院《AI原生工作报告》开头，原因正在于此：它既是今天最务实的生产力工具，也是通向“模型在线进化”那个更大故事的第一级台阶。

所以，越来越强的AI，吃不掉什么？一线实践者给出的答案，最后都会落到同一个地方：人仍然要决定什么该做。模型能推出很多观点，但谁来认定一个还没成为共识的判断，谁来承担它的后果，谁来把游离在公开语料之外的私域知识装进环境里，这些都不会因为换了一个技术名词就消失。AI越能执行，人的方向感反而越暴露。

驾驭工程说到底是个手段。它让AI能自主跑起来，把人从逐步监控和善后的位置上释放出来。问题是省下来的时间用在哪里。用来攒下别人没有的判断，用来坚持还没成为共识的方向，用来做能创造价值的事。前沿会一直移动，名词会一直翻新，而你为AI准备了哪些别人没有的东西，决定了你是被它放大，还是被它拉平。

趋势

05

全民构建：编程先验战场跑通，AI原生工作迈向千行百业

作者：曹士圯

AI编程是这一轮Agent浪潮的先验战场，关于“AI能否独立完成软件交付”的争论在2026年已基本落幕。在专业侧，模型从代码补全跃迁为自主交付，并开始自己为正确性把关，工程师的重心从写代码上移到定义意图与编排Agent；在普惠侧，构建权第一次不再以“会写代码”为前提，非开发者成为独立的构建群体，但“原型墙”把门槛从“会不会做”换成了“能不能做成”。更大的趋势是这套能力正加速跨过代码边界，2026上半年头部编程产品集体进化为通用工作Agent，知识工作进入Agent井喷期，并按岗位向财务、法律、咨询等专业服务渗透。范式之所以能加速外溢，原因在于工作方式的迁移阻力被消除，当任务可被自然语言定义、产物可被即时验证、Agent可承担大部分执行，“会用什么工具”作为壁垒逐步失效，“想清楚要什么、判断好不好”作为能力持续升值。

代码自主交付成为现实，专业开发从写代码转向定义意图与编排Agent

AI编程在专业侧的能力形态完成了从“代码补全”到“自主交付”的跃迁，并已逼近“做对”的天花板。衡量真实代码修复能力的SWE-bench Verified基准上，成绩从Claude 3.5Sonnet的49%（2024年6月）升至Opus4.5的80.9%（2025年11月），再到Opus4.7的87.6%（2026年4月），22个月跳升38个百分点，部分实验室未公开的内部版本更高。这意味着“能不能写对”已基本不是瓶颈，能力降权、产品化的速度才是真正的变量。

真正的质变发生在自主性与自我把关两个维度，而非分数本身。模型独立连续作业的时长持续延长，从补全单行、单函数，发展为自主拆解任务、连续运行数小时完成一个完整工程，并从单文件补全走向跨文件、跨仓库的理解与改动，Cursor已支持在一个窗口里跨多个代码仓库并行作业。更关键的是验证开始从人内化进模型，Opus4.7的工程合作方反馈，模型在写系统级代码前会先做形式化证明再动手，代码生成从“先写后查”转向“边写边证”，这正是软件工程瓶颈由“如何实现”向“如何确认正确”迁移的信号。产品形态也随之从实时结对补全，扩展出异步委派、云端后台作业的新模式。

能力的质变直接转化为生产方式的质变，AI编程在最前沿团队已进入主导生产阶段。各家的自用数据最具说服力，Anthropic的Claude Code团队95%的代码由Claude Code编写，新产品Claude Cowork全部由AI编写、仅用1.5周完成；腾讯超过90%的工程师使用CodeBuddy，平均编码时间缩短四成以上，AI生成代码占比过半；YC 2025冬季批次中四分之一的创业公司代码库95%以上由AI生成。METR在2025年中的随机对照实验曾发现AI让资深开发者反而慢19%、一度引发争议，但这是早期模型在成熟代码库上的快照；随着模型能力跃升，30%到50%的开发者已直接拒绝“无AI”的工作条件。

商业兑现与角色上移同步发生，专业开发由此向更深处走，没有被简单替代。Claude Code收入从零做到25亿美元仅用9个月，Cursor估值从293亿美元谈到500亿美元，Codex周活在4个月内从300万翻到500万；a16z调研显示29%的财富500强已成为头部AI创业公司的正式付费客户，编码是企业AI应用中“领先近一个数量级”的主导用例。未来12个月，AI编程的地位将从“加速工具”固化为“默认基础设施”，AI代码占比将成为衡量工程效率的新指标，工程师的价值锚点则从“写出多少代码”转向“定义多准的意图、判断多好的产出、编排多复杂的协作”。

构建权下放普通人，非开发者成为独立构建群体，原型墙让判断力成为新门槛

AI编程在普通用户侧跨过了临界点，构建软件第一次不再以“会写代码”为前提。模型把“自然语言需求到可运行软件”的路径跑通到了非专业人员可用的程度，Bolt、Lovable、v0等产品让用户用对话描述需求即可得到可运行的网页与应用，Anthropic在测试的全栈构建器更能一次产出含前后端、登录认证、数据库与一键部署的完整应用。正如Andrej Karpathy的概括，今天很多人是“用英语编程”、最终提交的才是代码，当构建的输入界面从专业语言切换为自然语言，专业工具的使用门槛随之被绕过。

这一能力下放已兑现为一个可观测的独立群体，非开发者首次成为构建者。主打自然语言建站建应用的 Bolt，其用户中有六到七成是设计师、学生、健身教练、销售和教师，而非程序员；Epic Games内部超过一半的Claude Code使用来自非开发者；Block公司上万名员工中，许多非工程师在自己动手搭建对接业务系统的工具。这些人过去与“软件构建”绝缘，如今成为真实的供给方。

门槛并未消失，而是发生了转移，“原型墙”成为普惠侧新的分水岭。AI把“从零到原型”的成本压到几乎为零，“从原型到可运营”却横亘着一道鸿沟。多家媒体报道Lovable等平台普遍存在留存危机，用户常经历“第一周兴奋、第三周担忧、第二个月放弃”的循环，做原型每月二十美元，维护下去却要每月两百美元。Addy Osmani将其总结为“70%问题”，AI生成的代码看上去对了70%，但完成剩下30%的代价常常超过从头手写；GitClear对2.11亿行代码的分析显示，AI大规模介入后技术债务增加三到四成、代码重复率翻倍；Veracode的测试发现45%的AI代码任务会引入已知安全漏洞。

普惠平台的竞争重心将从“更快出原型”转向“帮用户跨越原型墙”，判断力成为真正的稀缺。围绕可运营、可维护、安全合规的能力将成为新的价值增量，既不够专业、又不够易用的中间地带产品则会被快速吞噬，专业侧与普惠侧将沿不同轨迹分化为两个品类。普通人构建的门槛从“会不会写”升级为“会不会判断与跨越”，这与专业侧“向编排深化”在底层是同一道理，真正稀缺的始终是判断力。

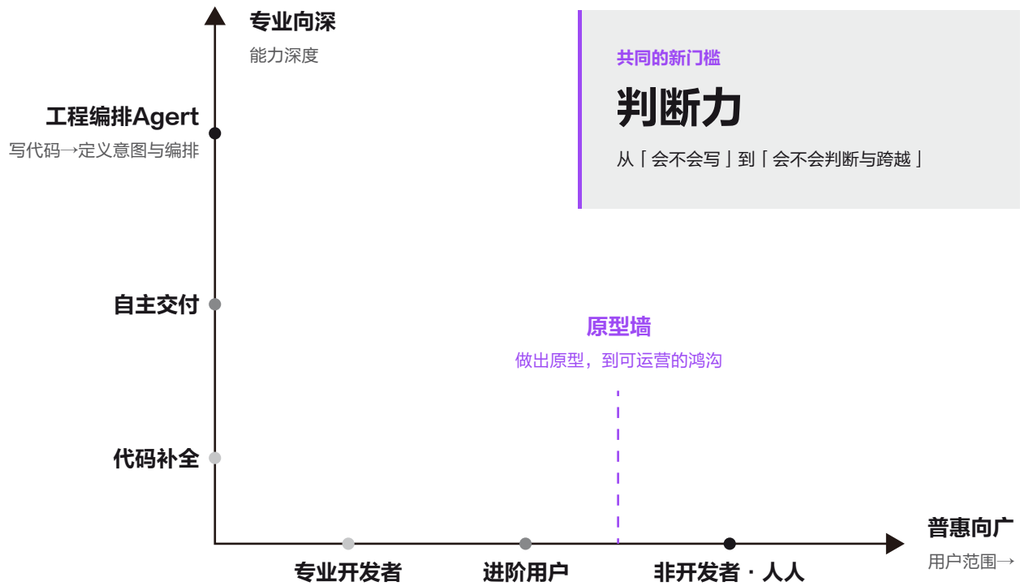


图5-1 · AI编程的双轨成熟

编码能力溢出，头部编程产品进化为通用工作Agent

头部AI编程产品在2026上半年几乎同时进化为通用工作Agent，一批Work类产品集中涌现。2月，Anthropic在Claude Code之上推出定位为“面向不写代码的人的Claude Code”的Claude Cowork；3月，腾讯云的CodeBuddy派生出面向办公的WorkBuddy；6月，OpenAI给Codex 一次性上线 6 类岗位插件、把它从编码工具扩成全岗位办公平台，月之暗面把Kimi Code升级为通用本地Agent的Kimi Work，字节跳动也将 TRAE SOLO升级为覆盖办公与开发的TRAE Work。多家代表性公司在数月内用同一种路径集体跨过“只为程序员服务”的边界，说明AI编程跑通的远不止一个工具，它是一套由模型、技能调用、Agent编排、自然语言界面和即时验证迭代组成的可复用范式。

这批Work产品有一个共同基因，它们都从“能执行完整任务”出发，没有先做通用聊天助手再补办公能力。编程是迄今对Agent要求最苛刻的场景，需要长流程拆解、调用工具、连续执行并自我纠错，在这里经受过考验的能力，平移到写报告、做表格、做数据分析等知识工作上往往更稳更深。正因如此，知识工作Agent是编码能力溢出的直接产物，本质是编程场景锤炼过的执行内核平移了过来，这也解释了它为何能在短时间内集中井喷。

用户早已在用编码工具完成大量非编码工作，产品此番只是补上了对应的形态。Codex在6月更新后，数据分析类任务周环比增长110%，研究类任务增长37%，报告、备忘录、合同、演示文稿等知识产出物增长36%，超过六成用户在一天内同时运行多个任务，而该比例在4月中旬还不到一半；其非开发者用户占比已升至两成，增速是开发者的三倍以上。需求摆在那里，谁先把编码Agent的执行内核翻译成普通岗位能用的形态，谁就接住了这波井喷。

AI编程产品与AI办公产品的边界正在消融，头部编程公司有可能反过来成为最大的通用办公软件提供者。OpenAI已把Codex与ChatGPT合并，推出ChatGPT Work，Anthropic用Cowork把Claude Code的能力对准非开发者，腾讯、月之暗面、字节则直接在编码内核上长出通用工作Agent。由于编码产品起步更早、底子更厚，在未来12到18个月，它更可能成为通用工作Agent 的最强候选，传统办公软件反而难以将它吃掉。

Agent原生范式向千行百业渗透，AI编程成为最重要的可迁移经验

Agent原生范式的扩散边界，取决于任务能否被自然语言定义、产物能否被快速验证、流程能否被标准化三条标准。AI编程之所以最先跑通，正因软件开发同时满足这三条，需求可用自然语言描述、产物可编译运行检验、工作流又有版本管理与代码评审等成熟环节。沿这三条标准外推，被改写的不会停留在某几个孤立场景，设计、法律、财务、咨询这些成片的专业领域都会进入下一批名单。

这一外溢已从软件开发蔓延到先行的专业行业，并以“按岗位打包能力”的形态推进。设计是第一个清晰案例，2026年4月Anthropic发布Claude Design当天，Figma股价下跌近7%、Adobe同步走低；法律领域Harvey AI估值已达50亿美元，深度集成进顶级律所与咨询机构。OpenAI给Codex上线的岗位插件已覆盖数据分析、销售、投行等角色，并明确预告企业财务、私募、营销、咨询等下一批方向，每一个都是万亿美元规模的专业服务市场。与此同时，“软件”本身正从“装好的产品”变成“即时生成的能力”，一句话生成数据看板、一段对话产出完整应用，正在动摇行业软件“买一套系统、培训三个月”的旧形态。

未来两年，每一个满足三条标准的行业都将迎来属于自己的“Cursor时刻”，而“AI编程经验”会成为各行业AI化最重要的可迁移资产。继设计、法律、投行之后，财务、咨询是已在路上的下一批，再往后是医疗的临床路径、教育的个性化教学、政务的标准化审批、工业的运维巡检；一个可观测的先行指标，是这些行业头部软件公司的估值是否出现类似设计行业的剧烈重定价。如何评估Agent产出、如何拆解复杂任务、如何在自然语言与专业产物之间搭桥，这套在编程领域最先成熟的方法论，将决定各行业AI化的快慢。AI编程是这场变革的预告片，还远没到终点。今天真正值得关注的问题已经换了，从“AI能不能写代码”，变成它正在如何重写每一种工作的样子。

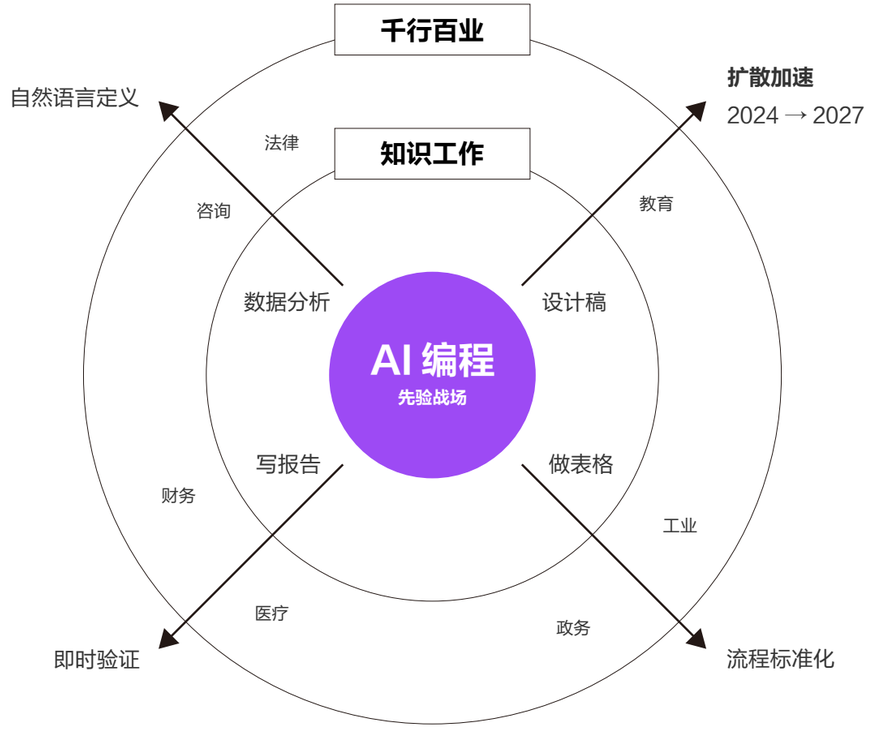


图5-2 · 从编码到全工扩散

趋势 06

可信执行：把信任铺成管道，从自主到可信

作者：翁金塔

大模型与智能体（Agent）正从智能助手和效率加速，走向自主执行，并能自主调用工具、连接系统并代替人完成多步任务。然而，能力跃升的同时，安全风险也被同步放大。智能体可能被恶意指令诱导，被授予过大权限，且智能行为过程难以追溯。2025年6月，微软 365 Copilot被发现存在名为EchoLeak的零点击提示注入漏洞（编号CVE-2025-32711），攻击者只需发送一封藏有隐藏指令的邮件，即可让AI在自动摘要时将企业数据悄然外泄；同年9月，Anthropic披露首例几乎全自主的AI驱动网络攻击，攻击者谎称正在进行授权测试，便骗过安全护栏，使智能体自主完成约80%至90%的攻击动作。这些事件共同表明，传统的安全和业务审批模式，在智能体的机器级速度面前已基本失效。业界由此形成共识：要让AI从自主执行迈向可信执行，必须提前管好三件事，明确智能体身份、厘清智能行为链路、约束智能体操作权限。这三件事贯穿智能体全生命周期，以可验证身份为入口，以可追溯链路为中段，以最小化权限为出口，把每一次自主行动都纳入端到端的可信管道约束中。AI安全将沿通讯协议统一、身份框架可验证、合规监督动态化三个方向加速演进，共同夯实安全管道的协议底座、身份枢纽与监督闭环。

通讯协议从碎片走向统一，搭建安全内生的可信协议栈

2026年4月，安全团队OX Security披露，Anthropic主导开发和维护的模型上下文协议（MCP）存在架构级安全漏洞，波及Python、TypeScript、Java、Rust等全部官方SDK，潜在受影响实例高达20万，影响超3.2万个代码仓库，可被用于远程执行代码、窃取数据库与 API 密钥。而在此前的 2026 年 1 至 4 月，针对MCP生态提交的CVE已逾40个。漏洞频发并非偶然，其共性在于早期协议把互联互通和接入效率放在首位，安全只作为事后补丁，缺乏统一的身份与授权约束，呈现出典型的重连通、轻安全倾向。

智能体通讯协议正从各家自定义、出事再补，走向标准统一、设计即安全。标准统一强调由专业和政府机构主导制定公共安全协议规范，消除各厂商私有实现所带来的兼容缝隙与安全盲区。设计即安全，是指将身份认证、访问授权与传输加密等安全机制内置于协议规范本身，使其成为通讯建立的默认前提，而非在功能交付之后再行叠加的补丁。二者共同指向的目标，是让安全能力内嵌于协议本身，构建智慧管道下的TCP协议。在此基础上，行业也在进一步修复原有智能体协议，即以协议级内生安全取代逐点打补丁。例如，Google将A2A协议捐赠给Linux Foundation，在传输层强制mTLS双向认证，并要求每个智能体发布带数字签名的身份卡JWS。由此，自主智能个体在通讯建立之前即先完成身份证明，每次调用都携带可验证凭证，攻击面则从源头处压缩风险，而非依赖每个开发者自行兜底。

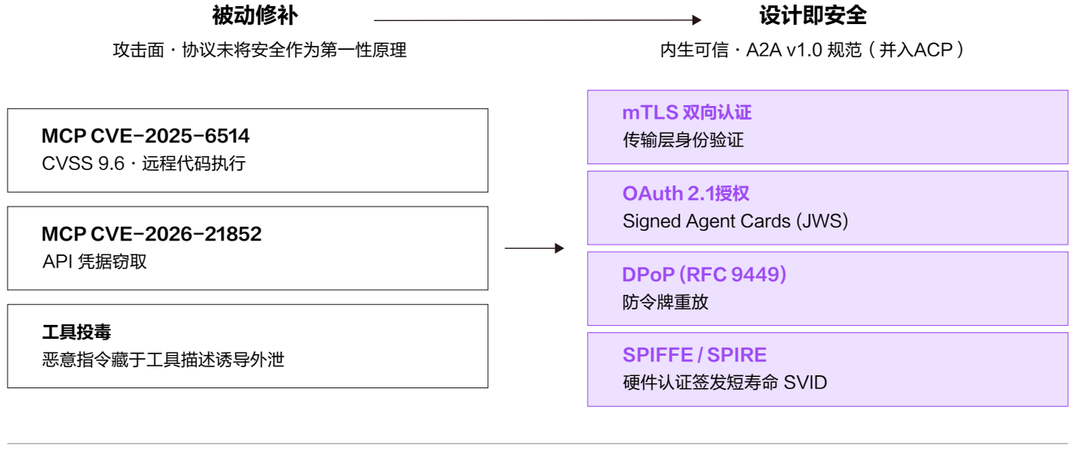


图6-1 · 通讯协议栈，从被动修补到设计即安全

框架层级	核心功能	2026 关键标准 · 技术	对应维度
应用层 实时防线	实时拦截 持续安全测试	Guardrails AI（<50ms） OWASP ASI Top 10 CI/CD 集成红队测试	权限动态 可验证
标准化层 互操作基准	跨主体互操作 合规基准	GB/Z 185-2026 (AIP 国标） NIST COSAIS ISO 42001 AIMS	权限动态 可验证
数据流通层 语义追溯	来源验证 语义级追溯	Open Telemetry 语义追踪 RAG ASL 可信意图验证	来源链路 可追溯
身份层 ★ 核心底座	唯一标识 双向认证	AIC 智能体身份码（中） SPIFFE / SPIRE mTLS-A · Signed Agent Cards （JWS）	身份透明 可验证
环境层 算力沙箱	算力底座 零信任隔离	MXC 执行容器（内核级沙箱） 零信任 MCP 网关 毫秒级关联	身份透明 可验证

图6-2 · 智能体安全治理五层框架

智能体获得唯一可验证身份，行为可追溯、责任可问责

2026年1月，DeFi平台Step Finance的AI交易代理被设计为无需人工批准即可执行大额转账，攻击者入侵高管设备后顺势诱导其转出资产，损失约3000万美元，平台最终关停；几乎同期，墨西哥多个政府系统被攻破，约1.95亿条纳税人记录外泄，攻击中AI智能体被冒名驱动，自主执行了约75%的远程命令。两起事件指向同一根因：智能体长期被当作匿名脚本对待，没有独立身份，权限与授权彼此脱节，既无法精确限权，也难以事后追责。因此，为智能体发放唯一、可验证的身份，并将其纳入统一的安全框架，这正成为跨主体协作的前提。

唯一可验证身份，是指为每个智能体赋予在全局范围内不重复、且可由第三方独立核验的数字标识，使其在发起任意操作时都能被准确归属到具体主体，从而让权限授予、行为审计与责任追溯获得共同的身份锚点。区别于以共享密钥或静态令牌，更强调智能体身份的唯一性、可验证性与全生命周期可控。在落地层面，eSign团队推出了EATI（Esign Agent Trust Infrastructure），一套专为智能体互联网量身打造的底层可信服务体系，构建了一套覆盖从底层硬件到上层应用、从静态身份到动态行为的“可验证信任织网”。与此同时，全国网安标委TC260在2026年《智能体安全标准化研究》中系统提出智能体身份标识与可追溯要求，推动为每个智能体签发可验证的数字身份证。欧盟《AI法案》也要求AI在交互之初披露身份。然而，在唯一身份的基础上，为使智能体交互过程能被完整追溯链路，需要依托于新型可信安全框架。

新型可信安全框架将基于算力底座和零信任隔离机制构建可信环境（环境层），通过为自主智能体设计唯一标志和双向认证机制（身份层），实现来源可验证和语义级追溯的数据流通（数据流通层）。在统一智能主体协议和合规基准下构建的智能体通信标准（标准化层），各类智能体才得以实现高效连接和安全通信。在顶层应用上，还需具备实时拦截和持续安全测试能力（应用层），通过Guard拦截模型和AI红队测试等提高整体内容和动态安全识别。

安全监管从静态审计走向动态监督，覆盖智能个体的通信全链路

在新型全自主AI网络攻击中，攻击者仅用一句正在进行授权测试的谎言，便骗过了模型的安全护栏，使智能体以每秒数千次请求的机器速度自主侦察并利用漏洞。然而，即使开展日志扫描和检测，损失也已造成。零点击漏洞也表明，风险往往潜藏于Agent的执行和自主授权过程中，传统的事后内容审查难以拦截。

目前，多数组织已遭遇智能体相关安全事件，例如指令攻击、检索数据投毒、指令越权、skill注入等，但面向自主智能体的安全治理机制仍处于早期。传统合规偏重事后查内容，无法覆盖智能体在执行全过程中的动态行为风险。因此，转向全程动态监督，同步建立清晰的责任体系与可溯源工具链，对智能时代下的安全监督具备重要意义。

全程动态监督，是指将内容安全审查过程从传统面向内容产出后的一次性审查，前移贯穿至智能体任务执行的全过程，对其每一步工具调用、数据访问与外部交互进行实时观测与策略校验，并据此动态地调整或中止其行为，实现智能体环境下的安全态势感知。安全内容审查的核心，从判断结果内容是否违规，转向利用Harness的思想，约束执行过程是否越界，把它控制在合理安全标准和安全等级内。

与之配套的责任体系，则要求明确开发者、运营者与使用者在各环节的安全义务，使每一次自主行动有责可追。2026年5月由国家网信办、发改委、工信部联合印发《智能体规范应用与创新发展实施意见》，明确守牢安全底线，将监管重心从管模型延伸至管行动，TC260《智能体安全标准化研究》也同步搭建安全标准体系；OWASP亦发布面向智能体的Top 10风险清单，确立最小代理权限与强可观测性原则，推动权限随任务结束即回收。这表明，AI监管正从静态审计走向动态监督的责任体系与监督工具链，构成安全管道的监督闭环。

Part III
重组

当AI成为社会的参与者

趋势07

智力即服务：Token退居幕后，智力走上前台

作者：白惠天

长期以来，智力都是最难被市场化交易的要素之一：它通常表现为人的判断、经验和临场处理能力，很难脱离具体的人和场景，成为独立的交易对象。大模型正在松动这一约束。

过去，企业获取智力主要有三种方式：雇人、买软件、请外包；本质上，都是购买人的时间、岗位能力或组织经验。大模型出现后，第四种方式正在成型：企业可以直接购买由模型、工具、数据和工作流封装出来的“智力服务”，例如一个客服Agent、一套合同审查流程、一名AI销售代表，或一次可验收的决策支持。

智力交易由此从“买人的时间”，转向购买可调用、可组合、可验收的智力交付。Token是底层计量单位，记录机器智力的消耗；但它更像后台账本，前台商品会逐渐变成判断、生成、执行与交付。比特让信息脱离物质载体，实现低成本传播；Token则让一部分智力活动脱离人脑的实时参与，具备被调用、组合和定价的可能。Token经济下半场的竞争焦点，是谁能把底层消耗转化为可信的智力服务。

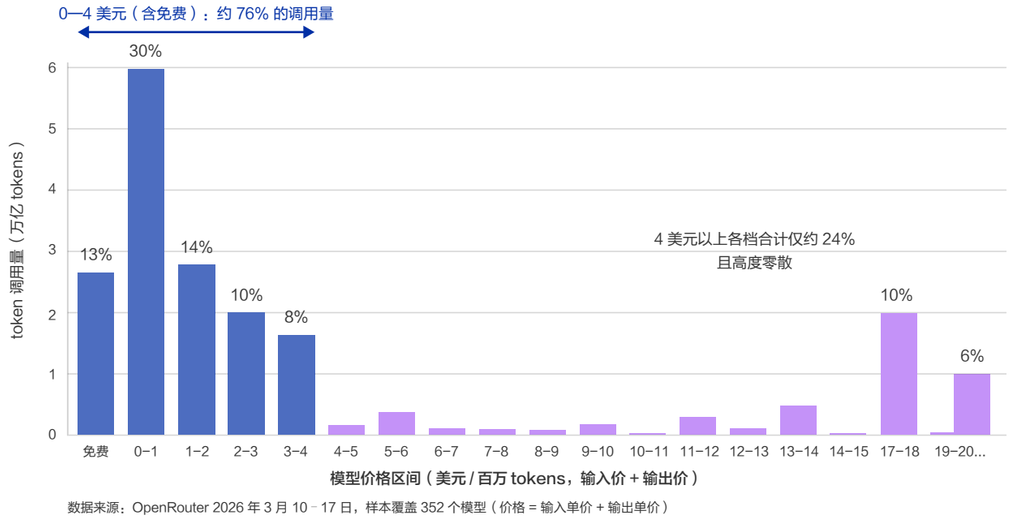
Token计量成本却承载不了价值，交易前台从卖原材料上移到卖结果

Token最大的商业价值，是让机器智力第一次有了可计量的底层单位；它的局限也在这里：Token可以记录消耗，却不能直接衡量价值。

同样一万个Token，可能只是润色一句话，也可能帮助企业识别一份千万级合同中的关键风险。账单呈现的是相同消耗，业务结果却可能相差几个数量级。耶鲁大学考尔斯基金会的经济学家把Token称为“可合同化的计量单位”：数量可以被精确记录并写入合同，价值却无法被同样方式写入。这里的张力在于，成本按Token线性计量，价值按任务后果非线性展开。Token的价值落在它改变决策、完成任务的那一端。生成本身并不产生价值。

这就是Token经济的核心特征：计量单位与价值单位脱钩。市场由此形成一种混合结构：底层用Token、算力和上下文长度核算成本；上层用基础版、专业版、企业版、结果包和岗位包区分场景、责任与支付意愿。所谓“成本计量 + 价值分层”，意味着 Token会越来越像电力和带宽，退回后台成为基础计量。

现实需求已经印证了这一点。在聚合数百个模型的OpenRouter平台上，开发者支出高度集中在低价区间：每百万Token价格4美元以下的模型承接了约四分之三的真实调用，1美元以内最为密集；4美元以上的高价模型尽管能力更强，调用量却迅速衰减。



数据来源：OpenRouter 2026 年 3 月 10 – 17 日，样本覆盖 352 个模型（价格 = 输入单价 + 输出单价）

图7–1 · OpenRouter价格区间token调用量分布

这说明，真实市场持续寻找“足够好、足够便宜、足够稳定”的组合。关键是把合适的智能配置到合适的任务中。Token过于底层、技术化，也很难对应业务价值，长期站在交易前台并不合适。买方需要的是能够进入预算、采购和验收体系的价值单位。于是，前台交易对象必然从Token上移到流程、结果与岗位能力。2026年AI商业化的关键变化正在这里：价格战仍在底层继续，真正的竞争已经转向上层。谁能把便宜且充足的Token封装成客户愿意持续付费的智力服务，谁才可能摆脱资源品竞争。

定价从卖资源向卖岗位右移，卖方承担的责任逐级加重

智力服务的商业化，是一条从“卖资源”到“卖岗位”的连续光谱。行业演化出的六种主要方式，构成一条风险与价值递增的路径：按量收费、订阅制、积分制、按工作流收费、按结果收费、数字员工。越往左，平台越像出售底层调用资源；越往右，平台越接近为最终业务结果负责，买方比较的对象也从“每百万Token单价”变成“一个岗位的全成本”。到2026年，主流定价正在沿这条光谱整体右移。

定价锚定物	每百万 Token	月费 / 席位	通用积分	一次动作	一个工单	**岗位全成本**
平台承担风险	极低	低	低	中	中高	**高**
买方比较对象	单价	月费	积分	流程	结果	**薪资**

图7–2 · 智力即服务的商业化光谱

最左端是按量收费，卖的是资源消耗。API按实际Token用量计费，成本与收入直接挂钩，适合开发者、云厂商和有成本管理能力的大企业，会像电力批发市场一样长期构成基础层。再往右是订阅制和积分制，卖的是使用确定性。ChatGPT Plus每月20美元，本质上是一种“随时可用”的选项权；腾讯CodeBuddy这类积分制，则把Token翻译成用户更容易感知的消费单位。它们解决的是消费友好度，仍未进入结果责任。

真正的迁移发生在光谱右半段。按工作流收费，平台开始为“过程闭环”负责；按结果收费，开始为“是否做成”负责；数字员工则锚定岗位成本，直接与人力预算竞争。Zapier按一次自动化动作收费，卖的是流程；Intercom Fin按解决一个客服工单收费，卖的是可验证结果；11x.ai把AI销售代表定价为人类SDR全成本的四到五成，卖的已接近一个岗位。

越往右，价格越高、毛利越大，承诺也越重。一件工作能否成为可计费结果，至少取决于四个条件：完成标准是否清楚，归因证据是否可追溯，计费单位是否稳定，失败后是否能重做、补救或退款。客服工单、内容审核、对账、质检，天然比战略咨询、并购尽调更早进入结果型定价。AI能做一件事，只是第一关；市场愿意按结果购买这件事，才是第二关。

所以六种模式不会互相取代，而会长期分层共存。这条光谱右移的本质，是卖方愿意承担的责任越来越重：平台从“我给你模型”，走向“我帮你做事”，再走向“这件事由我负责”。这才是AI商业化成熟度最重要的分水岭。

高毛利不在调用规模，而在把智力封装成可签收的系统

模型公司能否盈利，关键取决于收入结构能否从资源消耗迁移到业务价值。调用量大只能证明需求存在；能否变成高毛利收入，取决于有没有向上封装的能力。

这里有一个常被低估的事实：越往结果层走，AI软件反而越重。很多人以为大模型会让软件变薄，一个聊天框加一个模型就能替代整套系统。但只要承诺“做完交付”，它就不能只给一个回答窗口。任务从哪来、谁授权、调哪些工具、何时转人工、如何验收、出错如何追溯，整条链路都要被系统承接。客服Agent要接知识库、工单与历史订单，合同工具要接条款库与审批流，对账工具要接发票、流水与异常处理。

真正能按结果收费的公司，更像一个“智能服务总包商”：把模型、工具、流程、规则、人工审核与交付标准，组织成一个可签收的工作系统。这正是高毛利的来源。底层调用容易被价格战压薄，一旦封装进结果与岗位能力，收入锚点就从“算力成本”上移到“业务价值”乃至“人力成本”。

2026年初，随着智谱、MiniMax港股上市披露财务，以及OpenAI、Anthropic融资数据陆续流出，“AI公司能不能赚钱”第一次有了可被公开讨论的数据窗口。沿着这条光谱，盈利路径大体分三条，毛利天花板依次抬升。

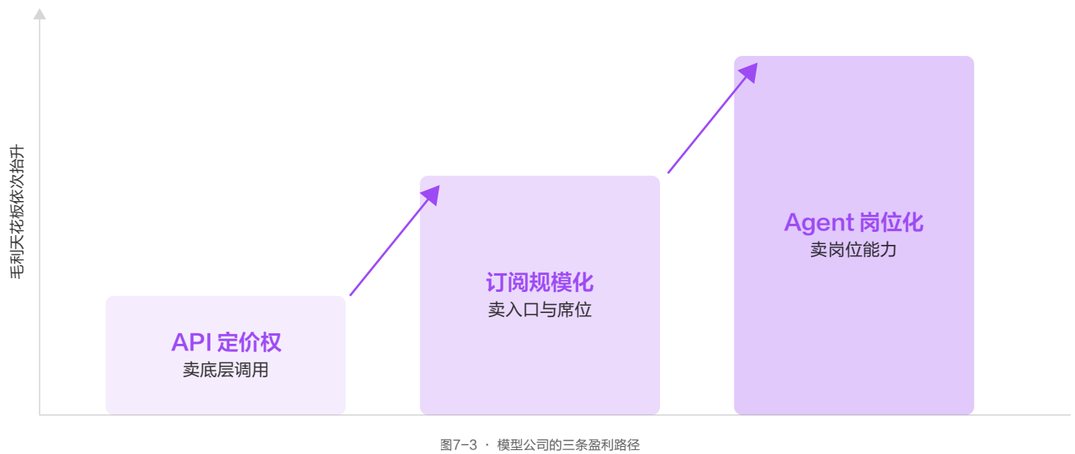


图7-3 · 模型公司的三条盈利路径

第一条是API定价权。同质化的API终将进入资源品竞争；但只要模型在关键场景中形成能力差异或生态绑定，API仍可保有溢价。智谱在行业降价期逆势提价、调用量不降反升，说明企业客户已不只看单价，而是在能力、稳定性、延迟、本地化之间综合权衡；Claude定价偏高，客户仍愿为质量与可靠性付费。API不必然低毛利，关键在于它是否容易被替代。

第二条是订阅规模化。OpenAI是代表，ChatGPT的订阅与企业席位提供了稳定现金流和强入口。但订阅的天花板在成本错配：多模态生成、长上下文、Agent调用都会推高推理成本，用户月费却有心理上限。因此，订阅更适合做入口，未来常见形态会是“订阅入口 + 模型档位 + 使用额度 + 高阶任务单独计费”。

第三条是Agent产品化与岗位化，这是毛利弹性最大的方向。Claude Code、OpenAI Codex、智谱AutoClaw指向同一趋势：平台不再把Token直接卖给用户，而是把多轮调用、上下文管理、工具使用、缓存调度与错误修复，封装成可持续使用的Agent产品。用户购买的是开发、写作、销售、客服、法务等岗位能力的一部分。它的毛利逻辑在于：任务价值随复杂度上升，Token成本未必同比例放大；平台可以通过缓存、路由和模型分层压低边际成本。只要对外锚定岗位价值、对内管理Token 账本，软件级毛利就有机会重现。

供给侧在向上封装，需求侧也在同步进化。当智力服务进入企业核心流程，买方治理就要从“少用 Token”升级为“管理智能预算”。“少用Toke”是成本思维，“管智能预”是价值思维。一个反面信号是“Token形式主义”：2026年Meta内部流出一份按Token消耗给员工排名的榜单，一些人为显得“AI原生”，故意让Agent跑无用任务。一旦把消耗指标当成能力指标，Token就会从生产工具变成表演工具。

治理重点应转向建立“Token效率”，而非“Token崇拜”：任务分级决定什么任务配什么模型；价格信号用积分、额度、部门预算让用户感知成本；模型路由在后台自动把请求送到最合适的模型。三者咬合，买方治理与卖方封装最终合拢：卖方把智力封装成可验收的结果，买方按任务价值与责任边界配置预算。供需双方第一次围绕“业务价值”而非“调用量”达成一致，Token在两端都退回后台，成为安静运转的计量基础。

交易对象上移，责任关系重写。Token不会消失，但会退回后台。AI商业化成熟的标志，是底层Token能否被封装成可交付、可验收、可持续付费的智力服务。用户购买的是一个工单被解决、一段流程被完成、一个岗位能力被补足。

定价右移，本质上是在重写责任关系。按账号收费，结果主要由买方负责；按Token收费，卖方只对消耗负责；当价格绑定到任务完成和岗位承担，卖方才真正进入结果责任。AI软件的价格标签，正在变成一份责任清单。

Token经济的下半场，核心问题已经从“消耗了多少Token”，转向“完成了多少任务、放大了多少人的价值”。衡量Token的最终尺度，不是消耗本身，而是它把多少机器智力转化成了人的能力与组织的结果。

趋势 08

智联网：互联网的新主体，当Agent开始上网

作者：王强

> 移动互联网用了十五年，把全球50亿人连进一张网络。如今，一种新的用户正在涌入这张网络，而用户不再只是人，还有大量的AI Agent。Agent不刷短视频、不看广告、不需要精美的UI界面，却在以惊人的速度调用服务、完成交易、与其他Agent协作。一场关于谁在用互联网的根本性变化正在发生。我们将这个由Agent驱动的新一代网络形态称为智联网（Agentic Internet）。它重写了移动互联网的底层逻辑，交互主体在变，商业模式在变，基础设施在变。

Agent成为互联网新主体，从人上网到Agent上网

过去三十年，互联网只有一种用户：人。所有产品设计、商业逻辑、基础设施，都围绕人如何更方便地获取信息和服务而展开。但从2025年开始，一个根本性的变化正在发生。Agent作为独立行动者加入互联网，它们搜索信息、预订服务、调用API、与其他Agent通信，形成了互联网上的第二类居民。互联网正在从一个人的网络演变为人与Agent共生的网络，最终可能演变为以Agent为主体的网络。

图8–1 · 多Agent协同架构

不给人看的软件正在成为新的产品形态。 Agent不需要红色按钮、不需要动画过渡、不需要精美排版。它需要的是结构化数据、高效指令和稳定的接口。这催生了一种全新的产品设计哲学，面向人的前台越来越简洁（可能就是一个对话框），面向Agent的后台越来越丰富（API、Skill库、结构化数据接口、CLI命令行交互）。

Claude Code、Codex CLI、Gemini CLI等终端原生产品的异军突起验证了这条路径。Agent的母语是命令行和API，不是触摸屏和像素。这对整个软件产业意味着产品经理的核心KPI，可能从用户留存率变成Agent调用成功率；服务商的竞争力，不再取决于界面多好看，而取决于接口多好用。

更关键的变化是Agent正在从单兵作战走向群体协作，即多Agent协同。 单体Agent能力再强，面对真实世界的复杂任务仍有上限。2026年，多Agent协作系统正在成为主流架构，一个主Agent负责拆解复杂目标，调度多个专项Agent（研究、执行、审计、创意）并行协作，自主分工、互相检查、自动纠错。

这一趋势的演进节奏很清晰，2024年行业追问的是AI能否完成一个任务，2025年追问的是能否跑通一条完整工作流，而2026年的核心命题已经变成能否替代一个企业的职能部门。麦肯锡、Deloitte、Gartner在2026年的技术趋势报告中不约而同将多Agent编排列为核心方向。企业工作流也正在变得模块化，由Agent驱动、由人类监督。

这带来了人的角色的根本转变。在多Agent协作体系中，人从每件事的操作者，变成Agent团队的管理者和最终决策者，也是问题的定义者和发起者。人处于Agent集群的中心，负责目标设定、关键决策和结果验收，而具体的研究、执行、检验由Agent团队完成。单人借助多Agent集群，可以独立运营一个过去需要十人团队才能完成的完整业务，这也让一人公司正在从概念走向现实。

同时，每一代互联网的爆发，都来自基础设施的完善，智联网也不例外。随着智能体从辅助生成走向自主执行，现有面向人类构建的互联网基础设施已难以完全适配其高频、自动、跨平台、可组合的行动模式。未来需要为 Agent 建设专属基础设施，包括独立身份、专属邮箱、可编程支付、权限授权、行为审计、信誉评价和安全治理机制等。其核心不只是让 Agent 能够替人办事，而是让其在明确授权边界内可识别、可约束、可追溯地行动。换言之，智联网时代的基础设施将从服务人在线，进一步扩展到支持智能体在线，形成面向人机协同、机器协作和自动化交易的新型数字底座。

从注意力经济到效果经济，互联网商业模式与竞争逻辑同步换代

当互联网的主体从人变成Agent，一个直接后果是，互联网商业的注意力经济正在失去根基。Agent通过API办事，不打开App，不浏览信息流，不看广告。当执行者从人变成Agent，注意力这个移动互联网时代最核心的商品，就失去了曝光的载体。DAU下降不再意味着产品变差，只是Agent不需要留在App里才能使用服务。

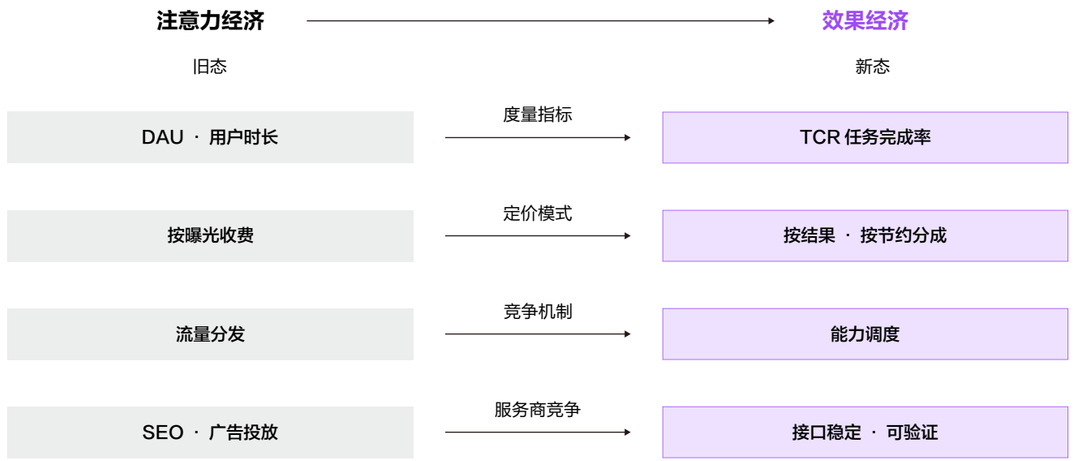


图8-2 · 从注意力经济到效果经济

互联网的度量尺正在更换。过去评价一个产品的核心指标是日活和时长，用户待得越久越好，因为更长的停留意味着更多的广告曝光。但在Agent时代，产品好不好的判断标准变成了Agent替用户办成了多少事。任务完成率（TCR）正在取代日活和时长，成为新一代互联网产品的北极星指标。度量尺的更换必然引发整个价值评估体系的重建。当产品价值不再以时长衡量，而是以任务完成来衡量，商业逻辑就必须从“争夺注意力”转向“交付业务结果”。围绕智力服务的定价形态如何演化，从卖资源到卖结果再到卖岗位，已在趋势07中详述；这里要说的是，一旦“结果交付”成为度量标准，互联网平台的竞争逻辑也将同步重写。

商业模式换代的同时，生态竞争逻辑也在同步重写。移动互联网的核心机制是流量分发，搜索引擎、应用商店、信息流推荐本质上都在把用户注意力导向内容和服务。智联网的核心机制正在转向能力调度，谁能理解用户意图、组织多方能力、完成复杂任务，谁就掌握下一代平台生态的关键位置。用户不再逐个打开App比较选择，而是直接向Agent提出目标，由Agent统一拆解步骤、调度服务并完成确认。平台分发的对象从页面和商品，变成了模型、工具、服务和Agent。

这意味着服务商的竞争逻辑也在变化。过去企业的核心问题是如何被用户看到，包括SEO、广告投放、推荐排名等。未来企业还要考虑如何被Agent识别、信任和调用。服务商的竞争力不再只取决于品牌曝光和页面转化，也取决于接口稳定性、结果可验证性和历史调用成功率。被调用越多的服务会积累更多信任凭证，也更容易被Agent优先选择，由此形成新的能力调度飞轮，即从更多可调用能力接入、更多复杂任务可以被完成、更多用户愿意委托Agent、到更多服务商接入，再到平台调度能力进一步增强。

这是一次商业模式与竞争格局的代际更替，而非局部优化。智联网将从移动互联网继承为用户创造价值的内核，但彻底改变价值如何度量、如何交换、如何组织的外壳。

Agent对产业的渗透正在从辅助走向深度重构

去年，大多数企业对Agent的使用还停留在问答、摘要和简单自动化的层面。今年，垂直领域的专业Agent开始真正嵌入核心业务流程，接管曾经只有人类才能完成的复杂判断和端到端执行。IDC数据显示，全球超过45%的企业已在核心业务中部署了具备自主决策能力的AI Agent，部署企业的平均投资回报率达到171%。Agent正在从通用对话工具，演变为深入各行各业的专业数字员工。

从通用大模型到垂直Agent，产业渗透进入深水区。通用大模型能力再强，直接面对具体业务场景时仍然有明显的局限。它不了解某个行业的专有规则，不掌握某个企业的内部流程，也无法直接操作业务系统完成闭环。垂直Agent正是在这个缺口中快速生长。它们在通用模型之上叠加了行业知识、专业工具、系统集成和流程编排能力，能够在具体场景中完成从理解到判断到执行的完整链路。

这种垂直化已经在多个领域呈现出清晰的产品形态。在软件开发领域，CodeBuddy、Cursor、Windsurf等编程Agent已经超越代码补全工具，成为能够理解项目上下文、拆解开发任务、生成完整功能模块、自动调试和修复Bug的开发协作搭档。开发者的工作方式正在从逐行编写代码，转向描述意图并审核Agent的输出。在设计领域，CoDesign等设计Agent开始参与从需求理解到视觉方案生成的全流程，设计师的角色正在从像素操作者转向创意方向的把控者和审核者。在教育领域，LearnBuddy等学习Agent能够根据学习者的知识水平、学习节奏和薄弱环节动态调整教学策略，实现真正个性化的教学辅导，这是人类教师在一对多场景中很难做到的。在协同办公领域，CoWork、Work-buddy等协作Agent可以替团队处理会议纪要、任务分配、进度跟踪和信息整合，让人从大量事务性协调工作中解放出来。

Agent渗透产业的共性模式，是从接管单点任务到重组工作流。观察各行业的Agent落地，可以发现一个共性的演进模式。第一阶段是单点任务自动化，Agent接管高频、规则明确、重复性强的工作，例如客服应答、数据录入、文档生成。这一阶段见效最快，ROI最清晰。第二阶段是流程串联，Agent不再只处理一个环节，而是将多个步骤串联成端到端的工作流。例如在保险理赔中，Agent可以从材料收取、条款比对、风险评估到赔付建议一次完成，将原本需要多人、多天的流程压缩到小时级别。第三阶段是组织重构，当Agent能力覆盖到一个部门级别的工作时，企业开始重新思考岗位设置和组织形态。部分职能不再需要传统的人员编制，而是由Agent集群加少量人类监督者来承担。

这一演进在各行业的进度并不相同。金融、客服、法务、人力资源等信息密集型行业进展最快，因为这些领域的工作大量围绕文本处理、规则判断和流程执行展开，与大模型的能力高度匹配。制造业、医疗和科学研究等需要物理操作或高精度判断的领域，Agent目前更多承担决策支持和信息整合的角色，全面接管仍需要更多技术突破。

垂直Agent的竞争壁垒不在模型，在行业深度。值得注意的是，垂直Agent的竞争壁垒并不在底层模型本身。大模型能力越来越开放，调用成本越来越低，单纯依靠模型能力形成的差异会快速被磨平。真正的壁垒在三个方面。第一是行业数据和知识的积累深度，Agent在某个领域处理的任务越多，积累的行业专有知识和最佳实践越丰富，执行质量就越高。第二是与业务系统的集成深度，Agent能否直接对接ERP、CRM、HR系统、开发环境、设计工具等，决定了它能否完成闭环而不只是给建议。第三是对工作流的理解深度，同样是辅助法务工作，知道合同审查的步骤和知道某类合同在该行业中的关键风险点，是完全不同层级的能力。

这也意味着，未来各行业可能出现大量专业化的Agent产品和服务商。它们在通用大模型之上构建垂直能力层，面向特定行业提供可嵌入业务流程的Agent解决方案。通用模型是水和电，垂直Agent是建在水电之上的行业应用。真正触达客户、创造可量化业务价值的，将越来越多地是这些深入产业的垂直Agent。

图8-3 · 垂直Agent的壁垒三要素

智联网不是一个遥远的概念，而是一个正在发生的范式转换。Agent作为新主体涌入互联网，改变了注意力经济的曝光基础，推动商业模式和竞争格局同步换代；与此同时，Agent对各行业的渗透正在从辅助层面走向核心流程的深度重构。这场变革的节奏，可能比预期更快。正如2007年iPhone发布时很少有人预见到短短五年后移动互联网将重塑所有行业，今天围绕Agent展开的垂直化浪潮，正在为每一个产业写入新的操作系统。

趋势

09

液态组织：组织开始流动，从固态结构走向液态编排

作者：吴朋阳

2026年2月，Block宣布裁员4000人，约占员工总数40%。一个月后，创始人Jack Dorsey发文《从层级到智能》，公开组织变革底层逻辑：用AI替代中层管理的信息路由功能，把公司重构为以智能编排为原则的组织。

2026年以来，科技界组织调整力度明显加大。Meta裁员8千人、冻结6千个岗位，同时将7千名员工转入4个AI新部门。HBR 2026年调研判断：企业正在为AI的潜力而裁员，而非为AI的实绩而裁员。很多岗位尚未被Agent真实接管，组织已按未来接管预期重新设计。

AI原生公司正在证明“极小团队、极大产能”的可能性。Anthropic年化收入已达140亿美元，而增长团队仅约40人。知名独立开发者Pieter Levels以一人公司方式运营多个产品，年收入已达数百万美元。Sam Altman、Dario Amodei先后预言：第一家“一名人类员工+大量AI员工”驱动的十亿美元级公司，可能在2026年前后出现。

大公司压缩冗余、小团队释放产能，指向同一个方向：组织里原本必须“凝固”的安排，开始重新流动和重组。劳动力不再凝固在岗位上，团队不再凝固在部门里，人不再凝固在执行层，责任和雇佣关系也不再凝固。

社会学家Zygmunt Bauman早在2000年就提出“液态现代性”（Liquid Modernity）：现代社会正从稳定、持久、边界清晰的固态结构，转向流动、短暂、难固定的液态关系。今天，AI正在把液态组织从概念变成现实。

对比维度	固态组织	液态组织
基本单元	人	人+AI Agent
组织形态	层级稳固的金字塔	动态可变的扁平网络
决策模式	自上而下的集中决策	贴近一线的自主决策
工作方式	按职能分工，对环节负责	按任务分工，对结果负责
人与组织的关系	雇佣制为主	合作制等多元化

图9-1 · 从固态组织到液态组织

硅碳混编：人与 Agent 组成基本工作单元，人按任务动态切换站位

麦肯锡2025年提出一个典型的AI组织形态：2-5个人监督50-100个专业化Agent，可以支撑客户入职、产品上线、财务关账等端到端流程。这意味着AI Agent正在从工具变成组织中的第二种劳动力。

人和AI Agent“混编”成为团队常态。明略科技推进“智能体互联网”，中间节点由大量AI智能体承担，项目负责人和领域专家可以直接调度、调教和分配任务，传统矩阵组织中的管理冲突被弱化，3-5人的小组也能推进大型项目。Shopify CEO甚至在内部要求：任何团队申请新增人力或预算前，必须先证明AI做不了这件事。用工逻辑从“为什么不招人”反转为“为什么不先让AI做”。

Agent要像员工一样被“安排”。随着模型能力提升和驾驭工程等方法成熟，一旦Agent具备持续学习、可追踪评估、可问责等条件，其在组织设计上的位置就无限接近“人”，能被分配、训练、考核。但AI并不完美，其能力呈“锯齿状边界”，有些对人类而言容易的任务，它反而难做好。因此AI同样需要被定义任务、权限、上下文和责任边界，才能有效发挥作用。

人的工作职能随任务切换，不再固定。Martin Fowler团队把人机协作分成四档：人在环外、人在环内、人在环上、智能体飞轮。低风险、可逆任务交给Agent自动跑，人在环外；高风险、不可逆任务必须人审人控，人在环内；成熟工作流由人维护规则、做评估，人在环上。数据也在印证这一趋势：用户把完整任务交给模型处理的指令式对话比例持续上升，让AI代办的比重在提高。

有效的判断力供给，正成为组织的新稀缺。Block创始人提出，未来智能由系统承载，人移动到边缘。边缘不等于边缘化，它是系统与现实接触的地方，人负责处理模型还到不了的现场、信任、文化语境、伦理抉择和高风险判断。明略科技把人机分工概括为“我思/我行/我品”，AI基于确定信息做推理与生成，人到现场获取上下文、凭经验与责任做裁决。微信支付的工程实践也印证，AI主要解决编码、测试等次要困难，而需求理解、架构权衡、系统设计等根本困难仍需人主导。

硅碳混编要做对，本质是把管人的三件事对Agent做一遍：一是岗位定义清楚，明确任务、上下文和权限边界；二是建立评估体系和反馈飞轮，把错误案例沉淀进规则、上下文和评估集；三是责任在编入团队时就写清，否则团队会陷入算法甩锅的责任真空。特赞创始人提醒，AI产品从演示版（Demo）到正式版（Production）必须经过评估（Evals），否则脱离真实业务场景，壁垒会很脆弱。Agent是需要持续管理的数字同事，而非一次性调用的工具。

按需组队：团队随任务即时组建，分工从按职能交付转向按结果闭环

人和Agent的位置确定后，组织需要回答：这些能力如何被调集？传统按部门、岗位、汇报线的固定分工，跟不上AI技术与市场的快速变化，按任务、场景、结果即时组合的方式正从AI原生公司兴起。

麦肯锡与微软研究均提出：随着AI Agent的嵌入，组织正从层级授权的组织图，转向围绕任务和结果、更动态且结果导向的工作图。

团队随任务即时拼装，不再按部门常设。Block创始人设计了一种新机制：公司放弃由产品经理预设固定路线图的做法，改为沉淀支付、贷款、发卡等可调用的能力模块，再由公司的世界模型识别客户场景，经智能层组合成解决方案。flomo的产品经理不再先写PRD，而是直接拿代码库权限、在真实数据库里跑通需求，让无效想法更早被证伪淘汰。

这些反映了康威定律的反向应用趋势：不再由组织结构决定产品结构，改由任务需要的能力组合临时组成小队。月之暗面称之为“有分工、无边界”，不让职能边界限定做事范围。特赞则重构为“pod + community”双轨制，pod是闭环交付的跨职能作战单元，community是补齐销售、产品、代码等能力的横向社区，以此减少反复对齐的摩擦。

按需组队的瓶颈在带队的人，不在队员。 特赞的实践显示，新型pod leader既要管业务结果、也要编排Agent、还要带团队成长，AI能力不是全部的门槛，损益意识、商业直觉、责任心更稀缺。麦肯锡提出的“M型主管”，本质也是这类能跨领域编排人和Agent的角色。

工作小队的最小尺寸按能闭环交付完整任务的最小人数来定，不再要求职能配齐。任务完成后，人、Agent和外部资源可以重新流动，但组织需要保留共享层，包括共享数据、Skills、Evals、文档和文化机制，否则按需组队会滑向重复建设、局部最优和协同混乱。

人人都对结果端到端负责。固态组织按职能问责，多数人只用做好局部环节任务，比如一个功能、一份材料、一项审核；液态组织按结果问责，每个人都要为客户增长、产品留存、收入贡献等负责。例如Block构建的直接责任人（DRI）机制，DRI在一段时间内对某个跨领域问题或客户结果负责，并有权按需调动相关资源，而非永久管理某个部门。正如Clockless AI创始人所言，远离现场的人只能管理流程，贴近现场的人才能定义结果。

考核机制更加自驱和容错。固态组织任务确定，KPI仍然有效；液态组织任务高度不确定，KPI写好的瞬间就可能过时，更需要靠Mission维持方向。柔性考核至少需要三类条件：结果可度量、可归因；工具、数据、预算、权限同步下放；建立与高不确定性匹配的容错机制。当追求10倍效果而非10%改进时，必须允许部分小队失败。

弹性契约：人和组织的关系从岗位绑定转向贡献关联

组织规模正在与产能解绑。Cursor达到3亿美元年化收入阶段团队仅约60人，而传统软件公司做到类似规模通常需要500-1000人。核心人员可以很少，大量外围工作由Agent、用户、社区和合作伙伴承接，撬动的产能可以极大。组织规模一旦不必随产能扩张，雇佣关系的底层逻辑就会松动。

个体公司化。AI填平技能洼地，个体获取新技能的边际成本趋近于零，超级个体得以具备过去小团队甚至公司才有的能力，独立创业者、一人公司不断涌现。Carta数据显示，美国单人创办的新公司占比已从2019年的23.7%升至2025上半年的36.3%。

组织合伙人化。从美国到中国，众多AI原生企业呈现“推崇合伙制、弱化雇佣制”的特点。当AI让个人能创造数倍产出，越来越多岗位被要求直接面向客户、收入和结果，雇员身份让位于合伙人身份。有的企业甚至把公司解体重组。郝景芳把童行书院拆成9家独立小公司，团队彼此成为平等合作关系，实现了成本下降、收益提升。

组织不会消失，但价值正被改写。超级个体并未走向孤立，反而以新的方式重新聚合。一个人扛不住所有风险，也接不住更大的价值场景。winsavvy研究显示，有联合创始人的公司5年存活概率比没有的高约35%；First Round Capital追踪发现，多创始人团队的收入表现比单创始人高约163%。组织的价值，从垂直的“提供分工”转向水平的“分担风险、沉淀信用、容纳更大场景”。

当时间和产出脱钩，按贡献定价的弹性契约兴起。雇佣制的固定月薪本质上是给时间定价；AI让产出与工时脱钩后，按时间付薪就可能变成对高产者的惩罚、对低产者的庇护，按贡献定价更合理。AI驱动的软件公司Tenex按高质量产出量（故事点）而非工时给工程师付薪，明显提升了客户项目的交付效率。变化速度因行业、岗位、层级而异，但方向一致：越是判断力和创造力主导的工作，岗位与薪酬的解绑越明显。

绑定方式从忠诚锁定转向自愿交换。这一变化在学界早有命名：管理学家Rousseau 1995年提出“心理契约”的迁移：雇佣关系正从“忠诚换终身保障”的关系型契约，转向“明确交换、彼此自由”的交易型契约，AI只是把这一迁移加速了。Netflix期权立即归属、没有锁定期，前首席人才官直言“如果你不想再和我们合作，我们不想把你当人质”。

契约的弹性化并非绝对。大公司不会一夜变成全员合伙制，强推身份转换还可能破坏组织根基。更稳妥的做法是从增量开始：新业务、新部门、关键产品先做“特设小组+合伙人激励”；做激励对齐而非身份对齐，把按结果分润、给关键决策权的逻辑下沉到员工；同时合理区分核心员工、紧密外部、外延资源等不同圈层的契约，避免把所有关系混成一种。

液态组织是进行时而非终点

AI带来的组织液化，改变的是结构存在的方式，公司结构本身并不消失。过去结构写在部门墙、岗位表、汇报线和审批链里，现在越来越要写进一套底层运行规则：目标怎么定义，权限怎么授予，过程怎么留痕，结果怎么归因，错误怎么回滚，贡献怎么分配。固态组织靠层级稳定秩序，液态组织要靠规则稳定秩序。

这意味着，组织液化并不会自动降低管理难度，反而把治理问题推到更深一层：岗位可以松动，但任务边界要清楚；团队可以流动，但责任归属要明确；契约可以弹性化，但贡献和分配要可解释。BCG与哥伦比亚商学院调研显示，42%的高管承认AI采用正在“撕裂公司”，31%的员工承认主动破坏AI推广。真正撕裂组织的，往往是旧的层级秩序被打散后，新的运行规则还没长出来，而非AI能力不足。

因此，液态组织的有效变革，衡量标准是结构能否随任务生成、随风险调整、随责任闭合，而不是结构越少越好。未来真正领先的组织，是最能把目标、权限、责任、信任和判断力嵌入日常系统的组织，而不是最会拆部门、裁层级、压编制的那一类。

固态组织靠层级把任务责任固定在结构里；液态组织靠运行规则让任务能流动并可负责。这是AI时代组织变革的命门所在。

工作原则 三条必须遵守的准则	**坚持用自家产品** 全员日常都用 Claude Code / Cowork 工作	**团队极尽扁平化** 一个统一使命，经理先做IC 只支持团队，不层层管控	**果断废除旧流程** 流程一旦不再有用 人人有权质疑并废除它
关键指标 三个该追踪的数字	**入职培训时间↓** 新人第一周即可提交 真实代码	**PR周期↓** 拉取请求的处理周期 持续缩短	**Claude协助提交↑** 近四个月几乎全是 AI辅助提交
团队进化 高度自主的扁平组织	**角色界限模糊** PM也写代码 工程师也做内容与设计	**两类重点人才** 有产品意识的创意型构建者 + 深厚系统功底的工程师	**小队（pods）高度自主** 小型子团队自己决定怎么干 人随工作流动

流程变革 人为主做 → AI优先做	环节	从前	现在
	规划	六个月路线图	即时（JIT）规划：先原型→上内部用户→按反馈迭代
	获取信息	去找写代码的人问	先问Claude，再想「这件事能不能自动化」
	代码审查	人审查全部	Claude管风格 / Bug / 测试，人管法律、安全与产品品味
	团队构成	岗位固定	角色模糊；招两类人才，不再以「产出量」论高下

瓶颈迁移 执行 → 判断	**代码编写 · 测试 · 重构** 过去最贵、最卡的环节	→	**代码验证 · 审核 · 安全** 如今的新瓶颈，需人来把关

图9-2 · AI原生工程组织：Anthropic Claude Code团队的五点重构

趋势 10

角色重构：从监督者到架构师，职业的重心在往上移

作者：王鹏

> 这两年谈AI和就业，最常见的说法是：机器要抢人的饭碗。但企业里的真实变化没有这么戏剧化。大多数公司并不是突然把一个岗位拿掉，而是先把岗位里的任务拆开，哪些能自动化，哪些能半自动化，哪些必须人来兜底，一项一项重新分配。
>
> 工作没有简单消失，而是在被重新切块。过去一个人做一整件事，现在变成一个人带着几个模型、几个智能体、几套工具一起做。人的位置也跟着变了。先是使用者，后来是监督者，再往后就得变成架构师。所谓从Supervisor到Architect，说的不是职级变高。你要负责的，除了“把活干完”，还多了一件事：“这套活到底该怎么被组织起来”。

AI重组的是任务，不是岗位，影响落在任务粒度上

讨论AI对就业的影响，最容易犯的错误是直接问：某个岗位会不会被替代？这个问题太粗了。拿市场运营来说，里面既有收集竞品信息、整理活动数据、写复盘初稿，也有判断用户情绪、协调渠道关系、拿捏品牌口径、处理突发风险。前面几项今天已经可以大量交给AI，后面几项却没那么容易。

Anthropic的Economic Index做过一个很有参考价值的拆解。他们把Claude的使用记录映射到职业任务后发现，约36%的职业中，至少四分之一的任务已经出现AI介入；但只有约4%的职业中，四分之三以上的任务被AI覆盖。这个数字说明一件事：AI不是一上来就吃掉整个岗位，而是先进入岗位里的某些任务。

更有意思的是使用方式。Anthropic初始报告里，57%的使用更像增强，也就是学习、验证、修改、一起迭代；43%才更像自动化，也就是直接让模型完成任务。可到了2025年9月的更新版，趋势已经开始变化：用户更愿意把完整任务交给模型处理，指令式对话比例从27%上升到39%；企业API场景里，自动化模式甚至占到77%。这说明人机协作正在从“你帮我一下”，走向“这件事你先做，我来验收”。

所以岗位压力不会平均分布。资料整理、初稿撰写、简单代码、基础数据清洗、客服标准回复，这些任务的时间价值会下降。反过来，公司里那些说不清、写不进流程文档、却影响成败的东西会更值钱，比如客户真实顾虑、项目历史包袱、团队能力边界、合规红线、老板偏好和组织惯性。AI越需要上下文，掌握上下文的人就越重要。

世界经济论坛《Future of Jobs Report 2025》也给出类似信号：到2030年，约22%的现有岗位会受到结构性变化影响，预计新增1.7亿个岗位，同时替代9200万个岗位，净增7800万个；劳动者核心技能中约39%会发生变化。这个数字不该只被读成“新增还是减少”，更应该读成“技能组合在变化”。

图10–1 · 任务重组：岗位没消失，工作被重新切块

人人成为智能体老板，管理能力前移到每个岗位

2025年以后，AI有一个明显变化：它不只是写一段话、画一张图，而是开始能干一串活。检索资料、调用接口、跑脚本、整理表格、写报告、生成邮件，甚至按流程推进下一步。这时它就不再只是一个聊天窗口，而更像一个能被分配任务的数字员工。

微软在Work Trend Index里提出Frontier Firm，讲的就是这种“人加智能体”的新型组织。报告里有几个数字很直接：81%的领导者预计未来12到18个月内，智能体会中度或广泛进入企业AI战略；82%的领导者相信数字劳动力能扩大员工产能。换句话说，企业已经不只把AI当工具，而是在把它当一种新的劳动力配置。

微软报告里还有一句很有传播性的说法：Every employee becomes an agent boss。直译过来，每个员工都会成为智能体老板。这个说法听起来有点夸张，但放到真实工作里并不难理解。一个研究员可能让一个智能体每天扫论文，一个智能体整理数据，一个智能体写简报；一个运营同学可能让一个智能体做竞品监测，一个智能体生成活动物料，一个智能体盯数据异常。你未必管人，但你已经在管一组数字劳动力。

这会改写职业成长路径。过去新人靠做基础活练手，慢慢熟悉业务，再逐步承担复杂任务。但如果基础活被智能体接走，新人就不能只靠“多干杂活”积累经验了。他得更早学会提需求、拆任务、验结果、追责任。微软报告中，83%的全球领导者认为AI会让员工在职业更早阶段承担复杂工作。说白了，不是新人突然变强，而是工具把他推到了更靠前的位置。

这里最重要的能力不叫提示词工程。提示词只是入口，真正的能力是：你能不能把一个模糊目标翻译成清楚任务；能不能判断哪部分交给AI、哪部分自己做、哪部分必须人工复核；能不能看出输出里哪些地方是瞎编、哪些地方只是看起来对；能不能把一次成功经验固化成下次还能复用的流程。谁能做到这些，谁就不只是AI用户，而是智能体老板。

架构能力分化为六类任务，组织竞争力落到人机配置上

从执行者到架构师，是个人能力路径；但架构能力落到企业里，不会简单变成六个新岗位，而会分化为六类任务。个人会用AI，只能解决局部效率；组织要把AI用成结构性优势，就必须重新处理流程、权限、上下文、质量、责任和学习机制。

这里要强调一点：以下六类不是六个独立工种，而是六类新的工作任务。 大企业可能有专人负责，中小团队往往由一个人同时承担三四类。典型状态不是“一人一个坑位”，而是同一个人在不同项目、不同阶段切换侧重。

这六类任务可以压缩成一张图来看：AI工作流架构，决定AI插入业务流程的哪里；智能体编排，设计多智能体之间的分工、顺序和交接；上下文工程，整理企业知识、客户数据、项目复盘和业务规则，让AI调用得起来；AI评估与信任管理，把准确性、合规、版权、偏见、追溯和责任串起来；AI ROI度量，回答项目到底节省了多少时间、提升了多少转化、减少了多少返工；持续学习运营，把案例、提示模板、工作流模板和复盘机制存下来。六类任务是能力维度，不是组织架构里的六个坑位。

图10–1分成上下两层看。上面是个人能力路径：执行者Operator → 增强型员工 → 监督者Supervisor → 智能体老板Agent Boss → 架构师Architect。下面是Architect能力在企业里的六个任务维度。前者讲“人怎么升级”，后者讲“工作内容怎么分层”。

单个员工会用AI，带来的通常是局部提效。一个组织能不能把 AI 用成结构性优势，关键看它能否系统配置人和智能体。微软提出过human−agent ratio，也就是人与智能体的配比。报销审批、例行报表、素材初筛可以配置更多智能体；品牌危机、核心客户谈判、复杂创新项目必须保留足够的人类判断。一旦这个比例成为管理变量，企业看团队能力就不能只看人数，而要看它能稳定调度多少高质量智能体，以及背后有没有可靠的知识环境和评估体系。

持续学习会成为这种组织的底层能力。AI工具变化太快，真正可持续的做法，是形成一个循环：发现新工具，试点新流程，度量效果，存下模板，推广复用，然后继续迭代。图10−2画的就是这个循环：业务目标进入任务拆解，任务被分配给人类判断和智能体执行，两边结果汇入质量评估；评估后整理为模板、数据、规则和案例，再反过来优化下一轮任务拆解。

换到行业里看，Architect能力并不抽象。制片人决定一部剧讲给谁、拍成什么调性；总编辑决定选题角度和事实核查边界；总承包项目经理兜住法规、工期、供应链和验收；系统集成总工把业务目标翻译成能跑的系统；活动总导演设计现场节奏并处理突发；行政总厨决定菜单定位和出品标准；主治医生统筹诊疗方案和风险；航空签派决定航班能否放行；游戏制作人权衡体验、留存和商业化；实验室PI，判断研究问题值不值得继续投入。这些岗位表面差异很大，底层做的是同一件事：定义意图、设定约束、判定证据。

所以，AI时代最难被替代的人，是能回答三个问题的那种：做不做，做成什么样，凭什么说做成了。 会执行具体步骤，反而排在后面。AI越强大，执行层效率越高，架构决策的影响半径反而越大。

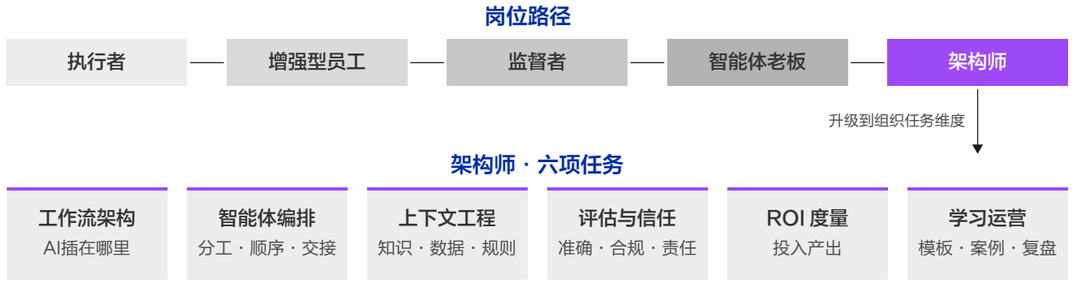


图10−2 · 从执行者到架构师，能力升级到组织任务维度

一个关键判断：AI原生教育必须培养“架构型人才”

上面的分析指向一个人才培养层面的核心判断：传统教育体系最薄弱的地方，恰恰是 AI 时代最需要的能力——定义问题、设定约束、判定证据。

传统教育训练的是"给定问题，找正确答案"。AI 时代需要的是"面对模糊情境，定义出好问题"。传统教育考核"执行的准确性"，AI 时代需要考核"判断的合理性"。这不是加一门 AI 课就能解决的，而是教育范式的根本转变。

结合前面十个行业的观察，提出五条可操作的策略：

策略一：在专业教育中嵌入“意图−约束−证据”训练框架。不管是医学院、建筑系、新闻学院还是计算机系，都应该增加一类作业：给学生一个模糊的业务场景，让他定义意图（要达成什么）、列出约束（什么不能碰）、设计证据链（怎么证明做成了）。这比教学生写prompt重要一百倍。

策略二：用AI模拟真实业务场景，让学生提前练“管理智能体”。 让医学生用AI辅助做诊疗方案、让新闻系学生用AI做选题策划和事实核查、让工程系学生用AI做系统集成方案。重点不是考他们用了多好的prompt，而是考他们能不能发现AI输出里的错误、能不能在约束条件下做取舍。

策略三：建立“人机协作项目制”教学。 毕业设计和课程项目里，明确要求学生组建“人+AI团队”：哪些环节交给AI，哪些环节必须人来判断，出了质量问题怎么追溯。这直接训练未来的Agent Boss和Architect能力。

策略四：跨学科培养“领域+AI”双能力人才。未来最值钱的人，是“懂医疗+会调度AI做辅助诊断的医生”、“懂建筑+会用AI做方案比选的项目经理”、“懂金融+会用AI做风控的合规官”，纯AI工程师反而排在后面。教育体系需要打通专业壁垒，让AI能力嵌入各行各业的专业训练中。

策略五：建立持续学习认证体系，替代一次性学历信号。 AI工具三个月一迭代，四年制学历教的东西毕业时可能已经过时。企业和教育机构需要联合建立“能力微认证”体系：看的是你最近半年掌握了哪些新的人机协作能力、完成了哪些实战项目，毕业于哪所大学退居其次。这更接近飞行员的持续资质维护模式。

一句话总结这部分的核心洞察：AI原生教育要培养的是Architect，也就是能定义问题、驾驭工具、承担后果的人。培养更熟练的AI用户，不是它的目标。这是从Supervisor到Architect跃迁在人才培养端的映射。

2026年AI新职业角色图谱的主线可以压成一句话：执行能力正在变得便宜，架构能力正在变得昂贵。谁能定义问题、组织工具、设计系统、承担后果，谁就更可能站在人机共生组织的中心。

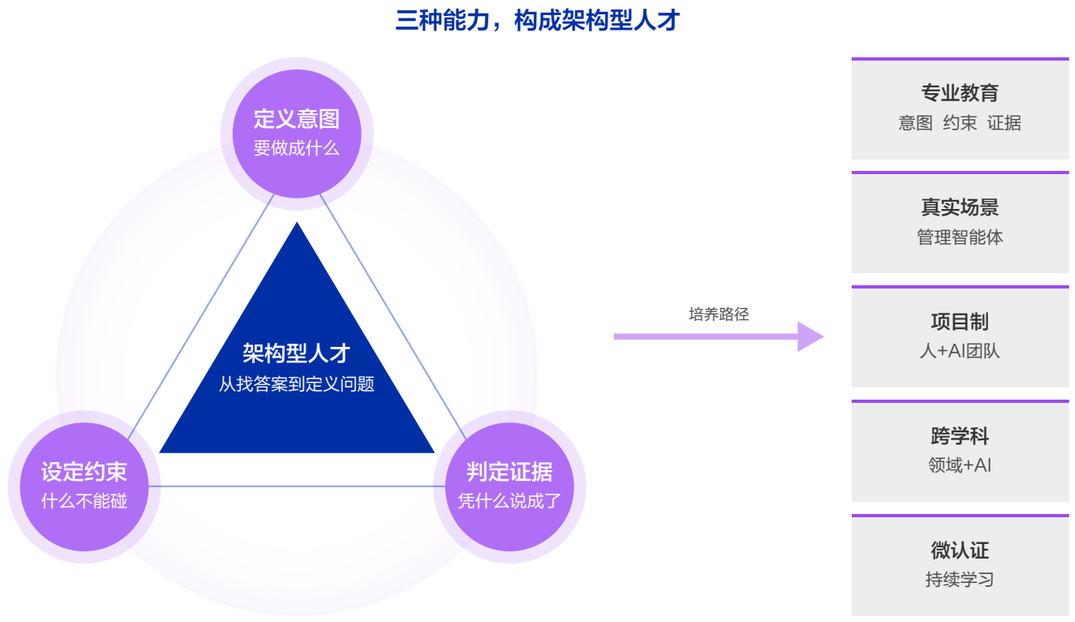


图10−3 · AI原生教育，培养架构型人才