DEEP RESEARCH · EVIDENCE-AUDITED EDITION

人工智能全景与前沿深度调研报告

分支 · 模型 · Agent · 世界模型 · 具身智能 · AI4Science

基础设施 · 行业 · 治理 · 人物思想与最新动向

阅读方式

第 2 页为可点击目录；点击条目可跳转，PDF 侧边栏同时提供章节书签。

事实状态分为：一手来源证实 / 独立报道 / 机构自述 / 有限证据 / 推断与争议。

人物职位、模型状态与法规信息截至 2026-08-23；高时效事项请回到页面来源复核。

信息口径：2026-08-23 · 中文 · 62 页 · AI 辅助研究与多代理交叉核验

目录

点击任一标题跳转到对应页面；侧边栏书签可用于快速浏览全部章节。

执行摘要：AI 正进入完整系统竞争 3
一张图读懂：系统栈与未单列的经典分支 4
容易被大模型叙事遮蔽的八条路线 5
研究方法、证据等级与局限 6
从符号系统到基础模型：七十年主线 7
学习范式矩阵：不要把模型和训练方式混为一谈 8
监督、无监督与自监督：表示学习的地基 9
深度网络的结构谱系 10
强化学习：从奖励到规划 11
因果、神经符号与可验证推理 12
持续学习、元学习与联邦学习 13
基础模型：从单任务模型到通用底座 14
LLM 的完整技术栈 15
预训练：数据、token、目标与规模 16
后训练：把“会续写”变成“会协作” 17
推理时计算：让模型在回答前搜索和验证 18
Agent：从回答问题到完成任务 19
记忆、RAG 与上下文工程 20
原生多模态：共同表示与跨模态行动 21
生成式 AI：图像、视频、音频与 3D 22
世界模型：预测环境与行动后果 23
空间智能：李飞飞与 World Labs … 24
VLA：把视觉、语言和动作接在一起 25
机器人学习的真正瓶颈 26
自动驾驶、无人机与移动智能体 27
AI4Science：从预测到闭环发现 28
生物、蛋白质、基因组与药物 29
材料、化学与逆向设计 30
天气、气候、地球与物理 31
数学、算法与自主实验 32
脑启发与神经形态计算 33
效率：蒸馏、量化、稀疏与边缘 AI 34
芯片与 AI 计算栈 35
分布式训练与推理软件 36
数据中心、能源与物理约束 37
开放权重、开源与封闭模型 38
评测：排行榜为什么不够 39
安全、对齐与系统可靠性 40
企业 AI 参考架构：从模型到可审计工作流 41
怎样读 28 个行业案例：案例不是平均效果 42
医疗与制药：最有价值，也最不能跳过验证 43
金融、保险与法律：权限和责任比文案质量更重要 44
制造、供应链与工业机器人 45
软件工程与网络安全：生产率证据为何相反 46
教育、研究与知识工作 47
传媒、零售与物流：内容生成只是入口 48
能源、农业、公共部门与国防边界 49
ROI、劳动与组织：为什么高采用不等于高价值 50
全球格局：美国、中国、欧盟与关键供应链 51
美国治理：创新、采购、行业法与州法并存 52
欧盟 AI Act：已进入执行期 53
中国治理：分层规则与综合立法推进 54
国际治理、标准与来源证明 55
人物图谱 I：奠基者与不同智能观 56
人物图谱 II：前沿实验室与 AGI 平台 57
人物图谱 III：世界模型、机器人与科学 58
人物图谱 IV：安全、开放生态、中国路线与… 59
人物图谱 V：2026 实验室领导层与区域… 60
2026—2030 趋势雷达与核心术语 61
最终判断、核查清单与 AI 披露 62

执行摘要：AI 正进入完整系统竞争

大语言模型仍是通用认知接口，但决定价值的单位已经从“模型”变成“模型 + 工具 + 数据 + 环境反馈 + 治理”。

1. 技术重心

预训练规模仍重要，但后训练、推理时计算、验证器、工具调用、长期记忆和多智能体编排正在成为能力增量的重要来源。2026 AI Index 显示产业界生产了 2025 年超过九成的显著前沿模型，组织采用率达到 88%；同时评测、透明度和治理并未同步成熟。[S01][S02]

2. 从语言到世界

原生多模态把文本、图像、视频、音频和动作放入共同表示；世界模型进一步学习环境如何变化以及行动会产生什么后果。V-JEPA 2、Genie 3、World Labs 的空间智能路线，以及 VLA 机器人模型，代表“感知—预测—规划—行动”的会合。[S11][S12][S13][S14]

3. AI4Science

科学 AI 正从预测单一结构或性质，转向生成候选、调用工具、运行模拟或实验、用可计算目标评价并继续迭代的闭环。AlphaFold、GNoME、MatterGen 与 AlphaEvolve说明这条路径已覆盖生物、材料和算法，但真实实验、因果验证与可重复性仍是瓶颈。[S18][S20][S21][S22]

4. 最大的不确定性

benchmark 提升不能直接推出真实世界可靠性；Agent 获得权限后，错误会从“答错”升级为“做错”。未来三到五年的核心竞争将围绕独有数据闭环、可信评测、软硬件协同、工作流入口和监管准入，而不是单一模型排行榜。[S02][S24]

本页主要来源

[S01] Stanford HAI（2026），The 2026 AI Index Report。https://hai.stanford.edu/ai-index/2026-ai-index-report
[S02] International AI Safety Report（2026），International AI Safety Report 2026。
https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S11] Meta AI（2025），Introducing V-JEPA 2。https://ai.meta.com/research/vjepa/
[S12] Google DeepMind（2025），Genie 3: A new frontier for world models。
https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/
[S13] World Labs（2026），About World Labs。https://www.worldlabs.ai/about
[S14] Google DeepMind（2025），Gemini Robotics。https://deepmind.google/models/gemini-robotics/
[S18] Nature / Google DeepMind（2021），Highly accurate protein structure prediction with AlphaFold。
https://www.nature.com/articles/s41586-021-03819-2
[S20] Nature / Google DeepMind（2023），Scaling deep learning for materials discovery。
https://www.nature.com/articles/s41586-023-06735-9
[S21] Nature / Microsoft Research（2025），A generative model for inorganic materials design。
https://www.nature.com/articles/s41586-025-08628-5
[S22] Google DeepMind（2025），AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms。
https://deepmind.google/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/
[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework

一张图读懂：系统栈与未单列的经典分支

术语之所以显得混乱，是因为它们处在不同层级；本页同时标出未逐章展开、但仍属于 AI 全景的经典方向。

关键原则

“LLM、Agent、世界模型、具身智能”不是互斥类别。视觉、语音、经典 NLP、知识表示、规划、推荐、时序、概率模型与 HCI 也没有消失，而是被基础模型重组或继续作为系统模块。本报告对这些经典分支提供定位，不宣称逐一达到教材级展开。[S40]

层级	回答的问题	典型方向	主要瓶颈
目标与价值	系统为什么存在？	产品、科研、公共服务、自动化	目标错配、责任与分配
应用与工作流	如何嵌入现实任务？	医疗、金融、制造、教育、软件	流程重构、数据权、监管
智能系统	如何持续完成多步任务？	Agent、RAG/记忆、工具、多智能体	可靠性、权限、可观测性
世界与行动	如何预测并改变环境？	世界模型、空间智能、VLA、控制	真实数据、长尾、硬件安全
基础模型	如何形成通用表示与生成能力？	LLM、多模态、生成模型、科学模型	数据、算力、可解释性
学习范式	机器怎样从经验中学习？	监督、自监督、强化、持续、因果	泛化、样本效率、遗忘
经典任务族	主要处理什么信号或决策？	视觉、语音/音频、NLP、知识表示/规划	长尾、歧义、可解释与评测
业务算法族	如何预测、排序和交互？	推荐、时序/异常、概率模型、人机交互	反馈回路、漂移、公平与隐私
系统基础	怎样训练和部署？	芯片、互连、编译器、分布式、推理	成本、能耗、供应链

本页主要来源

[S40] Russell & Norvig（2021），Artificial Intelligence: A Modern Approach。https://aima.cs.berkeley.edu/
[S01] Stanford HAI（2026），The 2026 AI Index Report。https://hai.stanford.edu/ai-index/2026-ai-index-report
[S02] International AI Safety Report（2026），International AI Safety Report 2026。
https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

容易被大模型叙事遮蔽的八条路线

经典 AI 没有消失；很多真实系统仍依赖因果、概率、搜索、端侧计算、隐私技术和人机协作。

组合原则

成熟产品通常把规则、传统机器学习、基础模型、数据库、验证器和人工流程组合起来。选择方法应从任务风险、可验证性、数据形态、延迟和成本出发，而不是先指定“必须用大模型”。[S40][S50][S51]

方向	为什么重要	当前边界
知识表示与规划	显式规则、约束、可审计推理	开放世界维护成本高
因果与概率模型	干预、反事实和不确定性决策	依赖可识别假设与可靠数据
神经符号方法	组合学习能力与形式约束	统一架构和规模化评测不足
持续与在线学习	系统部署后适应新分布	灾难性遗忘、污染和审计
边缘 AI	低延迟、隐私、断网可用	内存、功耗和安全更新
隐私保护学习	降低原始数据集中暴露	隐私预算、侧信道和效用损失
合成数据与仿真	覆盖稀有场景、降低采集成本	仿真偏差和反馈循环
评测科学与人因	把模型分数连接到真实任务与责任	动态任务、长期效果和行为适应难测

本页主要来源

[S40] Russell & Norvig（2021），Artificial Intelligence: A Modern Approach。https://aima.cs.berkeley.edu/
[S50] Google（2017），Federated Learning: Communication-Efficient Learning of Deep Networks。https://proceedings.mlr.press/v54/mcmahan17a.html
[S51] arXiv（2019），Continual Learning in Neural Networks。https://arxiv.org/abs/1910.02718
[S55] NeurIPS（2018），Neural Ordinary Differential Equations。https://arxiv.org/abs/1806.07366
[S67] ISO（2023），ISO/IEC 42001:2023。https://www.iso.org/standard/42001
[S68] C2PA（2025），C2PA Specification 2.2。https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification

研究方法、证据等级与局限

本报告是截至指定日期的二手证据综合，不是实验研究，也不声称预测具有确定性。

研究问题

截至 2026-08-23，如何用一个互不混淆的层级框架解释 AI 的主要学习范式、模型与系统形态、前沿路线、行业成熟度、治理结构及关键人物的路线选择？

检索与纳入

检索平台覆盖 arXiv、Nature/Science、机构论文页、法规与监管官网、大学/公司公告及模型卡；关键词按 taxonomy、learning paradigm、foundation model、agent、world model、VLA、AI4Science、industry、governance 与人物姓名组合。三支专题研究先汇集数百个候选链接，去重后纳入 109 个可定位到标题、机构和 URL 的核心条目；无法核实的新闻摘要不进入核心结论。历史材料不受日期限制，职位、法规和模型状态重点复核 2025—2026 年来源。

证据标签

A 级：同行评审论文、法规、官方统计；B 级：预印本、模型卡、技术报告；C 级：机构自述、产品案例；D 级：分析性推断。C 级数据只能说明“该机构公开声称”，不能自动外推为行业平均。

核验覆盖与复现

全部正文页均执行页内引文 ID—页底来源一致性检查；人物职位、法律时点与高时效公司变动优先回到本人或机构一手页，并以 Reuters/AP 等独立报道补足可追溯性；企业 ROI 与产品能力保留“机构自述”标签。最终 PDF 逐页渲染检查，并核对页数、文字可提取性、内部目录、书签、外部链接与桌面副本哈希。完整 109 项来源按首次使用分布在各页底部，可直接点击复查。

已知局限

前沿模型不公开完整训练数据与评测细节；机器人演示通常无法覆盖长尾；企业 ROI 存在选择性披露；人物观点会随技术变化；国际比较受汇率、采购和统计口径影响。报告使用公开证据，未访问商业机密或未发布研究，也没有外部人类领域专家签字认证。[S01][S02][S42]

本页主要来源

[S01] Stanford HAI（2026），The 2026 AI Index Report。https://hai.stanford.edu/ai-index/2026-ai-index-report
[S02] International AI Safety Report（2026），International AI Safety Report 2026。https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S42] Stanford CRFM（2025），Foundation Model Transparency Index。https://crfm.stanford.edu/fmti/

从符号系统到基础模型：七十年主线

AI 历史不是一条直线，而是“知识表示、统计学习、算力与数据、交互反馈”四股力量周期性重组。

1956—1980s：规则与知识

早期 AI 把智能视为符号搜索、逻辑推理和显式知识。优势是可解释、可组合；缺点是知识获取成本高、对噪声和开放世界脆弱。专家系统在封闭领域有效，却难以通过规则手工编码常识。[S40]

1990s—2012：统计学习

概率模型、支持向量机、集成学习等把重点从“写规则”转向“从数据估计规律”。互联网提供数据，GPU 与反向传播使深层网络在视觉、语音上突破。

2012—2020：深度表示与规模化

卷积网络、序列模型、注意力和 Transformer 让端到端学习成为主流。Transformer 的并行训练与长程依赖能力，叠加大规模预训练，最终形成 foundation model 范式。[S03][S45]

2020—2026：对齐、推理与行动

大模型从少样本语言接口扩展到多模态、代码、工具和 Agent；能力来源从预训练扩展到人类/AI反馈、强化学习、推理时计算、检索和环境反馈。世界模型与机器人把“生成文本”推进到“预测并改变环境”。[S04][S05][S07][S14]

本页主要来源

[S03] NeurIPS / arXiv（2017），Attention Is All You Need。https://arxiv.org/abs/1706.03762
[S04] OpenAI / NeurIPS（2020），Language Models are Few-Shot Learners。https://arxiv.org/abs/2005.14165
[S05] OpenAI（2022），Training language models to follow instructions with human feedback。https://arxiv.org/abs/2203.02155
[S07] OpenAI（2024），Learning to reason with LLMs。https://openai.com/index/learning-to-reason-with-llms/
[S14] Google DeepMind（2025），Gemini Robotics。https://deepmind.google/models/gemini-robotics/
[S40] Russell & Norvig（2021），Artificial Intelligence: A Modern Approach。https://aima.cs.berkeley.edu/
[S45] OpenAI（2020），Scaling Laws for Neural Language Models。https://arxiv.org/abs/2001.08361

学习范式矩阵：不要把模型和训练方式混为一谈

同一个 Transformer 可以接受监督学习、自监督预训练、偏好优化、强化学习和持续学习；“模型结构”不等于“学习范式”。

组合使用才是现实

典型 LLM 先用自监督学习预测 token，再用监督微调学习指令，通过偏好优化或强化学习改变行为；机器人模型还要加入专家轨迹、仿真数据和真实交互。评估时必须追踪每一阶段的数据与目标。[S05][S06][S16]

范式	反馈信号	适用场景	核心风险
监督学习	输入—标签	分类、预测、检测	标签成本、分布偏移
无监督/自监督	数据内部结构	预训练、表征、生成	目标与下游价值错位
强化学习	奖励与交互	控制、规划、后训练	奖励黑客、样本效率
模仿学习	专家轨迹	机器人、驾驶	覆盖不足、继承专家偏差
迁移/微调	源任务知识	领域适配	灾难性遗忘、负迁移
持续学习	时间序列任务	在线个性化、机器人	稳定—可塑性冲突
联邦学习	分布式本地更新	隐私敏感机构	通信、异质性、泄漏
因果学习	干预/结构假设	政策、科学、稳健决策	可识别性、假设错误

本页主要来源

[S05] OpenAI（2022），Training language models to follow instructions with human feedback。https://arxiv.org/abs/2203.02155
[S06] Stanford / arXiv（2023），Direct Preference Optimization。https://arxiv.org/abs/2305.18290
[S16] Physical Intelligence（2024），π0: A Vision-Language-Action Flow Model for General Robot Control。https://www.physicalintelligence.company/download/pi0.pdf
[S50] Google（2017），Federated Learning: Communication-Efficient Learning of Deep Networks。https://proceedings.mlr.press/v54/mcmahan17a.html
[S51] arXiv（2019），Continual Learning in Neural Networks。https://arxiv.org/abs/1910.02718

监督、无监督与自监督：表示学习的地基

现代基础模型的规模化关键，是把互联网上大量未标注数据转成可训练信号；这并不意味着数据“免费”或没有偏差。

监督学习

模型最小化预测与人工标签之间的误差，适合明确任务和稳定标签。但医学、法律、机器人等领域的高质量标签昂贵且存在专家分歧；训练集标签还可能固化历史偏差。

无监督与自监督

无监督学习寻找聚类、密度或潜在结构；自监督学习则从数据本身构造目标，例如遮蔽词恢复、下一个 token、图像—文本对齐或视频潜在空间预测。它把表示学习扩展到互联网规模，是 LLM 和多模态模型的共同地基。[S03][S10][S11]

生成目标与理解

预测下一个 token 能迫使模型压缩大量语言规律，但“能生成”不自动等于具备稳定世界模型、因果理解或可执行计划。V-JEPA 路线主张在抽象表示空间预测关键变化，减少对像素级细节的浪费；两条路线可能互补，而不是简单二选一。[S11]

数据质量

规模化数据包含重复、版权材料、错误和群体偏差。去重、来源记录、许可、数据配比与合成数据验证会直接影响模型行为。训练数据不透明也是当前 foundation model 透明度的薄弱环节。[S42]

本页主要来源

[S03] NeurIPS / arXiv（2017），Attention Is All You Need。https://arxiv.org/abs/1706.03762
[S10] OpenAI（2021），CLIP: Learning Transferable Visual Models From Natural Language Supervision。https://arxiv.org/abs/2103.00020
[S11] Meta AI（2025），Introducing V-JEPA 2。https://ai.meta.com/research/vjepa/
[S42] Stanford CRFM（2025），Foundation Model Transparency Index。https://crfm.stanford.edu/fmti/

深度网络的结构谱系

架构不是“谁取代谁”的单线进化，而是对数据结构、计算约束和任务需求的不同归纳偏置。

工程现实

许多前沿系统是混合架构：视觉编码器、Transformer 主干、Mixture-of-Experts、扩散/流动作头、检索模块与传统控制器共同工作。架构价值应以端到端任务、成本和可靠性衡量，而非只看参数量。[S03][S55]

结构	擅长	基本机制	仍在发挥作用的地方
MLP	通用函数逼近	全连接变换	Transformer FFN、表格模型
CNN	局部空间模式	卷积与权重共享	视觉、信号、边缘设备
RNN/LSTM	顺序与状态	递归更新隐状态	流式音频、小型时序系统
Transformer	全局依赖与并行	注意力 + 前馈网络	语言、多模态、科学、控制
GNN	关系和图结构	邻居消息传递	分子、材料、交通、推荐
Diffusion/Flow	高质量生成	逐步去噪或流匹配	图像、视频、3D、机器人动作
State-space	长序列效率	结构化状态更新	长上下文、时序、边缘推理
Neural ODE	连续动力学	微分方程定义网络	物理、生物、控制

本页主要来源

[S03] NeurIPS / arXiv（2017），Attention Is All You Need。https://arxiv.org/abs/1706.03762
[S16] Physical Intelligence（2024），π0: A Vision-Language-Action Flow Model for General Robot Control。https://www.physicalintelligence.company/download/pi0.pdf
[S55] NeurIPS（2018），Neural Ordinary Differential Equations。https://arxiv.org/abs/1806.07366

强化学习：从奖励到规划

强化学习不是“让模型更聪明”的魔法，而是在序列决策中优化长期回报；奖励设计决定系统学会什么。

基本问题

智能体观察状态，采取动作，环境返回新状态和奖励。价值函数估计长期回报，策略决定动作，模型式方法还学习环境动力学以便规划。深度 Q 网络证明神经网络可从像素学习多款 Atari 游戏；AlphaGo/AlphaZero 将策略、价值、搜索和自我博弈结合。[S52][S53]

Model-free 与 Model-based

无模型方法直接学策略或价值，通常简单但数据需求高；模型式方法学习或使用环境模型，通过想象未来进行规划。MuZero 不要求预先给出环境规则，却学习对决策有用的潜在动力学，连接了强化学习与世界模型。[S54]

LLM 后训练

RLHF/RLVR 把人类偏好或可验证结果作为奖励，改善指令遵循、推理和安全行为。但奖励代理可能被钻空子；如果验证器只覆盖最终答案，模型也可能学到投机过程。DPO 等方法把偏好学习重写为较稳定的监督式目标。[S05][S06][S07]

机器人与现实环境

真实交互昂贵、危险且速度慢，因此常用模仿学习、离线强化学习、仿真和少量在线微调组合。仿真到现实的差距、奖励稀疏和安全探索仍是核心难题。[S16]

本页主要来源

[S05] OpenAI（2022），Training language models to follow instructions with human feedback。https://arxiv.org/abs/2203.02155
[S06] Stanford / arXiv（2023），Direct Preference Optimization。https://arxiv.org/abs/2305.18290
[S07] OpenAI（2024），Learning to reason with LLMs。https://openai.com/index/learning-to-reason-with-llms/
[S16] Physical Intelligence（2024），π0: A Vision-Language-Action Flow Model for General Robot Control。
https://www.physicalintelligence.company/download/pi0.pdf
[S52] Nature / DeepMind（2015），Human-level control through deep reinforcement learning。
https://www.nature.com/articles/nature14236
[S53] Nature / DeepMind（2017），Mastering the game of Go without human knowledge。
https://www.nature.com/articles/nature24270
[S54] Nature / DeepMind（2020），Mastering Atari, Go, chess and shogi by planning with a learned model。
https://www.nature.com/articles/s41586-020-03051-4

因果、神经符号与可验证推理

统计相关性擅长预测“会发生什么”，因果模型试图回答“如果干预会怎样”；符号系统则提供约束、证明和组合结构。

因果推断

因果图、潜在结果和结构因果模型通过明确假设区分相关与干预。它们在政策评估、医疗决策和科学发现中重要，因为高预测准确率并不能保证改变一个变量会带来预期结果。难点在于不可观测混杂、可识别性和分布迁移。

神经符号

神经网络从高维数据学习表示，符号模块处理逻辑、规则、程序或知识图谱。组合方式包括让模型生成程序、调用定理证明器、把约束加入训练或用检索提供结构化知识。优势是可验证与组合；代价是接口复杂、知识维护和搜索爆炸。

验证器的重要性

在数学、代码和科学任务中，单纯生成候选往往不够。执行单元测试、形式证明、物理约束或实验评价，可以把开放式生成转化为“搜索 + 评价 + 迭代”。AlphaEvolve 正是把语言模型提出算法与自动评价器、进化式选择结合。[S22]

边界

可验证不等于正确：测试可能不完整，形式化规范也可能描述错目标。混合系统必须同时审计生成器、工具、规范与数据链。

本页主要来源

[S22] Google DeepMind（2025），AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms。https://deepmind.google/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/
[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S40] Russell & Norvig（2021），Artificial Intelligence: A Modern Approach。https://aima.cs.berkeley.edu/

持续学习、元学习与联邦学习

真正长期运行的智能体必须适应变化，却不能在学习新任务时忘掉旧能力，也不能把敏感数据集中复制。

持续学习

模型顺序接触任务或数据分布，目标是在保持旧知识的同时吸收新知识。典型方法包括经验回放、重要参数正则化、动态扩展网络和参数隔离；根本矛盾是稳定性与可塑性。[S51]

元学习

元学习试图“学会如何学习”，在任务分布上训练可快速适应的新初始化、优化器或策略。它适合少样本个性化和机器人，但如果新任务脱离训练分布，快速适应也可能失效。

联邦学习

参与方在本地计算更新，仅汇总梯度或参数，适用于手机、医院和金融机构。它减少原始数据集中化，却不自动提供隐私：更新可能泄漏信息，还需要安全聚合、差分隐私和攻击检测。[S50]

大模型中的现实做法

2026 年多数生产系统仍通过 RAG、外部记忆、定期微调和模型版本管理实现“持续更新”，而不是让基础模型在线改权重。外部记忆易审计、可撤销，但只能补充知识，不能完全替代技能学习。

本页主要来源

[S50] Google（2017），Federated Learning: Communication-Efficient Learning of Deep Networks。https://proceedings.mlr.press/v54/mcmahan17a.html
[S51] arXiv（2019），Continual Learning in Neural Networks。https://arxiv.org/abs/1910.02718
[S09] Meta AI / NeurIPS（2020），Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks。https://arxiv.org/abs/2005.11401

基础模型：从单任务模型到通用底座

Foundation model 的核心不是“大”，而是先在广泛数据和任务上形成通用表示，再被大量下游系统适配。

为什么出现

规模化自监督预训练把一次昂贵训练的能力复用于搜索、写作、编程、视觉、语音和科学任务。GPT-3 展示了参数与数据规模带来的少样本能力；随后适配方法把通用底座变成具体产品。[S04]

规模定律

经验规模定律描述损失随模型、数据和计算增长而下降，但最优配比并非“参数越大越好”。Chinchilla 工作强调在固定计算下增加训练 token、减小过大的模型可获得更好效果，推动数据—模型—计算协同设计。[S45][S46]

能力并非一个数

通用模型在知识、推理、代码、多模态、工具使用和语言覆盖上呈锯齿状；某项 benchmark 领先不代表在真实业务中稳定。模型卡、系统卡和分域评测比单一总分更有价值。[S02][S41]

权力与透明度

基础模型集中训练使少数机构掌握算力、数据和部署入口；开放权重可以促进研究和本地化，也使撤回与监控更难。透明度索引显示训练数据、算力、风险评测和下游使用信息仍常不充分。[S42]

本页主要来源

[S02] International AI Safety Report（2026），International AI Safety Report 2026。https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S04] OpenAI / NeurIPS（2020），Language Models are Few-Shot Learners。https://arxiv.org/abs/2005.14165
[S41] Google DeepMind（2026），Model Cards。https://deepmind.google/models/model-cards/
[S42] Stanford CRFM（2025），Foundation Model Transparency Index。https://crfm.stanford.edu/fmti/
[S45] OpenAI（2020），Scaling Laws for Neural Language Models。https://arxiv.org/abs/2001.08361
[S46] Google DeepMind（2022），Chinchilla: Training Compute-Optimal Large Language Models。https://arxiv.org/abs/2203.15556

LLM 的完整技术栈

产品里的“一个聊天框”，背后至少有数据、token、预训练、后训练、推理、检索、工具、安全和观测九个层。

判断产品的正确单位

比较 LLM 产品时应问：底座版本是什么、是否检索、有哪些工具权限、如何记忆、怎样评测、失败由谁复核。只报模型名无法解释真实表现。[S08][S09][S43]

层	作用	常见方法	失败模式
数据	提供语言与知识分布	清洗、去重、配比、合成	污染、版权、偏差
表示	把文本变成离散/连续表示	tokenizer、embedding	跨语言不均、长尾切分
预训练	学习通用预测能力	Transformer、MoE	成本高、知识滞后
后训练	形成指令与偏好行为	SFT、RLHF、DPO、RLVR	奖励错配、过度拒答
推理	生成与搜索候选	采样、CoT、test-time compute	成本、幻觉、脆弱
外部系统	连接知识和行动	RAG、MCP、API、Agent	权限、注入、执行错误
治理	监控质量和风险	eval、日志、红队、审批	覆盖不足、责任模糊

本页主要来源

[S03] NeurIPS / arXiv（2017），Attention Is All You Need。https://arxiv.org/abs/1706.03762
[S08] Anthropic（2024），Building effective agents。https://www.anthropic.com/research/building-effective-agents
[S09] Meta AI / NeurIPS（2020），Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks。https://arxiv.org/abs/2005.11401
[S43] Anthropic（2024），Model Context Protocol。https://www.anthropic.com/news/model-context-protocol

预训练：数据、token、目标与规模

预训练把世界压缩成参数，但压缩过程受数据选择和目标函数约束；它不会自动保留来源、时间和因果结构。

Token 与上下文

Tokenizer 决定模型处理的基本单位，不同语言的字符—token 比例不同，会影响成本和上下文效率。位置表示与注意力机制决定模型如何利用序列，但宣称的上下文长度并不等于能在任意位置稳定检索和推理。

训练目标

自回归模型最大化下一个 token 的似然，擅长开放式生成；遮蔽建模恢复被隐藏部分，更偏向双向表示。多模态模型还可能使用对比学习、重建、扩散或联合 token 预测。[S03][S10]

数据工程

高质量预训练需要去重、质量过滤、语言和领域配比、个人信息与安全过滤、许可策略，以及针对数据污染的 benchmark 隔离。合成数据能补足稀缺样本，但循环使用未经验证的合成内容可能放大错误。

规模与稀疏化

Mixture-of-Experts 只激活部分专家，扩大总参数而控制每 token 计算；并行、低精度和检查点使训练跨越大量加速器。规模定律提供资源分配参考，却无法保证涌现能力按计划出现。[S45][S46]

本页主要来源

[S03] NeurIPS / arXiv（2017），Attention Is All You Need。https://arxiv.org/abs/1706.03762
[S10] OpenAI（2021），CLIP: Learning Transferable Visual Models From Natural Language Supervision。https://arxiv.org/abs/2103.00020
[S45] OpenAI（2020），Scaling Laws for Neural Language Models。https://arxiv.org/abs/2001.08361
[S46] Google DeepMind（2022），Chinchilla: Training Compute-Optimal Large Language Models。https://arxiv.org/abs/2203.15556

后训练：把“会续写”变成“会协作”

后训练改变模型在同样知识基础上的行为分布；它能增强指令、推理和安全性，也可能牺牲多样性或引入新的迎合。

监督微调 SFT

用高质量指令—回答样本教模型遵循格式、角色和任务。它稳定、易控制，但覆盖面取决于示范，容易学到示范风格而非深层目标。

RLHF 与 RLAIF

偏好数据训练奖励模型，再优化策略使输出更符合人类或 AI 评审偏好。InstructGPT 证明小型对齐模型在人类偏好上可胜过更大但未对齐模型；局限是标注者代表性、奖励模型可被利用和训练不稳定。[S05]

DPO 与偏好优化

DPO 直接利用成对偏好数据优化策略，绕开显式奖励模型和在线强化学习，工程更简洁。它仍依赖偏好数据质量，不能消除目标定义问题。[S06]

可验证奖励与推理

数学、代码和某些科学任务可通过正确答案、单元测试或形式约束给出奖励，支持 RLVR 和大规模自我改进。优势是信号客观；风险是模型只优化可测部分、利用测试漏洞，或在不可验证领域迁移失效。[S07][S22]

宪法式方法

Constitutional AI 用一组原则指导模型自我批评和 AI 反馈，减少部分人工有害样本标注。原则的选择仍是治理问题，不能替代外部审计。[S44]

本页主要来源

[S05] OpenAI（2022），Training language models to follow instructions with human feedback。https://arxiv.org/abs/2203.02155
[S06] Stanford / arXiv（2023），Direct Preference Optimization。https://arxiv.org/abs/2305.18290
[S07] OpenAI（2024），Learning to reason with LLMs。https://openai.com/index/learning-to-reason-with-llms/
[S22] Google DeepMind（2025），AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms。https://deepmind.google/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/
[S44] Anthropic（2022），Constitutional AI: Harmlessness from AI Feedback。https://arxiv.org/abs/2212.08073

推理时计算：让模型在回答前搜索和验证

能力提升不只发生在训练阶段；生成多个候选、分解问题、调用验证器和动态分配计算正在重塑“推理”。

核心机制

模型可显式产生中间步骤，使用树搜索或采样生成多个轨迹，再由价值模型、规则、执行器或第二个模型选择。推理时计算允许难题花更多资源，而简单题快速回答。[S07]

为什么有效

预训练模型内部包含许多潜在解题路径，单次贪心生成只访问其中一条。搜索扩大候选空间，验证器把概率最高转化为满足目标；在数学和代码中，外部执行尤其有价值。

不等于真实思维

可见的 chain-of-thought 可能是不完整或事后合理化的解释，不能直接当作内部因果过程。隐藏推理也增加审计难度；安全评估需要观察最终行为、工具调用、日志与环境后果，而非只读文字理由。[S02]

成本与控制

test-time compute 会增加延迟、token 和能耗；系统必须按风险分配预算，例如医疗决策要求多模型复核和证据检索，创意改写则不需要昂贵搜索。

本页主要来源

[S02] International AI Safety Report（2026），International AI Safety Report 2026。
https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S07] OpenAI（2024），Learning to reason with LLMs。https://openai.com/index/learning-to-reason-with-llms/
[S22] Google DeepMind（2025），AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms。
https://deepmind.google/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/

Agent：从回答问题到完成任务

Agent 是带有目标、状态、工具和循环控制的系统，不是某个单独模型；自主程度应由风险和可逆性决定。

最小闭环

一个 Agent 至少包含：理解目标、维护状态、规划或选择下一步、调用工具、观察结果、判断是否完成。复杂系统还加入长期记忆、角色分工、预算、异常恢复和人类审批。[S08]

Workflow 与 Agent

固定工作流按照预定义步骤执行，容易测试；Agent 动态决定步骤，适合路径不确定的任务，但行为空间更大。Anthropic 建议先从简单可组合模式开始，只有在灵活性收益明显时才增加自主性。[S08]

工具协议

函数调用与 MCP 等协议把模型连接到文件、数据库、浏览器和企业系统。协议解决接口标准化，不解决授权、数据最小化、提示注入和业务责任。[S43]

主要失败

目标解释错误、规划循环、工具返回被误读、网页提示注入、权限过大、写入操作不可逆、多智能体相互放大错误。高风险 Agent 需要最小权限、沙盒、干运行、事务回滚和人工确认。[S02][S24]

本页主要来源

[S02] International AI Safety Report（2026），International AI Safety Report 2026。https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S08] Anthropic（2024），Building effective agents。https://www.anthropic.com/research/building-effective-agents
[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S43] Anthropic（2024），Model Context Protocol。https://www.anthropic.com/news/model-context-protocol

记忆、RAG 与上下文工程

把模型参数之外的信息组织好，往往比继续扩大模型更能提升企业任务；但检索只提供证据，不保证模型忠实使用。

三种“记忆”

参数记忆来自训练，更新慢且来源难追踪；上下文记忆存在当前窗口，精确但容量和注意力有限；外部记忆保存在数据库、向量库或事件日志，可更新、可撤销、可审计。

RAG

检索增强生成先根据问题从文档库取回片段，再让模型在证据条件下回答。它适合最新、私有和可引用知识，并降低重新训练成本。[S09]

真正的难点

文档切分、元数据、权限过滤、查询改写、混合检索、重排序、版本管理和引用对齐共同决定效果。检索到错误文档会制造“有引用的错答案”；模型也可能忽略检索内容或把不可信文本当指令。

长期 Agent 记忆

事件日志、摘要、技能库和用户偏好可支持跨会话任务，但必须设计遗忘、纠错和隐私机制。把所有历史无限累积会增加噪声、成本和敏感信息暴露。

本页主要来源

[S09] Meta AI / NeurIPS（2020），Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks。https://arxiv.org/abs/2005.11401
[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S43] Anthropic（2024），Model Context Protocol。https://www.anthropic.com/news/model-context-protocol

原生多模态：共同表示与跨模态行动

多模态不只是“给 LLM 加一个看图插件”，而是对齐、融合和生成文本、视觉、音频、视频乃至动作。

三种结构

早期融合在输入层联合 token；中期融合用交叉注意力或适配器连接编码器；统一模型把不同模态离散化或映射到共同序列，使用一个主干预测。选择取决于数据规模、实时性与生成需求。

对比学习

CLIP 使用大规模图像—文本对进行对比学习，把匹配图文拉近、非匹配拉远，得到可迁移的视觉语义表示。它推动开放词汇分类和跨模态检索，但继承网页配对偏差。[S10]

原生交互

实时语音、视频理解和屏幕操作要求模型处理连续流、时间顺序、说话轮次和低延迟。模型卡应说明输入限制、语言覆盖、人物识别、隐私和危险场景，而不是只展示样例。[S41]

走向动作

当输出空间从文字扩展为机器人动作或软件操作，多模态模型变成 VLA 或 computer-use Agent；安全边界随执行权限扩大。[S14][S15]

本页主要来源

[S10] OpenAI（2021），CLIP: Learning Transferable Visual Models From Natural Language Supervision。https://arxiv.org/abs/2103.00020
[S14] Google DeepMind（2025），Gemini Robotics。https://deepmind.google/models/gemini-robotics/
[S15] Google DeepMind（2023），RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control。https://arxiv.org/abs/2307.15818
[S41] Google DeepMind（2026），Model Cards。https://deepmind.google/models/model-cards/

生成式 AI：图像、视频、音频与 3D

生成模型正从一次性内容合成走向可编辑、可控制、具有时间和空间一致性的世界构建。

主要模型族

VAE 学习可采样潜在空间；GAN 通过生成器与判别器对抗；扩散模型从噪声逐步还原数据；flow matching 学习连续概率流。不同方法在质量、速度、覆盖和可控性上权衡。

从文本到可控创作

产业重点已从“生成一张好图”转向角色一致、镜头控制、局部编辑、参考素材绑定、声音同步、版权与内容溯源。视频模型还要学习时间连续性和物理规律，但逼真的输出不等于内部拥有可行动的世界模型。

3D 与世界

NeRF、Gaussian Splatting、3D diffusion 和 world model 让系统从多视角重建或直接生成可导航空间。World Labs 把持续一致、可交互的 3D 世界视为空间智能产品基础。[S13]

风险

身份仿冒、非自愿亲密影像、虚假证据、训练许可和创作者补偿是核心治理问题。水印与检测只能提供部分信号，发布平台还需来源记录、投诉和快速处置。[S02]

本页主要来源

[S02] International AI Safety Report（2026），International AI Safety Report 2026。
https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S12] Google DeepMind（2025），Genie 3: A new frontier for world models。
https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/
[S13] World Labs（2026），About World Labs。https://www.worldlabs.ai/about
[S41] Google DeepMind（2026），Model Cards。https://deepmind.google/models/model-cards/

世界模型：预测环境与行动后果

世界模型至少应学习环境怎样随时间和动作变化；能生成逼真视频，只是可能的实现，不是充分证据。

四个层次

观察预测模型预测下一帧或未来表示；动作条件模型预测候选动作后果；任务模型还估计奖励、终止或价值；生成式环境模型构造可交互世界供人或 Agent 训练。只有后三类直接服务于规划。

V-JEPA 2

Meta 的路线不重建所有像素，而在潜在表示空间预测被遮蔽或未来内容。V-JEPA 2 用大规模视频预训练，再以少量机器人数据学习动作条件模型，展示了视觉理解、预测和零样本机器人规划；结果仍主要来自官方论文设定。[S11]

Genie 3

Google DeepMind 把 Genie 3 定义为可实时导航的通用世界模型，官方披露 720p、24fps、数分钟一致性，并明确列出动作空间、多智能体交互、真实地点准确性和长时持续等限制。[S12]

如何验证

更强测试应包含动作干预、反事实、对象恒常、长时一致性、分布外物理、可重复性和真机闭环。视觉逼真度、可控性、物理正确性与规划效用必须分别报告。

本页主要来源

[S11] Meta AI（2025），Introducing V-JEPA 2。https://ai.meta.com/research/vjepa/
[S12] Google DeepMind（2025），Genie 3: A new frontier for world models。https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/
[S70] Yann LeCun（杨立昆） / Dædalus（2026），Learning Abstractions。https://www.amacad.org/sites/default/files/publication/downloads/daedalus_wi-sp26_03_lecun.pdf
[S72] World Labs（2026），A Functional Taxonomy of World Models。https://www.worldlabs.ai/blog/taxonomy-of-world-models

空间智能：李飞飞与 World Labs 的路线

空间智能把视觉从“识别画面”推进到“理解、生成、推理并交互于持续存在的三维世界”。

核心主张

World Labs 认为 AI 若要在虚拟和物理世界行动，需要几何、物体、场景、遮挡、持续性和可交互性，而不仅是图像标签或自然语言描述。公司由李飞飞、Justin Johnson（贾斯汀·约翰逊）、Christoph Lassner（克里斯托夫·拉斯纳） 与 Ben Mildenhall（本·米尔登霍尔） 创立。[S13]

功能分类

其 2026 年公开分类把 world model 按功能分成渲染器、模拟器、规划器与统一模型。这个框架有助于避免把视频生成器、物理模拟器和机器人规划器混称为同一件事。[S72]

产品与证据

Marble 能从文本、图片、视频或粗略 3D 输入生成可探索、空间连贯的世界，首先面向创作、设计和沉浸式体验。公司展示证明产品存在，不证明模型已经掌握通用物理、因果或机器人控制。

与具身智能的交点

空间世界可作为 real-to-sim-to-real 数据和训练环境；但迁移到实体机器人仍需传感器标定、接触动力学、动作空间、安全控制和真实反馈。空间智能是必要能力集合之一，不是具身智能的同义词。

本页主要来源

[S13] World Labs（2026），About World Labs。https://www.worldlabs.ai/about
[S72] World Labs（2026），A Functional Taxonomy of World Models。https://www.worldlabs.ai/blog/taxonomy-of-world-models
[S11] Meta AI（2025），Introducing V-JEPA 2。https://ai.meta.com/research/vjepa/

VLA：把视觉、语言和动作接在一起

Vision-Language-Action 模型把语义知识转成动作序列，是具身智能的重要架构，但不是机器人全栈。

从 RT-2 到 π 系列

RT-2 把机器人动作表示为 token，与互联网视觉语言任务共同微调，证明网页语义知识能部分迁移到未见物体和指令。[S15] Physical Intelligence 的 π0 使用预训练 VLM 与 flow-matching 动作专家生成连续动作，目标是跨任务、跨机器人策略。[S16]

Gemini Robotics

Google DeepMind 在 Gemini 基础上增加动作输出，并提供强调空间与具身推理的 ER 模型。2026 年新版本继续扩展全身动作和多 embodiment，但这些仍是开发方报告，需要公开失败率和独立复现。[S14][S41]

什么没有被模型取代

状态估计、运动规划、碰撞检测、力控、稳定控制、急停和硬件冗余仍是机器人系统核心。高层 VLA 适合选择技能和给出目标，低层安全控制不能依赖自然语言模型偶发正确。

评价标准

至少报告任务分布、每项试验次数、成功定义、人工干预、未见环境、恢复能力、耗时、硬件损耗与安全事件。精选 demo 只能证明某次成功，不能证明可靠性。

本页主要来源

[S14] Google DeepMind（2025），Gemini Robotics。https://deepmind.google/models/gemini-robotics/
[S15] Google DeepMind（2023），RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control。https://arxiv.org/abs/2307.15818
[S16] Physical Intelligence（2024），π0: A Vision-Language-Action Flow Model for General Robot Control。https://www.physicalintelligence.company/download/pi0.pdf
[S17] NVIDIA（2025），NVIDIA Isaac GR00T N1。https://research.nvidia.com/labs/gear/gr00t-n1/
[S41] Google DeepMind（2026），Model Cards。https://deepmind.google/models/model-cards/
[S77] Physical Intelligence（2025），π0.5: a VLA Model with Open-World Generalization。https://www.physicalintelligence.company/download/pi05.pdf

机器人学习的真正瓶颈

网页 token 可复制，真实机器人轨迹却昂贵、慢、异构且有损坏风险；数据闭环比模型名字更稀缺。

数据来源

遥操作和示范提供高质量动作；自主执行产生成功与失败经验；仿真与生成式 world model 扩大覆盖；互联网视频提供语义与人类行为先验。不同来源的动作标签、频率、传感器和物理真实性不一致。

跨本体迁移

机械臂、人形、移动底盘具有不同关节、动力学和坐标系。共享表示可迁移“做什么”，但“怎样做”常需本体适配、动作分块、控制器和少量真机数据。π0/π0.5 与 GR00T 都在探索通用策略加本体适配。[S16][S17][S77]

长程与恢复

十秒桌面取放与一小时家庭任务是不同问题。误差累积、遮挡、物体滑动、工具损坏和人类介入要求在线状态估计、重规划、故障检测与可恢复策略。

近期可行路径

结构化仓储、工厂、实验室和特定服务场景会先规模化，因为环境可设计、任务可重复、风险可隔离。家庭通用机器人需要更高的长尾覆盖和硬件可靠性，进展速度大概率慢于纯软件 Agent。[S84]

本页主要来源

[S16] Physical Intelligence（2024），π0: A Vision-Language-Action Flow Model for General Robot Control。https://www.physicalintelligence.company/download/pi0.pdf
[S17] NVIDIA（2025），NVIDIA Isaac GR00T N1。https://research.nvidia.com/labs/gear/gr00t-n1/
[S62] International Federation of Robotics（2025），World Robotics 2025 — Industrial Robots。https://ifr.org/news/global-robot-demand-in-factories-doubles-over-10-years/1st-quarterly-newsletter-2011
[S77] Physical Intelligence（2025），π0.5: a VLA Model with Open-World Generalization。https://www.physicalintelligence.company/download/pi05.pdf
[S84] Stanford HAI（2026），Technical Performance — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance

自动驾驶、无人机与移动智能体

自主移动已经在限定运行设计域形成商业服务，但这与“任何地点、任何天气、无需远程支持”相距很远。

自动驾驶系统

感知、定位、预测、规划、控制、地图、远程协助和安全运营共同构成系统。端到端神经策略可减少手工接口，却增加可解释、覆盖和验证难度；模块化系统也会出现接口误差。

商业成熟度

Waymo 等在限定城市和气象条件提供大规模 robotaxi 服务，并发布对比人类驾驶的安全影响数据。[S31] 这类数据应按运营里程、道路类型、碰撞定义与报告边界审读，不能外推到未运营区域。

无人机与移动机器人

仓储 AMR、农业无人机、矿区车辆和巡检机器人受益于可控路线和基础设施；城市开放空域还需通信、避障、天气、监管和责任制度。

世界模型的作用

模拟可生成罕见天气、交通冲突和危险场景，减少真车试错；但模型错误会把虚假物理带入策略。安全关键训练必须组合实车、仿真、形式约束和独立测试。

本页主要来源

[S31] Waymo（2026），Waymo Safety Impact。https://waymo.com/safety/impact/
[S12] Google DeepMind（2025），Genie 3: A new frontier for world models。https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/
[S17] NVIDIA（2025），NVIDIA Isaac GR00T N1。https://research.nvidia.com/labs/gear/gr00t-n1/
[S84] Stanford HAI（2026），Technical Performance — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance

AI4Science：从预测到闭环发现

最强成果通常出现在领域结构清晰、评价函数明确、模拟或实验能验证候选的任务。

四种作用

第一，替代昂贵模拟的 surrogate model；第二，从序列、图或场学习表示；第三，生成分子、材料、算法或实验方案；第四，以 Agent 连接文献、代码、仪器与评价器形成闭环。

为什么与通用聊天不同

科学结论需要不确定性、单位、边界条件、可重复实验和因果机制。语言流畅不能替代测量；论文检索与实验数据中的负结果也不能被忽略。

闭环结构

提出假设 → 生成候选 → 运行模拟/实验 → 计算目标与约束 → 更新策略。若目标可自动计算，可用搜索、强化学习或进化方法；若实验昂贵，则需要主动学习选择最有信息量的下一次实验。[S22]

证据边界

2026 AI Index 指出科学 AI 论文和模型快速增长，但论文级复现、真实实验验证和跨实验室泛化明显落后于知识问答 benchmark。[S80]

本页主要来源

[S18] Nature / Google DeepMind（2021），Highly accurate protein structure prediction with AlphaFold。
https://www.nature.com/articles/s41586-021-03819-2
[S20] Nature / Google DeepMind（2023），Scaling deep learning for materials discovery。
https://www.nature.com/articles/s41586-023-06735-9
[S22] Google DeepMind（2025），AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms。
https://deepmind.google/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/
[S80] Stanford HAI（2026），Science — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/science

生物、蛋白质、基因组与药物

结构预测是里程碑，但生物功能、动力学、毒性、临床终点与制造仍需独立模型和实验。

AlphaFold 2

以注意力和几何表示从序列预测蛋白结构，在 CASP14 取得突破，随后数据库扩大结构可访问性。[S18] 它显著改变结构生物学工作流，却不自动解决构象动力学、无序区域和细胞环境。

AlphaFold 3

将预测范围扩展到蛋白质与 DNA、RNA、小分子、离子等生物分子相互作用，采用扩散式结构生成。[S19] 论文结果不能被改写为“已能准确预测任何药物作用”。

生成与设计

蛋白生成、分子图模型和基因组基础模型可提出候选，降低搜索成本。候选仍必须经过合成、体外、动物与临床验证；进入人体试验也不等于最终获批。

医疗部署

科研模型与临床器械属于不同监管对象。临床 AI 需要目标人群外部验证、数据漂移监控、人机责任和上市后监督。[S61][S81]

本页主要来源

[S18] Nature / Google DeepMind（2021），Highly accurate protein structure prediction with AlphaFold。
https://www.nature.com/articles/s41586-021-03819-2
[S19] Nature / Google DeepMind（2024），Accurate structure prediction of biomolecular interactions with AlphaFold 3。
https://www.nature.com/articles/s41586-024-07487-w
[S61] U.S. FDA（2026），Artificial Intelligence-Enabled Medical Devices。
https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices
[S81] Stanford HAI（2026），Medicine — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/medicine

材料、化学与逆向设计

AI 把材料搜索从“预测给定结构的性质”推进到“按目标生成候选”，但可合成性和实验稳定性决定实际价值。

GNoME

Google DeepMind 用图网络和主动学习扩大稳定晶体材料候选空间，并通过高通量第一性原理计算筛选。[S20] 核心贡献是计算候选与数据库，不等于所有预测材料都已被实验制造。

MatterGen

Microsoft Research 的生成模型按化学组成、对称性和目标性质生成无机材料候选，体现逆向设计：先给性能目标，再寻找结构。[S21]

自驱动实验室

机器人合成、自动表征、贝叶斯优化和主动学习可形成“设计—制造—测量—更新”闭环。实验通量、仪器校准、失败记录和样品污染是系统级问题。

评价

必须分别报告计算稳定、可合成、实验成功、长期性能、成本与环境影响。只比较预测误差会忽视从候选到材料的巨大落差。

本页主要来源

[S20] Nature / Google DeepMind（2023），Scaling deep learning for materials discovery。
https://www.nature.com/articles/s41586-023-06735-9
[S21] Nature / Microsoft Research（2025），A generative model for inorganic materials design。
https://www.nature.com/articles/s41586-025-08628-5
[S22] Google DeepMind（2025），AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms。
https://deepmind.google/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/
[S80] Stanford HAI（2026），Science — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/science

天气、气候、地球与物理

神经天气模型已在部分指标上达到或超过传统数值系统，但业务预报仍需要集合不确定性、极端事件和物理约束。

GraphCast

GraphCast 从再分析数据学习全球中期天气演化，在论文的大量变量和提前期上优于当时业务基线，并以远低于传统数值模式的计算生成预报。[S23]

混合系统

AI 可作为完整预测器、后处理器、参数化模块或高分辨率降尺度器。完全数据驱动模型速度快，物理模型提供守恒、可解释机制和异常情形支撑；混合路线更符合业务可靠性要求。

气候与地球

遥感基础模型用于土地覆盖、灾害、农业和环境监测；气候投影需要几十年尺度、政策情景和耦合过程，不能由短期天气准确率直接推出。

不确定性

单一确定性预报不能表达风险。集合预测、校准、极端事件回测、区域公平性与故障降级是公共服务部署的关键。[S80]

本页主要来源

[S23] Science / Google DeepMind（2023），GraphCast: Learning skillful medium-range global weather forecasting。https://www.science.org/doi/10.1126/science.adi2336
[S80] Stanford HAI（2026），Science — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/science
[S56] International Energy Agency（2025），Energy and AI。https://www.iea.org/reports/energy-and-ai

数学、算法与自主实验

当候选结果能被证明、编译、运行或实验测量时，语言模型可以成为搜索器，而不是最终裁判。

生成器 + 评价器

FunSearch 与 AlphaEvolve 的共同思想是：模型提出程序，自动评价器运行并给出可比较分数，搜索机制保留改进并继续变异。这比要求模型一次性写出正确答案更稳健。[S22]

算法落地

AlphaEvolve 官方披露其发现的算法被用于 Google 基础设施和数学/计算问题；2026 年进一步展示跨领域应用。[S22] 这些结果有可运行评价支撑，但通用“自主科学家”仍需文献理解、实验设计与因果判断。

形式证明

定理证明器提供严格验证环境，适合神经模型搜索策略；困难在于把自然语言数学形式化、设计辅助引理和处理超大搜索空间。

自动化实验

科学 Agent 可编排文献、代码、仪器和实验日志。必须保存完整 provenance、负结果和人工修改，防止自动系统只报告成功轨迹。Discovery Loop 也把自动化实验闭环作为创业方向。[S34]

本页主要来源

[S22] Google DeepMind（2025），AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms。https://deepmind.google/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/
[S34] Discovery Loop（2026），Discovery Loop — Continuous Exploration。https://www.discoveryloop.com/
[S80] Stanford HAI（2026），Science — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/science

脑启发与神经形态计算

脑启发研究既包括从神经科学提取学习原则，也包括用事件驱动硬件追求低功耗；它不是“复制人脑”这一条路线。

研究层次

计算神经科学建模感知、记忆与决策；脉冲神经网络用事件和时间编码；神经形态芯片把存储与计算靠近、支持稀疏事件；预测编码与主动推断提供不同的学习理论。

潜在优势

对稀疏传感流、低延迟闭环和极低功耗边缘设备，事件驱动计算可能比密集 GPU 更合适。神经形态相机也可在高速和高动态范围场景减少冗余。

现实限制

训练工具、软件生态、可比 benchmark、通用任务性能与制造规模仍弱于 GPU/TPU 生态。脉冲网络在特定任务节能，不等于在 LLM 或通用推理上已经更优。

与主流 AI 的关系

注意力、记忆、世界模型和持续学习可受脑启发，但现代深度网络并非生物神经系统的忠实模拟。工程价值应由任务、能耗、延迟和可靠性共同评估。

本页主要来源

[S39] Richard Sutton（理查德·萨顿）（2019），The Bitter Lesson。http://www.incompleteideas.net/IncIdeas/BitterLesson.html
[S51] arXiv（2019），Continual Learning in Neural Networks。https://arxiv.org/abs/1910.02718
[S83] Stanford HAI（2026），Research and Development — 2026 AI Index。
https://hai.stanford.edu/ai-index/2026-ai-index-report/research-and-development

效率：蒸馏、量化、稀疏与边缘 AI

高效 AI 不只是把模型变小，而是联合优化参数、内存、带宽、批处理、缓存、编译和任务质量。

蒸馏与小模型

知识蒸馏让学生模型匹配教师分布或中间表示，可显著降低部署成本；学生受容量和训练数据限制，也可能继承教师错误。[S48]

参数高效适配

LoRA 冻结底座，仅训练低秩增量，降低领域适配的显存与存储。[S49] 适配器便于多租户，但不能把新事实可靠写入所有内部推理路径。

量化与稀疏

INT8/INT4 降低权重和激活内存，稀疏化减少计算；精度损失受模型、硬件与任务影响。Mixture-of-Experts 降低每 token 激活参数，却增加路由和通信。

推理系统

KV cache、连续批处理、speculative decoding、FlashAttention、张量并行和请求调度共同决定吞吐与延迟。[S47] 单看每秒 token 会忽视首 token 延迟、并发、上下文长度和能耗。

边缘部署

隐私、网络可靠性和实时性推动端侧模型；常见架构是端侧感知/过滤 + 云端复杂推理 + 本地安全控制。

本页主要来源

[S47] Stanford（2022），FlashAttention。https://arxiv.org/abs/2205.14135
[S48] Google（2015），Distilling the Knowledge in a Neural Network。https://arxiv.org/abs/1503.02531
[S49] Microsoft（2021），LoRA: Low-Rank Adaptation of Large Language Models。https://arxiv.org/abs/2106.09685
[S83] Stanford HAI（2026），Research and Development — 2026 AI Index。
https://hai.stanford.edu/ai-index/2026-ai-index-report/research-and-development

芯片与 AI 计算栈

GPU 只是起点：训练和推理能力取决于加速器、HBM、互连、网络、封装、编译器和电力的协同。

计算层

GPU/TPU/NPU 通过矩阵单元和低精度提升吞吐；高带宽内存决定参数和 KV cache 可否高效供给；先进封装把芯粒与 HBM 靠近；互连支持跨卡同步。

训练 vs 推理

训练强调大规模同步、稳定性和 checkpoint；推理强调低延迟、批处理、内存复用和成本/请求。Agent 与视频模型提高生成长度和多模态算量，使推理逐渐成为主要资本与能源负担。

供应链

芯片设计、EDA、先进代工、EUV、HBM、封装、服务器与云分布在多个国家。出口管制和主权 AI 推动本地化，但完全闭环成本高。

正确指标

FLOPS 不能单独代表可用性能；还要看模型 FLOP 利用率、内存带宽、通信、良率、软件成熟度、单位 token 成本和可获得数量。[S83]

本页主要来源

[S01] Stanford HAI（2026），The 2026 AI Index Report。https://hai.stanford.edu/ai-index/2026-ai-index-report
[S47] Stanford（2022），FlashAttention。https://arxiv.org/abs/2205.14135
[S56] International Energy Agency（2025），Energy and AI。https://www.iea.org/reports/energy-and-ai
[S83] Stanford HAI（2026），Research and Development — 2026 AI Index。
https://hai.stanford.edu/ai-index/2026-ai-index-report/research-and-development

分布式训练与推理软件

前沿模型是分布式系统工程：任何一个通信、存储或调度瓶颈都可能吞掉理论算力。

并行方式

数据并行复制模型、分配批次；张量并行切分矩阵；流水线并行切分层；专家并行分布 MoE。现实训练常组合多种并行并使用 ZeRO/FSDP 切分优化器、梯度和参数。

故障与数据

数千加速器运行数周必然遇到硬件、网络和数值故障，需要 checkpoint、弹性恢复、数据顺序追踪和异常检测。数据加载与存储吞吐也可能成为瓶颈。

推理服务

连续批处理把到达时间不同的请求拼接；paged attention 管理 KV cache；路由可把简单请求交给小模型、困难请求升级到大模型或推理模型。

Agent 成本

Agent 一次任务可能调用模型数十次、检索多个数据库并执行工具。成本控制应以“成功任务”而非“单次 token”计算，并记录重试、失败和人工接管。

本页主要来源

[S47] Stanford（2022），FlashAttention。https://arxiv.org/abs/2205.14135
[S56] International Energy Agency（2025），Energy and AI。https://www.iea.org/reports/energy-and-ai
[S83] Stanford HAI（2026），Research and Development — 2026 AI Index。
https://hai.stanford.edu/ai-index/2026-ai-index-report/research-and-development

数据中心、能源与物理约束

AI 已把软件扩展问题变成电网、冷却、土地、芯片制造和资本开支问题；效率提升可能被使用量增长抵消。

能源规模

IEA 估计全球数据中心 2024 年用电约 415 TWh、约占全球 1.5%，基准情景到 2030 年约 945 TWh；这是所有数据中心，不等同于纯 AI 用电。[S56]

局部约束

电网接入、变压器、输电、冷却水和建设周期可能比服务器采购更慢。高密度训练集群还需要电力质量、备用系统和低延迟网络。

反弹效应

每 token 更高效会降低价格并扩大用量；多模态、推理时计算和 Agent 又增加每任务调用。判断环境影响必须看总任务量、能源结构、硬件生命周期和实际利用率。

披露

厂商很少公开训练/推理完整能耗边界。报告应把测量、估算、情景预测和营销目标分开，并避免用一次查询的粗略数字代表整个系统。[S56][S83]

本页主要来源

[S56] International Energy Agency（2025），Energy and AI。https://www.iea.org/reports/energy-and-ai
[S83] Stanford HAI（2026），Research and Development — 2026 AI Index。
https://hai.stanford.edu/ai-index/2026-ai-index-report/research-and-development
[S01] Stanford HAI（2026），The 2026 AI Index Report。https://hai.stanford.edu/ai-index/2026-ai-index-report

开放权重、开源与封闭模型

“可下载权重”不等于完整开源；开放程度至少包含权重、代码、数据、训练配方、许可证和评测六个维度。

政策权衡

开放模型促进本地语言、教育、研究和竞争，但发布后难以撤回，安全层可被移除。封闭模型便于监控和更新，却集中权力并限制独立审计。[S02][S42]

采购问题

许可证是否允许商用？能否离线运行？训练数据和安全评测是否透明？供应商是否支持版本固定、日志、数据驻留与退出迁移？这些问题比“开源”标签更可操作。

模式	优势	代价/风险	适用
API 封闭	快速、托管、安全更新	锁定、黑箱、数据边界	通用企业服务
开放权重	本地部署、微调、审计	数据/配方不透明、不可召回	研究、主权、定制
较完整开源	可复现、社区创新	维护成本、许可与滥用	公共基础设施
专用小模型	低成本、低延迟、可控	能力窄、维护多版本	边缘和确定任务

本页主要来源

[S02] International AI Safety Report（2026），International AI Safety Report 2026。
https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S42] Stanford CRFM（2025），Foundation Model Transparency Index。https://crfm.stanford.edu/fmti/
[S78] Stanford CRFM（2021），On the Opportunities and Risks of Foundation Models。https://crfm.stanford.edu/report
[S83] Stanford HAI（2026），Research and Development — 2026 AI Index。
https://hai.stanford.edu/ai-index/2026-ai-index-report/research-and-development

评测：排行榜为什么不够

模型评测、任务评测、系统评测和业务结果评测属于不同层；任何一层缺失都会产生错误结论。

五层评测

模型：知识、推理、代码、多模态；任务：真实完成率与质量；系统：工具、恢复、延迟、成本；风险：偏差、隐私、越权、攻击；结果：生产率、客户/患者结局、公平和环境。

benchmark 污染

公开题目可能进入训练数据，模型还可针对排行榜调优。2026 AI Index 指出常用 benchmark 存在无效题和平台偏好问题；新鲜隐藏集、动态任务和人类基线更可信。[S84]

Agent 评测

必须统计端到端完成、部分完成、重试、人工接管、不可逆错误和资源消耗。单步成功率 99% 在 100 步链上也可能只剩约 36.6% 全链成功，因此长任务需要检查点和恢复。

机器人评测

仿真得分不能替代真机；精选视频不能替代重复试验。应报告未见场景、不同本体、长期稳定性和安全干预。

本页主要来源

[S01] Stanford HAI（2026），The 2026 AI Index Report。https://hai.stanford.edu/ai-index/2026-ai-index-report
[S41] Google DeepMind（2026），Model Cards。https://deepmind.google/models/model-cards/
[S84] Stanford HAI（2026），Technical Performance — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance
[S85] Stanford HAI（2026），Responsible AI — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/responsible-ai

安全、对齐与系统可靠性

安全不是让模型永远拒答，而是让系统在威胁、错误和不确定性下保持可控、可恢复、可追责。

三类风险

恶意使用包括诈骗、网络攻击和危险知识；失灵包括幻觉、越权工具、欺骗性行为和事故；系统性风险包括劳动、信息生态、集中化和关键基础设施依赖。[S02]

技术措施

数据过滤、偏好/宪法式训练、能力和风险评测、红队、监控、输出过滤、最小权限、沙箱、网络隔离、速率限制和事故响应需要层层组合。[S24][S44]

不可消除性

越狱和提示注入说明输入可绕过单层防护；开放权重模型无法召回；未知能力和真实世界分布使上线前测试不完备。目标应是 defence-in-depth 与社会韧性，而不是宣称零风险。[S02]

治理责任

开发者、部署者、工具提供者和最终组织分别控制不同风险。企业不能以“模型供应商负责”替代自身数据、权限、流程和人员责任。

本页主要来源

[S02] International AI Safety Report（2026），International AI Safety Report 2026。https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S44] Anthropic（2022），Constitutional AI: Harmlessness from AI Feedback。https://arxiv.org/abs/2212.08073
[S67] ISO（2023），ISO/IEC 42001:2023。https://www.iso.org/standard/42001
[S85] Stanford HAI（2026），Responsible AI — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/responsible-ai

企业 AI 参考架构：从模型到可审计工作流

真正可持续的企业 AI 不是“接一个聊天框”，而是数据、权限、工具、审批、评测与运营共同构成的产品系统。

九个层

入口与身份 → 任务编排 → 模型路由 → RAG/记忆 → 工具/API → 权限与策略 → 人工审批 → 评测/日志 → 反馈改进。高风险写操作应支持预览、双人确认、幂等和回滚。

数据与权限先行

检索必须继承源系统的行列权限和保密级别；模型看到的内容应按任务最小化。把内部文档全部复制进统一向量库，可能破坏原有访问边界和删除义务。

模型路由

简单分类交给小模型，复杂推理升级到前沿模型；敏感任务可本地部署；同一任务用规则、传统 ML 和 LLM 组合。路由依据应包含质量、延迟、成本、数据驻留和可用性。

上线门槛

离线基准通过后还需影子运行、A/B 或分阶段试点；建立错误分类、人工接管、供应商版本变更监控和停机方案。NIST AI RMF 与 ISO 42001 可作为流程骨架，但不保证单个模型安全。[S24][S67]

本页主要来源

[S08] Anthropic（2024），Building effective agents。https://www.anthropic.com/research/building-effective-agents
[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S43] Anthropic（2024），Model Context Protocol。https://www.anthropic.com/news/model-context-protocol
[S57] McKinsey（2025），The state of AI in 2025。https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
[S67] ISO（2023），ISO/IEC 42001:2023。https://www.iso.org/standard/42001

怎样读 28 个行业案例：案例不是平均效果

配套附表中至少 18 个案例明确属于公司自述、公司案例或供应商自述；它们适合发现用例，不适合直接估算行业 ROI。

使用规则

把每个案例拆成任务、基线、样本、时间窗、模型版本、人工复核、总拥有成本和失败代价；缺少这些字段时，只把它当作“候选用例”。软件生产率研究出现正反结果，正说明任务与人群比宣传口号更重要。[S58][S59]

证据层	能回答什么	不能回答什么	升级验证
公司/供应商自述	是否公开部署、宣称的用途与结果	独立因果效果、失败率、行业平均	客户抽样、合同/成本、独立复测
监管数据库	是否获准、登记或进入监管流程	长期临床/社会结局	上市后监测、真实世界研究
同行评审研究	特定样本与方法下的效果	跨机构、跨版本自动泛化	外部验证、复现、长期跟踪
随机/准实验	更可信的任务级因果效应	所有任务与组织的永久效果	多场景重复、机制与成本分析

本页主要来源

[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S57] McKinsey（2025），The state of AI in 2025。https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
[S58] NBER（2023），Generative AI at Work。https://www.nber.org/papers/w31161
[S59] METR（2025），Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity。https://metr.org/Early_2025_AI_Experienced_OS_Devs_Study-paper.pdf
[S61] U.S. FDA（2026），Artificial Intelligence-Enabled Medical Devices。https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices

医疗与制药：最有价值，也最不能跳过验证

成熟用例集中在影像、工作队列、临床文书和运营；诊断、治疗和药物发现必须保留专业责任与前瞻验证。

已落地

医学影像/心电/病理辅助识别、分诊、临床记录、编码与理赔、患者联络和药物警戒已有规模部署。FDA 持续维护 AI 医疗器械清单，但“获准上市”不等于已证明改善患者长期结局。[S61]

环境式记录

ambient scribe 把对话转成病历，能减少文书负担；医院公开的时间节省和 ROI 受采购价格、科室、编码和基线影响，应标记为机构观察而非通用因果效果。[S81]

前沿

生物基础模型、生成式分子设计、虚拟细胞、临床试验匹配和多智能体诊断快速发展。AI 设计候选进入临床仍要经过毒性、制造、试验终点和监管全链。

部署原则

本地人群外部验证、临床责任人、置信与证据展示、模型漂移监控、患者隐私、事故报告和人工覆盖是底线。通用聊天模型不应被包装为无监督医生。

本页主要来源

[S18] Nature / Google DeepMind（2021），Highly accurate protein structure prediction with AlphaFold。
https://www.nature.com/articles/s41586-021-03819-2
[S19] Nature / Google DeepMind（2024），Accurate structure prediction of biomolecular interactions with AlphaFold 3。
https://www.nature.com/articles/s41586-024-07487-w
[S61] U.S. FDA（2026），Artificial Intelligence-Enabled Medical Devices。
https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices
[S81] Stanford HAI（2026），Medicine — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/medicine

金融、保险与法律：权限和责任比文案质量更重要

生成式 AI 先进入检索、摘要、客服和代码，最终授信、交易和法律结论仍受模型风险与审慎义务约束。

成熟任务

反欺诈、反洗钱、交易监控、信用风险、文档抽取、合规审阅和客服长期使用机器学习；LLM 更适合内部知识、投研检索、会议摘要和代码。

为什么谨慎

金融机构面临模型验证、记录保存、消费者公平、解释、隐私和第三方风险。FSB 还关注少数云/模型供应商集中、相似模型导致羊群行为以及生成式欺诈。[S60]

Agent 风险

能读取账户、发起支付或下单的 Agent 必须使用交易限额、分权授权、异常检测、双因素确认和不可否认日志。自然语言“确认”不应替代正式授权。

法律服务

检索、合同对比和初稿能提效，但虚构判例与保密泄露会产生职业责任。所有引用应回到原始法条或判决，且最终判断由持证专业人员承担。

本页主要来源

[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S57] McKinsey（2025），The state of AI in 2025。https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
[S60] Financial Stability Board（2024），The Financial Stability Implications of Artificial Intelligence。https://www.fsb.org/2024/11/fsb-assesses-the-financial-stability-implications-of-artificial-intelligence/
[S67] ISO（2023），ISO/IEC 42001:2023。https://www.iso.org/standard/42001

制造、供应链与工业机器人

工业 AI 的壁垒通常是传感器、工艺知识、设备接口和持续反馈，而不是某个通用模型名称。

成熟任务

机器视觉质检、预测性维护、过程控制、排产、数字孪生、需求预测、能耗优化和工业机器人已广泛使用。生成式 AI 主要成为工程知识助手、故障排查和设计/仿真入口。

规模

IFR 报告 2024 年全球新装工业机器人约 54.2 万台、连续第四年超过 50 万台；中国占当年新装量约 54%。这是工业自动化规模，不等于全部机器人都使用基础模型。[S62]

物理约束

实时控制、功能安全、停机成本、旧协议、网络分区和认证决定上线速度。把云端 LLM 直接接到工控系统会扩大攻击面；高层建议与低层确定性控制必须分离。

通用化方向

工业基础模型、跨产线迁移、自然语言配置、合成数据和具身机器人会逐步扩展，但最先成功的是流程明确、可隔离、回报可测的单元。

本页主要来源

[S17] NVIDIA（2025），NVIDIA Isaac GR00T N1。https://research.nvidia.com/labs/gear/gr00t-n1/
[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S57] McKinsey（2025），The state of AI in 2025。https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
[S62] International Federation of Robotics（2025），World Robotics 2025 — Industrial Robots。
https://ifr.org/news/global-robot-demand-in-factories-doubles-over-10-years/1st-quarterly-newsletter-2011

软件工程与网络安全：生产率证据为何相反

AI 对代码的价值高度依赖任务、代码库上下文、开发者经验与复核成本；“平均提速”不能替代本团队试验。

正向结果

多家企业的现场随机实验发现编码助手与完成任务数增加相关，尤其帮助经验较少者和样板任务。

反向结果

METR 对熟悉大型开源仓库的资深开发者开展随机研究，允许使用 2025 年初 AI 工具时，完成任务反而平均更慢；样本有限但足以否定“任何场景必然提速”。[S59]

解释

陌生 API、测试生成和小型函数容易受益；隐含约束、遗留系统、复杂调试和高质量审查会产生提示、等待、验证与返工成本。应测端到端交付、缺陷率、维护性和安全，而不只测生成速度。

网络安全

模型可辅助漏洞发现、恶意代码和社会工程，攻防同时升级。生产 Agent 需隔离凭证、限制网络、记录命令、审查补丁；提示注入本身就是新的供应链攻击面。[S02][S95]

本页主要来源

[S02] International AI Safety Report（2026），International AI Safety Report 2026。
https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S59] METR（2025），Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity。
https://metr.org/Early_2025_AI_Experienced_OS_Devs_Study-paper.pdf
[S95] UC Berkeley（2026），Dawn Song（宋晓冬）。https://people.eecs.berkeley.edu/~dawnsong/

教育、研究与知识工作

AI 可以降低解释和反馈成本，也可能让学习者跳过形成能力所需的困难；评价目标必须从“交作业”转向“真正学会”。

应用

个性化辅导、备课、题目生成、语言练习、反馈、行政服务、文献检索和数据分析已普及。2026 AI Index 汇总称美国高中及大学学生约五分之四使用 AI 完成学校任务，制度与教师培训滞后。[S82]

教学设计

推荐流程是：先独立尝试 → 使用 AI 获取提示或比较方案 → 要求解释和验证 → 迁移到新题。口试、现场任务、过程档案和来源披露比简单检测“是否 AI 写的”更可靠。

科研辅助

AI 可整理文献、写代码和生成假设，但会虚构引用、遗漏负结果、误解统计。所有引用回到原文；关键分析可复现；作者负责最终解释。

公平

付费工具、带宽、语言覆盖和教师能力不均会扩大差距。公共教育需要开放资源、无障碍、隐私和对弱势学生的效果评估。

本页主要来源

[S02] International AI Safety Report（2026），International AI Safety Report 2026。https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S80] Stanford HAI（2026），Science — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/science
[S82] Stanford HAI（2026），Education — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/education
[S94] Stanford University（2026），Percy Liang（梁柏西）。https://cs.stanford.edu/people/pliang/

传媒、零售与物流：内容生成只是入口

商业价值来自与目录、库存、分发和运营数据连接；同时必须处理版权、身份、品牌安全与内容来源。

传媒创意

文案、分镜、概念图、翻译、配音和后期可加速迭代。核心争议转向训练许可、风格/肖像/声音权、创作者补偿和虚假证据。

来源与标识

C2PA Content Credentials 用签名记录来源和编辑历史，能验证声明链未被篡改，却不能证明内容本身为真，也无法标记没有凭证的所有生成内容。[S68]

零售

搜索、推荐、购物助手、客服、商品目录、需求预测和库存调拨可形成可测 KPI。Klarna 的客服数据是企业自报，应只作为案例，不应外推为所有企业的 ROI。[S32]

物流

仓库机器人、路径优化和需求预测成熟；下一步是 Agent 协调订单、库存、运输与异常。自动写操作必须设置预算、供应商规则和人工例外处理。

本页主要来源

[S32] Klarna（2024），Klarna AI assistant handles two-thirds of customer service chats in its first month。https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/
[S65] 中国国家互联网信息办公室（2025），人工智能生成合成内容标识办法。https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm
[S68] C2PA（2025），C2PA Specification 2.2。https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification
[S79] Stanford HAI（2026），Economy — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/economy

能源、农业、公共部门与国防边界

公共价值可能巨大，但错误成本、采购透明和双重用途使这些领域需要更严格的可追责设计。

能源

负荷预测、电网调度、风光预测、设备维护、材料与勘探可受益；同时数据中心成为新增大负荷，必须与电网、储能和低碳电源共同规划。[S56]

农业

遥感、病虫害识别、精准施肥/灌溉、育种与机器人采摘提高效率，但小农数据、设备成本、网络和模型跨地区泛化限制普及。

公共服务

税务、福利、城市运维和政务助手可缩短处理时间；涉及权利的自动决策必须提供通知、解释、申诉、人工复核和歧视测试。采购合同应保障审计和退出。

国防与安全

情报分析、后勤、仿真和自主系统具有明显双重用途。本报告不提供作战或攻击方法；核心治理问题是有意义的人类控制、国际人道法、升级风险与事故责任。[S02][S66]

本页主要来源

[S02] International AI Safety Report（2026），International AI Safety Report 2026。https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S56] International Energy Agency（2025），Energy and AI。https://www.iea.org/reports/energy-and-ai
[S66] United Nations（2024），Global Digital Compact。https://www.un.org/pact-for-the-future/en/annex-i-global-digital-compact

ROI、劳动与组织：为什么高采用不等于高价值

AI 价值呈现“高采用、低规模化、强分化”；短期收益最常出现在窄任务、可测输出和人在回路。

采用与规模化

2026 AI Index 汇总的调查显示 2025 年 88% 组织使用 AI；另一调查中近三分之二仍未全企业规模化。一个测“是否用”，一个测“是否重构并产生价值”，并不矛盾。[S57][S79]

生产率异质性

客服现场研究发现生成式 AI 提高生产率并更多帮助经验较少者。[S58] 资深开源开发者研究却显示特定场景变慢。[S59] 任务、技能和工作流设计比“有没有 AI”更重要。

真正成本

许可证/API、推理、数据工程、安全、评测、集成、人员培训、错误返工和供应商切换共同构成总拥有成本。只计算 token 或席位费用会系统性低估。

劳动分配

AI 往往先改变任务组合而非立刻消灭职业：例行生成被自动化，人类转向监督、异常处理、关系和责任。收益可能集中于资本和高技能群体，因此培训、参与式设计和分配政策重要。

本页主要来源

[S57] McKinsey（2025），The state of AI in 2025。https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
[S58] NBER（2023），Generative AI at Work。https://www.nber.org/papers/w31161
[S59] METR（2025），Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity。https://metr.org/Early_2025_AI_Experienced_OS_Devs_Study-paper.pdf
[S79] Stanford HAI（2026），Economy — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/economy

全球格局：美国、中国、欧盟与关键供应链

国家竞争力不是一个模型分数，而是人才、研究、资本、芯片、云、电力、制造、应用市场和治理能力的组合。

供应链现实

美国设计与云、台湾先进代工、荷兰 EUV、韩国/美国 HBM、日本材料设备和中国制造/市场相互依赖。出口管制推动冗余与本地化，但完全脱钩成本极高。

模型差距

AI Index 指出美中顶级模型在 2025—2026 年多次换位，单一日期和 benchmark 的 2—3 个百分点不代表完整国家实力。[S01]

地区	相对优势	主要短板/风险
美国	前沿模型、资本、云、芯片设计、数据中心	集中化、能源、人才与州法碎片
中国	论文/专利量、制造、机器人、应用市场、开放模型	先进芯片约束、国际互操作与数据治理
欧盟	工业、科研、设备、规则与基本权利	超大云/前沿模型资本较弱、实施复杂
其他地区	语言、行业、主权数据和特定人才	算力、能源、资本和公共采购能力不均

本页主要来源

[S01] Stanford HAI（2026），The 2026 AI Index Report。https://hai.stanford.edu/ai-index/2026-ai-index-report
[S56] International Energy Agency（2025），Energy and AI。https://www.iea.org/reports/energy-and-ai
[S83] Stanford HAI（2026），Research and Development — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/research-and-development
[S86] Stanford HAI（2026），Policy and Governance — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/policy-and-governance

美国治理：创新、采购、行业法与州法并存

截至 2026-08，美国没有与 EU AI Act 等同的统一横向联邦 AI 法；义务来自多层制度组合。

政策转向

2025 年 EO 14179 撤销上一届部分政策方向，强调移除创新障碍；America's AI Action Plan 侧重基础设施、竞争力、政府采用与国际影响。[S26][S64]

风险框架

NIST AI RMF 与生成式 AI profile 提供自愿风险管理语言，适合采购和内部治理，但不是合规认证或安全保证。[S24]

实际约束

医疗、金融、就业、消费者保护、知识产权、隐私、生物识别、网络安全、出口管制和政府采购继续适用；州法快速增长造成跨州产品合规差异。

写作纠错

不能继续把已撤销或改变的 EO 14110 当作 2026 年联邦政策主线；也不能写成“美国完全无监管”。应描述为创新/采购导向 + 行业与州法 + 国家安全/出口控制。

本页主要来源

[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S26] The White House（2025），White House Unveils America's AI Action Plan。https://www.whitehouse.gov/releases/2025/07/white-house-unveils-americas-ai-action-plan/
[S64] The White House（2025），Removing Barriers to American Leadership in Artificial Intelligence。https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/
[S86] Stanford HAI（2026），Policy and Governance — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/policy-and-governance

欧盟 AI Act：已进入执行期

欧盟以风险分级、基本权利和统一市场为核心，并对通用模型及系统性风险模型设专门义务。

框架

禁止部分不可接受实践；高风险系统承担数据、技术文档、日志、人类监督、准确稳健与合格评定；聊天机器人和深伪承担透明义务；GPAI 提供者承担文档、版权和训练内容摘要等义务。[S25]

时间线

禁止性实践和 AI 素养自 2025-02-02 适用；治理与 GPAI 义务自 2025-08-02；多数条款与第 50 条透明义务自 2026-08-02 适用。

2026 变更

Regulation (EU) 2026/1744 调整部分高风险义务时间表，Annex III 部分用例推迟至 2027-12-02，嵌入受监管产品的 Annex I 系统推迟至 2028-08-02。[S63]

实践

企业需要先判断角色：provider、deployer、importer 或 distributor；再做系统清单、风险分类、GPAI/高风险分析、数据与日志、供应链合同和事件响应。

本页主要来源

[S25] European Union（2024），Artificial Intelligence Act。https://eur-lex.europa.eu/eli/reg/2024/1689/oj
[S42] Stanford CRFM（2025），Foundation Model Transparency Index。https://crfm.stanford.edu/fmti/
[S63] European Union（2026），Regulation (EU) 2026/1744。https://eur-lex.europa.eu/legal-content/EN-ES/TXT/?uri=CELEX%3A32026R1744
[S67] ISO（2023），ISO/IEC 42001:2023。https://www.iso.org/standard/42001

中国治理：分层规则与综合立法推进

中国形成“发展战略 + 算法推荐 + 深度合成 + 生成式服务 + 内容标识”的分层制度；综合 AI 法截至截止日仍在推进。

生成式服务

《生成式人工智能服务管理暂行办法》适用于向境内公众提供生成式服务，涉及训练数据、个人信息、内容安全、安全评估和备案，并同时强调促进创新与分类分级监管。[S27]

内容标识

《人工智能生成合成内容标识办法》自 2025-09-01 生效，要求显式和文件元数据隐式标识，平台核验提示，禁止恶意删除、篡改或伪造。[S65]

安全框架

人工智能安全治理框架提供风险分类和技术/管理措施参考，属于治理框架，不应被误写为一部统一强制法律。[S28]

关键纠错

国务院 2026 年度立法工作计划明确推进人工智能综合立法；截至 2026-08-23 应写“推进中”，不能写“已经颁布综合 AI 法”。产品仍需按场景核对数据、网络安全、算法和行业规则。

本页主要来源

[S27] 中国国家互联网信息办公室（2023），生成式人工智能服务管理暂行办法。https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
[S28] 中国国家互联网信息办公室（2025），人工智能安全治理框架 2.0。https://www.cac.gov.cn/2025-09/15/c_1759484392681349.htm
[S65] 中国国家互联网信息办公室（2025），人工智能生成合成内容标识办法。https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm
[S86] Stanford HAI（2026），Policy and Governance — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/policy-and-governance

国际治理、标准与来源证明

全球尚无统一 AI 法，但科学评估、管理体系、内容来源和跨国对话正在形成互操作层。

安全报告

International AI Safety Report 2026 由 100 多名独立专家参与，聚焦通用 AI 能力、风险和缓解，明确区分已发生伤害与高度不确定但潜在严重的风险，并提出“证据困境”。[S02]

联合国与欧洲委员会

Global Digital Compact 是推动科学小组、治理对话、能力建设和人权导向的政治承诺，并非统一全球强制法。欧洲委员会《人工智能与人权、民主和法治框架公约》则是首个该领域具法律约束力的国际条约框架，具体义务仍取决于缔约和国内实施。[S66][S104]

标准

NIST AI RMF 提供风险流程；ISO/IEC 42001 认证的是 AI 管理体系，不是模型零风险；C2PA 记录内容来源链，不判断内容真假。[S24][S67][S68]

组织做法

建立适用法地图、模型/系统清单、影响评估、供应链问卷、独立评测、事故分级和跨境数据机制。标准可以统一语言，最终责任仍在部署者与主管机关。

本页主要来源

[S02] International AI Safety Report（2026），International AI Safety Report 2026。https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S24] NIST（2023），AI Risk Management Framework 1.0。https://www.nist.gov/itl/ai-risk-management-framework
[S66] United Nations（2024），Global Digital Compact。https://www.un.org/pact-for-the-future/en/annex-i-global-digital-compact
[S67] ISO（2023），ISO/IEC 42001:2023。https://www.iso.org/standard/42001
[S68] C2PA（2025），C2PA Specification 2.2。https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification
[S104] Council of Europe（2024），Framework Convention on Artificial Intelligence。https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

人物图谱 I：奠基者与不同智能观

荣誉和引用量说明历史贡献，不证明今天的路线必然正确；理解人物要看其问题定义、证据和组织位置。

核心分歧

Yann LeCun（杨立昆） 更强调从感知流学习世界模型；Yoshua Bengio（约书亚·本吉奥）/Geoffrey Hinton（杰弗里·辛顿） 更重视前沿系统风险；Richard Sutton（理查德·萨顿） 强调可扩展搜索与学习；Judea Pearl（朱迪亚·珀尔） 强调因果结构。它们回答的是不同问题，不能用一句"谁对谁错"取代实证。

人物	当前公开角色（2026-08）	核心路线 / 谨慎点
Geoffrey Hinton（杰弗里·辛顿）	多伦多大学荣休教授；风险倡议	深度表示；近年强调失控与滥用风险，概率是专家判断
Yoshua Bengio（约书亚·本吉奥）	Mila 顾问；LawZero 科学负责人	Scientist AI、非代理监督；仍属架构提案
Yann LeCun（杨立昆）	AMI Labs 联创、Executive Chair	JEPA、抽象世界模型；Meta 旧员工页已滞后
Richard Sutton（理查德·萨顿）	强化学习学者	"苦涩教训"：通用计算方法长期胜出；不等于忽视结构
Judea Pearl（朱迪亚·珀尔）	UCLA 教授	因果图与干预；依赖可识别性和结构假设
David Silver（戴维·席尔瓦）	UCL 教授；长期参与 DeepMind RL	搜索、自我对弈、强化学习；开放世界迁移仍难
Fei-Fei Li（李飞飞）	Stanford 教授；World Labs 联创	空间智能；生成 3D 世界不等于完整物理理解

本页主要来源

[S13] World Labs（2026），About World Labs。https://www.worldlabs.ai/about
[S30] Yann LeCun（杨立昆）（2026），Yann LeCun（杨立昆） — official home page。https://yann.lecun.com/
[S36] Yoshua Bengio（约书亚·本吉奥）（2026），Yoshua Bengio（约书亚·本吉奥） — Biography and work。https://yoshuabengio.org/en
[S39] Richard Sutton（理查德·萨顿）（2019），The Bitter Lesson。http://www.incompleteideas.net/IncIdeas/BitterLesson.html
[S69] AMI Labs（2026），Advanced Machine Intelligence Labs。https://amilabs.xyz/
[S70] Yann LeCun（杨立昆） / Dædalus（2026），Learning Abstractions。https://www.amacad.org/sites/default/files/publication/downloads/daedalus_wi-sp26_03_lecun.pdf
[S71] LawZero（2025），Yoshua Bengio（约书亚·本吉奥） launches LawZero。https://lawzero.org/en/news/yoshua-bengio-launches-lawzero-new-nonprofit-advancing-safe-design-ai
[S87] Nobel Prize（2024），Geoffrey Hinton（杰弗里·辛顿） — Facts。https://www.nobelprize.org/prizes/physics/2024/hinton/
[S101] UCLA（2026），Judea Pearl（朱迪亚·珀尔） — UCLA faculty profile。https://samueli.ucla.edu/people/judea-pearl/
[S102] UCL（2026），Intelligent Systems Group。https://www.ucl.ac.uk/engineering/computer-science/research/research-groups-and-centres/intelligent-systems-group

人物图谱 II：前沿实验室与 AGI 平台

公司领导者同时是技术路线倡导者和商业组织负责人；其时间表与社会影响判断必须标注利益相关性。

关键纠错

Jeff Dean（杰夫·迪恩） 于 2026-08-05 确实离开 Google；Demis Hassabis（戴密斯·哈萨比斯） 同日改任 GDM Chair 与 Alphabet Chief Scientist。任何写“Jeff Dean（杰夫·迪恩） 仍任 Google Chief Scientist”或“Demis Hassabis（戴密斯·哈萨比斯） 仍负责 GDM 日常运营”的资料都已过时。[S33][S34]

人物	截至 2026-08 的动向	核心路线 / 需谨慎
Demis Hassabis（戴密斯·哈萨比斯）	GDM Chair、Alphabet Chief Scientist；领导 Isomorphic	AGI + AI4Science；2026-08 已交出日常运营
Jeff Dean（杰夫·迪恩）	离开 Google，创办 Discovery Loop	自动化 ML/科学实验闭环；公司愿景待成果验证
Ilya Sutskever（伊利亚·苏茨克维）	SSI 联创兼 CEO	只做安全超级智能；技术细节极少
Dario Amodei（达里奥·阿莫迪）	Anthropic CEO	能力扩展 + 可解释/安全；AGI 时间是预测
Sam Altman（萨姆·奥尔特曼）	OpenAI CEO	算力、模型、产品分发与普遍受益；治理需独立检验
Mustafa Suleyman（穆斯塔法·苏莱曼）	Microsoft AI CEO；转向超智能团队	human-centered AI；目标不等于当前能力
Elon Musk（埃隆·马斯克）	xAI 创始人；xAI 并入 SpaceXAI 体系	Grok、算力与科学愿景；公司数据属自报

本页主要来源

[S33] Google（2026），The next chapter of our AI momentum。https://blog.google/company-news/inside-google/message-ceo/next-chapter-ai-momentum/
[S34] Discovery Loop（2026），Discovery Loop — Continuous Exploration。https://www.discoveryloop.com/
[S35] Google DeepMind（2026），About Google DeepMind。https://deepmind.google/about/
[S37] SSI（2026），Safe Superintelligence Inc.。https://ssi.inc/
[S38] Dario Amodei（达里奥·阿莫迪）（2024），Machines of Loving Grace。https://darioamodei.com/machines-of-loving-grace
[S73] Safe Superintelligence Inc.（2026），SSI Updates。https://ssi.inc/updates
[S74] Dario Amodei（达里奥·阿莫迪）（2026），The Adolescence of Technology。https://darioamodei.com/essay/the-adolescence-of-technology
[S75] xAI（2026），xAI joins SpaceX。https://x.ai/news/xai-joins-spacex
[S76] Microsoft（2026），Announcing Copilot leadership update。https://blogs.microsoft.com/blog/2026/03/17/announcing-copilot-leadership-update/
[S100] OpenAI（2026），Built to benefit everyone: our plan。https://openai.com/index/built-to-benefit-everyone-our-plan/

人物图谱 III：世界模型、机器人与科学

这些人物的共同点不是“都做机器人”，而是分别从空间、动作、搜索、视觉和科学验证切入。

措辞规则

“本人宣布加入”不等于公司公开职衔；“产品 demo”不等于通用能力；“结构预测”不等于完整药物研发。对 2026 年高频变化职位必须写日期和证据类型。

人物	机构 / 方向	正在做什么 / 边界
Chelsea Finn（切尔西·芬）	Stanford；Physical Intelligence	元学习、VLA、π 系列；通用性仍受真实数据限制
Sergey Levine（谢尔盖·莱文）	Berkeley；Physical Intelligence	强化学习、机器人控制与经验数据闭环
Pieter Abbeel（彼得·阿比尔）	Berkeley；Amazon AGI	机器人学习与基础模型汇合；团队结果尚未完整公开
Jitendra Malik（吉滕德拉·马利克）	Berkeley；Meta FAIR Robotics	视觉到机器人感知—行动；职位以本人 CV 为准
John Jumper（约翰·江珀）	宣布离开 GDM、加入 Anthropic	AlphaFold；Anthropic 具体职衔和项目未公开
Andrej Karpathy（安德烈·卡帕西）	本人曾宣布加入 Anthropic 预训练团队	教育、autoresearch；截至截止日精确状态需保留不确定
Kaiming He（何恺明）	MIT；GDM 兼职 Distinguished Scientist	ResNet/MAE 到生成建模；没有离开 MIT

本页主要来源

[S16] Physical Intelligence（2024），π0: A Vision-Language-Action Flow Model for General Robot Control。https://www.physicalintelligence.company/download/pi0.pdf
[S77] Physical Intelligence（2025），π0.5: a VLA Model with Open-World Generalization。https://www.physicalintelligence.company/download/pi05.pdf
[S89] Reuters（2026），US scientist John Jumper（约翰·江珀） to leave Google DeepMind for Anthropic。https://sg.finance.yahoo.com/news/us-scientist-john-jumper-leave-204039201.html
[S90] TechCrunch（2026），OpenAI co-founder Andrej Karpathy（安德烈·卡帕西） joins Anthropic's pre-training team。https://techcrunch.com/2026/05/19/openai-co-founder-andrej-karpathy-joins-anthropics-pre-training-team/
[S91] Stanford University（2026），Chelsea Finn（切尔西·芬） — Stanford Profile。https://profiles.stanford.edu/chelsea-finn
[S92] UC Berkeley（2026），Sergey Levine（谢尔盖·莱文） — Faculty Profile。https://www2.eecs.berkeley.edu/Faculty/Homepages/svlevine.html
[S93] Amazon（2025），Andy Jassy（安迪·贾西） shares leadership updates。https://www.aboutamazon.com/news/company-news/andy-jassy-peter-desantis-amazon-leadership-update
[S96] MIT（2026），Kaiming He（何恺明）。https://people.csail.mit.edu/kaiming/
[S103] UC Berkeley（2026），Jitendra Malik（吉滕德拉·马利克） — curriculum vitae。https://people.eecs.berkeley.edu/~malik/malik-cv-full.pdf

人物图谱 IV：安全、开放生态、中国路线与基础设施

AI 前沿还包括评测、网络安全、开放研究、认知架构、模型复用和全栈计算，不应只追逐模型公司 CEO。

看人物的方法

先确认职位日期，再读论文和机构项目；把技术主张、预测、商业愿景和已验证结果分开；最后检查反例与利益相关。

人物	公开角色	路线与谨慎点
Shane Legg（谢恩·莱格）	Google DeepMind Chief AGI Scientist	AGI 评测与安全准备；AGI 定义仍不统一
Percy Liang（梁柏西）	Stanford CRFM Director	foundation model 评测、透明和开放构建
Dawn Song（宋晓冬）	UC Berkeley 教授	Agent/网络安全、形式方法；典型双重用途
Jensen Huang（黄仁勋）	NVIDIA 创始人兼 CEO	AI factory、数字孪生、physical AI；也是平台营销叙事
Song-Chun Zhu（朱松纯）	BIGAI 院长；北大/清华教授	小数据大任务、认知架构；统一 AGI 是纲领非成品
Zhi-Hua Zhou（周志华）	南京大学教授、副校长；LAMDA	集成学习与 learnware；不等于 foundation model
Kaiming He（何恺明）	MIT 教授；GDM 兼职	视觉表征和生成模型；以论文复现评估

本页主要来源

[S42] Stanford CRFM（2025），Foundation Model Transparency Index。https://crfm.stanford.edu/fmti/
[S78] Stanford CRFM（2021），On the Opportunities and Risks of Foundation Models。https://crfm.stanford.edu/report
[S88] Google DeepMind / Shane Legg（谢恩·莱格）（2025），Taking a responsible path to AGI。https://deepmind.google/blog/taking-a-responsible-path-to-agi/
[S94] Stanford University（2026），Percy Liang（梁柏西）。https://cs.stanford.edu/people/pliang/
[S95] UC Berkeley（2026），Dawn Song（宋晓冬）。https://people.eecs.berkeley.edu/~dawnsong/
[S97] BIGAI（2026），关于北京通用人工智能研究院。https://www.bigai.ai/about/
[S98] Nanjing University（2026），Zhi-Hua Zhou（周志华）。https://www.nju.edu.cn/en/info/2352/11111.htm
[S99] NVIDIA（2026），Jensen Huang（黄仁勋） — Management Team。https://www.nvidia.com/en-eu/about-nvidia/governance/management-team/jensen-huang/

人物图谱 V：2026 实验室领导层与区域生态

除了创始学者，还要关注负责模型、算力、组织和产品落地的当代实验室领导者。

区域平衡

全球前沿仍高度集中，但欧洲的 Mistral/AMI、法国和英国研究网络，中国的 DeepSeek 与产业生态，以及加拿大、日本、韩国、印度和中东的算力与人才计划都在改变版图。人物表用于建立入口，不等于影响力排名。

人物	截至 2026-08 的公开角色	路线与观察
Koray Kavukcuoglu（科雷·卡武克乔奥卢）	SVP, Google DeepMind；Google Chief AI Architect	负责 Gemini、前沿研究与产品团队的日常领导
Jakub Pachocki（雅库布·帕霍茨基）	OpenAI Chief Scientist	大规模强化学习、推理模型与自动化 AI 研究
Mira Murati（米拉·穆拉蒂）	Thinking Machines Lab 联创兼 CEO	可定制、可理解、协作式前沿模型与大规模基础设施
Alexandr Wang（亚历山大·王）	Meta Chief AI Officer	模型、数据与组织整合；公司目标需与独立评测分开
Liang Wenfeng（梁文锋）	DeepSeek 创始人	开放权重、训练效率与原创研究；公开组织资料有限
Arthur Mensch（阿瑟·门施）	Mistral AI 联创兼 CEO	欧洲前沿模型、开放与企业/主权部署

本页主要来源

[S33] Google（2026），The next chapter of our AI momentum。https://blog.google/company-news/inside-google/message-ceo/next-chapter-ai-momentum/
[S105] OpenAI（2024），Ilya Sutskever（伊利亚·苏茨克维） to leave OpenAI, Jakub Pachocki（雅库布·帕霍茨基） announced as Chief Scientist。https://openai.com/index/jakub-pachocki-announced-as-chief-scientist/
[S106] Thinking Machines Lab（2026），Thinking Machines Lab and NVIDIA Announce Long-Term Gigawatt-Scale Strategic Partnership。https://thinkingmachines.ai/news/nvidia-partnership/
[S107] Meta（2026 update），Empowering Parents, Protecting Teens: Meta's Approach to AI Safety。https://about.fb.com/news/2025/10/teen-ai-safety-approach/
[S108] Associated Press（2025），Upstart Chinese AI company DeepSeek's founder started out as a low-key hedge fund entrepreneur。https://apnews.com/article/0673d5c39d90108189cc31b88d85b9f8
[S109] Mistral AI（2026），About Mistral。https://mistral.ai/about/

2026—2030 趋势雷达与核心术语

以下趋势是基于当前证据的条件性推断，不是确定预测；影响取决于成本、监管、数据闭环与社会选择。

雷达

高概率：小模型与模型路由、合规工程产品化、推理成本优化。中概率：受控 Agent 大规模进入业务、机器人在结构化环境泛化。高影响但高不确定：自主 AI 研发加速、开放世界通用机器人、快速 AGI 跃迁。[S01][S02]

趋势/术语	简明定义	关键观察
受控 Agent	低权限、可审计、可回滚的任务执行系统	先在窄流程扩散，再逐步提高自主性
世界模型	预测环境状态及行动后果的模型	必须区分视频逼真、物理正确与规划效用
VLA	视觉—语言—动作模型	连接语义与控制；真实数据和安全是瓶颈
AI4Science	用 AI 进行科学建模、设计和实验	有自动评价器的闭环最可靠
推理时计算	回答单题时追加搜索、采样和验证	提高难题能力，也提高延迟和成本
RAG	检索外部证据再生成	可更新、可引用；不保证忠实使用
持续学习	顺序学习新任务且保持旧能力	不同于外部记忆和长上下文
开放权重	可下载参数但未必开放数据/配方	开放程度需多维度描述
主权 AI	国家/地区控制关键算力、数据和模型	推动本地化，也可能增加碎片
AI factory	把数据中心视为生产智能的基础设施	算力扩张与能源/电网深度耦合

本页主要来源

[S01] Stanford HAI（2026），The 2026 AI Index Report。https://hai.stanford.edu/ai-index/2026-ai-index-report
[S02] International AI Safety Report（2026），International AI Safety Report 2026。https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S56] International Energy Agency（2025），Energy and AI。https://www.iea.org/reports/energy-and-ai
[S84] Stanford HAI（2026），Technical Performance — 2026 AI Index。https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance

最终判断、核查清单与 AI 披露

判断任何 AI 项目时，先问它解决哪一层问题、证据是什么、失败会造成什么、谁能停止和纠正它。

十项核查

1 任务与成功定义；2 数据来源与权限；3 模型/版本；4 工具和写权限；5 隐藏测试；6 人工接管；7 失败恢复；8 总拥有成本；9 法规与责任；10 上线后监控。缺少其中任何一项，都不应以 demo 或 benchmark 直接扩张。

三条总判断

第一，AI 从单模型竞赛进入完整系统竞赛。第二，越接近物理世界和高风险决策，可靠性、硬件、流程和治理越重要。第三，长期壁垒更可能来自独有数据闭环、专业验证、软硬件协同、工作流入口与准入能力，而不是某一周的排行榜。

已纠正事项

Jeff Dean（杰夫·迪恩） 离职与 Discovery Loop 经 Google 和公司官网双重确认；Demis Hassabis（戴密斯·哈萨比斯）与 Yann LeCun（杨立昆） 的角色使用机构或本人一手页；John Jumper（约翰·江珀） 与 Andrej Karpathy（安德烈·卡帕西） 的动向改用可定位的 Reuters/TechCrunch 报道并保留职衔不确定性；美国行动计划链接已更新，EU、美国和中国政策更新到截止日。[S26][S30][S33][S34][S63][S89][S90]

AI 辅助研究披露

本报告于 2026-08-23 使用 OpenAI Codex 桌面代理（基于 GPT-5；客户端未向报告暴露精确构建号）及三个专题子代理，协助检索、证据摘录、综合、写作、排版与独立反方/伦理审阅。程序核验覆盖全部 59 个正文页的引文—来源映射、PDF 结构与逐页渲染；人物和政策等高时效事实回到一手来源。没有外部人类领域专家签字认证，因此封面采用“证据审阅版”而非“人工审校/专家认证”。编制流程对报告的准确性与解释负责；读者在医疗、法律、金融、安全或截止日后的高时效事项上应重新核验。

责任使用声明

本报告仅提供高层研究与决策框架，不提供攻击、监控、武器化或规避安全控制的操作步骤。网络安全、国防、自主系统、生物和医疗等双重用途方向，应遵循最小权限、人工监督、分阶段评测、可回滚、事故报告和适用法律。

核心来源索引

总览：[S01] Stanford AI Index 2026；[S02] International AI Safety Report 2026。技术：[S03] Transformer、[S11] V-JEPA 2、[S14] Gemini Robotics、[S18—S23] AI4Science。治理：[S24—S29] NIST/EU/美中/OECD、[S63—S68][S104] 国际规则与标准。人物：[S30][S33—S39]、[S69—S103][S105—S109]。109 项来源按首次使用列在各页底部并带可点击原始链接。

本页主要来源

[S01] Stanford HAI（2026），The 2026 AI Index Report。https://hai.stanford.edu/ai-index/2026-ai-index-report
[S02] International AI Safety Report（2026），International AI Safety Report 2026。https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[S26] The White House（2025），White House Unveils America's AI Action Plan。https://www.whitehouse.gov/releases/2025/07/white-house-unveils-americas-ai-action-plan/
[S30] Yann LeCun（杨立昆）（2026），Yann LeCun（杨立昆） — official home page。https://yann.lecun.com/
[S33] Google（2026），The next chapter of our AI momentum。https://blog.google/company-news/inside-google/message-ceo/next-chapter-ai-momentum/
[S34] Discovery Loop（2026），Discovery Loop — Continuous Exploration。https://www.discoveryloop.com/
[S63] European Union（2026），Regulation (EU) 2026/1744。https://eur-lex.europa.eu/legal-content/EN-ES/TXT/?uri=CELEX%3A32026R1744
[S89] Reuters（2026），US scientist John Jumper（约翰·江珀） to leave Google DeepMind for Anthropic。https://sg.finance.yahoo.com/news/us-scientist-john-jumper-leave-204039201.html
[S90] TechCrunch（2026），OpenAI co-founder Andrej Karpathy（安德烈·卡帕西） joins Anthropic's pre-training team。https://techcrunch.com/2026/05/19/openai-co-founder-andrej-karpathy-joins-anthropics-pre-training-team/