2025 十大AI技术趋势

BAAI
智源研究院

IntelMining
智能矿业

科技POWER
智能矿山

IntelMining 矿业科技综合服务平台

“IntelMining智能矿业”致力于打造矿业科技综合服务平台，创始团队毕业于中国矿业大学、北京科技大学等矿业知名学府，拥有央企工作背景和服务中关村高新技术企业的从业经历，2020年获得中关村高新技术企业认证。平台主要依托三大阵地开展科技服务：

■品牌自媒体，打造“IntelMining智能矿业”和“InMi硬米”两大品牌公众号、视频号、抖音号等自媒体矩阵；

■网站平台，打造集行业资讯、在线展厅、首发平台、活动管理、智库人才、技术交易等为一体的综合服务平台；

■品牌活动，打造线上线下相结合的供需对接、项目路演、首次发布、行业沙龙、高峰论坛、空中宣讲等活动会议，帮助各级各类主体直接触达行业伙伴。

持续输出行业咨询、渠道拓展、成果转化、技术服务等能力，充分为行业科技发展应用赋能。

www.intelmining2018.com

媒体矩阵

◀业务微信
■联系方式：
张晓宏 18911270075
郎老师 18101060076

卷首语

岁月不居，时节如流。站在新旧交接的十字路口回望，一系列前所未见的技术突破正在重塑机器智能的定义，引发着深层次的变革，预示着更新、更美好的智能图景。

大模型的持续进化，如同蝴蝶振翅般颠覆了我们对人工智能的传统理解。从初次尝试新架构到发现新的普适定律，从能力泛化到模态无缝融合，这些突破性进展正在不断刷新机器智能的边界。大模型逐步拥抱文本、视觉、音频、乃至 3D 数据，实现了感知与认知能力的全面升级，机器具备了更加细腻丰富的理解能力，人机交互焕发了全新的活力。与此同时，人工智能正在向着另一个关键维度挺进——对真实物理世界的模拟与适应。在这一主题下，机器不仅能够自主感知和推理复杂场景，更能够主动规划行动、做出决策。而具身智能的加速落地，又进一步塑造了机器的物理形态。从感知到决策再到控制执行，端到端的智能系统正在崛起，机器的适应性和灵活性持续突破。

令人振奋的是，这些趋势正互为助力、相得益彰。基座模型能力迭代，为世界模拟和具身智能注入了更精准的感知与认知基础，应用落地数据又反哺着基座模型的成长；大模型的惊人能量，撬动着基础科学的浩瀚宇宙，大模型本身又作为科研对象，静候研究者揭晓它更深的奥秘。在这些力量的驱动下，Agentic AI 与新时代的超级应用应运而生，悄然渗入每个人的工作和生活中，春风化雨般改变着人机交互的形态。

光明总是与黑暗共存。技术和应用正在高歌猛进，重塑人类社会的方方面面，而安全隐患在暗处滋生。我们必须建立起与日益智能的机器系统相称的安全技术框架，探索具备最大共识的治理之道，才能最大限度地释放人工智能的无穷潜能，让技术以负责任的方式造福人类社会。

于是，在 2025 年的开端，我们提出十个人工智能技术及应用趋势。通过深入剖析科技的演进轨迹，更清晰地洞察未来几年的科技发展方向，预测哪些核心技术将成为关键驱动力、哪些新兴技术将蓬勃发展，它们将如何以创新之力指引人类社会迈向更加智能、美好与互联的未来。

科技的曙光将照耀人类前行的路途。这些技术将在激烈的竞争与协作中相互促进，共同谱写人与智能系统共生共荣的磅礴篇章。而我们作为亲历者，将见证科技为人类文明注入澎湃动能，推动人类能力的边界向更高更远处延伸。站在科技的肩膀上，身可高百尺，手可摘星辰。

目录

趋势一

科学的未来： AI4S 驱动科学研究范式变革 p04

趋势二

“具身智能元年”： 具身大小脑和本体的协同进化 p06

趋势三

“下一个 Token 预测”： 统一的多模态大模型实现更高效 AI p08

趋势四

Scaling Law 扩展： RL + LLMs，模型泛化从预训练向后训练、推理迁移 p10

趋势五

世界模型加速发布，有望成为多模态大模型的下一阶段 p12

趋势六

合成数据将成为大模型迭代与应用落地的重要催化剂 p14

趋势七

推理优化迭代加速，成为 AI Native 应用落地的必要条件 p16

趋势八

重塑产品应用形态，Agentic AI 成为产品落地的重要模式 p17

趋势九

AI 应用热度渐起，Super App 花落谁家犹未可知 p18

趋势十

模型能力提升与风险预防并重，AI 安全治理体系持续完善 p19

参考文献 p20

科学的未来：AI4S驱动科学研究范式变革

- 2024 年度的诺贝尔物理学奖、化学奖均颁发给了 AI 领域科学家。大模型引领下的 AI4S，已成为推动科学研究范式变革的关键力量。
- 科学研究的范式带有其所处时代的认知水平、价值取向、工具先进性、科研资源等因素的深刻烙印。自人类开始记录自然现象以来，科学研究经历了经验观察、理论建构、仿真模拟、数据驱动的科学发现四个阶段。

来源：2024 年诺贝尔奖官方插画——物理奖及化学奖获得者

- 大模型时代，AI4S（AI for Science）展现出的赋能效果与小模型时期大相径庭。传统人工智能在科学研究中多聚焦于特定任务的优化，如数据挖掘算法辅助科研数据处理，或基于既有模式进行推理预测，但其模型规模与泛化能力有限，难以解决复杂问题。而大模型以海量数据训练，具备强大的跨领域知识整合能力；模型架构赋予其多层次的学习和处理能力，能够捕捉高维数据中的复杂结构和模式，并对复杂科学问题进行整体理解与全局综合分析。大模型还能通过生成式能力提出创新性假设，为科学研究开辟新方向。

在高性能算力的支持下，大模型正以前所未有的广度与深度重塑科学研究格局，成为推动科研突破、解决实际问题的强劲动力。2024 年 2 月，俄亥俄州立大学发布用于执行化学任务的 LlaSMol 大模型，在名称转换、特征预测、分子描述、化学反应知识等任务上取得较优成绩；同时，研究团队发布了包含 14 个任务、300 多万个高质量样本的数据集 SMolInstruct，为后续相关研究提供宝贵资源。5 月，DeepMind 和 Isomorphic Labs 团队联合发布 AlphaFold 3，能够准确预测蛋白质与其他分子的相互作用，相较上一代模型，应用范围取得巨大突破。

国内方面，2024 年 6 月，中国气象局发布“风清”“风顺”“风雷”三个人工智能气象大模型系统，具有大气强物理融入和可解释性，在实现高效计算

的同时，可为预测结果提供物理可解释性依据，自动挖掘包括天气系统内在的物理演变。12 月，北京智源人工智能研究院提出的 BAAIWorm 天宝被选为 Nature Computational Science 期刊封面；BAAIWorm 是一个全新的、基于数据驱动的生物智能模拟系统，首次实现秀丽线虫的精细神经系统、身体与环境的闭环仿真，为探索大脑与行为之间的神经机制提供重要研究平台。此外，智源研究院正在研发 OpenComplex 平台，该平台建立了将蛋白质结构预测、RNA 结构预测和蛋白质-RNA 复合物结构预测三类任务统一的端到端生物大分子三维结构预测深度学习框架，以期逐步构建能够模拟生物过程的“数字孪生系统”。

2025 年，多模态大模型将进一步融入科学研究，赋能多维数据的复杂结构挖掘，辅助科研问题的综合理解与全局分析，为生物医学、气象、材料发现、生命模拟、能源等基础与应用科学的研究开辟新方向。

David Baker

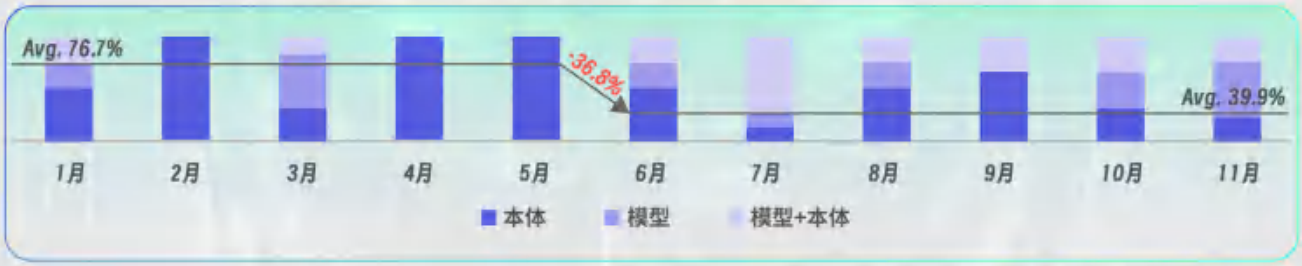
华盛顿大学蛋白质设计研究所所长
2024年诺贝尔化学奖得主

（正是因为 AI 的影响），我们看到原本如同黑魔法一般的蛋白质疗法（给动物免疫，让自然免疫系统找到解决方案）转变为实际合理的设计。

"具身智能元年"：具身大小脑和本体的协同进化

- 2024 年，全球范围内的具身智能竞争日渐白热化。从融资规模、产品发布等多维度来看，中美两国在该领域执行业牛耳。以国内为例，根据智源研究院数据统计，截至 2024 年底，国内已发布或在研人形机器人接近 100 家，融资规模超 100 亿元，称之为"百机大战"并不为过。

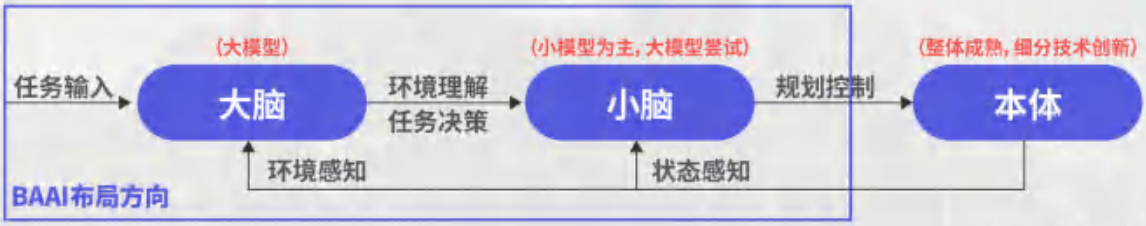
- 从厂商类型来看，专注本体 / 零部件、具身脑、具身脑和本体并重等三类厂商主导具身智能行业。自 2024 年 5 月以后，获得融资的"专注本体"的具身初创企业融资事件数平均占比较前五月下降了 36.8%。可以认为，具身赛道的创业和资本热度已从本体扩展到具身大小脑。



来源：BAAI 行研组

- 相较于整体成熟、更注重细节创新的本体，大模型目前在具身大脑应用较多。在具身小脑方向，大模型的应用尝试将将起步。

- 本体方面，作为一个相对成熟的领域，在本轮具身智能热度中，更倾向于在细分领域有所创新。比如灵巧手为代表的末端执行器、触觉传感器为代表的传感器、面向具身专门设计的感知芯片等细分领域，在 2025 年均可能迎来更新迭代。



来源：BAAI 行研组

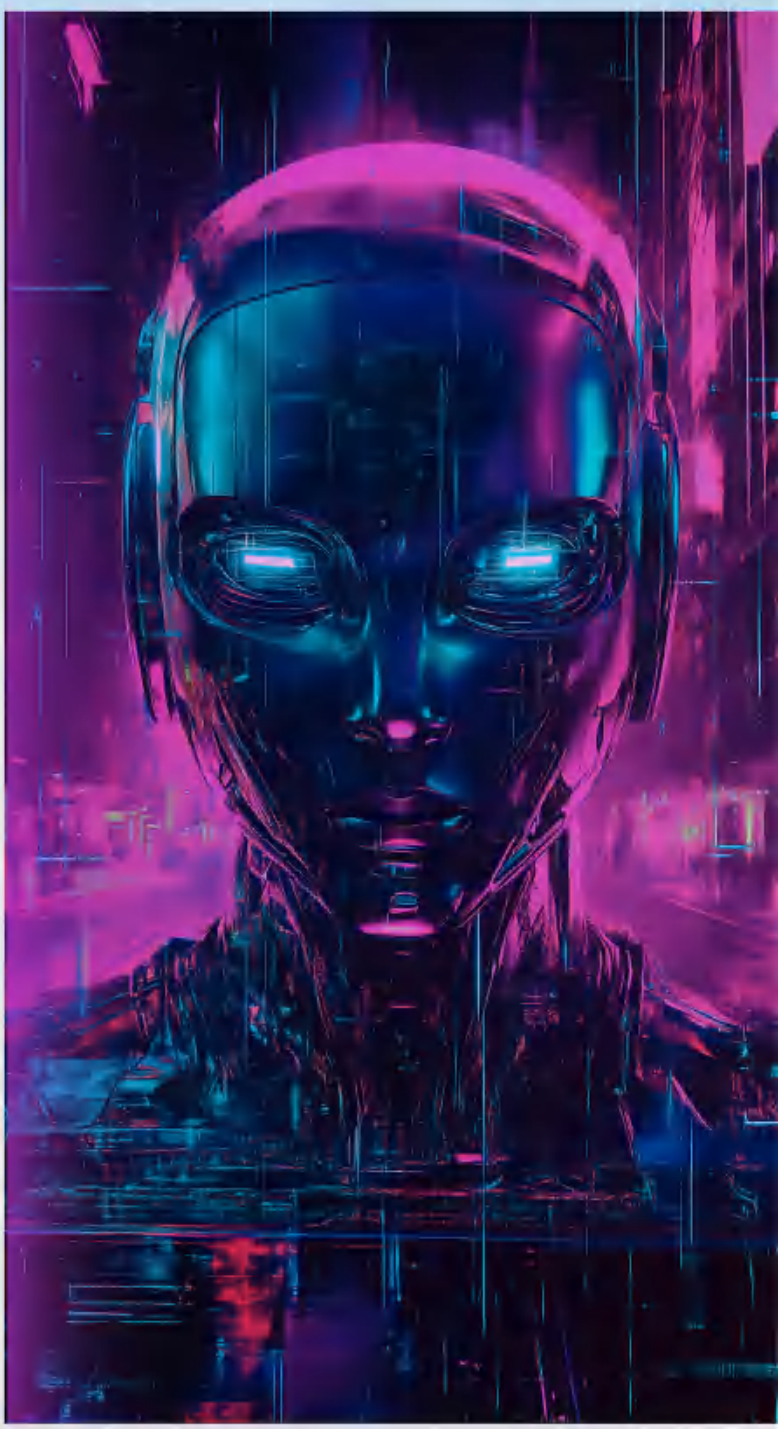
- 至于具身大模型，目前已形成两条主流技术路线：端到端模型和分层决策模型。分层模型方面，LLM、VLM 等已成为具身大脑的主流范式，而小脑侧仍以传统控制方法为主。端到端模型，作为近两年的研究热点，覆盖感知 - 决策 - 控制全流程，理论上可获取的信息量最为丰富，端到端的输出效果最优。就模型赋能效果来看，具身大模型已在感知决策端实现了较好的多任务迁移和处理，但控制执行侧的泛化仍需要技术路径的持续迭代和模型规模的 Scaling up，这或可成为 2025 年的突破方向。

国内外科技大厂及研究机构在近两年时间内密集推出具身模型成果。

海外方面，Google 联合 DeepMind 发布的 RT 系列模型。其中，RT-1 首先将 Transformer 应用到机器人领域，表现出较好的长时序任务执行能力。RT-2 则是首个端到端视觉语言动作模型（VLA，Vision-Language-Action Models），实现了感知信息输入 - 动作控制信息输出。RT-X 基于自采的大规模、多样化数据集训练，支持在多机器人平台、泛化任务和环境间迁移，通用性进一步提升。斯坦福大学在 2023 年发布的多模态视觉模型 VoxPoser（LLM+VLM），可根据感知到的环境信息与用户指令，指导合成机器人所需执行的操作轨迹。Physical Intelligence 公司发布π⁰通用机器人基础模型，将互联网规模的视觉 - 语言预训练与实际机器人交互数据相结合，在五项机器人任务的评估中优于其他的基线模型。

国内方面，银河通用尝试利用三维视觉小模型 + 基础大模型的技术组合解决具身模型泛化能力差，响应速度慢的问题。目前，银河通用的具身大模型机器人 Galbot 已落地应用于美团 24 小时无人值守药房，承担补货、取货等任务；星海图持续推动在具身本体及核心模组、端到端 AI 算法以及场景解决方案的研发及落地；北京智源人工智能研究院基于快系统和慢系统的设计路线，将快系统用于产生快速直觉的动作，当通过快系统执行任务失败时，再通过慢系统检测、定位任务失败节点，并进行纠正。

2025 年的具身智能，将继续从本体扩展到具身脑的叙事主线，我们可以从三方面有更多期待。在行业格局上，近百家的具身初创或将迎来洗牌，厂商数量开始收敛；在技术路线上，端到端模型继续迭代，小脑大模型的尝试或有突破；在商业变现上，我们也必将看到更多的工业场景下的具身智能应用，部分人形机器人迎来量产。



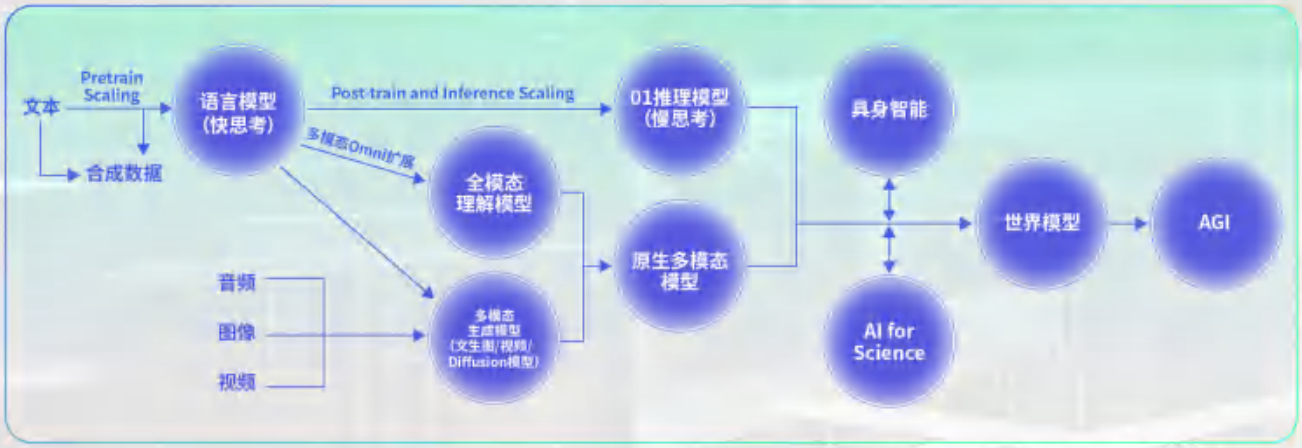
黄仁勋

英伟达创始人

人工智能的下一个浪潮将是具身智能，即能理解、推理、并与物理世界互动的智能系统。

“下一个Token预测”：统一的多模态大模型实现更高效AI

- 2023 年以来的大模型热度肇始于 LLM 在多任务中的涌现，但囿于 LLM 所学习的模态单一，模型能力很难向高维的真实世界拓展。而人工智能的本质在于对人的思维的信息过程的模拟，人类对于信息的交互和处理，总是呈现多模态、跨模态的输入输出状态。当前的语言大模型、拼接式的多模态大模型，在对人类思维过程的模拟上存在天然的局限性。
- 以传统多模态大模型为例，Diffusion Transformer（DiT）和 LLM+CLIP 是当前主流的多模态构建路线，但这两条路径中数据的“后融合”方式会造成多模态信息的损失，各模态信息表征本质上是相互孤立的，大模型对多种模态数据理解的不充分会导致多种模态生成的割裂和误差增大。因此，从训练之初就打通多模态数据，实现端到端输入和输出的原生多模态技术路线给出了多模态发展的新可能。
- 基于此，训练阶段即对齐视觉、音频、3D 等模态的数据，实现多模态的统一，构建原生多模态大模型成为多模态大模型进化的重要方向。

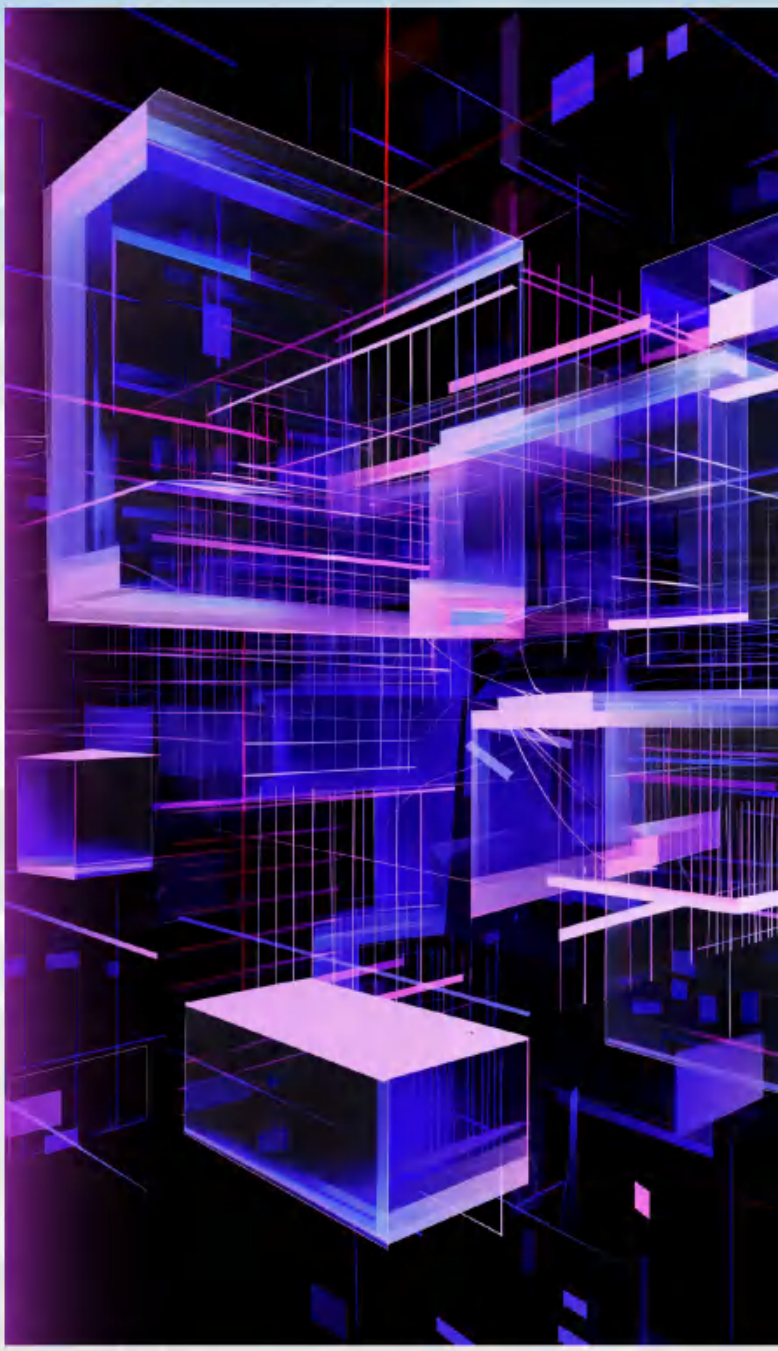


来源：BAAI

2024 年，海外头部模型厂商积极布局原生多模态模型，在性能泛化上也得到初步证明。2024 年 5 月，OpenAI 发布了新一代原生多模态基础模型 GPT-4o，这款模型的创新之处在于放弃了 GPT-4 等前代模型使用独立神经网络处理不同输入数据的做法，采用单一统一的神经网络来处理所有输入，这一创新使得 GPT-4o 在多模态融合能力显著提升，OpenAI 团队称其为首个原生多模态模型。同月，Meta 团队发布原生多模态大模型 Chameleon，模型同样采用了统一的 Transformer 架构，使用 10 万亿 token 文本、图像和代码混合模态数据完成训练，34B 参数模型性能接近 GPT-4V，并且同时生成两种模态。12 月，OpenAI 发布 o1 正式版，更侧重复杂问题的解决和更强大的推理能力，在 STEM 方面表现出色，尤其是科学、编程、数学模型等方面能力更为突出；同月，Google 发布原生多模态大模型 Gemini 2.0，支持图像、视频、音

频等多模态输入和输出，可调用 Google 原生的代码、搜索以及第三方工具。

相较于海外原生多模态大模型的如火如荼，国内原生多模态大模型目前处于探索阶段。2024 年 9 月，北京智源人工智能研究院发布完全自研的自回归原生多模态大模型 Emu3-8B，成为国内首发全自研原生多模态大模型。



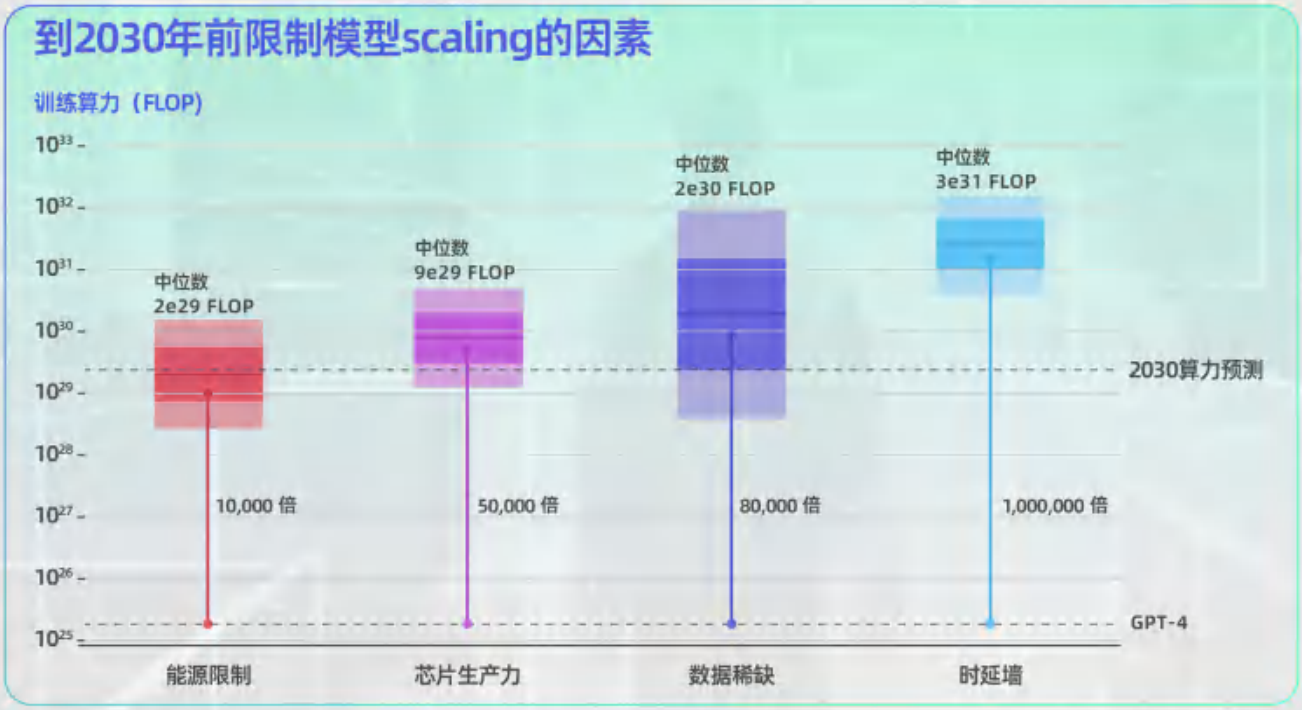
李沐

前亚马逊首席科学家

目前存在一种趋势，即多模态。现如今，多模态技术的发展趋势在于整合不同类型的模态信息。

Scaling Law扩展:
RL + LLMs，模型泛化从预训练向后训练、推理迁移

- Scaling Law(规模定律)作为大模型领域的"第一性原理",主要强调模型性能与模型参数量、数据量和计算量的正相关关系,在 GPT-4、Claude 3.5 等基础模型训练中发挥了不可或缺的指引作用。
- 时至 2024 年末,通过基于 Chinchilla 或 OpenAI Scaling Law,扩大模型参数量和数据量带来的模型性能提升已有所放缓。尽管根据 Epoch AI 对电力、芯片、数据获取及处理能力等预训练关键要素的增长空间测算,预训练 Scaling Law 仍在生效,海外头部厂商也仍在大力投入超大规模集群的建设,我们还可期待在 2025 年看到下一代基础模型的到来。但不得不承认的是,由于预训练 Scaling Law 亚线性的幂律关系客观存在,通过预训练实现模型性能提升的门槛在不断加高,距离 GPT-4 发布已过去近两年时间。

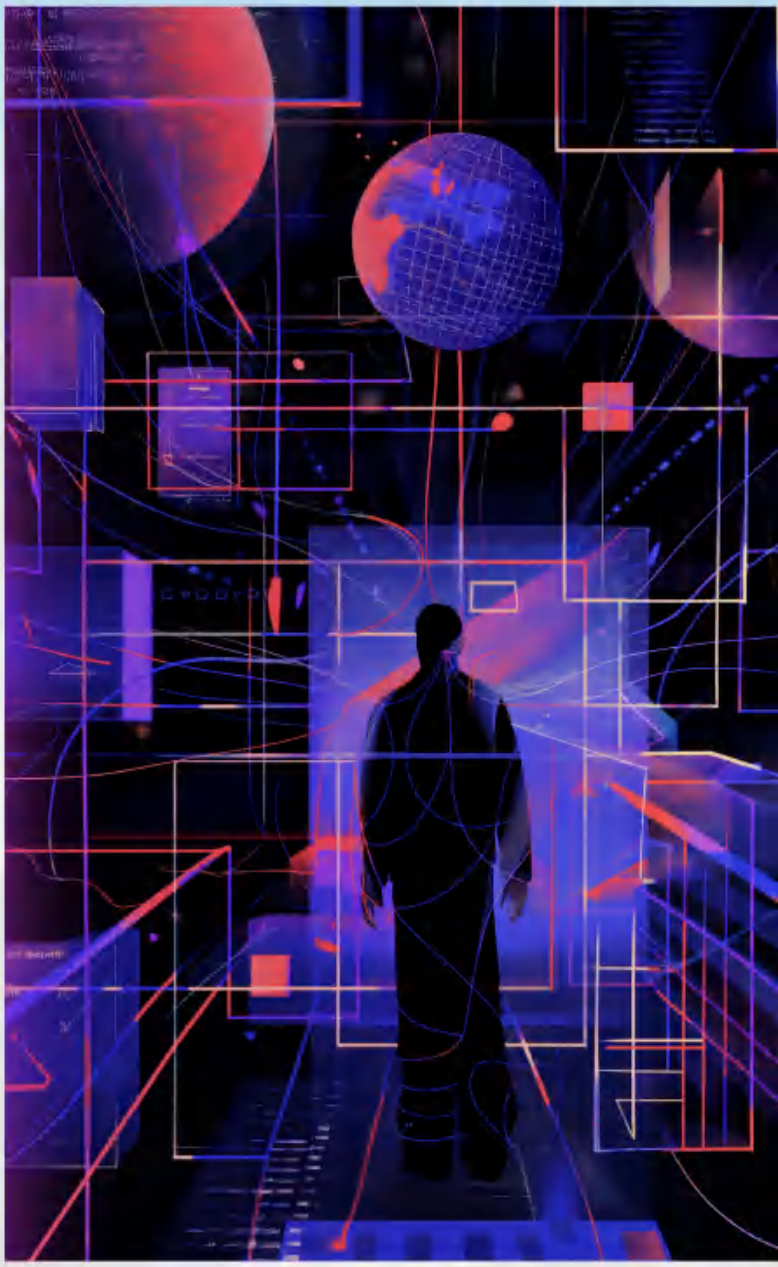


来源：Epoch AI

以 OpenAI o1 的发布为标志，Scaling Law 扩展到后训练、推理等其他阶段。大模型训练的共识逐渐从"资源获取"转向"资源分配",算力和数据从预训练向包括微调、对齐在内的后训练以及推理阶段倾斜。而在 Scaling Law 迎来扩展的过程中,强化学习在其中所起的重要作用格外凸显。GPT-4 时代,RLHF(基于人类反馈的强化学习）的进展已彰显强化学习对提升模型实用性的关键作用。在 Test-Time Compute（推理计算时）等新 Scaling Law 路径获得突破的当下,强化学习的思想正被应用到后训练、推理等更多阶段。

比如，OpenAI 发布的 o1 和 o3 正是通过利用强化学习在训练和推理时的规模定律，提高找到最佳推理路径的可能性和效果。在该趋势的推动下，国内如 Moonshot 将强化学习技术应用于搜索场景，并发布以逻辑思考和深度推理为核心功能的数学模型 K0-Math；DeepSeek 使用强化学习训练，充分挖掘和激活模型潜力，发布 DeepSeek R1 模型，探索释放长思维链潜力；蚂蚁技术研究院设立了强化学习实验室，也围绕该方向展开探索。

整体来说，在即将到来的 2025 年，我们会看到 Scaling Law 作为大模型训练的黄金经验法则，往模型训推的全流程，往特定的行业场景的不断被再次发现。在这过程中，强化学习作为发现后训练、推理阶段的 Scaling Law 的关键技术，也将会得到更多的应用和创新使用。



张宏江

北京智源人工智能研究院创始理事长
美国国家工程院外籍院士

即使在 pre-training(预训练)中有放缓趋势，但 GPT-o1 的发布，让我们看到另外一个天地，就是相对于预训练模型的‘快思考’模式，推理模型 o1 可以给更多的思考时间，Scaling Law 的推理性能已出现‘拐点’，有一个指数级增长。

世界模型加速发布，有望成为多模态大模型的下一阶段

- 通过构建对外部世界的模拟，AI 系统能够完成对世界的内部表征，在复杂多变的环境中实现更为精准的决策与预测。作为赋予 AI 更高级别的认知、适应和决策能力的技术，世界模型不仅能推动 AI 在自动驾驶、机器人控制及智能制造等前沿领域的深度应用，更有望突破传统的任务边界，探索人机交互的新可能。
- 至于世界模型的范式演变，目前仍处于早期阶段。一方面，随着 Sora、Genie 的发布，大模型显露出其蕴含常识的潜力，从语言、图像，到视频、3D 数据，刻画世界的数据角度越全面，离世界纷繁的运行法则就越近；另一方面，JEPA 对信息的高度抽象，又直指每个事物的核心特征，暗合了客观规律的简洁性。与此同时，多模态大模型的推理能力延展到三维空间以来，空间智能与其发生交汇，激发了机器智能对真实世界里更复杂场景的理解与交互的新灵感。关于如何构建世界模型的路线之争无疑将在 2025 年持续，或许随着不同路线的性能泛化程度在新的一年出现分野，我们会看到世界模型技术路线的收敛。

2024 年 12 月，国外世界模型成果发布频频。Google 发布 Genie 2，能够基于单张照片或文字描述，快速生成复杂、可交互的虚拟 3D 环境；World Labs 发布了能从单张图片生成可交互 3D 世界的 AI 系统，且可以用键鼠自由控制视角；Meta 随即推出导航世界模型 NWM，能从单张图像生成连续一致的视频，基于该模型，智能体能够根据过去的观测和导航动作预测未来的视觉观测，从而实现在各类环境中自主导航；之后不久，纽约大学、耶鲁大学、斯坦福大学的联合研究团队发布成果，在多模态大模型内部发现了其空间推理能力的潜力和短板。该研究将多模态大模型、空间智能在推理层面统一了起来，有助于更完整的世界模型的构建。

国内方面，2024 年 10 月、12 月，北京智源人工智能研究院推出并开源了国内首个原生多模态世界模型 Emu3、首个利用大规模无标注的互联网视频学习的 3D 生成模型 See3D。值得一提的是，Emu3 只基于下一个 token 预测，无需扩散模型

或组合式方法，便能把图像、文本和视频编码为一个离散空间，在多模态混合序列上从头开始联合训练一个 Transformer，展现了其在大规模训练和推理上的潜力，实现了视频、图像、文本三种模态的统一理解与生成。See3D 则通过视频中的多视图信息，让模型像人类一样，学习并推理物理世界的三维结构，而非直接建模其几何形态。

在世界模型仍处于性能泛化尚未充分验证，而资源投入已居高不下的当前，如何平衡商业变现压力和前沿技术投入，将是国内外 AI 厂商和机构在新一年里在世界模型方向的重要议题。

李飞飞

斯坦福大学教授

空间智能是视觉智能的未来方向，解决空间智能问题将是迈向全面智能的基础性和关键性一步。3D 空间智能将改变生活，在 2025 年，空间智能的界限很可能会再次突破。

合成数据将成为大模型迭代与应用落地的重要催化剂

- 高质量数据将成为大模型进一步 Scaling up 的发展阻碍。合成数据已经成为基础模型厂商补充数据的首选。根据 Epoch AI 报告，在 2026 年以前，AI 训练将用尽互联网上包含音视频在内的高质量数据，而现存真实世界数据集或将在 2030 年至 2060 年之间耗尽。合成数据已经成为基础模型厂商补充数据的首选。在大模型训练方面，合成数据可以降低人工治理和标注的成本，缓解对真实数据的依赖，不再涉及数据隐私问题；同时，合成数据可以提升数据的多样性，有助于提高模型处理长文本和复杂问题的能力。在大模型产业化方面，合成数据可以缓解通用数据被大厂垄断，专有数据存在获取成本等问题，促进大模型的应用落地。随着真实数据的耗尽，合成数据在模型训练的占比将持续提高，成为大模型性能迭代与应用落地的重要催化剂。

2024 年 12 月，微软发布语言模型 Phi-4，该模型使用了不少于 50 个合成数据集来训练，共约 4000 亿 Token，该模型在 GPQA and MATH 两个 BenchMark 上击败了 GPT-4o 和 Llama3.3，Phi-4 的参数规模是 Llama3.3 的五分之一，但性能却高于后者 5%。

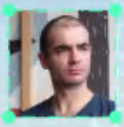
OpenAI 最新发布的 o1 大模型在复杂推理能力上显著提升，研发团队相应设置了对思维链（CoT）输出结果的欺骗性检测，该方案利用 ChatGPT 合成提问数据，评估并监测 o1 模型的回复是否有意或无意地忽略重点事实和人类要求。

根据 Semianalysis 数据，Anthropic 在多个环节利用合成数据训练模型，Claude 3.5 Opus 模

型训练完成后并不急于发布，而主要用于内部数据合成以及强化学习奖励建模推进包括 Claude 3.5 Sonnet 在内的大模型训练。

2024 年 6 月及 8 月，智源研究院推出千万级指令微调数据集 Infinity-Instruct 及迭代版本，该数据集 50% 以上均为合成数据。Opencompass 测试结果显示，经过在 Infinity-Instruct-7M 数据集上的微调，Llama3.1-70B、Mistral-7B-v0.1 综合能力评价可基本对齐官方自己发布的对话模型。

2024 年 12 月，清华、智谱 AI 团队利用文本语料库提取 6000 亿文本 - 语音合成数据，将预训练扩展到 1 万亿个 token，在语音语言建模和口语问题解答方面取得了 SOTA，将语音问答任务方面的性能从之前的 13%（Moshi）提高到 31%。



Ilya Sutskever

OpenAI前首席科学家

正如我们所知的那样，预训练毫无疑问将会终结，与此同时我们也不会再有更多数据了。原因在于，我们只有一个互联网，训练模型需要的海量数据即将枯竭，唯有从现有数据中寻找新的突破，AI 才会继续发展。以后的突破点，就在于智能体、合成数据和推理时计算。

推理优化迭代加速，成为AI Native应用落地的必要条件

随着大模型在各类生成任务上的表现愈发突出，其应用外延持续扩展，催生出各类人工智能应用。大模型硬件载体也从云端向手机、PC等端侧硬件渗透，在这些资源受限（AI算力、内存等）的设备上，大模型的落地应用会面临较大的推理侧的开销限制，对部署资源、用户体验、经济成本等均带来巨大挑战。在此背景下，模型推理优化技术日益成为产研侧关注重点。

对该领域的研究大致可分为算法加速和硬件优化两个方向。前者多集中在数据层、模型层和系统层三个维度，通过对输入提示词、输出内容的优化，模型结构及压缩技术的设计，推理引擎和服务系统的升级，来降低模型推理过程中的计算开销、访存开销、存储开销。目前，以模型量化、知识蒸馏、模型稀疏等为代表的技术已大量应用，并初步取得成效。后续如何继续在保障输出序列长度和输出质量的基础上，降低推理开销成为持续提升的关键方向；后者则关注硬件端加速，针对大模型在推理过程中自回归的序列生成方式，专门设计芯片方案，带来显著的推理速度收益。

当前国内外厂商围绕长文本、复杂交互、边缘部署等应用场景，持续推动推理优化技术迭代，以在成本开销与终端用户体验侧寻求最佳平衡点。Meta与麻省理工团队通过对模型层的智能化裁剪，在去除多达一半的模型层数下，依然维持了问答基准测试性能；微软推出的BitNet架构使用"BitLinear"层替代标准线性层，通过降低参数精度的方式，在性能具备竞争力的前提下，显著节省内存消耗。无问芯穹发布的FlashDecoding++通过对注意力和线性算子的针对性优化和计算图层面的深度算子融合技术，大幅提高大语言模型推理效率；潞晨科技推出的Colossal-Inference推理引擎通过张量并行、分块式KV缓存、KV缓存量化、分页注意力算法等优化技术实现推理速度的提升和计算资源的有效利用。硬件加速方面，Cerebras设计的Wafer-Scale Engine（WSE）将计算单元和内存单元高度集成，其第三代WSE相比英伟达H100可获得获得数千倍的带宽速度提升。

Jeff Dean
谷歌首席科学家

在机器学习推理领域，降低成本和延迟是一个核心挑战，这直接关系到先进模型能否惠及更多用户。

重塑产品应用形态，Agentic AI成为产品落地的重要模式

2025年，更通用、更自主的智能体将重塑产品应用形态，进一步深入工作与生活场景，成为大模型产品落地的重要应用形态。

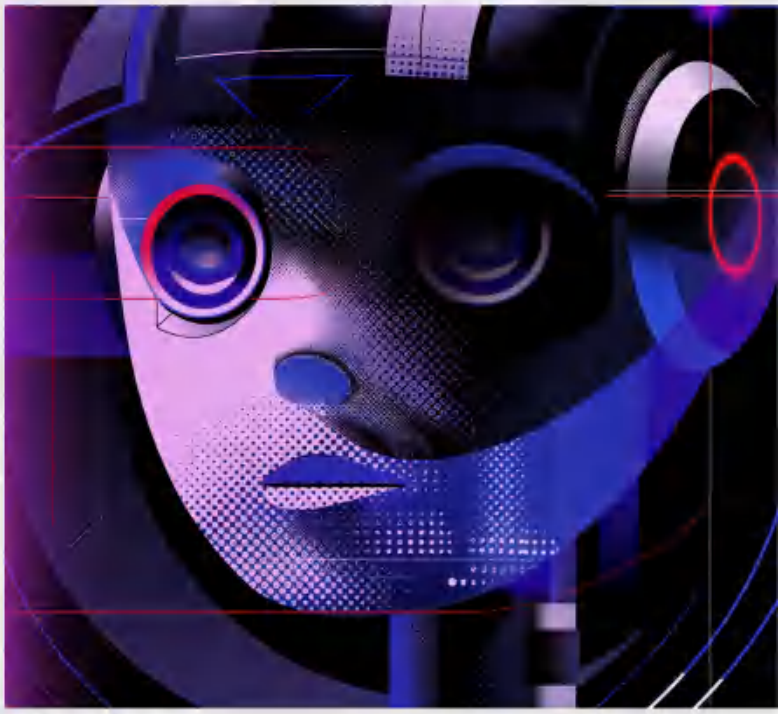
从 Chatbot、Copilot 到 AI Agent、Agentic AI，2023 年以来行业对于 AI 应用形态的理解越发深入。

2024 年，OpenAI、Anthropic 等头部模型厂商积极布局智能体产品与技术；全球已出现 300 多家智能体初创公司。微软的研究显示，近 70% 的财富 500 强企业员工已开始使用 AI 工具处理繁琐任务，如筛选电子邮件、记录会议纪要等。

在理论发展方面，2023 年 12 月，OpenAI 提出了在有限直接监督下，长时间自主行动以实现计划目标的系统，“Agentic AI Systems”，并提出了评估该系统“Agenticness”程度的四个指标；2024 年 3 月，吴恩达在红杉资本（Sequoia Capital）的人工智能峰会（AI Ascent）上进一步阐释了“Agentic”是对智能体智能程度的描述。2024 年 6 月，吴恩达进一步提出“Agentic workflow”是构建适应性更强智能体的重要方法，健全了构建智能体的理论体系。至此，业内对智能体的术语使用，更多地从 AI Agent 迁移到 Agentic AI，其背后标志着从判断产品是否属于 Agent，到探讨产品的智能化程度这一更有落地意义的转变。

在技术发展方面，2024 年 10 月，Anthropic 发布能够解读计算机屏幕信息、自主操作软件和实时浏览互联网的 Computer Use，同时发布帮助大模型系统快速接入多种外部数据资源的上下文协议 MCP。智谱 AI 推出可以执行超 50 步复杂操作，且支持跨应用执行任务的 AutoGLM 升级版，以及可操作桌面和自主处理文档、浏览网页的智能体 GLM-PC。OpenAI 也计划将在 2025 年 1 月发布可以独立浏览网络并完成如旅行预订等在线交易的智能体“Operator”。国内外头部模型厂商在构建更通用、更自主智能体的探索与尝试已蔚然成风。

从更强调产品概念的 Agent，到更强调应用智能程度的 Agentic AI，我们在 2025 年将看到更多智能化程度更高、对业务流程理解更深的多智能体系统在应用侧的落地。



吴恩达

deeplearning.ai（AI教育平台）创始人
百度前首席科学家
Coursera的现任董事长兼联合创始人

LLM 正在从主要优化消费级问答体验，转向优化支持智能体工作流（如工具使用、计算机操作、多智能体协作等），这将大幅提升智能体的工作性能！

AI应用热度渐起，Super App花落谁家犹未可知

近一年时间，生成式模型在图像、视频侧的处理能力得到大幅提升，叠加推理优化带来的降本，Agent/RAG 框架、应用编排工具等技术的持续发展，为 AI 超级应用的落地积基树本。大模型应用从功能点升级，渗透到 AI 原生的应用构建及 AI OS 的生态重塑。尽管从用户规模、交互频次、停留时长等维度来看，C 端 AI 应用仍未出现爆发式增长，但超级应用的可能形态或已初现端倪：

终端设备厂商基于硬件设备重构 AI OS 生态，基础模型及垂直应用赛道厂商深度结合大模型能力打造 AI APP。

AI OS 方面，苹果在 10 月正式发布 Apple Intelligence，从系统层级对手机应用进行重构，覆盖 AI 写作、照片处理及语音助手等功能，得益于其软硬生态的强耦合，有望深度整合系统级体验，带来交互形态的再升级。AI APP 方面，以 ChatBot、生活服务为代表的 AI 应用经过 1 年多时间的业务验证，已有大量产品落地。Chat 类如 OpenAI 发布的 ChatGPT，月活接近 6 亿，年预估收入约 37 亿美元；国内方面字节跳动的人工智能应用豆包处于头部，达到了 7116 万月度活跃用户数（截至 2024 年 12 月，数据来自 AI 产品榜），百度发布的文小言，Moonshot 发布的 Kimi 分列其后。生活服务类如蚂蚁集团发布的系列个人管家产品，包括生活管家支小宝、金融管家蚂小财、AI 健康管家等，可根据用户习惯和使用场景，智能推荐专属服务。

虽然 Super APP 花落谁家尚未尘埃落定，但从用户规模、交互频次、停留时长等维度来看，AI 应用热度持续攀升，已到应用爆发的黎明前夕。

李彦宏

百度公司创始人
董事长兼首席执行官

在移动互联网时代，出现了许多用户量达数亿甚至十亿的超级应用，然而在 AI 时代，这样的超级应用尚未出现。无论是在美国、欧洲还是中国，都正在探索能够发挥生成式 AI 能力、且能吸引数十亿人使用的应用形态。

模型能力提升与风险预防并重，AI安全治理体系持续完善

- 作为复杂系统，大模型的 Scaling 带来了涌现，但复杂系统特有的涌现结果不可预测、循环反馈等特有属性也对传统工程的安全防护机制带来了挑战。基础模型在自主决策上的持续进步带来了潜在的失控风险，如何引入新的技术监管方法，如何在人工监管上平衡行业发展和风险管控?这对参与 AI 的各方来说，都是一个值得持续探讨的议题。
- 与此同时，在信息传播速度日益加快的当下，由 AI 系统引发的偏见、深度伪造、隐私泄露、版权争议问题丛生，社会对 AI 安全的关注度急剧上升。近年来，各个国家、组织在 AI 安全上持续投入，并进行了技术研究、治理框架、国际合作等多种形式的探索，后续有望构建起与智能水平相匹配、合乎伦理、可靠，可控和尊重知识产权的 AI 安全治理体系。

2024 年 5 月，OpenAI 在第二届全球人工智能安全峰会上公布公司正在实施的 10 大 AI 安全措施。6 月，Google 发布了 SAIF（Secure AI Framework）安全 AI 框架旨在帮助减轻 AI 系统特定的风险，如窃取模型、训练数据污染、注入恶意信息和提取训练数据中的机密信息等，确保组织能够负责任地部署人工智能技术。10 月，Anthropic 更新其制定的《安全责任扩展政策（RSP）》以构建一种灵活的动态 AI 风险治理框架。

国内方面，2024 年 4 月，联合国科技大会发布了两项大模型安全标准，其中《大语言模型安全测试方法》由蚂蚁集团牵头。该标准率先给出了四种不同攻击强度的攻击手法分类标准，提供了严格的评估指标和测试程序等，为大模型本身的安全性评估提供了一套全面、严谨且实操性强的结构性方案。此外，蚂蚁集团自研的大模型安全一体化解决方案“蚁天鉴”，旨在打造 AI 大模型的安全铠甲，确保大模型技术在安全可靠的环境中发挥效能。目前，蚁天鉴的检测与防御产品已开放给 20 家外部机构和企业使用，为通用大模型及医疗、金融、政务等垂直领域行业大模型应用安全保驾护

航。同月，华为提出 L4 级 AI 安全智能体，针对恶意软件变异快、加密攻击难发现、海量事件难处置等问题，通过大模型和图 AI 等新技术实现自主防御新型攻击和全网自动化运营。北京智源人工智能研究院持续推进 AI 安全底层关键技术研究，并提出泛化的 AI 防御大模型和 AI 监管大模型；同时，智源研究院积极锻造 AI 安全中国力量，组织或参与 AI 安全国际合作：2024 年 3 月，发起并承办我国首个 AI 安全国际对话高端闭门论坛，与全球 AI 领袖学者及产业专家联合签署《北京 AI 安全国际共识》；9 月，参与第三届国际 AI 安全对话，签署《AI 安全国际对话威尼斯共识》；同月，参与筹备中国 AI 安全网络，将在国际安全会议上发出中国声音；10 月，与英国 AISI（AI 安全研究所）建立沟通，商讨《前沿 AI 风险承诺》，积极参与国际 AI 开发者社区安全讨论；11 月，联合多家高校及科研院所，推进迭代新版本中国 AI 安全治理框架，共建中国 AI 安全方案。

翁荔

OpenAI前研究副总裁（安全）

在西游记中，孙悟空有紧箍咒约束行为，我们应该给 AI 模型也带上紧箍咒，也就是教会 AI 安全基本准则约束和道德标准，让其遵守行为规范，以人类利益为先，成为我们贴心的伙伴，而不是冰冷的机器人。

参考文献

[1] Zhang, Qiang et al. "Scientific Large Language Models: A Survey on Biological & Chemical Domains." ArXiv abs/2401.14656 (2024): n. pag.

[2] Lei, Ge et al. "Materials science in the era of large language models: a perspective." ArXiv abs/2403.06949 (2024): n. pag.

[3] Yu, Botao et al. "LlaSMol: Advancing Large Language Models for Chemistry with a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset." ArXiv abs/2402.09391 (2024): n. pag.

[4] Abramson, J., Adler, J., Dunger, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024). https://-doi.org/10.1038/s41586-024-07487-w.

[5] Zhao, M., Wang, N., Jiang, X. et al. An integrative data-driven model simulating C. elegans brain, body and environment interactions. Nat Comput Sci 4, 978–990 (2024). https://-doi.org/10.1038/s43588-024-00738-w.

[6] Brohan, Anthony et al. "RT-1: Robotics Transformer for Real-World Control at Scale." ArXiv abs/2212.06817 (2022): n. pag.

[7] Brohan, Anthony et al. "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control." Conference on Robot Learning (2023).

[8] Padalkar, Abhishek et al. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models." ArXiv abs/2310.08864 (2023): n. pag.

[9] Huang, Wenlong et al. "VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models." ArXiv abs/2307.05973 (2023): n. pag.

[10] Team, Chameleon. "Chameleon: Mixed-Modal Early-Fusion Foundation Models." ArXiv abs/2405.09818 (2024): n. pag.

[11] Wang, Xinlong et al. "Emu3: Next-Token Prediction is All You Need." ArXiv abs/2409.18869 (2024): n. pag.

[12] Bi, DeepSeek-AI Xiao et al. "DeepSeek LLM: Scaling Open-Source Language Models with Longtermism." ArXiv abs/2401.02954 (2024): n. pag.

[13] Liu, Yixin et al. "Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models." ArXiv abs/2402.17177 (2024): n. pag.
Bruce, Jake et al. "Genie: Generative Interactive Environments." ArXiv

[14] Bruce, Jake et al. "Genie: Generative Interactive Environments." ArXiv abs/2402.15391 (2024): n. pag.

[15] Bar, Amir et al. "Navigation World Models." (2024).

[16] Achiam, OpenAI Josh et al. "GPT-4 Technical Report." (2023).

[17] Yang, Jihan et al. "Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces." (2024).

[18] Ma, Baorui et al. "You See it, You Got it: Learning 3D Creation on Pose-Free Videos at Scale." (2024).

[19] Villalobos, Pablo et al. "Will we run out of data? Limits of LLM scaling based on human-generated data." (2022).

[20] Abdin, Marah et al. "Phi-4 Technical Report." (2024).

参考文献

[21] Assran, Mahmoud, et al. "Self-supervised learning from images with a joint-embedding predictive architecture." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[22] Yang, Jihan et al. "Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces." (2024).

[23] Zhou, Zixuan et al. "A Survey on Efficient Inference for Large Language Models." ArXiv abs/2404.14294 (2024): n. pag.

[24] Epoch AI (2023), "Key Trends and Figures in Machine Learning". Published online at epoch.ai. Retrieved from: 'https://epoch.ai/trends' [online resource].

[25] Jaime Sevilla et al. (2024), "Can AI Scaling Continue Through 2030?". Published online at epoch.ai. Retrieved from: 'https://epoch.ai/blog/can-ai-scaling-continue-through-2030' [online resource].

[26] Gromov, Andrey et al. "The Unreasonable Ineffectiveness of the Deeper Layers." ArXiv abs/2403.17887 (2024): n. pag.

[27] Wang, Hongyu et al. "BitNet: Scaling 1-bit Transformers for Large Language Models." ArXiv abs/2310.11453 (2023): n. pag.

[28] Hong, Ke et al. "FlashDecoding++: Faster Large Language Model Inference on GPUs." ArXiv abs/2311.01282 (2023): n. pag.

[29] Liu, Xiao et al. "AutoGLM: Autonomous Foundation Agents for GUIs." ArXiv abs/2411.00820 (2024): n. pag.

[30] Gunter, Tom et al. "Apple Intelligence Foundation Language Models." ArXiv abs/2407.21075 (2024): n. pag.

[31] BAAI 行研组，具身智能软硬件及应用研究报告.

[32] BAAI 行研组，2023 年 LLM-based Agent 行业研究报告.

[33] OpenAI o1 system card, https://cdn.openai.com/o1-system-card-20241205.pdf.

[34] OpenAI GPT-4o system card, https://cdn.openai.com/gpt-4o-system-card.pdf.

[35] Google, Secure AI Framework - SAIF, https://www.saif.google/secure-ai-framework.

[36] Anthropic, Anthropic's Responsible Scaling Policy, October 15, 2024, https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf.

[37] World Digital Technology Academy, Large Language Model Security Testing Method, blob:https://wdtacademy-.org/31f5f7ee-b4f2-4265-8d5b-027fe649c747.

[38] Semianalysis, Scaling Laws-o1 Pro Atchitecture, Reasoning Training Infrastructure, Orion and Claude 3.5 Opus "Failures", https://semianalysis.com/2024/12/11/scaling-laws-o1-pro-architecture-reasoning-training-infrastructure-orion-and-claude-3-5-opus-failures/.

[39] Google DeepMind, Genie 2: A large-scale foundation world model, https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/.

[40] 王鹤，三维视觉小模型 + 基础大模型实现具身智能, https://www.techwalker.com/2024/0401/3156920.shtml.

参考文献

[41] 仉尚航，具身基础模型最终状态可能是“4D 世界模型”，wz.my/nzt9l.

[42] Air Street Capital, State of AI Report, https://www.stateof.ai/ .

[43] 腾讯研究院，为什么这家公司的芯片推理速度比英伟达快 20 倍? https://www.tisi.org/30434/.

[44] Dan H, TW123, Complex Systems for AI Safety, https://www.alignmentforum.org/s/FaEBwhhe3otzYKGQt/p/n767Q8HqbrteaPA25#Resources_on_Complex_Systems.

[45] 潞晨科技, Colossal-Inference, https://www.luchentech.com/#inference

[46] Cerebras, product-chip, https://cerebras.ai/product-chip/.

[47] 微软 Ignite 2024 技术大会, https://is.gd/HcKSBk

[48] Anthropic, Introducing the Model Context Protocol, https://www.anthropic.com/news/model-context-protocol

[49] Bloomberg, OpenAI Nears Launch of AI Agent Tool to Automate Tasks for Users, https://www.bloomberg.com/news/articles/2024-11-13/openai-nears-launch-of-ai-agents-to-automate-tasks-for-users.

[50] CNBC, OpenAI gets new $1.5 billion investment from SoftBank, allowing employees to sell shares in a tender offer, https://www.cnbc.com/2024/11/26/openai-gets-1point5-billion-investment-from-softbank-in-tender-offer.html.

[51] AI 产品榜, https://dnipkggqxh.feishu.cn/wiki/YTIUwM6Vmij4IQkSm9PctPWunIb.

编写委员会：

指导组： 黄铁军 王仲远 林咏华 曹岗

编写组： 倪贤豪 靳虹博 陈泓伊 殷靖东

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」