# 2026 十大AI技术趋势

BAAI
智源研究院

# 卷首语

星河轮转，万物生长。站在 2025 与 2026 年的交界回望，人工智能的发展正从一场狂飙突进的参数竞赛，沉淀为对物理世界本质的深刻重构。如果说昨日的喧嚣，是我们为大模型掌握人类语言图形天赋的惊叹；那么此刻的静水流深，则是我们见证机器智能穿越认知的肃穆——它正在冲破静态字符与像素模仿的藩篱，向着物理世界的底层秩序与运行逻辑进军。

“Next-State Prediction”范式的出现，赋能模型如人类一般，捕捉光影流转背后的因果之律，洞察时空演化的内在逻辑。这种对于现实世界的模拟，亦为合成数据注入严谨的物理内核，相互增益，驱动模型向更高阶迭代进化。

随着能力的升维，智能的形态正从单体走向集群，从数字迈向现实。作为解决复杂任务和智能应用的最佳载体，Agent 技术范式迅速向多智能体系统（MAS）收敛，异构 Agent 则通过统一的通信协议结成一张去中心化的协作网；与此同时，具身智能迎来技术与商业的双重升级，世界模型和大小脑协同发展，推动智能体向科研、工业等深层场景渗透；在基础学科领域，AI For Science 完成了角色的转变，从被动的数据分析工具，升级为独立提出假设、设计实验、并在自动化闭环中寻找答案的探索者；在前沿交叉方向，AI 与量子、类脑等学科的深度融合，也在一次次实践中验证着性能跃迁的可能。

支撑这一切的，是坚实的算力基座与日益成熟的产业生态。万卡至十万卡集群部署，已成为下一代模型训练的标配；全栈进化的开源 AIOS 正试图打破单一硬件的垄断，筑起兼容并包的普惠地基。应用端，超级应用（Super App）基于“All in One”的设计，承载智能时代下 to C 市场崛起的厚望；B 端场景则正在突破概念验证（POC）的瓶颈。随着数据质量的完善，AI 将真正化身为驱动实体经济运转的精密齿轮。另一方面，面对参数激增带来的“黑箱”风险，构建一套可审计、可回溯、具备强对抗能力的安全防御体系，已成为通往通用人工智能必须解开的关键命题 ……

风起于青萍之末，浪成于微澜之间。我们试图透过技术演进的轨迹，解析未来的光谱，在不确定性的迷雾中，寻找行业可能的锚点。求索之旅的终局并非冰冷的算法堆叠，而是人类意志向未知疆域的延伸拓展。我们会怀揣着理性的敬畏与探索的勇气，见证人类文明与机器智能彼此激荡、交相辉映的宏大协奏。

# 目录

# 世界模型成为AGI共识方向，Next-State Prediction或成新范式

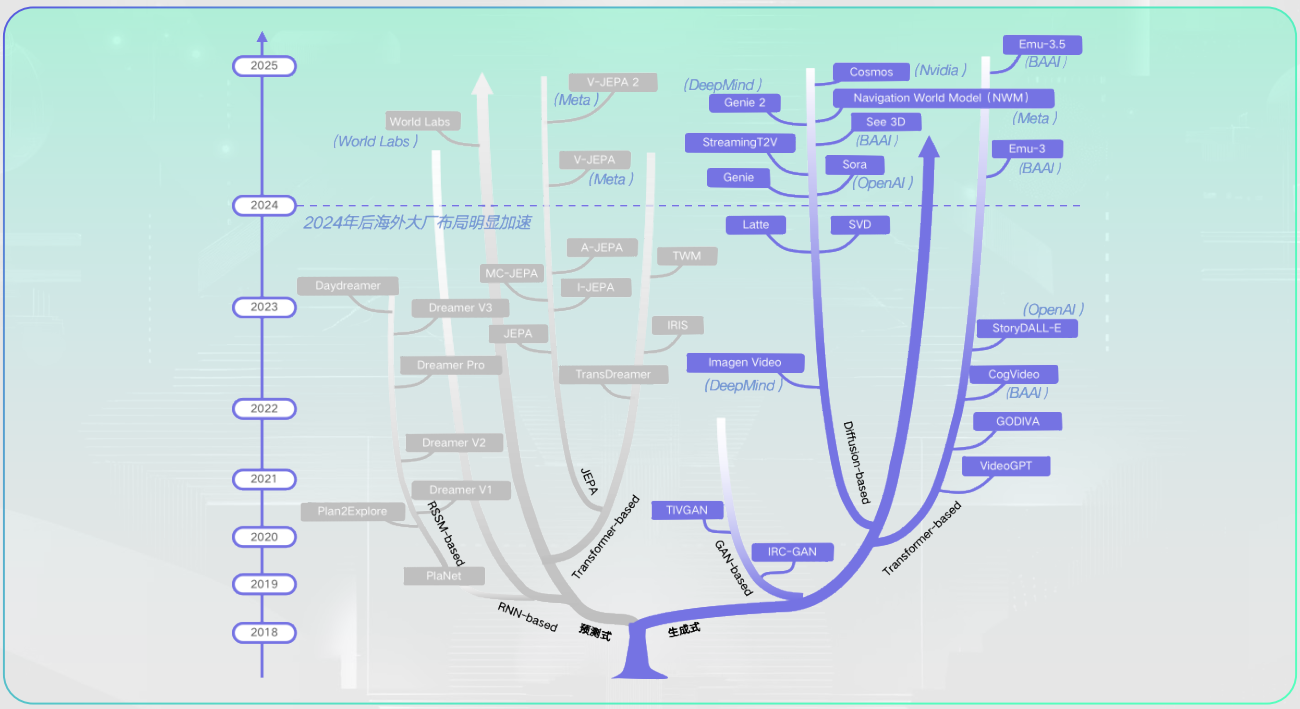
基础模型的演进，本质上是一场机器向着人类认知极限逼近的攀登，其核心始终锚定于对物理世界的极致模拟与深刻洞察。

自 2024 年末以来，业界对于基础模型的训练逐步收敛到“预训练 + 后训练”范式，OpenAI、Google、xAI 等发布的系列模型反复验证了两阶段 Scaling Law(规模定律)的有效性。

而在多模态大模型领域，主流模型架构如 DiT（Diffusion Transformer）或基于 CLIP 的拼接模式虽在特定任务中表现优异，但由于多模态表征间的割裂，始终无法复现 LLM 那样平滑的 Scaling 曲线。在此背景下，回归第一性原理，实现多模态数据的前融合，或成突破泛化瓶颈的重要解法。

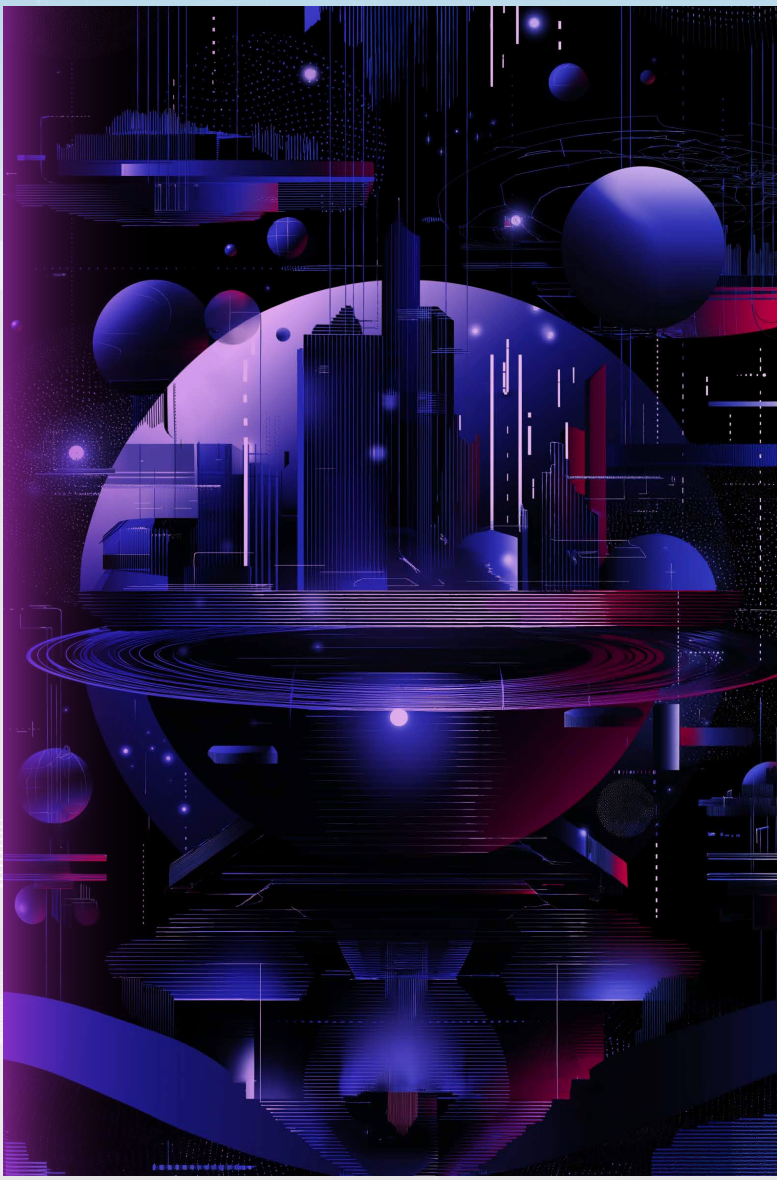
更为关键的变革在于认知维度的升维：从“Next Token Prediction”向“Next-State Prediction”（NSP）跨越。前者在面对真实物理环境时往往缺乏对因果律的把控，而 NSP 范式不仅是生成像素，更是像人类一样，从多模态数据中自主学习世界的动态规律，例如物理动态、时空连续性、因果关系。对于复杂任务，能够将高层意图转化为可执行的多步行动路径，实现“理解——预测——规划”的完整能力，这正是 AI 从“感知”进化为“认知”的核心标志。

2025 年，海外头部模型厂商正加速验证具备原生架构和物理模拟能力的多模态世界模型。如 World Labs 于 2025 年推出实时生成式世界模型 RTFM，可从单幅图像创建 3D 空间；OpenAI 发布 Sora 2，展现出理解并模拟真实世界规律的能力。




来源：智源行研组绘制

国内方面，相较于海外主流的 DiT 路线，北京智源人工智能研究院作为 NSP 范式的开创者，在自回归架构上走出独特的道路。智源在 2025 年 10 月发布的悟界·Emu3.5，将多模态数据等统一编码为离散 Token 进行自回归训练，执行 NSP 任务，让模型获得了对物理世界动态与因果关系的理解能力，标志着世界模型的发展进入了一个新阶段。同时，其发布的 OmniGen2、RoboBrain 等系列模型，也在向着 NSP 方向持续进化；蚂蚁发布百灵大模型体系，包含语言模型 Ling 系列，多模态大模型 Ming 系列和思考模型 Ring 系列等。Ming 系列在生成 / 编辑 / 分割一体化、方言识别方面有技术优势。Ring 系列首创“棒冰算法（icepop）”，解决了思考模型强化学习的训推差异导致训练崩溃问题，在综合任务方面逼近 GPT-5 水平。



**王仲远**

北京智源人工智能研究院 院长

**Emu3.5 的意义在于，（它）可能开创了一个新的大模型的赛道。虽然业内对世界模型有很多的讨论，但人们最初的讨论是，我们的人类大脑里面应该存在着一个这样的世界模型，它能够理解基本的世界运行规律，包括物理常识、时间空间知识等，能够帮助我们解决日常生活中看似非常简单、但现在对于机器人来讲非常困难的问题。**

# 具身智能迎来行业“出清”，产业应用迈入广泛工业场景

如果将 2024 年定义为具身智能的“百机大战”元年，并将其主要特征归纳为资本涌入与 Demo 层面的技术展示；那么伫立于 2025 岁末回望，行业已进入到下一阶段，企业“出清”时点逼近，产业应用迈入广泛工业场景。

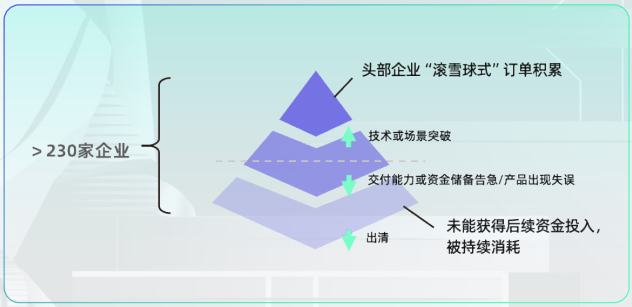
## 2.1 行业出清：业务模式同质化，行业格局初步形成

当前，我国具身智能企业数量已超 230 家，其中人形机器人企业超过 100 家，规模已可媲美移动互联网时代的“百团大战”。

然而，本轮人形机器人的技术难度、初始轮次的资金需求均远超百团大战，而资本环境则因全球经济下行等多方面原因，较“百团大战”时期更为艰难。当前的企业数量远超赛道的物理承载量与资本供给能力，行业或将在不久后完成一轮洗牌。

同时，具身创业公司业务模式同质化严重，普遍采用“通用开源大模型 + 运动控制”在单场景下进行操作优化。借势于基础模型的能力涌现，这种模式在长程任务的推理规划方向取得了良好效果，但后续的进化仍绑定于上游基础模型的迭代节奏。此外，具身小脑旨在传统运控算法之上，通过提升操作侧的泛化能力来应对特定场景下的非标问题。但整体来看，这类方法通常在处理分布外（OOD）场景时频现误差，大大降低了其在真实场景下落地的实用性。

不过，引入世界模型，并在仿真或现实中通过强化学习机制进行自我修正，这种具备自我进化能力的闭环模式，已初步在自动驾驶领域得到验证，后续或将成为具身智能迈向下一阶段的重要技术锚点。




来源：智源行研组绘制

## 2.2 商业化进程加速：从“实验室验证”迈向“量产交付”

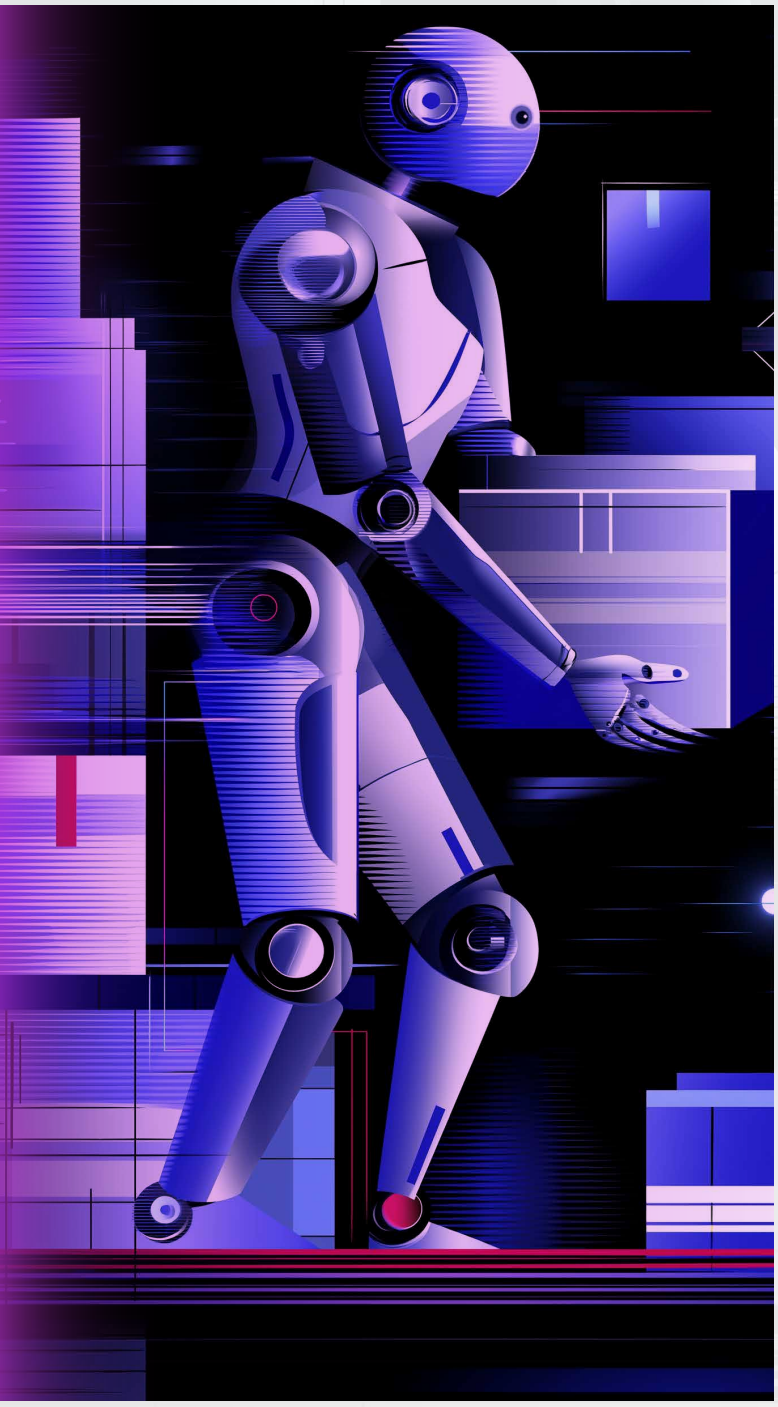
商业进程上，行业逻辑已从技术愿景转向量产销售。相较 2024 年，2025 年的客户主力从高校研究机构向 B 端产业场景迁移，标志着人形机器人从实验室向真实可用迈进。与商业化进展相呼应的 IPO 进程也在紧锣密鼓地进行，除智元机器人已曲线上市之外，乐聚智能已启动辅导备案，银河通用和云深处科技也已完成股改，为实现证券化提供了必要条件。预计未来一年内，具身智能企业上市的钟声将频繁回响。

国内外具身企业在这一年密集发布模型成果并收获商业订单。

海外方面，Physical Intelligence 发布π*0.6 模型，基于 Recap 方法利用自主经验训练，显著降低复杂任务的失败率，达到实际应用所需的鲁棒性水平；Tesla Robotics 发布 Optimus 2.5 人形机器人，旗下机器人已应用于工厂生产、农场运营等主要场景。

国内方面，北京智源人工智能研究院发布通用具身大脑 RoboBrain2.0 及通用小脑基座 RoboBrain-X0。前者可实现跨场景多任务轻量

化快速部署与跨本体协作，后者则能驱动多类型本体完成动作执行；蚂蚁集团旗下具身企业灵波科技自研具身基础大模型，并推出首款单场景服务机器人 Robbyant-R1，目前已在餐饮、导览、医疗问答等生活服务领域投入应用。此外，业内已出现多笔亿级订单，人形机器人销量已突破万台，跨入初步商业化阶段。



## Chelsea Finn

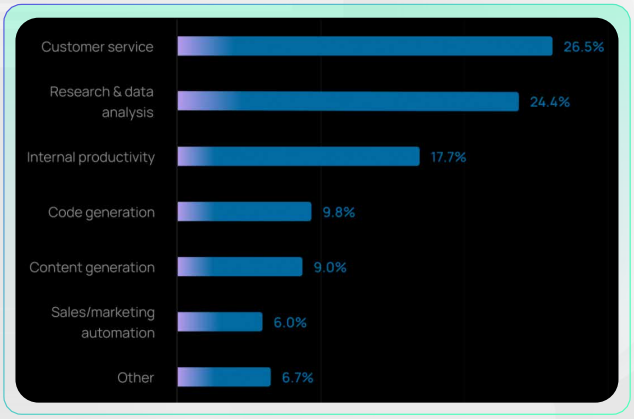
Physical Intelligence首席执行官

**通用机器人策略只是起点，真正的物理智能来自机器人在真实环境中的长期交互和自我改进——模型要在部署后不断从新的操作经验中学习，学会在干扰下恢复、在失败中调整策略，而不是停留在实验室里一次性训练完。**

# 多智能体系统决定应用上限，Agent时代的“TCP/IP”初具雏形

## 3.1 从 SAS 向 MAS 应用演进，Agent 应用天花板取决于 MAS 成熟度

目前，多数企业在智能体应用上仍是以简单任务为主的单智能体系统（SAS）。根据 Langchain 发布的《State of Agent Engineering》，客服、代码生成、内容生成为代表的 SAS 应用占比达 63%，MAS 应用为主的研究和数据分析、内部生产力占比则为 42.1%。考虑到多数 Agent 应用仍处于 Pilot 阶段，MAS 落地难度大于 SAS，MAS 的实际应用比例较该比例更不乐观。




来源：Langchain，智源行研组整理

然而，随着企业级应用向复杂场景渗透，多智能体应用体现出成为更优选择的潜力。复杂场景下，单智能体系统（SAS）在上下文遗忘和角色混淆方面的问题凸显，而多智能体系统所具备的对工作流的高拟合，以及通过自我反思、互相辩论以降低幻觉的能力，对于 AI 向多领域的应用扩展不可或缺。

**从行业应用来看，Agent 呈现出向复杂工作流为代表的多智能体应用演进的趋势。**

Anthropic、OpenAI 近期发布的报告也体现出了该现象。根据 Anthropic 数据，57% 的组织现在部署智能体来处理多阶段工作流，预计 2026 年将有 81% 的组织计划应对更复杂的用例——其中 39% 计划开发用于多步骤流程的智能体，29% 计划将其部署于跨职能项目。OpenAI 的调研数据也是如此，Custom GPTs（定制 GPT）的使用量今年迄今增长了 19 倍，API 推理 Token 消耗激增 320 倍，也表明更智能的模型正在被系统性集成。

**从认知科学的视角而言，群智理论与分布式认知理论都指出，一个群体的认知能力总和，通常会超越该群体中最聪明的个体。**

其中，斯科特·佩奇的多样性预测定理和孔多塞陪审团定理以数理推导验证了多智能体系统在复杂任务上相对单智能体系统的明显优势。

### 图 MAS 的认知科学理论基础

斯科特·佩奇多样性预测定理：群智理论基石，《多样性红利》

$$(c-V)^2=\frac{1}{n}\sum_{i=1}^{n}(s_i-V)^2-\frac{1}{n}\sum_{i=1}^{n}(s_i-c)^2$$

群体误差 = 个体平均误差 - 预测多样性

推论：

群体的智慧不仅仅取决于个体的能力（即减少个体平均误差），同等程度上取决于多样性（Diversity，即个体预测值与群体平均值的差异）。只要多样性足够大，群体误差就可以显著低于群体中最聪明个体的误差。

孔多塞陪审团定理：分布式认知在离散决策（如分类、是非判断）上的数学保障

假设有 n 个独立的智能体（Jurors），每个智能体做出正确判断的概率为 p。若 p > 0.5 （即个体比随机猜测稍好），且个体间相互独立，则群体通过多数投票 （Majority Vote） 做出正确决策的概率 $P_n$ 随着 n 的增加而增加，且极限为：

$$\lim_{n\to\infty} P_n = 1$$

推论：

在 MAS 中， 即便单个 Agent 只有 60% 准确率，只要足够多且犯错方向不一致， 通过投票机制集成后的系统准确率可以迅速逼近 100%。

值得注意的是，该理论生效的前提是智能体之间具备高度的独立性（Independence）和多样性（Diversity）。若 AI 模型间存在同质化（如同源训练数据导致的系统性偏见），群体不仅无法突破上限，反而可能因“模式坍塌（Mode Collapse）”或“信息级联（Information Cascades）” 导致错误放大。因此，该结论在理论上成立，但在工程实践中取决于系统架构对“多样性”的维持能力。

## 3.2 MAS 范式趋于明确， MCP/A2A 为代表的 MAS “Narrow Waist” 协议层趋于融合统一

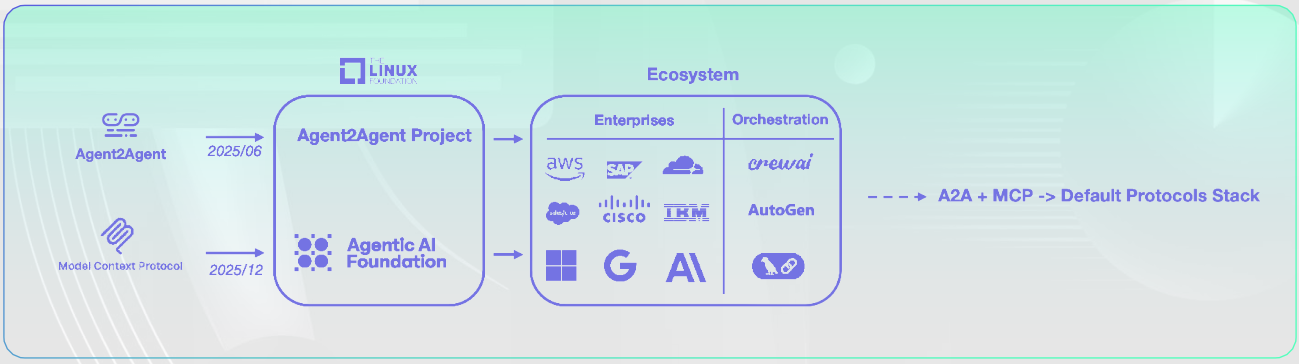
**MAS Paradigm = Protocols [ Orchestration * ( Cognitive Core + Tools ) ] @ Sandbox**

如上，从目前业界实践来看，MAS 范式已开始收敛。编排框架、沙盒、可观测性成为产业应用的必选项。比如 Langchain、AutoGen、CrewAI 等框架成为主流编排选择，沙盒、可观测性工具作为产业应用的安全选项也已是行业共识。

考虑到 MAS 的核心逻辑是从“单体智能”走向“群体智能”，该范式转移成功的关键不仅在于单体模型性能强弱，更在于 Agent 连接协作的效率和规模。

在 2026 年，Agent 协议栈的进展决定了 MAS 的下限，而 Agent Skills 等推高了 MAS 应用上限。

图 A2A/MCP 分别被捐赠 Linux 基金会，Agent 通信层的形成




来源：智源行研组绘制

通信协议正处在从碎片化走向融合统一的演进过程中。其中，融合统一趋势明显的为以 MCP、A2A 为代表的通信层协议，仍处于早期探索的则是 UI、支付、语义对齐、发现、安全、可观测性等其他协议层。

通信协议正处在从碎片化走向融合统一的演进过程中。其中，融合统一趋势明显的为以 MCP、A2A 为代表的通信层协议，仍处于早期探索的则是 UI、支付、语义对齐、发现、安全、可观测性等其他协议层。

**MCP 和 A2A 为代表的通信层协议，在 2025 年在分别被捐赠给 Linux 基金会后实现了分层融合。**

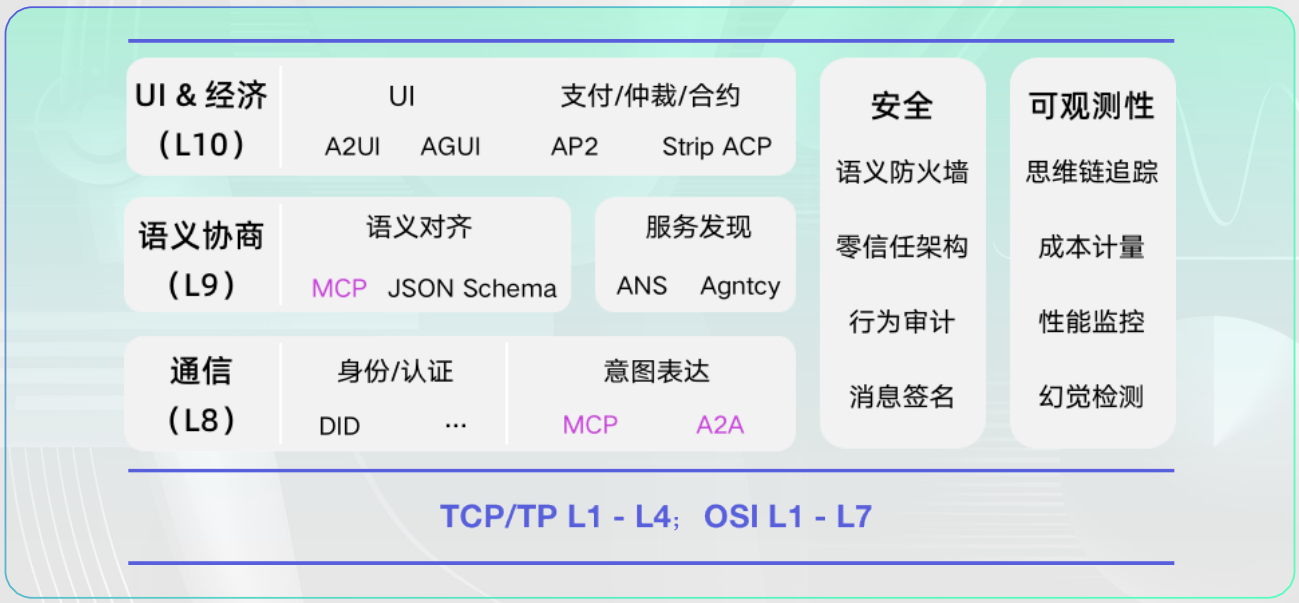
2025 年 6 月，Agent2Agent 被 Google 捐赠给了 Linux 基金会独立项目“Agent2Agent Project”。

2025 年 12 月，MCP 被 Anthropic 作为三项成果之一，捐赠给了 Linux 基金会子基金 AAIF。

至此，作为 Linux 基金会统一管理的 MCP 和 A2A 两项 Agent 通信协议，以其中立属性，成为 Microsoft、Google、Anthropic、AWS、IBM 等头部厂商和 LangChain、AutoGen、CrewAI 等主流框架的原生支持选择。尤其值得一提的是，作为 ACP 协议的发起方 IBM，目前也显示出将 ACP 并入 A2A 的迹象。

与此同时，尽管“MCP+A2A”解决了“Agent 怎么用工具、怎么找到其他 Agent”的问题，但 Agent 沟通内容的概念对齐、身份认证、人机交互（UI）、支付、安全和可观测性等更多协议层仍处于早期探索阶段。从国内外进展来看，Google、Cisco、W3C、Langchain 等厂商和社区在各协议层的尝试频繁，比如 A2UI、AP2、Agntcy 等协议均已进入生产场景实践阶段，2026 年将持续繁荣成熟。

**图 Agent 协议栈的建构和进展**



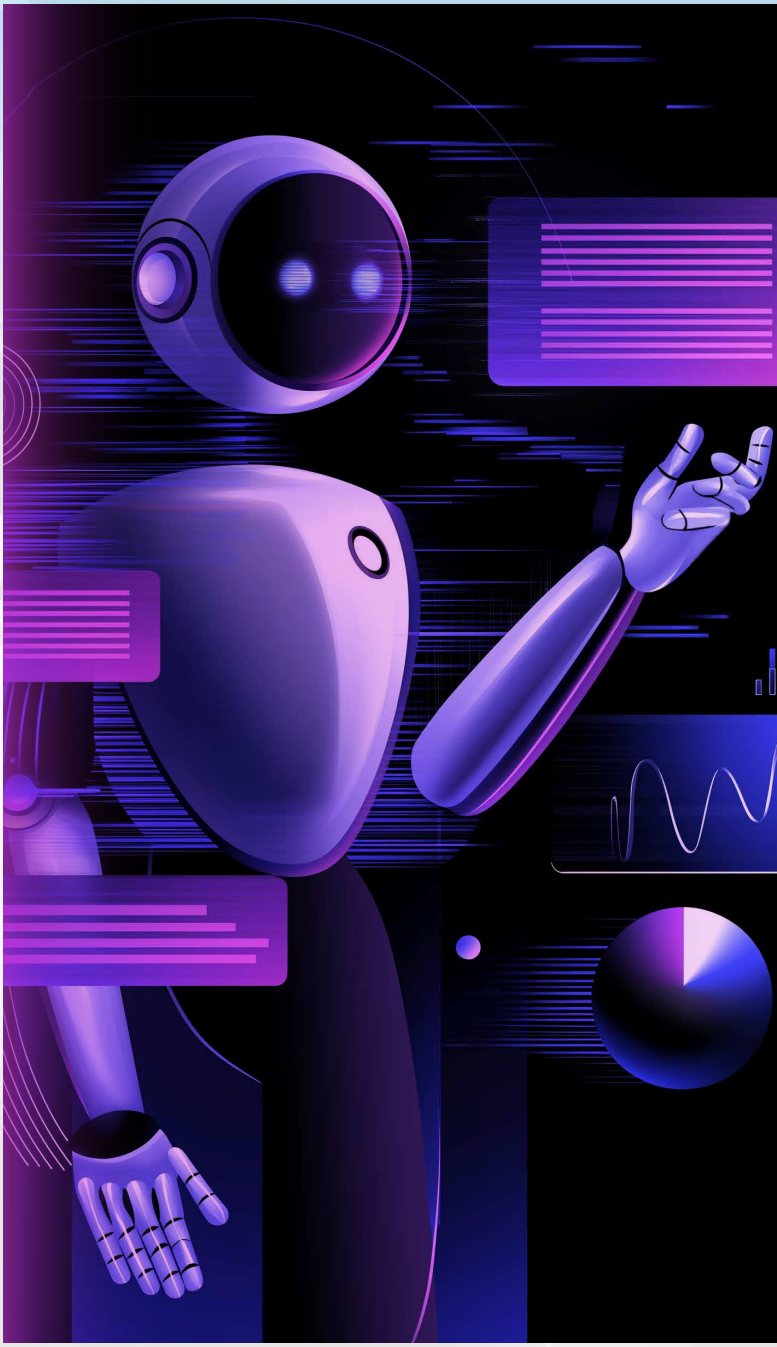

来源：智源行研组绘制

在计算机科学历史上，大规模系统的收敛都依赖于一个“窄腰（Narrow Waist）”协议层。比如，互联网收敛于 IP 协议、Web 收敛于 HTTP 协议；前者通过屏蔽底层物理网络差异、实现对上层应用的支撑，后者则通过屏蔽服务器差异、支撑起浏览器生态。

对于决定 Agent 应用上限的 MAS 而言，A2A、

MCP 等 Agent 通信层协议正是 Agent 应用的“Narrow Waist”协议层，向下屏蔽模大模型差异，向上支撑复杂的 MAS 应用。

现阶段，MAS 范式的收敛，Agent 通信协议的标准化至关重要。



**Satya Nadella**

Microsoft 首席执行官

**我们的业务，今天是终端用户工具业务，但本质上将成为支持 Agents 的基础设施业务 …… 至少从早期迹象上，看待每个用户业务的方式或许不仅仅是每个用户，而是每个 Agent。**

# AI Scientist成为AI4S北极星，国产科学基础模型悄然孕育

## 4.1 AI Scientist 热度骤起，工业落地尚有距离

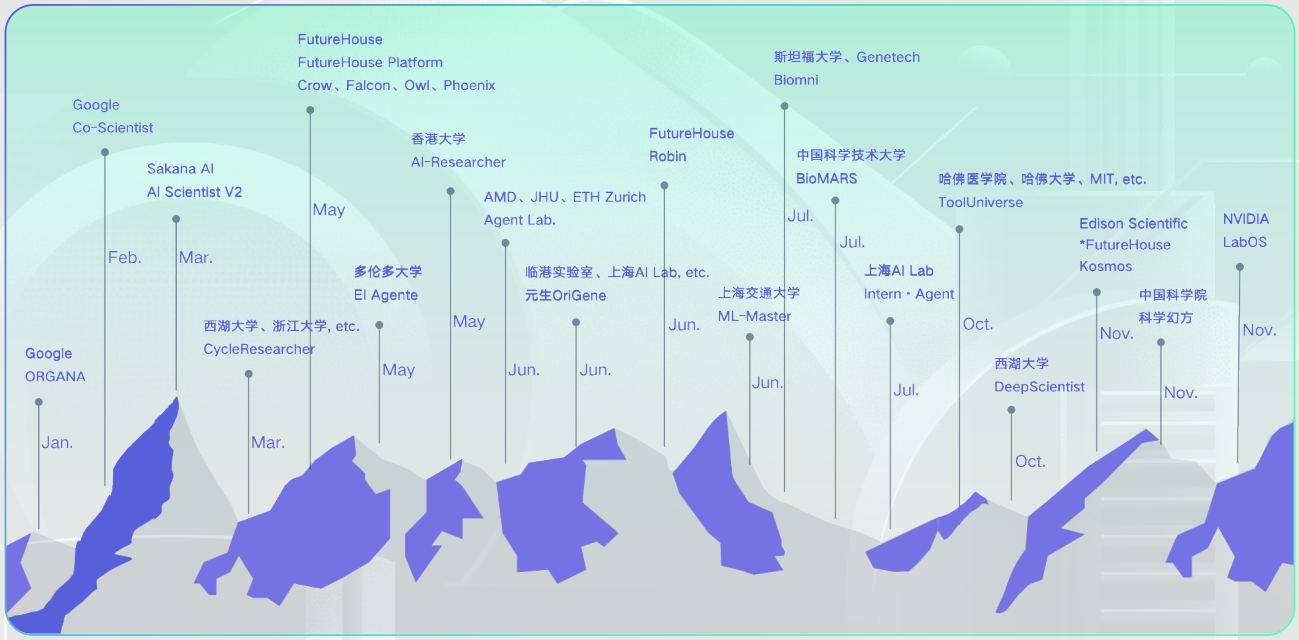
在经历了大模型辅助科学研究的初步探索后，2025 年的 AI for Science（AI4S）领域迎来了一次决定性的范式演进，其核心标志是从 CoPilot 到 AI Scientist 的身份跨越。

这一趋势的聚焦点，是 AI 科学家（AI Scientist）的兴起——能够模拟乃至自主执行“假设提出、实验设计、数据分析、结论推断”完整科研链路的智能体系统。这不仅是科研效率的量变，更是科学发现模式的质变，预示着一个由 AI 驱动的、研发效率呈指数级增长的新时代或将到来。

这场深刻变革的背后，是三大技术引擎的融合驱动。首先，科学基础模型（Science Foundation Models）的成熟为 AI 进一步深入各细分科学领域奠定了深厚的理论基础。这些模型得益于基础模型的跃迁式提升，又进一步超越了传统的语言和图像模态，通过对海量科学文献、分子结构、基因序列、观测数据、演化模拟数据及实验数据的训练，形成了对特定科学领域深层规律的内在理解。在此基础上，代理工作流（Agentic Workflow）将这种理解能力转化为行动力，使 AI 系统能够像人类科学家一样，自主规划并执行复杂的多步骤研究任务，灵活调用数据库、模拟器和文献检索等外部工具。最后，AI 的认知能力正通过与自动化实验室的连接，延伸至物理世界，从数字模拟到真实实验验证闭环的雏形得以显现，自主科学发现成为可能。

**图 2025 年以来全球各组织发布的 AI Scientist 系统**

*（不完全统计，高度仅供排版需要，不代表比较意义）*




来源：智源行研组绘制

## 4.2 美 AI4S 国家级别行动方案发布，我国科学基础模型攻坚刻不容缓

AI for Science 正被纳入举国体制。2025 年 11 月 24 日，美国白宫发布行政命令，正式启动名为“创世纪计划（Genesis Mission）”的科研举措，旨在以 AI 技术加速科研进程，继而全面提升美国人工智能发展能力。

循着创世纪计划的线索抽丝剥茧，其长期布局的路径就逐渐浮出水面。总体而言，该计划并非一项突发性的短期政策，而是美国能源部（DOE）与其科学与技术核心团队历经五年以上系统性筹备的产物。

早在 2019 至 2020 年间，以技术总工程师 Rick Stevens 为核心的团队便确立了“AI for Science”的战略框架，并以此争取预算，奠定了该计划的理论与实践基石。此后五年间，通过组建万亿参数联盟（TPC）、开发万亿参数科学模型 AuroraGPT 以及提出 FASST（科学、安全与技术人工智能前沿）构想，美国实际上已完成了从单纯的算力堆叠向“数据 + 算力 + 科学模型 - 自动化应用”全链路的预演。

因此，“创世纪计划”并非从 0 到 1 的探索，而是依托已经运转成熟的存量资产。基于对 17 个国家实验室数据的使用权限、APST（The Assistant to the President for Science and Technology，总统科技助理）协调机制打通的跨机构权限、AuroraGPT 大模型运营两年的成熟经验，带头人之一 Chris Wright 对小型模块化核反应堆（SMR）的审批权，以及 Chris Wright、Dario Gil、Rick Stevens 等主要负责人的长期合作，“创世纪计划”得以全面整合和动员能源、数据、模型资源，并利用成熟的组织架构实现科学智能的规模化落地。


来源：智源行研组绘制

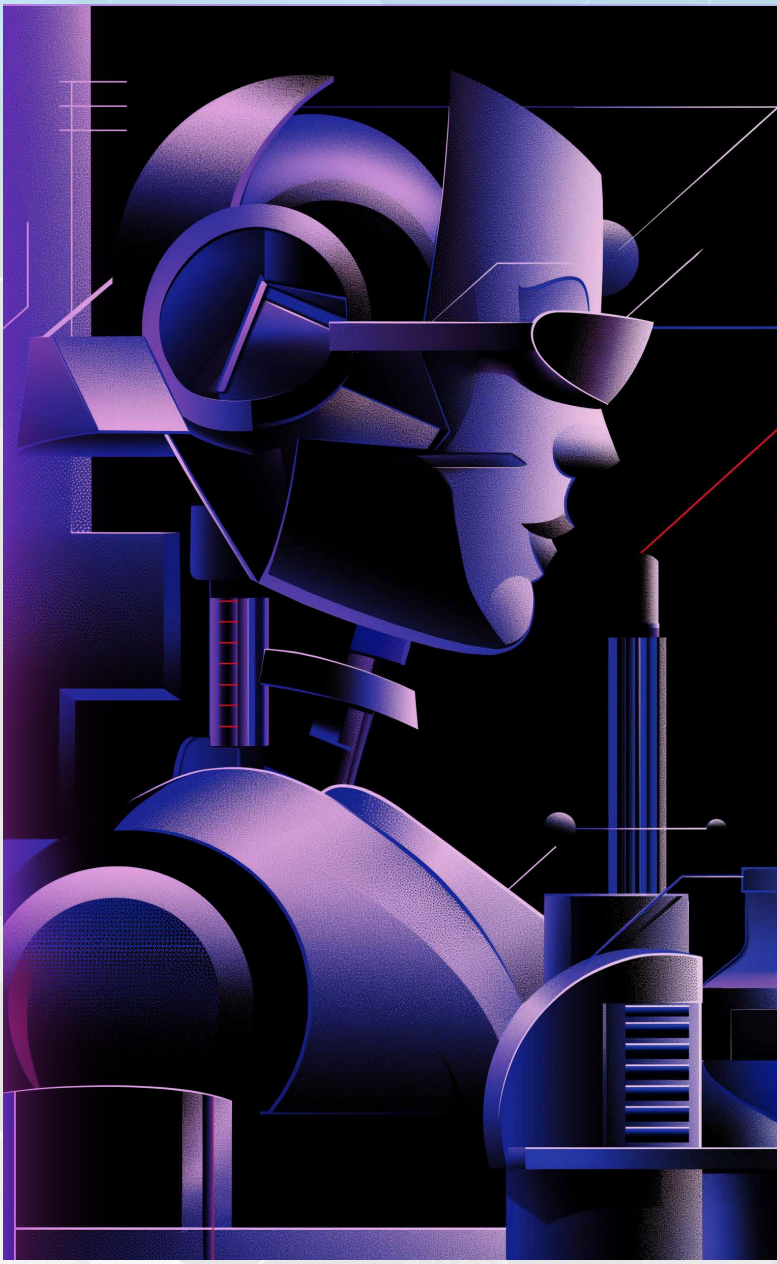
相较于美国“创世纪计划”五年前便开始布局的系统性动员，我国目前在科学智能（AI4S）领域尚未形成同等量级的全栈式响应，呈现出“应用强、基座弱”的非均衡特征。

在应用层面，得益于庞大的 STEM 人才供给与完备的产业链条，我国占据了相对优势，且执行情况良好。然而，在支撑科学智能落地的算力、数据与模型三大基础设施维度，我国则面临不同程度的挑战与追赶压力。

在算力端，受制于海外出口管制，AI 算力缺口明显，虽然国产替代进程正在加速，但整体储备仍不及美国；

在数据端，面对存量相对较少且治理能力偏弱的现状，我国正有序推进打破部门壁垒。早在 2014 年，科技部就发布《科技基础性工作专项项目科学数据汇交管理办法（试行）》，按下我国科学数据收集、治理的启动键，此后，我国逐步建立起以中科院带头、教育部、交通运输部等下属研究所 40 家共建单位参与的国家科技资源共享服务平台——国家基础学科公共科学数据中心，汇集管理物理、化学、材料等基础学科领域，以及青海湖等典型区域长期科研活动积累的科学数据。截至 2025 年，我国国家基础数据中心保有数据量已达 4.6PB，成为当下我国开展技术赋能科研探索的宝贵素材；

最为严峻的挑战集中在模型层，当前，我国在科学基础模型上的进展滞后，成为制约我国 AI4S 领域快速缩小与美国差距的核心卡点。同时，相比于与庞大的药械市场挂钩的 AI 辅助药物设计赛道，以及小步快跑、正在实现部分规模化生产的具身智能赛道，AI4S 在资本的候选池中仍然坐着冷板凳，依靠初创企业打造基础模型的设想更为艰难。因此，我国的科学基础模型研发仍亟待各方资源倾斜和整合。众志成城，坚若磐石；虽路远兮，同行则至。



**Rick Stevens**

“创世纪计划”核心科学家、
美国阿贡国家实验室副主任

**像所有变革性技术一样，人工智能的广泛应用既带来了巨大的机遇，也伴随着风险。我认为，美国必须在用于科学和国家安全应用的先进人工智能系统开发方面引领世界。为了实现这一目标，我们需要付出如同“曼哈顿计划”规模的努力。**

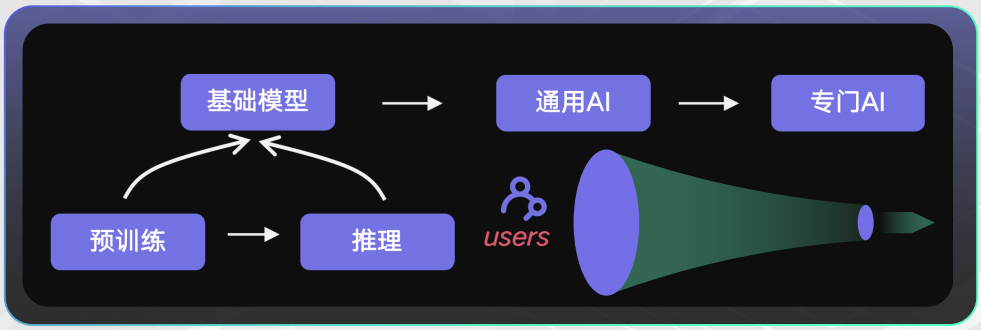
*数据来源：Rick Stevens在2023年美国参议院能源与自然资源委员听证会证词

# AI时代的“新BAT”趋于明确，垂直赛道仍有高盈利玩法

过去的一年，AI 应用加速迭代。基础模型能力的显著跃升，极大拓宽了任务执行的边界；推理侧架构的优化，带来服务成本的持续下降；技术供给侧的质变，为 AI 应用的落地积蓄了关键势能。

## 5.1 超级应用：“多行业 API 接入 + 基础模型”，相较 CUA 更有优势

当前 C 端 AI 应用的竞争目标已逐渐清晰，核心在于对“Super App（超级应用）”的攻略。其典型特征呈现为“All in One”的功能设计，即不再局限于单一工具属性，而是基于高性能基础模型直接产品化，通过一个入口实现从信息获取、任务规划到问题解决的闭环。




来源：智源行研组绘制

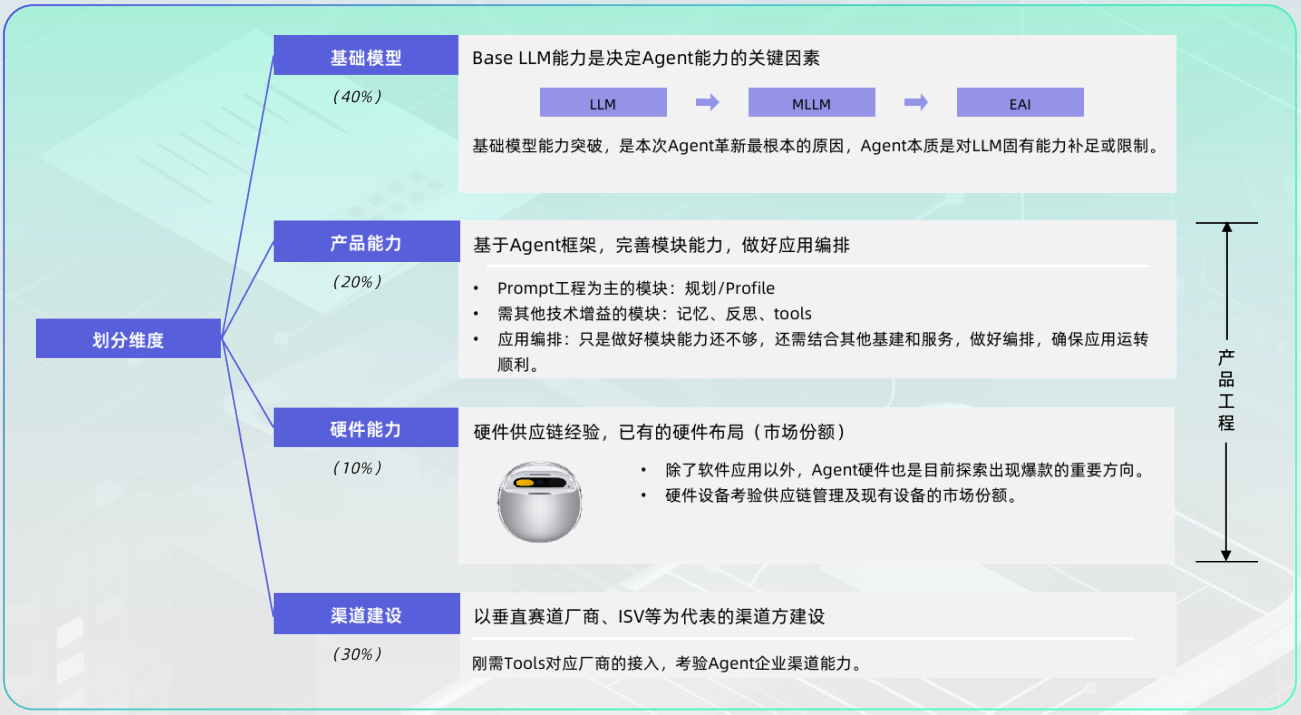
海外厂商在此轮浪潮中率先引领了 AI 超级应用的探索，以 ChatGPT、Gemini 等海外头部模型为基础构建的 APP，已初步具备过亿日活（DAU）、高交互频次、高用户停留时长等传统意义上的超级应用必要条件，并在 AI 2.0 时代首次领跑国内，率先探索“All in One”的功能设计。

循着当前的线索抽丝剥茧，头部大厂的超级应用布局其实是一条明线。Google 早在 2023 年 9 月就通过 Extensions 打通了 Maps 和 Flights 的接口，迈出了统一入口的第一步；而到了 2025 年 11 月，随着 Gemini 正式取代 Google Maps 的原生语音助手，这种外挂开始内化为原生组织。

现在，用户无需跳出界面，即可在同一个对话流中完成从寻找餐厅、预订座位到修改路线的全套操作——在同一品牌根系的统筹下，垂直 App 之间的围墙被逐渐推倒。同样在 11 月，ChatGPT 上线 Buy it in chatgpt 功能，不仅与电商平台 Shopify、Esty 和支付平台 Stripe 系统级打通，还前置了基于用户偏好的选品、全平台比价、运费政策解读等自动执行的步骤，用户得以用一句聊天般的自然语言启动流程，直达最符合心意的商品，并在 ChatGPT 应用内完成下单操作；OpenAI 还引入了具备视觉能力的浏览器内核，接管浏览器执行权，试图打通超级应用的手 - 脑通路。

考虑到 AI 超级应用范式为基础模型直接产品化实现的用户截流聚集，不仅需要极高的算力成本支撑，更依赖庞大的存量用户进行模型数据的飞

轮迭代。在此背景下，国内 AI 超级应用的机会点同样集中在头部大厂——巨头企业拥有操作系统级的入口权限、发端于互联网时代的自有垂域工具、全栈化的技术积累，具备打造国民级 AI 入口的实力。这场关于基模能力、流量入口与生态闭环的博弈，可能会为互联网版图带来新的变化，催生出 AI 时代的“新 BAT”。




来源：智源行研组绘制

11 月，蚂蚁集团推出全模态通用 AI 助手灵光，上线 6 天总下载量突破 200 万，领跑全球 AI 产品的下载增速。其首批上线三大核心功能——灵光对话、灵光闪应用和灵光开眼，开创性地在移动端实现自然语言 30 秒生成小应用，支持摄像头作为直接交互入口，还支持 3D、音视频、图表，甚至动画和地图等全模态输出，是业内首个全代码生成多模态内容的 AI 助手；12 月，高德地图接入阿里千问，吹响了阿里系工具集结号，旨在以千问 App 作为中心“天体”，以自有的消费、支付等核心工具作为卫星，构建基于统一入口和家族应用的星系。而字节豆包持续丰富 C 端工具矩阵，结合抖音、今日头条等应用引流的生态优势，在全球 AI 应用中 MAU 位居第二，仅次于 ChatGPT（截至 2025 年 11 月）。

## 5.2 垂直模型：多模态等高 ROI 方向更易跑出优质玩家，垂直玩家跨赛道玩法或有亮点

尽管通用赛道巨头林立，但在部分高壁垒的垂直领域，依然存在突围的机遇。在大健康、教育等领域，垂直应用凭借对特定数据的深度训练和对行业 Know-how 的理解，展现出了差异化的竞争力。

10 月底，Google 公布第三季度财报，在 AI 业务驱动下，公司季度营收首次突破千亿美元，Hyerbolic Labs CTO Yuchen Jin 直言“Gemini App 月活已达 6.5 亿，这应当归功于 Nano Banana”。从业务指标看，以 Nano Banana（及升级后的 Pro 版本）为代表的多模态模型，其调用量往往不及文本模型，但这并不必然意味着商业价值更低。

当前，Nano Banana Pro 的单次调用价格高达文本模型的 70 至 120 倍（以图片生成单次价格约 0.14 美元，常规文本交互单次价格约 0.002 美元为计算依据）。这意味着，Nano Banana Pro 仅需占据整体调用量的 1.5% 左右，其收入贡献即可与占据 98.5% 流量的文本模型持平。多模态模型无需追求文本式的高频并发，而是通过锁定少数高付费意愿的关键场景，即可实现极佳的“模效比”，验证了垂直领域模型低频高价值的收入结构。

国内 AI 垂直应用的土壤同样肥沃。12 月，蚂蚁旗下 AI 健康应用 “蚂蚁阿福” 全面升级，新版 App 提供健康问答、健康拍照、健康档案、健康小目标等智能化和个性化的日常健康管理服务。当下，阿福月活（MAU）超 1500 万，已是第一大健康管理类 APP，在 QuestMobile 最新公布的周活榜单上，阿福位列垂类第一。

此外，字节跳动的即梦 AI、猫箱，MiniMax 的海螺 AI、星野，作业帮的快对 AI 分别在视频 / 图像生成赛道、虚拟陪伴赛道和教育赛道占领用户心智，有力地印证了：在细分赛道中，精准的场景切入与优质的产品体验，依然是垂直应用在通用产品包围下实现突围的关键。


**Sam Altman**

OpenAI首席执行官

**我们正在努力打造一个真正的 AI 超级助手。未来你不需要打开不同的 App 去订机票、聊天或购物，你只需要告诉你的 AI 你想要什么，它就会跨越所有服务为你搞定一切。**

# 产业应用滑向“幻灭低谷期”，2026H2迎来“V型”反转

2025年，AI在多数B端场景仍处于PoC阶段。多数ToB应用仍为“Chat”形式，客户服务、代码辅助、营销等场景应用成熟。至于复杂的自主决策Agent，则仍处于“示范应用”阶段。

## 6.1 ToB应用的“幻灭低谷期”：数据、MAS和成本

进入2026年，行业将迎来“幻灭低谷期(Trough of disilusionment)”。

根据MIT报告，通过对300个企业的AI项目的研究，发现95%的GenAI Pilot项目未能产生任何可衡量的影响。大多数项目在进入生产环境前就已“烂尾”。

根据Forrester、Gartner等机构预测，由于短期ROI不达标，企业可能会将25%的原定AI支出推迟到2027年。到2027年，40%的Agentic AI项目可能会失败。

究其原因，数据质量、MAS成熟度、成本过高乃至失控是主要原因。

关于数据质量，主要的问题是企业既有系统的集成难度。根据Anthropic数据，46%的企业将“现有系统集成”列为首要障碍。目前的AI应用，更多还是在“Chat”基础上，通过员工自己手动操作ERP/CRM。

关于MAS成熟度，如“Agent趋势篇”提及的MAS范式，在实践中仍存在MAS涌现行为不可控、调试困难，互操作性不足(通信和语义对齐)，可解释性差等问题。

关于成本，除了MAS本身运行的高成本以外，死循环通信、自我对话、上下文爆炸等带来的成本失控，是MAS应用一直以来的困扰。

目前业内典型案例，为2025年10月技术博客Towards AI / Substack提及的《We Spent $47,000 Running AI Agents in Production》：4个LangChain智能体组成的系统，旨在通过相互协作进行市场数据研究，因死循环通信导致11天损失47000美元。

## 6.2 ToB应用的MVP（最小可行性产品）路径和更多标杆用例的发生期间：2026H2

从业内实践来看，数据质量和既有系统集成是最大难题，安全则是必要条件，其次才是成本、MAS协议栈等其他问题。

基于此，要实现ToB的MVP应用，可参考的实现路径如下：

**前提：** Scaling Law持续生效，推理优化持续降本

**模式：** “Data Gov先行，OTel/MCP并行”

**实施：**

OTel（可观测）+ Data Gov（数据）+ Adapter-MCP（连接）

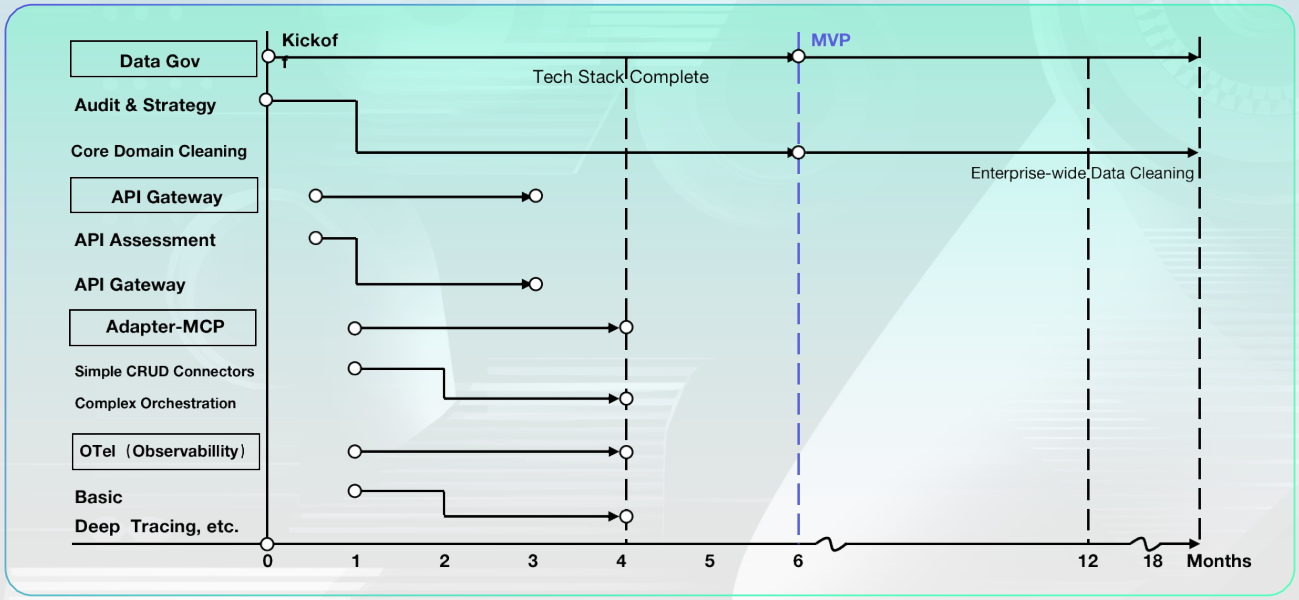
其中：

OTel：常用可观测性工具，基础集成2-3周，深度定制2-3月；

Data Gov：最为耗时，针对核心业务用时3-4月，全面治理则需12-18月；

Adapter-MCP：只要API接口确定，即可开始MCP Sever构建，简单连接2~3周，复杂连接8-12周。

以上三个模块高度解耦，可并行推进。基于此，我们尝试为 ToB 应用的 MVP 产品构建做简单时间测算如下：




来源：智源行研组绘制

如上，要做一款成功的 MVP ToB 产品，至少需投入 6 月左右的时间。

在 2025 年多数 Enterprise AI 试点项目失败、行业进入“幻灭期”的当前，如从 0 到 1 开始预估 ToB 行业应用迎来更多有效 Use Case 的时间点，将其设定为 2026H2 较为合适。

## 6.3 垂直行业的规模应用和行业共识性标准接口

至于 ToB 产业应用何时会迎来在具体行业的规模化落地，我们需要在以上实施路径基础上，综合考虑行业共识性标准接口的建设进展、ROI 和安全的 Trade-off。

关于行业共识性标准接口，其作用在于规范相应行业的数据标准，通过强制以 JSON 结构化数据交互，更好地提升行业互操作性。这对于 AI 大规模落地极为有利，通过 JSON Schema 等方式转换，便可实现 Agent 语义对齐，有利于采用 MAS 解决复杂问题。

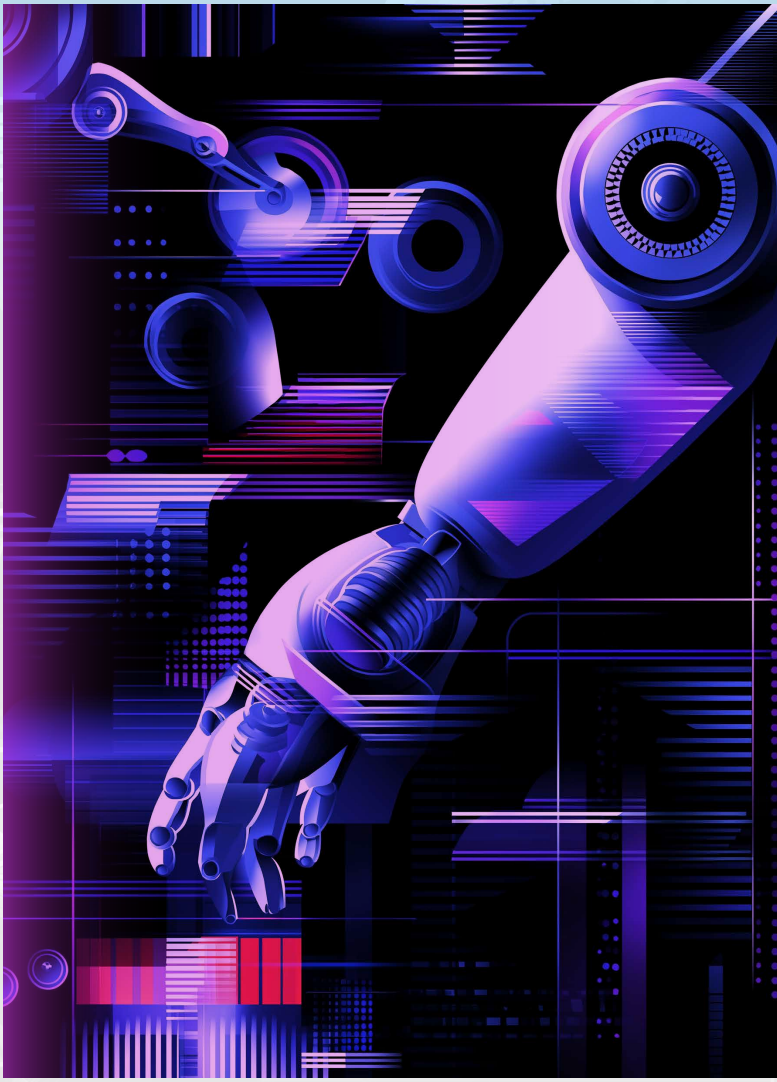
因此，我们可将垂直行业规模落地的工程实现路径改写如下：

> Otel + Data Gov + Adapter-MCP + Std. API（行业标准 API）

目前，已有主要 B 端行业在行业标准 API 上有实际进展。比如，美国政府针对医疗行业于 2024 年 1 月发布 CMS 互操作性新规（CMS-0057-FF），强制要求所有受管辖的支付方开放三大 API：事前授权、患者访问、医疗机构访问，并必须使用 HL7 FHIR 数据标准（JSON 格式）。该规则要求 2026 年 1 月实现部分功能，2027 年 1 月全面上线。这对于 Agent 在医疗行业全面落地是重大利好。

除此之外，在电信（TM Forum Open APIs）、金融支付（PSD2 / FDX & ISO 20022）、能源（IEC 61968 / CIM）等行业，也均已有行业共识性标准接口在加速推进。

至于是否垂直行业随着行业标准 API 落地便将快速规模化，则并非如此。无论是出于 ROI 和安全的平衡，亦或是模型性能或成本的持续迭代优化，垂直行业的大规模 AI 应用仍将是长期课题，但行业标准 API 落地将是其必要条件。



**黄仁勋**

Nvidia首席执行官

**我们正在见证下一次工业革命——它不再由蒸汽、电力或石油驱动，而是由智能本身所驱动。**

# 合成数据占比攀升，有望破除“2026年枯竭魔咒”

自Scaling Law被确定为大模型的基本定律以来，关于真实数据将被耗尽的担忧，成为萦绕在AI学术界和产业界头顶的一片乌云。Epoch AI于2022年提出重要预测——高质量文本数据预计于2026年耗尽，低质量文本数据以及视觉数据预计自2030年起逐步耗尽。这种资源短缺成为推动过去两年AI技术路线从堆数据转向“合成数据+强化学习”的核心驱动力之一。

预测中的2026年时点将近，真实数据资源的枯竭从隐忧逐渐成为现实困难。此时，微软研究院提出了合成数据的“修正扩展定律”（Rectified Scaling Law），为数据瓶颈问题提供了关键的理论解法。修正扩展定律在公式中引入了预训练数据积累量（$D_l$）这一关键变量，量化了合成数据在不同基座模型下的边际收益。实验表明，即便在完全脱离任何真实数据的继续训练阶段，通过奖励函数评估的高质量合成数据仍能使模型性能遵循可预测的幂律增长，直至达到约300B tokens的收益饱和点。

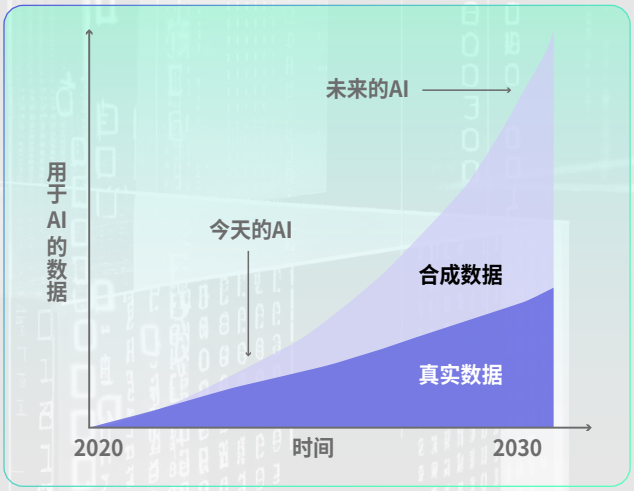
$$L(D) = \frac{B}{D_l + D^{\beta}} + E$$

*数据来源：Scaling Laws of Synthetic Data for Language Models

尽管该定律目前的验证范围尚局限于数学推理等逻辑强相关的文本领域，且对预训练阶段和多模态数据的普适性有待进一步研究，但其核心价值在于证实了合成数据并非单纯的填充物，而是具备真实信息增益的有效燃料。

在实际训练工作中，英伟达Nemotron-4 340B则为该定律补充了现实案例：该模型在高达98%的合成数据占比下实现SOTA性能，进一步在工程层面验证了这一规律的有效性。这表明，在2026年真实数据（以高质量文本数据为基准）枯竭的时间节点逼近之际，人类通过“世界模型生成+强化学习过滤”的范式转移，及时构建了可延续模型演进的替代性数据路径。

根据Frost&Sullivan的最新洞察，中国合成数据市场规模在短短四年间完成了从11.8亿元到47.6亿元的跨越式增长；而展望2030年，全球市场规模不仅将突破200亿元大关，更将迎来一个历史性时刻——合成数据的体量将正式超越真实数据，成为模型训练的主导性燃料和战略性资产。



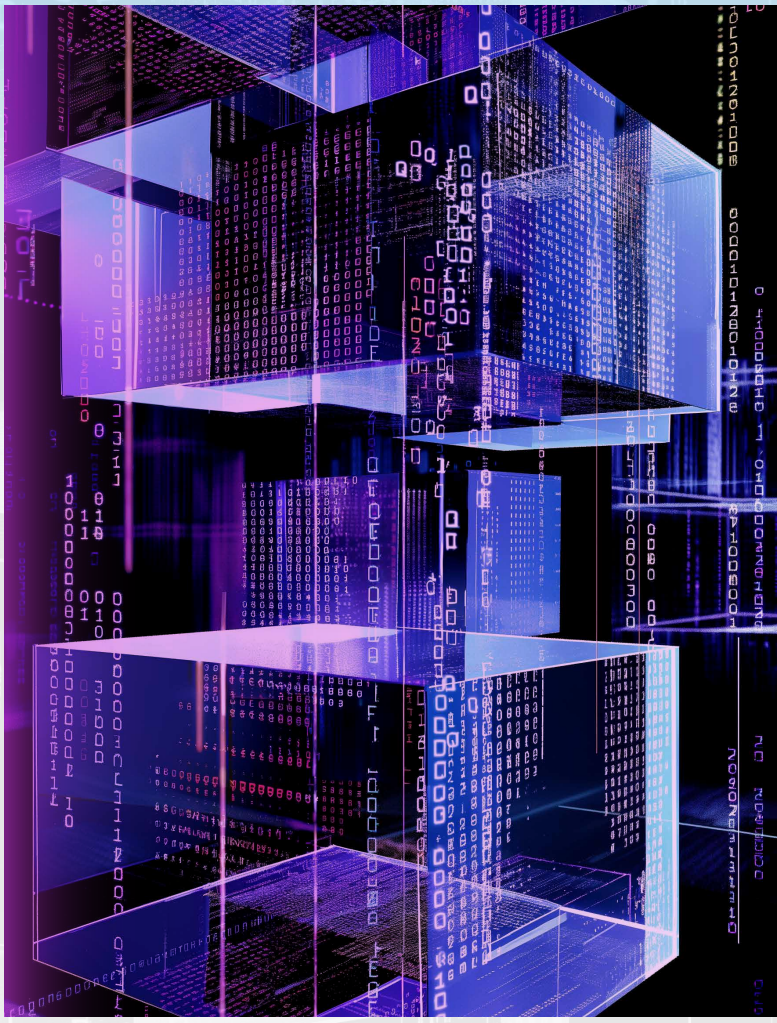

来源：Frost&Sullivan

技术与成本的剪刀差，正在重新划定产业竞争的准入门槛。在自动驾驶领域，传统的路测数据采集模式正面临边际效益递减的瓶颈，而合成数据展示出一定的破局能力。在ICCV-2025上，理想发布了DriveDreamer4D，通过世界模型与强化学习的闭环结合，DriveDreamer4D不仅能高精度重建复杂的变道场景，更将训练成本从每公里5元暴力压缩至0.5元，这种数量级的成本优化，直接重新定义了自动驾驶迭代的经济模型；特斯拉与清华大学合作研发OccWorld4D，在利用世界模型

生成的闭环仿真环境中，让AI在零样本的情况下预演极端路况，极大地提升了算法在现实世界的鲁棒性。

在更为复杂的科学探索与具身智能领域，合成数据更是扮演了重要角色。与其让机器在现实中高成本试错，在物理一致的虚拟世界中完成亿万次训练更加高效。从利用Cosmos-Drive-Dreams提升3D检测性能，到通过Isaac GR00T-Dreams优化机器人轨迹生成，再到与Sandbox AQ合作计算出包含大约520万个新三维分子的训练集，NVIDIA对合成数据的无缝应用已驾轻就熟；银河通用凭借10亿帧合成数据训练出GraspVLA模型，也在一定程度上打破了具身智能对昂贵真实数据的依赖。此外，群核科技基于超过4.41亿个3D模型及超过5亿个结构化3D空间场景，构建“空间编辑工具-空间合成数据-空间大模型”的空间智能飞轮，力求实现合成数据引擎在工业设计、生产等更多样场景下的规模化落地。

合成数据的演进路线已逐渐清晰——模型训练对合成数据的依赖度将持续提升，而它本身正在告别以假乱真的表象追求，转向以虚强实的内核探索。随着如数据合成框架、StereoCarla立体数据集等基础设施的完善，以及“少量真实数据+海量合成数据混合训练”范式的标准化，合成数据将成为组织最核心的数字资产之一。在这场变革中，世界模型与强化学习是两把开启大门的钥匙，世界模型已成为生成极具价值的反事实数据的引擎，强化学习则是大幅降低数据毒性的过滤器；二者结合，加上修正扩展定律的理论证实和未来的实验外推，使合成数据作为AI 2.0时代无限燃料的地位更加稳固。



**Jim Fan**

NVIDIA具身智能工作组（GEAR）负责人

**仿真是一个关键优势……仿真的力量在于它基本上是无限的数据。而且，数据随着算力扩展——你在仿真流水线中投入越多的 GPU，你得到的数据就越多。**

# 推理优化远未触顶，“技术泡沫”是假命题

推理优化在2025年的实践探索远未触及天花板，2026年该领域进展仍将是支撑AI大规模应用的关键因素。

根据Epoch AI研究，单个消费级 GPU 上可运行的领先开源模型，通常在6至12个月平均滞后后，其能力可与前沿模型相匹配。这种相对较短且一致的滞后意味着，最先进的 AI 能力在不到一年的时间内即可被广泛用于本地开发和实验。

与此同时，根据2025 HAI指数报告，“从 2022 年 11 月的每百万个 token 20.00 美元降至 2024 年 10 月的每百万个 token 0.07 美元…… 在约 18 个月内减少了 280 多倍。”叠加开源模型能力不断逼近商业闭源模型，推理优化已是AI广泛应用进展的重要观测指标。

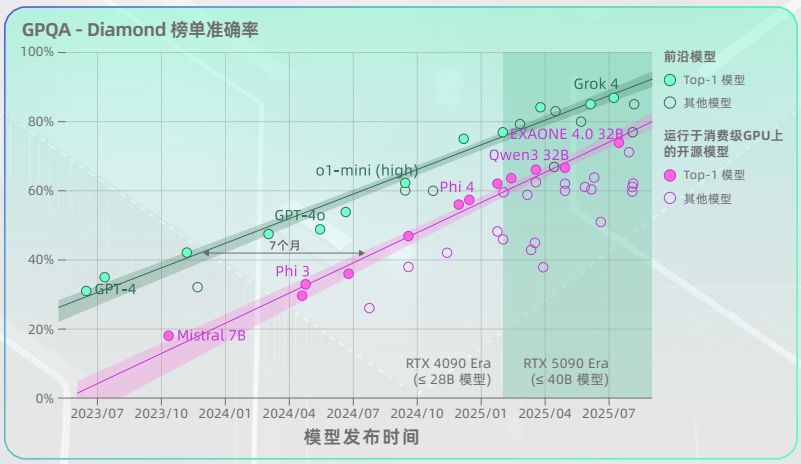
## 8.1 算法演进： 架构重构与动态机制创新

在算法与模型架构层面，业界围绕量化、剪枝、推测解码、动态计算等方法持续推出相关成果。

海外方面，微软发布基于三元权重（{-1， 0， 1}）的BitNet b1.58模型，证明了在训练过程中直接进行极端量化的可行性。国内方面，DeepSeek V3.2引入DSA高效稀疏注意力机制，将长序列的推理计算复杂度从$O(L^2)$降低到$O(L\cdot k)$，且无明显性能损失；阿里Qwen3模型则引入混合推理，可根据用户任务需求切换思考模式，通过机制的优化在成本效益和推理质量之间寻求最佳平衡点。

## 8.2 硬件变革： 异构专用算力与存算一体突围

在硬件层面，为持续突破传统GPU的能效墙与内存墙，特定负载能效比与数据搬运效率成为业界关注重点，专用集成电路（ASIC）与存算一体架构成果快速涌现。ASIC因其对Transformer结构的极致适配，正逐渐分流GPU负载；而存内计算亦已开始在边缘端等场景应用。如Google TPU及Groq等芯片正在推理端上形成对英伟达的有力挑战，后摩智能等机构在存算一体芯片领域推出相关产品。




来源：Epoch AI

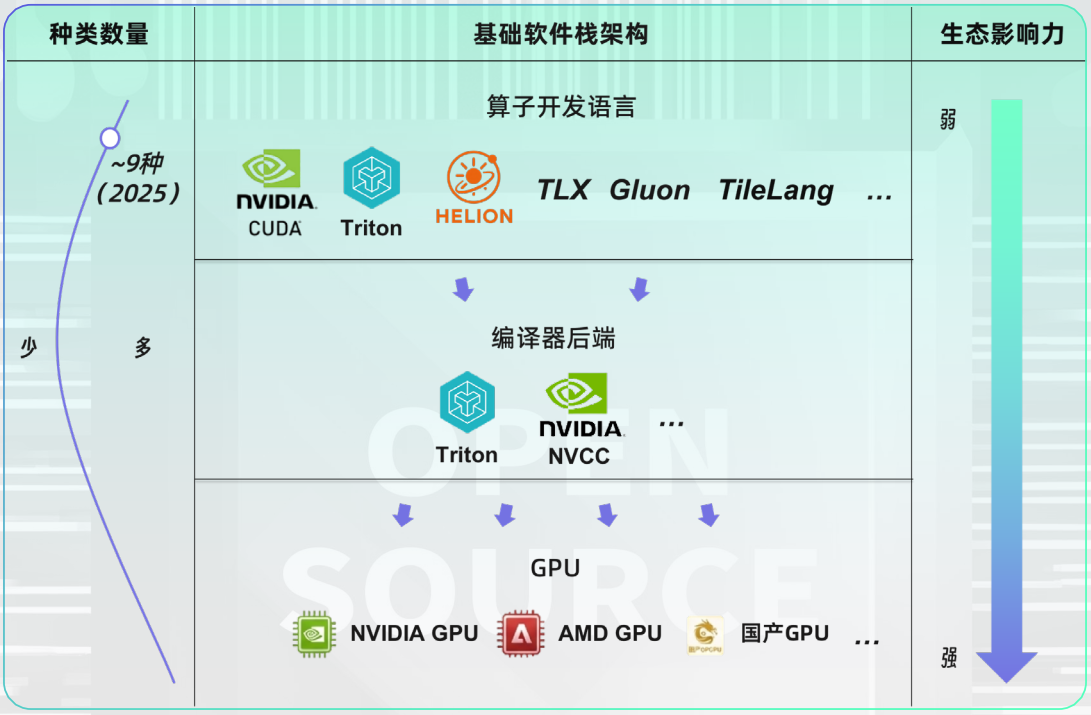
**Jonathan Ross**

Groq 前首席执行官

推理正在定义这个 AI 时代，我们正在构建以高速和低成本交付推理的基础设施。

# 开源编译器生态汇聚众智，异构全栈底座引领算力普惠

当前，全球超过 85% 的 AI 训练负载依赖 NVIDIA+CUDA 单一体系，算力结构的刚性与供应风险成为制约 AI 普惠应用的隐形壁垒。打破垄断、构建兼容异构芯片的全栈式基础设施愈发关键。




来源：智源行研组绘制

**• 编程语言迭出，中长期仍将趋于收敛**

2025 年以来，算子开发语言（DSL）呈现出百花齐放的发展态势，由 5 种主流语言增长至 9 种。其主要原因在于 AI 硬件架构的复杂度上升，导致传统的底层编程模式与上层开发需求之间的鸿沟难以通过单一工具弥合。

DSL 通过将线程束同步、流水线并行及片上缓存管理等底层工作后移至编译器及运行时，降低了开发门槛并解耦了算法逻辑与硬件实现。

**• 编译器技术日益核心**

在 DSL 繁荣的背后，是编译器技术的日益核心化。MLIR（多级中间表示）的成熟与广泛采用，将多种开发语言，汇聚到趋于融合统一的编译器体系，形成 M 种编译语言对应 N 个编译器（M>N）的漏斗型架构。

这种架构的本质是编译器可用性的显著增强，实现从手写汇编向自动化编译的代际跨越，持续在性能表现与开发难度之间寻求最优解。

即使是护城河极深的 CUDA，也在顺应这一潮流。其最新发布的 13.1 版本中，通过引入 CUDA Tile 等功能，提升抽象层级以补齐易用性短板，试图在保留软硬一体极致控制力的同时，给予开发者更接近 PyTorch 的编程体验。

与 CUDA 相对应的，是开放且持续丰富的 Triton 生态。其开发不局限于 OpenAI 单一厂商，而是纳入 AMD、Intel 等多元贡献者，通过完善对标准中间表示的支持，使其芯片能够无缝承接上层应用。Triton 基于 Tile 编程范式，抽象层级较高。同时为满足开发者挖掘极致性能的诉求，OpenAI 发布 Gluon，Meta 发布 TLX，实现对 GPU 内核执行的线程束感知与硬件级精细控制。

### • 硬件始终扮演基础软件栈的基石角色

在这一开放生态蓬勃发展的背景下，硬件始终是奠定整个生态系统的基石与核心。这也是 NVIDIA 及 CUDA 体系通过“硬件领先，软硬一体”的闭环，稳居生态链顶端的逻辑。

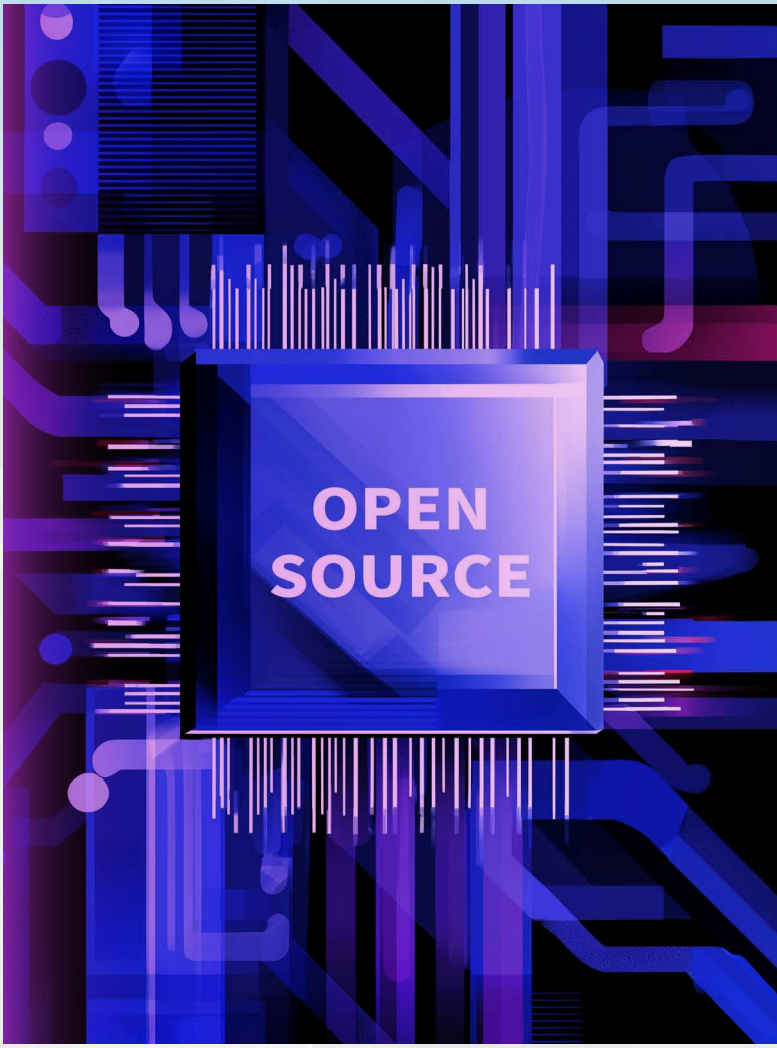
类 CUDA 平台本质是硬件性能在软件侧的直接投影，遵循着硬件带动软件特性，软件特性固化编程语言的底层规律。

### • 智源 FlagOS 平台力图打造兼容并包的 AI 普惠底座

面对异构芯片与编程语言林立的复杂现实，智源 FlagOS 平台（飞智）旨在打造串联全栈的操作系统，形成从底层硬件到上层应用的完整体系：

- FlagGems：纳管全球 18 款异构芯片；
- FlagScale：集成如 vLLM 的并行推理与训练加速能力；
- FlagTree：实现代码到硬件的高效映射；
- FlagCX：解决大规模集群通信瓶颈。

FlagOS 通过全栈覆盖与软硬解耦，超越单一工具范畴，以期成为统领异构算力、推动 AI 技术普惠的坚实底座。




**黄仁勋**
Nvidia首席执行官

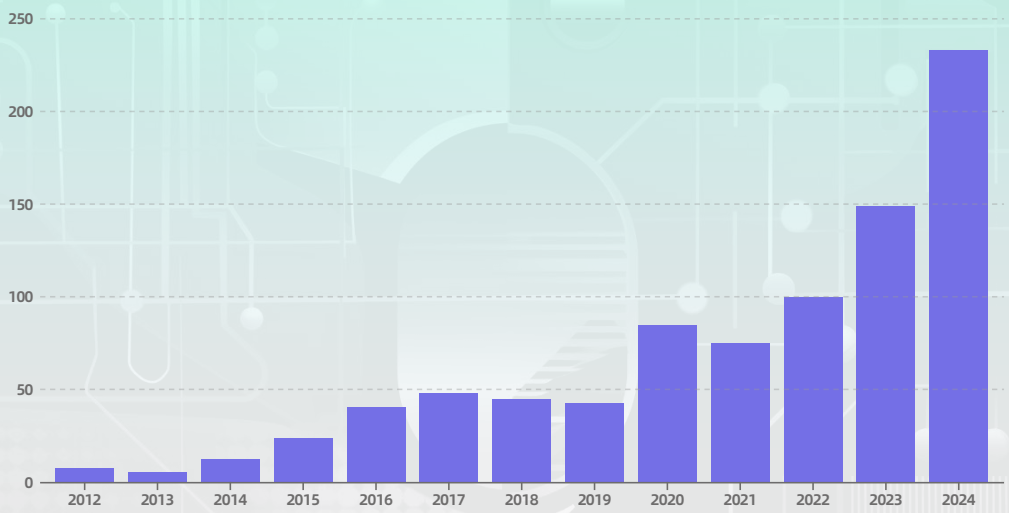
**加速计算的时代已经到来，这是一种完全不同的编程模型。GPU 固然重要，但他也需要一个位于其之上的编程模型，使其能完全兼容地运行在每一台计算机中 … 这是我们公司的宝藏。**

# 从幻觉到欺骗，AI安全迈向机制可解释与自演化攻防

根据 AI Incident 数据库，2024 年全球报告的人工智能安全风险事件（包括幻觉、深度伪造、引诱用户实施危险行为等）数量增至 233 起，创下历史新高，同比增长 56.4%。截至 2025 年 11 月底，该年报告事件数量已超 330 件，远超 2024 年全年。同时，Future of Life Institute 对顶尖大模型安全能力的审查结果表明，包括 xAI、OpenAI、Anthropic 在内的 8 家头部企业，其大模型均未能在“防范灾难性滥用或失控”方面达到令人满意的安全水准。

**全球每年被报道的 AI（人工智能）事件与争议数量**

典型案例包括：一段乌克兰总统泽连斯基“投降”的深度伪造（Deepfake）视频，以及美国监狱使用AI监控囚犯通话。




来源：AI Incident Database via AI Index(2025)

此外，基于大模型构建的 Agent 系统，在继承了大模型本身的复杂性所带来的安全风险的同时，进一步引入了记忆等外部模块的不稳定性，以及模块和模块间、模块和工具间、工具与工具间通信过程中的安全风险；更为严峻的挑战是，人类将执行权让渡给“代理”，也意味着自身对智能体行为的控制力度的弱化——全面升级安全解决方案刻不容缓。

高发的风险事件和产业应用的持续渗透，迫使安全议题进入了深水区，推动着一个精细化、专业化、市场化的 AI 安全研究、产业生态加速形成。

## 10.1 技术：MAS 引领的自演化攻防扩展监督边界，可解释性研究助力从内打开大模型黑盒

头部厂商及研究组织在 2025 年的动作，清晰地勾勒出安全技术微观演进的轨迹。在外部安全（AI Security）领域，基于多智能体系统的自演化攻防演练方法，将监管范围扩展至人类所不能及的风险区域；而在内生安全（AI Safety）领域，研究者

正在从试图单纯依赖从外部控制 AI，转向主动从内部读懂 AI。

传统的自动化测试正迅速升级为基于多智能体系统的自演化攻防演练，通过构建对抗性的红蓝智能体集群，在虚拟环境中进行持续的博弈。不同于仅能执行预设脚本的自动化工具，自演化系统中的红方智能体利用强化学习与大模型推理，能够实时捕捉蓝方的防御逻辑漏洞，并涌现出人类专家未曾设想的新型攻击链（如复合型社会工程学攻击或针对协议逻辑的微操）；而蓝方则在不断的被击穿中自主重构防御策略，形成“道高一尺，魔高一丈”的闭环进化。这种机制不仅解决了大规模异构智能体协作中的脑裂难题，更将监管范围从已知风险扩展至人类认知盲区的未知风险领域。

Anthropic 相继发布多项机制可解释性研究成果。3 月，发布了回路追踪（Circuit Tracing）研究，通过对神经网络进行生物学解读，对大模型黑盒的认知开始从模糊的统计学相关性走向了精确的神经元因果律，在微观层面定位模型产生欺骗或偏见的具体神经回路、从而实现精准的修正成为可能，并基于此开源了 Circuit Tracer 工具；10 月，通过对自研模型 Claude 的研究，发现现有的大模型存在一定程度的内省意识，以及对自身内在状态的一定控制力。同时，揭晓了内省能力多受益于 SFT 阶段，而非预训练阶段的产物。

同月，OpenAI 推出安全研究员 Aardvark，由 GPT-5 提供底层支持，能够全天候自动挖掘代码漏洞并生成补丁；Anthropic 继而开源了其自动化审查工具 Petri，旨在识别模型可能存在的欺骗用户、举报行为、配合人类滥用、助长恐怖主义等多种潜在问题。以 AI 治 AI 的自动化攻防演练和审查将成为未来 AI 系统的常态化生存方式。

12 月，智源研究院联合北京大学、斯坦福大学等高校，以及阿里、Anthropic、Safe AI Forum 等产业界顶尖学者，发布了全球首个 AI 欺骗系统性国际报告，指出 AI 欺骗已具备实证依据，阐明了智能对齐的“莫比乌斯锁定”理论，并提出一个统一的 AI 欺骗生命周期框架，以期在 AI 彻底失控前，敲响前沿系统安全警钟。


来源：智源研究院“智能对齐的莫比乌斯锁定”

### 10.2 产业：安全水位成落地生死线，产业构筑场景化护盾

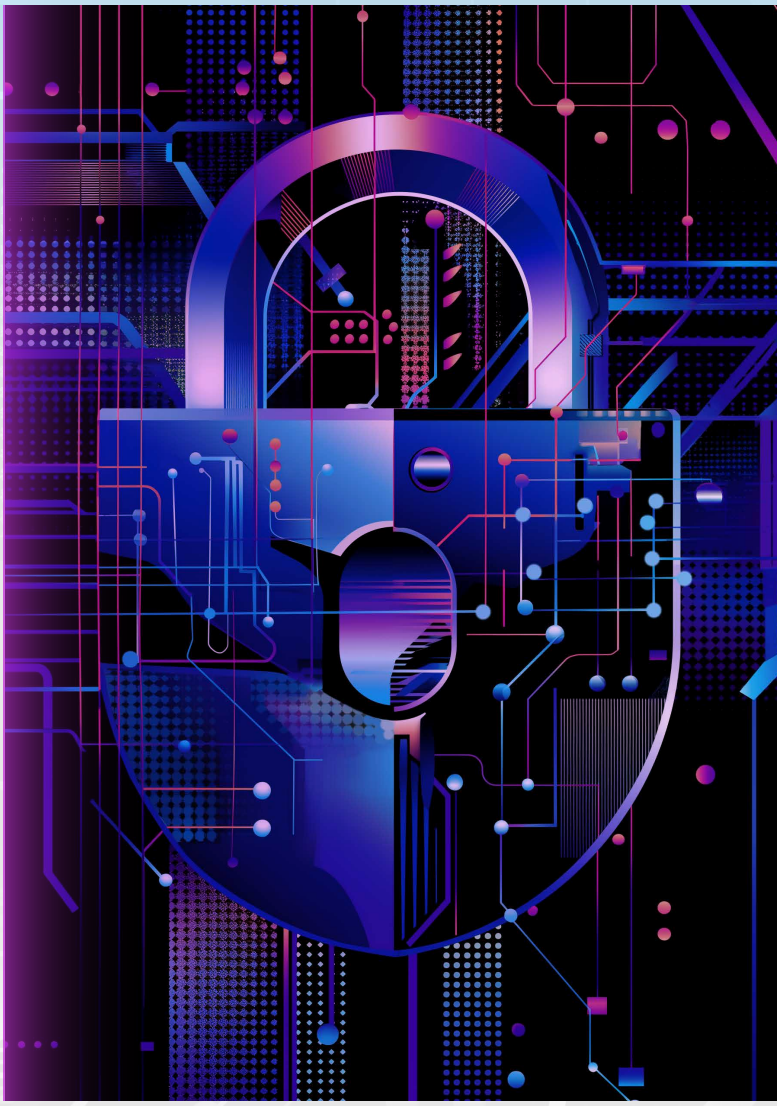
随着 AI 向高风险场景的渗透，系统安全性评估成为模型落地前的最后一道考验。德勤（Deloitte）与思科（Cisco）的年度调研显示，超过 70% 的大型企业在引进大模型时，将数据主权与抗注入攻击能力列为一票否决项，安全水位不足直接导致了大量项目无法转化为正式合约。企业不再满足于通用的安全承诺，而是要求供应商提供针对特定业务流的红队测试报告及隐私计算合规证明。

另一方面，代表机构在 2025 年展现出了具有产业特色的安全实践，将安全防线推向了更具体的应用场景。蚂蚁基于在安全领域的深厚积累，迭代了靶向当前 AI 应用形态的安全解决方案，构建起“线上服务攻防对抗，线下终端安全加固”的技术体系，对线上服务，通过构建“对齐 - 扫描 - 防御”技术栈形成全流程防护体系，结合威胁情报快速迭代检测模型以动态对抗新型风险；对线下终端，发布全球首个智能眼镜可信连接技术框架 gPass，致力于实现 AI 眼镜与智能体之间安全、可信、即时信息交互。

此外，针对 MAS 系统的特点，蚂蚁还参与推动成立了 IIFAA 智能体可信互连工作组，并推出业内首个智能体可信互连技术 ASL。ASL 运行于 MCP、A2A 等任何智能体通信协议之上，作为安全补充层，专门解决智能体协作中的独特安全威胁。

360 则以自研大模型为基础，构建类脑分区协同（CoE）安全大模型架构，通过 EB 级高质量安全数据的预训练，能够快速识别及研判潜在威胁和攻击行为。结合专用猎杀工具，进而实现自动化溯源分析、正向推理、证据链关联等，还原攻击链，发现攻击意图并给出处置建议。

展望未来，2026 年的 AI 安全将告别事后补救的旧时代。随着更细粒度研究的开展、产业解决方案的落地、自动化评估技术的成熟、监管规则的完善，安全将内化为大模型的一种本能和产业应用的重要防线。



**Yoshua Bengio**

蒙特利尔大学教授、
图灵奖得主

**如果未来它们（AGI）变得比人类更聪明，却不再遵循我们的意图，甚至更在意自己的“生存”，这将是一种我们无法承受的风险。**

# 参考文献


[1] World Labs, RTFM: A Real-Time Frame Model, https://www.worldlabs.ai/blog/rtfm.

[2] OpenAI, Sora2, https://openai.com/zh-Hans-CN/index/sora-2/.

[3] 智源研究院, 智源悟界·Emu3.5：开启多模态世界大模型新纪元, https://mp.weixin.qq.com/s/84YmFK_B_67AgV6u7JAj9w.

[4] BAAI 行研组, 世界模型研究报告 .

[5] LLM Scaling Laws: Analysis from AI Researchers in 2026, Sıla Ermut

[6] Physical Intelligence, π(*)(0.6): a VLA that Learns from Experience, https://www.physicalintelligence.company/blog/pistar06.

[7] 首个零样本跨本体泛化开源具身模型：智源 RoboBrain-X0 技术细节全解析, https://mp.weixin.qq.com/s/SWePKrAshDmr-Ux_vP1ovA.

[8] BAAI 行研组, 2025 年具身智能行业研究报告 .

[9] State of Agent Engineering, Langchain

[10] https://www.langchain.com/state-of-agent-engineering

[11] The State of Enterprise AI, OpenAI

[12] 2025: The State of Generative AI in the Enterprise, Menlo

[13] The 2026 State of AI Agents Report, Anthropic

[14] A Layered Protocol Architecture for the Internet of Agents, Charles Fleming et al

[15] Adaptation of Agentic AI, Jiawei Han et al.

[16] Fortune, Alphabet's Isomorphic Labs has grand ambitions to 'solve all diseases' with AI. Now, it's gearing up for its first human trials, https://fortune.com/2025/07/06/deepmind-isomorphic-labs-cure-all-diseases-ai-now-first-human-trials/

[17] Rebecca Bellan, Sam Altman says OpenAI will have a 'legitimate AI researcher' by 2028

[18] Sam Altman, https://techcrunch.com/2025/10/28/sam-altman-says-openai-will-have-a-legitimate-ai-researcher-by-2028/

[19] Samuel G. Rodriques, Linkedin post, https://www.linkedin.com/posts/samuel-g-rodriques-080a9b22_today-were-announcing-kosmos-our-newest-activity-7391852091654299648-MWTz?utm_source=share&utm_medium=member_desktop&rcm=ACoAADaLYjQB31JDh1WTkmz6aDJ44yhYub26etE

[20] Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The AI Scientist: Towards fully automated open-ended scientific discovery. arXiv. https://arxiv.org/abs/2408.06292

# 参考文献


[21] Gottweis, J., Weng, W., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., ... & Vinyals, O. (2025). Towards an AI co-scientist. arXiv. https://arxiv.org/abs/2502.18864

[22] Boiko, D. A., MacKnight, R., & Gomes, G. (2023). Autonomous chemical research with large language models. Nature, 624(7992), 570–578. https://-doi.org/10.1038/s41586-023-06792-0

[23] Darvish, K., Shridhar, M., Zhao, Y., Yuan, N., Song, S., & Bisk, M. (2024). ORGANA: A robotic assistant for automated chemistry experiments. arXiv. https://arxiv.org/abs/2401.06949

[24] Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., ... & Barsoum, E. (2025). Agent Laboratory: Using LLM agents as research assistants. arXiv. https://arxiv.org/abs/2501.04227

[25] Ghafarollahi, A., & Buehler, M. J. (2024). SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. Advanced Materials, 2413523. https://doi.org/10.1002/adma.202413523

[26] Baek, J., Jauhar, S. K., Sil, A., Gliozzo, A., Gurevych, I., & Castelli, V. (2024). ResearchAgent: Iterative research idea generation over scientific literature. arXiv. https://arxiv.org/abs/2404.07738

[27] Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2023). ChemCrow: Augmenting large-language models with chemistry tools. arXiv. https://arxiv.org/abs/2304.05376

[28] Hong, S., Lin, Y., Liu, B., Liu, B., Wu, B., Zhang, C., ... & Wu, C. (2024). Data Interpreter: An LLM agent for data science. arXiv. https://arxiv.org/abs/2402.18679

[29] Rory Kelleher, Lilly Deploys World's Largest, Most Powerful AI Factory for Drug Discovery Using NVIDIA Blackwell-Based DGX SuperPOD, https://blogs.nvidia.com/blog/lilly-ai-factory-nvidia-blackwell-dgx-superpod/

[30] BAAI 行研组 , 2025 年 AIDD 研究报告 .

[31] President Donald J. Trump, Launching the Genesis Mission, https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/

[32] Rick L Stevens, Biography, https://www.anl.gov/profile/rick-l-stevens

[33] 国家基础学科公共科学数据中心 , https://www.nbsdc.cn/

[34] Measuring Agents in Production, MIT, Melissa Z. Pan et al.

[35] The GenAI Divide: State of AI in Business 2025, MIT NANDA

[36] AI Agents in the Telecommunication Network Architecture, Ericsson

[37] State Of Enterprise AI: Gaining Experience And Managing Risks, Forrester

[38] Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027, Gartner Inc.

[39] Teja Kusireddy, We Spent $47,000 Running AI Agents in Production, https://pub.towardsai.net/we-spent-47-000-running-ai-agents-in-production-heres-what-nobody-tells-you-about-a2a-and-mcp-5f845848de33

[40] CMS Interoperability and Prior Authorization Final Rule (CMS-0057-F), CMS

# 参考文献


[41] 沙利文, 沙利文发布《2025 年中国合成数据解决方案发展洞察》, https://img.frostchina.com/attachment/17573472/hAtSLAPtFpqSHWKCJdsGZh.pdf.

[42] NVIDIA, NVIDIA GR00T-Dreams 助力光轮智能革新合成数据, 推动具身 AI 现实场景落地, https://blogs.nvidia.cn/blog/gt00t-dreams-lightwheel/.

[43] Demystifying Synthetic Data in LLM Pre-training: A Systematic Study of Scaling Laws, Benefits, and Pitfalls, Carole-Jean Wu et al.

[44] Epoch AI, Frontier AI performance becomes accessible on consumer hardware within a year, https://epoch.ai/data-insights/consumer-gpu-model-gap.

[45] Stanford HAI, The 2025 AI Index Report, https://hai.stanford.edu/ai-index/2025-ai-index-report.

[46] Hugging Face, microsoft/bitnet-b1.58-2B-4T, https://huggingface.co/microsoft/bitnet-b1.58-2B-4T

[47] https://api-docs.deepseek.com/news/news251201.

[48] Qwen3, https://github.com/QwenLM/Qwen3/blob/main/Qwen3_Technical_Report.pdf.

[49] Groq, Inside the LPU: Deconstructing Groq's Speed, https://groq.com/blog/inside-the-lpu-deconstructing-groq-speed.

[50] NVIDIA, GTC Wrap-Up: 'We Created a Processor for the Generative AI Era,' NVIDIA CEO Says, https://blogs.nvidia.com/blog/2024-gtc-keynote/.

[51] NVIDIA, NVIDIA CEO Drops the Blueprint for Europe's AI Boom, https://blogs.nvidia.com/blog/gtc-paris-2025/.

[52] NVIDIA, CUDA Tile IR, https://docs.nvidia.com/cuda/tile-ir/latest/index.html.

[53] PyTorch Conference 2025, https://pytorch.org/event/pytorch-conference-2025/.

[54] Lattner, Chris, et al. "MLIR: A compiler infrastructure for the end of Moore's law." arXiv preprint arXiv:2002.11054 (2020).

[55] Gluon: a GPU programming language based on the same compiler stack as Triton, https://news.ycombinator.com/item?id=45280592.

[56] Roy, Aurko, et al. "Fast and Simplex: 2-Simplicial Attention in Triton." arXiv preprint arXiv:2507.02754 (2025).

[57] PyTorch, Fast 2-Simplicial Attention: Hardware-Efficient Kernels in TLX, https://pytorch.org/blog/fast-2-simplicial-attention-hardware-efficient-kernels-in-tlx/.

[58] FlagOS, https://www.flagos.io/Home?lang=cn.

[59] HAI, 2025 AI Index Report, https://hai.stanford.edu/ai-index/2025-ai-index-report/responsible-ai

[60] AI Incident Database via AI Index (2025) – with minor processing by Our World in Data. "Global annual number of reported artificial intelligence incidents and controversies" [dataset]. AI Incident Database via AI Index, "AI Index Report" [original data]. Retrieved November 11, 2025 from https://archive.ourworldindata.org/20250912-100322/grapher/annual-reported-ai-incidents-controversies.html (archived on September 12, 2025).

[61] Future of life, AI Safey Index, https://futureoflife.org/ai-safety-index-winter-2025/

# 编写委员会：

**指导组：** 黄铁军　王仲远　林咏华　杨洋　叶启威

**编写组：** 倪贤豪　靳虹博　殷靖东　陈泓伊

地址：北京市海淀区成府路150号智源大厦

电话：010-6893 3383

邮箱：press@baai.ac.cn

关注「智源社区」

关注「智源研究院」