他们为什么都去了 AI Lab？顶尖头脑看见的，不只是钱

横纵分析法深度研究报告

研究时间：2026-07-24

作者: Koutian Wu (https://ktwu01.github.io/)

顶尖头脑看见的，不只是钱

研究时间：2026-07-24
作者：Koutian Wu；GitHub: ktwu01
研究领域：人工智能、科研制度、人才迁徙、技术权力
研究对象：1943—2026 年 AI 范式演变，以及 2026 年前沿实验室的人才流动
方法论：横纵分析法（Horizontal-Vertical Analysis）

2026 年 2 月 9 日，我在《颠覆性创新者的思维模式》里问：乔布斯、马斯克这类人，到底比其他人多看到了什么？

约五个半月后，这个问题换了一批主角。

UC Berkeley EECS 的总系主任 Jelani Nelson 暂离学校，进入 Anthropic 的预训练团队；OpenAI 联合创始人、前 Tesla AI 负责人 Andrej Karpathy 暂停自己的教育创业，再次回到前沿模型研发；AlphaFold 核心负责人、诺贝尔化学奖得主 John Jumper 离开 Google DeepMind，转入 Anthropic；Google 的 Gemini 联合负责人 Noam Shazeer 加入 OpenAI。数学家、物理学家、经济学家、哲学家也出现在前沿实验室的名单上。

新闻连在一起，很容易生成一个宏大故事：最聪明的人已经看到了普通人没有看到的未来；他们不再满足于论文、公司或财富，而是在追逐最高级别的智能，进而追逐最高级别的权力。

这个直觉抓住了一部分真相，也把另一部分真相压扁了。

一句话结论

> 他们看到的，是 AI 正在从一种技术变成“生产新技术的技术”，从研究对象变成决定研究速度、研究方向与研究准入权的基础设施。顶尖人才争夺的不是一份更体面的工作，而是自己站在这个反馈回路的哪一侧。

AI 确实可能带来前所未有的认知杠杆、行动杠杆与议程设置权，但它不会自动兑换成“无上权力”。控制前沿模型的人仍依赖芯片、能源、资本、组织、法律、国家与社会信任。研究员进入实验室，只是更靠近权力机器，不等于个人拥有它。

研究口径：先把新闻里的“都去了”拆开

这份报告把“顶尖人才”限定为四类人：

- 曾经主导前沿模型、关键算法或大型技术组织的人；
- 在数学、物理、生物、经济、哲学等学科拥有公认研究地位的人；
- 大学系主任、研究中心负责人、CTO 等拥有科研或技术议程设置权的人；
- 其职业选择已由本人、学校、公司或可靠媒体确认的人。

“加入前沿 AI 实验室（AI lab）”也不是一种状态。它至少包括永久离职、industrial leave、学术休假、兼职、visiting researcher、创业团队并入平台。把这些都写成“辞职”，会夸大迁徙的不可逆性。

最典型的例子正是 Jelani Nelson。Berkeley 的官方公告说，他卸任 EECS 主席是为了开始 industrial leave；Anthropic 确认他进入预训练团队。准确的描述应是“前 Berkeley EECS 系主任暂离学校加入 Anthropic”，不是“永久辞去教授职位”。

本报告将公开事实、当事人陈述与分析推断分开。人才流动能证明实验室具有吸引力，也能反映这些人的主观信念；它不能证明他们的技术预测必然正确。会加入前沿实验室的人，本来就更可能相信转折点正在临近，这里存在明显的选择效应。

纵向分析：智能怎样从哲学问题变成工业过程

今天的迁徙不是 2026 年突然出现的风潮。它是一条延续八十多年的曲线，在近几年越过了几个相互叠加的阈值。

1943—1956：智能第一次成为可制造的对象

1943 年，McCulloch 与 Pitts 把神经活动抽象为逻辑计算。1950 年，Alan Turing 没有继续纠缠“机器究竟有没有心灵”，而是把问题改写成一个可以观察的模仿游戏。1955 年的 Dartmouth 提案更直接：学习与智能的各个方面，原则上都可以被精确描述，从而让机器模拟。

这三步完成了一次概念迁移：

旧问题	新问题
心灵是什么	哪些行为可以计算
人为什么聪明	哪些机制可以学习
智能是否神秘	智能能否被描述、训练与制造

AI 从出生起就不是软件业里的一个小分支。它把逻辑、语言、神经科学、控制、学习、创造和科学发现统合进同一个研究计划。对数学家、物理学家、哲学家而言，它处理的正是各自学科最深处的问题：世界是否可被压缩为表示，推理是否可被机械化，知识怎样形成，行动怎样从判断中产生。

这一时期的研究中心主要在大学、政府资助项目与少数企业研究院。稀缺资源是理论、人才与耐心；一名研究者拿着纸笔和有限计算设备，仍有机会推动前沿。

1960—1989：两次寒冬留下的不是失败，而是筛选标准

专家系统曾经证明，人类知识可以被编码成规则。DENDRAL 能帮助分析化学结构，MYCIN 能在封闭医学场景中给出诊断建议。但每进入一个新领域，都要重新访谈专家、整理规则、补充例外；能力大致随人工录入的知识线性增加，维护成本却可能更快上升。

感知机、机器翻译、通用机器人也经历过高预期与能力边界的碰撞。资金退潮形成两次所谓“AI 寒冬”。这段历史给后来者留下一个极重要的判断标准：

> 震撼的演示不等于可扩展的范式。真正值得长期下注的系统，必须能随着数据、计算与训练而持续改善，而不是靠人无限书写规则。

今天评价大模型时，同一个标准仍然有效。一次漂亮的数学答案、一个能操作浏览器的 demo，不能单独证明通用智能已经到来。更可检验的观察对象，是能力曲线能否跨任务复现、能否通过反馈继续增长、能否在真实工作中节省的时间多于制造的麻烦。

1986—2011：从“编写智能”转向“训练智能”

1986 年，Rumelhart、Hinton 与 Williams 发表反向传播研究，展示多层网络如何从误差中调整参数，并在隐藏层形成完成任务所需的内部表示。

这次变化看似只是算法改进，实际改写了人的角色。过去，工程师要告诉机器“边是什么”“语法是什么”“什么规则对应什么结论”；现在，人更多地定义目标、数据与训练过程，让表示在优化中生成。智能的生产方式从手工作坊向可重复训练移动。

从卷积网络、LSTM 到深度置信网络，许多基础积累发生在深度学习并不热门的年代。Hinton、LeCun、Bengio、Sutton 等研究者长期维持一个低声望、低资金却高度连贯的共同体。2012 年以后公司争抢的并非几篇论文，而是一个学派数十年积累的研究品味、工程直觉与失败经验。

这也提醒我们，“看见未来”很少是凭空顿悟。乔布斯的产品判断依赖图形界面、触控、材料、芯片与供应链几十年的成熟；AI 的跃迁也依赖长期积累突然在数据、GPU 和算法上汇合。远见更像是比别人更早读懂约束何时松动，而不是从虚无中得到神谕。

2012—2017：第一波教授进公司，交换的是 GPU、数据与部署

2012 年，AlexNet借助 GPU 和大规模标注数据显著赢得 ImageNet 竞赛。Hinton、Alex Krizhevsky 与 Ilya Sutskever 随后成立 DNNresearch，并被 Google 收购。Hinton 当时仍保留多伦多大学身份，构成一种早期交换：大学提供理论与人才，公司提供数据、算力和产品反馈，论文仍然广泛公开。

2013 年，Yann LeCun 进入 Facebook 组建 FAIR。2015 年，Uber 从 Carnegie Mellon 的机器人中心招走约四十人，引发大学与产业之间第一次广为人知的“抽空”争议。同年 OpenAI 成立时仍承诺鼓励研究者发表论文，并分享代码与专利。

2016 年 AlphaGo 击败李世石，2017 年 AlphaGo Zero 仅从规则和自我对弈出发，发现了人类没有直接教授的策略。同年，Google 研究者发表 Transformer，让序列模型的训练高度并行化。

这一阶段出现了三条后来持续放大的信号：

1. 同一套学习机制能跨领域迁移，而不只是解决一个手工定义的任务；
2. 自我对弈、搜索和自动评估可以生成新训练反馈，能力不再完全受人类样本上限约束；
3. 完整实验越来越依赖公司掌握的数据、算力与工程系统。

2013 年的企业实验室仍像大学的富裕邻居。研究者可以双重任职，论文是声望货币，关键方法常被公开。2026 年的实验室已经更接近一种新制度：研究院、超级计算中心、产品公司与安全机构叠在一起。

2019—2022：Scaling 把探索性科研改造成可融资的工业计划

2019 年，Rich Sutton 在《The Bitter Lesson》中总结：长期看，能够利用不断增长计算的一般方法，往往超过依赖人类手工知识的方法。同年，OpenAI 在解释新组织结构时直言，前沿系统可能需要数十亿美元的云计算、人才与超级计算基础设施。

2020 年的 Scaling Laws研究显示，在被观察的范围内，语言模型损失会随模型规模、数据与计算呈经验性的幂律变化。GPT-3 展示了同一个大模型在少量示例下适配多种任务的能力；Chinchilla 等研究又开始回答怎样在参数和训练数据之间更有效地分配资源。

这里发生了一次制度级转折。若能力增长与资源投入之间存在部分可预测关系，智能研究就不再只等待偶发的天才突破，而可以被写进资本预算、数据中心计划和多年基础设施合同。研究问题从“某个想法是否有用”扩展成“投入十倍计算、改进数据和训练方法后，曲线会走到哪里”。

Scaling law 不是自然定律，也不保证能力永远增长。数据质量、能源、芯片、算法和可靠性都可能形成新的瓶颈。但它足以改变组织行为：风险资本愿意下注，云厂商愿意造集群，实验室愿意为稀缺人才支付远超大学的价格，研究者也愿意去唯一能完成某类实验的地方。

一项基于美国人口普查局雇主—雇员数据、追踪约 4.2 万名 AI 研究者的研究发现：在其美国样本中，到 2019 年约 68% 的研究者任职于产业界。永久转入企业后，研究者的收入和专利显著增加，论文产出则明显下降。知识生产中心的移动早于生成式 AI 热潮，2026 年只是它最醒目的阶段。

2020—2024：AI 从研究对象变成科学仪器

AlphaFold2 是另一条曲线。它没有只是把已有工作做快一点，而是跨过了蛋白质结构预测的长期瓶颈。2024 年，Demis Hassabis 与 John Jumper 因相关工作分享诺贝尔化学奖。这为工业前沿实验室提供了类似现代 Bell Labs 的制度合法性：一家公司的模型研究可以直接进入基础科学最高荣誉体系。

类似信号随后出现在算法、材料、天气、数学与物理中：

- AlphaDev 通过强化学习寻找更快的排序程序；

- GraphCast 用学习系统生成快速天气预测；
- FunSearch 把语言模型与自动评估器结合，搜索新的数学构造；
- AlphaEvolve 把模型、程序生成和自动验证接入算法与芯片优化；
- 理论科学家开始让模型生成推导、代码和候选证明，再由人类检查。

Harvard 理论物理学家 Xi Yin 对《Harvard Crimson》描述过一种强烈的个人体感：原本可能耗费多年编程的工作，被 AI 大幅压缩。UCLA 数学家 Terence Tao 在 2026 年的判断更克制：模型还会浪费时间，但已经进入“节省的时间多于浪费的时间”的阶段。

这两种说法并不冲突。AI 作为科学仪器已经有真实价值，却仍不等于自动科学家。Anthropic 自己在科学博客中承认，模型在部分科研流程上表现很强，也会幻觉、迎合用户，并卡在领域专家觉得简单的问题上。真正的瓶颈正在从“执行所有步骤”转向“选择问题、设计验证、识别伪结果并承担责任”。

对于数学家和物理学家，这种变化的吸引力非常直接。他们没有离开科学去做另一个行业；他们在靠近一台可能放大所有科学的仪器。

2025—2026：最诱人的阈值，是 AI 开始参与改进 AI

前沿实验室如今公开谈论一个更强的反馈回路：用 AI 帮助研究、评测和训练下一代 AI。

Karpathy 加入 Anthropic 后负责的团队，目标之一就是用 Claude 加速预训练研究。他解释自己的决定时说，未来几年对前沿大模型格外关键，自己想回到研发。OpenAI 在 2026 年 6 月公布的计划中，把“自动化 AI 研究员”列为三大目标之一，并称内部判断是到 2028 年 3 月，可能有相当一部分研究由 AI 系统与人类研究者共同完成。

Anthropic 对内部研发的分析也给出相似方向：模型已经生成大量工程代码，在定义清楚、反馈可验证的实验优化任务上进步很快；但人类在选题、研究品味、判断是否可信等环节仍占优势，递归自我改进尚未实现，也并非注定发生。

这正是“现在加入”的时间价值。若 AI 只是一代更好的软件，晚三年加入也许只是错过一轮产品周期；若 AI 能加速 AI 研究，哪怕只是组织层面的加速，早期优势也可能通过模型、人才、数据和反馈反复累积。顶尖人才看到的是一个可能正在收窄的参与定义期。

这条纵轴可以压缩成五次转换：

阶段	智能的角色	稀缺资源	研究中心
1943—1985	可描述的哲学与工程问题	理论、规则、长期资助	大学、政府、企业研究院
1986—2011	可从数据训练的系统	算法、数据、GPU	大学与公司合作

阶段	智能的角色	稀缺资源	研究中心
2012—2018	可跨任务扩展的能力	大数据、集群、工程人才	大型科技公司
2019—2024	可用资本规模化的认知生产	超级计算、能源、训练系统	少数前沿实验室
2025—2026	可能加速科学与自身研发的元工具	前沿模型、验证回路、研究品味、治理	“实验室—平台—基础设施”复合体

横向分析：2026 年的人才究竟流向了哪里

先看人，不看口号

截至 2026 年 7 月 24 日，公开证据支持以下代表性任职关系。表格把 2026 年的新流动与更早形成、延续到 2026 年的关系明确分开：

人物	原位置	时间与关系类型	最能说明的问题
Jelani Nelson	Berkeley EECS 系主任、理论计算机科学教授	2026-07，industrial leave；加入 Anthropic 预训练	大学管理层进入模型核心训练，同时保留回归期权
Andrej Karpathy	Eureka Labs 创始人；前 OpenAI、Tesla AI 负责人	2026-05，加入 Anthropic 预训练	他公开判断未来几年格外关键，因而暂停教育创业、返回研发
Peter Bailis	Workday CTO；前 Stanford 教授、Google Cloud VP	2026-03，离开 Workday，以 Member of Technical Staff 加入 Anthropic	管理头衔可以换成一线强化学习工程；具体薪酬与个人动机未公开
John Jumper	DeepMind AlphaFold 负责人、诺贝尔化学奖得主	2026-06，离开 DeepMind、加入 Anthropic；职责未公开	AI for Science（AI 驱动科学）人才开始在实验室之间流动
Noam Shazeer	Google VP、Gemini 联合负责人、Transformer/MoE 先驱	2026-06，离开 Google、加入 OpenAI	顶级算力不能消除实验室之间的人才流动；他将领导 AI 架构研究
Weijie Su	Wharton 统计学教授	2026-05，学术休假期间加入 OpenAI	数学与统计人才进入模型训练现场
Alex Lupsasca	Vanderbilt 黑洞理论物理学家	2025-10 起，兼任 OpenAI 研究员与 Vanderbilt 教授	“教授或公司”不是唯一选项，混合身份正在增加
Anca Dragan			

人物	原位置	时间与关系类型	最能说明的问题
	Berkeley 教授、机器人与人机协作研究者	此前转入；2026 年领导 Google DeepMind 安全与对齐研究	担心风险的人也要进入内部，才能接触前沿模型、数据与预算
Chad Jones、Anton Korinek	Stanford、UVA 经济学教授	2026 年，休假加入 Anthropic Institute	实验室需要的不只是模型工程，也包括经济与制度推演

表中任职信息综合 Berkeley 公告、Karpathy 任职报道、Bailis 任职报道、Jumper 任职报道、Shazeer 任职报道、Weijie Su 个人公开帖镜像、Lupsasca 的个人履历、Anthropic Institute 公告、Chad Jones 的个人公开帖镜像与《The Atlantic》的跨机构调查。对未公开具体岗位的人，本文不推断其职责。

这份表有意不追求“名单越长越好”。Harvard 物理学家 Xi Yin 被报道与 OpenAI 有关联，但截至研究日缺少本人、Harvard 或 OpenAI 对任职状态的清晰确认，因此本文只引用他公开讨论 AI 科研体验的内容，不把“加入 OpenAI”写成事实。

The Atlantic统计 OpenAI、Anthropic、Meta、Google DeepMind 等机构至少拥有八十多位现任或前任教授，并认为仍是低估。这个数字能说明迁徙的规模，却不能把每个人的动机归成同一种。有人为了训练前沿，有人为了安全，有人为了科学工具，有人研究 AI 对经济和社会的冲击，也有人只是休假一年。

同一批人才面对的并不是同一种实验室。各去向的核心交换可以放在一张矩阵里：

去向	对人才的核心承诺	独有筹码	主要代价或不确定性
Anthropic	在关键窗口同时推进前沿、安全与社会研究	预训练内部数据；安全使命；跨学科 Institute	商业目标与安全使命的张力；私人治理
OpenAI	让 AI 参与制造下一代 AI，并快速部署	产品分发；架构、科学与自动化研发	治理与商业化变化；研究公开边界
Google DeepMind	做 AlphaFold 式长期科学与通用模型研究	TPU、Google 工程与数据；成熟科研传统	大公司协调成本；个人议程受组织战略影响
Meta	用高资本和大规模分发快速组建团队	社交产品触达；整队吸收；基础设施投入	路线与组织调整快；团队稳定性
xAI			

去向	对人才的核心承诺	独有筹码	主要代价或不确定性
	以超级计算和工程速度追求“理解宇宙”	Colossus、工程协同、强烈使命叙事	创始团队流失；治理与研究文化仍待时间检验
独立实验室与大学	保留路线控制、公开研究或批判距离	创始人自治；学术共同体；公共问责	融资与算力依赖；前沿准入较弱

Anthropic：把“关键时刻”变成组织叙事

Anthropic 对跨学科人才的吸引力来自三层组合。

一层是前沿训练。预训练团队掌握 checkpoint、训练曲线、失败模式、数据配方和昂贵训练运行；外部研究者只能看到发布后的产品切片。Nelson 与 Karpathy 进入的正是这一层。

一层是安全使命。Anthropic 在 Claude’s Constitution（Claude 宪法）中把自身位置称为一种“经过计算的赌注”：如果强大 AI 无论如何都会出现，让强调安全的实验室留在前沿，比把前沿完全交给较少关注安全的开发者更好。这样的叙事对认为转折点临近、又担心失控与集中权力的人很有吸引力。

一层是把实验室扩展成小型大学。Anthropic Institute 吸收经济学家与社会科学家，科学项目连接物理、生物、化学和数学。公司在 2026 年提出“压缩的二十一世纪”——让数十年科学进步在更短时间内发生。从公开陈述与选择可以推断，对部分研究者来说，岗位还附带一种身份：参与解释和塑造一次可能的文明级转换。

这套叙事也有内在矛盾：安全研究要接近最强模型，所以最担心集中风险的人反而向最集中的机构聚集；商业成功为安全研究提供资源，也让发表边界与研究议程受私人治理影响。

OpenAI：最强吸力是“让 AI 参与制造下一代 AI”

OpenAI 的组织承诺更直接：自动化 AI 研究员、加速科学与经济、为每个人提供个人 AGI。Noam Shazeer 进入架构研究，数学、统计和物理研究者进入科学与安全团队，反映出模型竞争已经从“招更多机器学习工程师”扩展为“把各学科的推理结构吸收到模型研发中”。

物理学家 Alex Lupsasca 是一个能看见动机形成过程的案例。他曾对模型保持怀疑。据 OpenAI 对他的访谈，GPT-5 Pro 在一次由他设计的测试中快速完成研究生级推导，并复现了他此前已经得到的隐藏对称生成元。这个案例提供了有价值的专家观察，但来源于公司对其员工的访谈，尚不能等同于经过独立验证的新科学发现。

OpenAI 还拥有另一种大学没有的资源：产品分发。论文可能几年后影响一个领域，模型更新可以立刻进入数亿用户和大量组织的工作流。研究、产品、用户反馈和下一轮训练处在同一个系统里，实验室因此同时获得知识生产权与部署权。

Google DeepMind、Meta、xAI：同一磁场里的三种组织答案

Google DeepMind 最接近现代 Bell Labs 的成熟版本。AlphaGo、AlphaFold、天气和材料研究已经证明公司实验室可以产出基础科学结果；Google 的 TPU、数据、工程平台与分发也提供完整实验装置。Anca Dragan 解释自己转入实验室的部分原因时，强调了在前沿推进安全所需的数据、算力和预算。

但成熟平台也有协调成本。Shazeer 与 Jumper 的离开说明，拥有顶级算力并不足以消除实验室之间的人才流动；公开信息不足以把两人的决定归结为研究自主、团队结构、薪酬或任何单一动机。

Meta 的答案是资本、分发和整队吸收。Dawn Song 等学者先从大学创办 Virtue AI，团队再进入 Meta Superintelligence Labs。这不是简单的“大学输给大厂”，而是“大学—创业—平台”三段式流动。Meta 可以让智能体触达数十亿用户，也能提供极端优厚的待遇；代价是组织路线集中、调整速度快，个人议程更容易被平台战略改写。

xAI 则把超级计算、SpaceX 式工程速度与“理解宇宙”的口号放在一起。它对物理学家和第一性原理工程师的审美吸引力很强，但创始团队的大量离开至少说明：算力和宏大使命本身不足以保证核心人才留任。具体离职是否源于治理、组织信任或研究文化，公开证据仍有限。

仍在大学、选择创业的人同样重要

“顶尖人才现在都不创业了”不符合事实。

Yann LeCun 离开 Meta 创办 AMI Labs，原因之一正是他不同意主流大语言模型路线；Ilya Sutskever 创办 SSI，把安全超级智能设为单一目标；Mira Murati 与 John Schulman 等人选择 Thinking Machines Lab。许多所谓“创业公司”只是形态不同的前沿 AI 实验室：需要数十亿美元资本、长期研究周期、芯片与云合作，也必须与巨头争夺同一批人才。

Terence Tao 仍在学界工作，同时公开使用与评估 AI。他支持 2026 年的 Leiden Declaration，这份由数学共同体推动、获得国际数学联盟支持的声明要求公开 AI 工具使用、保留人类对正确性的责任、建设独立于产业的公共计算设施，并警告企业存在夸大能力的商业激励。

四种职业路径代表不同的权利组合，不应按“进步”与“保守”排序：

路径	获得什么	放弃或承担什么
加入前沿实验室	最强模型、算力、人才密度、部署反馈	发表自由、议程自主、公共问责较弱
创办前沿 AI 实验室	方向与文化控制权、巨大股权上行	融资与算力依赖、组织生存风险

路径	获得什么	放弃或承担什么
留在大学或公共机构	长期问题、公开发表、人才培养、批判距离	难以接近最大训练运行，反馈更慢
休假、兼职、联合任职	保留回归期权，连接两种制度	利益冲突、时间分裂、边界不透明

真正发生的不是“实验室战胜创业”，而是前沿创业正在实验室化，顶级学术研究正在基础设施化，两者的边界越来越模糊。

横纵交汇（一）：为什么是现在

1. 算力已经像粒子加速器，不再像一台个人电脑

Stanford 2026 AI Index显示，2025 年超过 90% 的重要 AI 模型来自产业界；自 2022 年起，全球 AI 计算能力约每年增长 3.3 倍，达到约 1,710 万张 H100 等效卡。美国拥有 5,427 个数据中心，领先芯片又高度依赖少数设计公司、云厂商与一家主要晶圆代工厂。

这种实验装置无法靠一位教授的经费独立复制。外部研究者可以调用 API，却看不到权重、训练数据、内部 checkpoint、故障日志与下一代模型。对研究前沿能力和安全的人，进入实验室会让某些原本无法完成的实验变得可能。

2. 模型已经进入专家自己的工作，而不再只是公共 demo

顶尖学者并非只看排行榜。他们会用模型攻击自己最熟悉、最难被营销蒙混的任务。

当数学家看到研究探索提速数倍，物理学家看到多年积累的推导被模型快速重建，生物学家看到 AlphaFold 改变结构预测，他们获得的是“局部但高可信”的私人证据。模型可能仍会在常识题上失败，却已经在某些高价值工作流中跨过净正收益线。

这种不均匀性解释了外界的困惑：普通用户看到的也许只是更流畅的聊天，领域专家看到的却是一个能生成代码、搜索假设、调用工具、接受自动验证的初级研究合作者。

3. 人才密度形成自我强化

优秀研究者想和优秀研究者工作。每当一个实验室聚集更多预训练专家、系统工程师、数学家、安全研究者与科学家，重要工作更可能在那里发生；重要工作越集中，下一位人才离开大学或其他公司进入该实验室的理由又越强。

这不只是一场薪酬竞标，也是隐性知识的聚集。大规模训练的许多判断无法完整写进论文：何时终止一次运行、怎样识别数据污染、哪个异常预示能力跃迁、什么评测正在被优化过头。参与真实运行的人更容易快速积累这些经验。

4. 反馈速度重写了“科研效率”

大学研究需要申请经费、排队使用计算资源、招学生、投稿与同行评审，这些机制保护开放性和质量，也让完整周期以月或年计。前沿实验室可以在一天内让研究者、工程系统、模型评测和产品数据往返多轮。

速度本身不是正确性。一个高速运行的封闭团队也可能集体走错方向。但当问题有清晰自动反馈——代码能否通过测试、证明能否形式验证、芯片布局能否改善指标——快速迭代会形成巨大优势。

5. “AI 帮助研究 AI”提高了早期位置的期权价值

若模型能承担更多编码、实验和搜索，研究者单位时间内能尝试更多假设；更多实验又产生训练和评测数据，帮助下一代模型。这个回路不必达到科幻式“智能爆炸”，也足以让领先组织按组织效率复利。

前沿实验室的人知道回路尚不完整。人仍在设定目标、选择评价函数、判断结果是否值得信任。恰因答案未定，研究者更愿意现在进入：如果关键规范、架构和安全习惯会在未来三年固化，晚十年再讨论伦理与治理，影响空间可能已经很小。

6. 使命、风险、身份与金钱共同定价

一位已经功成名就的教授、诺贝尔奖得主或 CTO 仍会在意金钱。高薪和股权补偿职业风险，为家庭提供安全，也给个人保留未来创业、资助研究或退出组织的能力。把钱从解释中删除，会把现实人物写成圣徒。

但金钱无法独立解释所有选择：

- 有人从 C-suite 头衔转成普通技术岗位；
- 有人暂停自己的公司，回到预训练研发；
- 有人选择强调安全的实验室，而不是报价最高的机构；
- 有人保留教授身份，用休假试探前沿；
- 也有人放弃巨头资源去创办路线不同的新实验室。

从公开陈述和职业路径只能推断一组共同变量：算力准入、研究杠杆、同侪密度、历史窗口、使命认同、薪酬股权、自主权损失与组织风险。它们不是一条可计算的公式，而是相互制约的门槛。若研究问题必须使用内部模型，算力准入接近零，其他条件再好也很难弥补；若组织失去信任，再多 GPU 也未必留得住人。

横纵交汇（二）：他们到底看到了什么

纵轴说明 AI 怎样变成可规模化的认知生产；横轴说明人才为什么向少数能运行这套生产系统的组织集中。两条线交汇后，可以看见五个更具体的判断。

他们看见了“认知活动的工业化”

工业革命不是发明了一台更强壮的手臂，而是把能量转换、机器、工厂和资本组织成可复制的生产系统。前沿 AI 正在对一部分认知劳动做相似的事：阅读、编码、搜索、比较、生成候选、调用工具与接受反馈，被装入同一个可扩展流程。

这里的关键不是模型是否像人，而是认知工作第一次可以被复制、并行、测量和持续更新。一个优秀研究员的时间每天只有二十四小时；一个模型可以同时帮助数千个团队。若质量达到可用阈值，哪怕没有全知全能，也会改写科研和组织的成本结构。

乔布斯看到的不是一块更好的手机屏幕，而是手指、软件、内容与供应链即将合成新的个人计算接口。这批人才看到的也不是“聊天机器人还会涨多少分”，而是模型、工具、验证、算力与分发正在合成新的认知基础设施。

他们看见了“所有科学的上游”

传统科研工具服务一个领域：望远镜观察宇宙，测序仪读取基因，粒子加速器探索高能物理。通用模型的特殊之处在于，它可以同时读取论文、写程序、设计分析、调用模拟、提出候选解释，并在数学、物理、生物和工程之间迁移。

它还远未成为完整科学家，却可能成为多种科学共同的上游工具。谁能改进这个工具，就可能同时提高许多学科的研究速度。对想理解世界的人而言，这比在单一问题上继续前进一小步更具诱惑。

这也是数学与物理人才被争夺的原因。数学提供可自动检查的反馈环境，也是训练一般推理的试验场；物理把符号推理、模拟、实验和世界模型连接起来。实验室需要的不只是他们已有的答案，更需要他们定义什么叫深问题、强证据和可信推理。

他们看见了“参与定义期”

一项技术刚进入社会时，很多默认设置还没有固定：模型追求什么目标，怎样服从人类，什么内容拒绝，研究如何公开，收益怎样分配，谁能审计，政府与公司如何协调。

研究者现在加入实验室，得到的不只是观察能力曲线的前排座位，也得到影响这些默认设置的机会。安全研究者面临一个悖论：站在外部可以保留批判距离，却很难接触最强系统；进入内部可以观察和干预真实风险，却会受雇主议程、保密制度与商业压力约束。

“我必须在里面才能改变方向”既可能是真诚使命，也可能成为自我合理化。判断它是否可信，要看研究者是否仍能公开异议、组织是否允许独立评测、治理是否有外部制衡，而不是只听使命口号。

他们看见了“能力的复利”，但还没有看见确定的 AGI

人才、算力、数据、部署和 AI 辅助研发可以构成复利式优势。这个判断有现实依据：模型已经能减少部分研发时间，实验室也把自动化研究列入明确路线。

更强的结论——递归自我改进必然发生、数年内一定出现超人通用智能——仍属于预测。2026 年的国际 AI 安全报告仍强调，现有系统在长程自主行动、可靠性和现实世界控制上存在明显限制。真实软件开发评测也发现，经验丰富的开发者有时会高估 AI 帮助，甚至因验证和修错而变慢。

人才流入是信号，不是证明。它说明一批最有资格观察前沿的人认为概率足够高，值得用职业生涯下注；它不说明下注已经赢了。

他们看见了“离权力源头更近”

“知识就是力量”只说了一半。知识要变成现实影响，还需要执行、资源、制度和合法性。AI 把知识生成与规模化行动之间的距离缩短了，所以它会提高多种权力的上限：

权力层	前沿实验室能做什么	仍受什么约束
认识论权力	决定哪些问题更快被搜索、验证和解释	事实、同行审查、领域专家、开放复现
基础设施权力	分配算力、模型、数据和内部工具的准入	芯片、能源、云供应链、资本
经济权力	把模型接入产品和组织，重组劳动与利润	市场、竞争、反垄断、客户信任
规范权力	通过训练、模型宪法与政策定义模型行为	法律、文化差异、公共监督
议程权力	决定研究预算、公开范围和风险优先级	董事会、员工、政府、公众

权力层	前沿实验室能做什么	仍受什么约束
强制权力	AI 可增强网络、情报、军事与行政能力	国家合法暴力、法律程序、国际关系

由此可以把“AI 等于权力”写得更准确：

> AI 最先放大的是 **power to**——做成事情的能力；当这种能力与模型控制权、部署权限、资本、强制资源或制度职位结合时，也可能转化为事实上的 **power over**。AI 不会自动赋予合法性，但可能显著降低支配、操纵和集中控制的成本。

一个研究员即使拥有最强模型，也不能独自制造先进芯片、调度电网、合法征税或要求社会服从。实际权力属于一组相互依赖的参与者：实验室治理层、研究团队、资本、云与芯片公司、能源系统、国家、部署机构和用户。

但也不能因此低估集中风险。当前沿模型能代替更多人的协调与执行时，小团体可能减少对大量合作者的依赖。Anthropic 的 Claude 宪法甚至明确把“AI 或包括 Anthropic 自身在内的一小群人借 AI 非法夺取权力”列为最严重风险之一；OpenAI 的 2026 计划也承认，转型技术既能集中权力，也能扩散权力。建造者自己都在用“权力”语言讨论问题，说明用户的直觉不是空想。

横纵交汇（三）：实验室化的公共代价

大学失去的不只是教授数量

教授离开一年，不只少一篇论文。他可能少带一届博士生、少开一门高阶课程、少参与同行评审，也少维护一个允许失败十年的研究方向。企业会优先投资能提高模型、产品、安全或政策优势的问题；那些无法快速进入训练与部署回路的学科，可能更难得到资源。

Berkeley 计算学院院长 Jennifer Chayes 对《The Atlantic》的担忧很精准：大学系所也许能存活，但开放创新体系能否存活并不确定。若最强模型、训练数据和验证结果都被少数公司掌握，其他科学家只能看到公司愿意公开的切片，科学共同体就从共同生产者变成受控接口的用户。

私有研究会加快发现，也会缩窄可见范围

产业研究不等于低质量研究。AlphaFold 已证明公司实验室可以做出划时代成果；集中工程能力也能完成大学难以协调的巨大项目。

问题在于选择权：哪些负结果不发布，哪些安全发现被保密，哪些数据无法审查，哪些模型只向付费客户开放。Stanford AI Index 指出，最强模型同时变得最不透明；参数量、训练数据、代码和训练周期常不再披露。研究能力越强，外部验证反而越弱，这会形成认识论上的单点故障。

数学界的警告不是反技术，而是争夺制度设计

Leiden Declaration 并没有要求数学家拒绝 AI。它要求披露工具使用、保留人类对正确性与引用的责任、坚持同行审查、建设公共计算设施，并提醒政府不要只听企业发布会。

这是一条重要反线：有些顶尖人才选择进入实验室，有些人选择留在外部建立检查机制。两者都可能是在认真回应同一变化。若所有批评者都进入公司，社会失去独立验证；若所有谨慎者都拒绝接触前沿，他们的判断又可能落后于真实能力。

一个健康体系需要三种角色同时存在：

1. 在内部建造并理解前沿系统的人；
2. 在大学和公共机构独立复现、批判与训练下一代的人；

3. 在政府、媒体和社会组织中把技术能力翻译成规则与公共选择的人。

只有第一种角色，AI 会强但不一定可问责；只有第二种角色，公共研究可能正确却没有实验装置；只有第三种角色，治理容易围绕过时想象立法。

情景推演：三种未来与可观察信号

基准情景：实验室—大学混合体长期存在

最可能的近中期形态是双向流动增加，而非大学消失：教授休假进入实验室，研究员回大学任教，公司与公共机构共同建设算力，学术界承担基础理论、人才训练和独立评测，企业承担最大训练与部署。

可以观察的信号包括：

- industrial leave 和联合任职是否多于永久离职；
- 公司是否继续发表可复现研究，而不只发布能力宣言；
- 大学能否获得公共计算资源与前沿模型审计权；
- 博士培养是否仍能产生不依附单一公司的独立议程。

乐观情景：认知杠杆扩散，而不是只向中心聚集

模型能力通过开放权重、低成本接口、公共算力和透明评测扩散；个人与小团队获得过去只有大机构拥有的研究能力。实验室保留竞争优势，却接受外部审计、事故报告和公共治理。AI 加快科学，同时让更多人进入科学。

早期信号会是：

- 前沿与开放模型之间的能力差距持续缩小；
- 公共科研云和国际计算设施实际投入运行；
- 自动化成果能被独立复现，训练数据与工具使用清楚披露；
- AI 提高新研究者的产出，而不是只放大原有明星和平台。

危险情景：认知基础设施成为私人关卡

少数实验室控制最强模型、芯片合同、能源、人才与分发，并把更多 AI 研发成果留在内部。大学无法验证能力声明，政府又依赖同一批公司提供技术意见和系统。模型帮助中心组织更快行动，却削弱员工、公众与其他机构的议价能力。

警报信号包括：

- 重要模型的训练与评测信息继续减少；
- 关键安全结果只以摘要出现，独立研究者无法检查；
- AI 辅助研发的收益主要转化为更快封闭迭代；
- 教授流失导致课程、导师与公开论文持续下降；
- 实验室的安全治理依赖创始人承诺，缺少可执行的外部制衡；
- 政府把模型采购和 AI 政策长期绑定给极少数供应商。

这三个情景可以同时发生在不同层面。开放模型可能扩散日常能力，最前沿训练仍高度集中；科学工具可能普惠，军事和情报能力仍被严格封闭。真正该追踪的不是一句“AGI 来没来”，而是谁拥有模型、谁能验证、谁能退出、谁承担失败，以及权力是否有制衡。

回到最初的问题：他们看到了什么？

他们看到的，不是一个已经完成的答案，而是一组正在同时改变的约束：

- 智能的一部分已经能被训练、复制和规模化；
- 自然语言正成为连接知识、代码、工具与行动的通用接口；
- AI 已在部分专家工作里越过净收益阈值；
- 前沿实验越来越依赖少数实验室掌握的完整基础设施；
- AI 参与 AI 研发，可能让组织优势累积得更快；
- 未来几年仍是模型目标、安全规范、科研制度和利益分配的参与定义期。

他们去前沿实验室，不一定因为那里已经拥有“最高级智能”，而是因为那里最接近制造下一代智能的机器；不一定因为每个人渴望统治，而是因为理解世界、改变世界、避免错误和获得影响力，在这个位置上突然叠到了一起。

钱是真实的，使命也是真实的，权力欲可能存在，对错尚未揭晓。最值得认真对待的不是他们的光环，而是他们用职业选择传出的概率判断：即使 AGI 并非确定事件，AI 成为科学、组织和国家的新认知基础设施，概率已经高到足以让最有选择权的一批人重新安排人生。

乔布斯式的“看见”，不是知道一切，而是在多数人仍按旧分类思考时，看见几个原本分开的系统即将连起来。2007 年前，电话、音乐、互联网与触控仍像不同产品；今天，模型、代码、科学、资本、算力与治理也仍被分在不同新闻栏目里。

这批人押注的是：它们已经属于同一个故事。

主要信息来源

历史与范式

1. McCulloch & Pitts, A Logical Calculus of the Ideas Immanent in Nervous Activity, 1943。
2. Alan Turing, Computing Machinery and Intelligence, 1950。
3. Dartmouth, A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, 1955。
4. Rumelhart, Hinton & Williams, Learning representations by back-propagating errors, 1986。
5. Krizhevsky, Sutskever & Hinton, ImageNet Classification with Deep Convolutional Neural Networks, 2012。
6. University of Toronto, Google acquires University of Toronto deep learning startup, 2013。
7. Vaswani et al., Attention Is All You Need, 2017。
8. OpenAI, Introducing OpenAI, 2015；OpenAI LP, 2019。
9. Kaplan et al., Scaling Laws for Neural Language Models, 2020。
10. Nobel Prize, The Nobel Prize in Chemistry 2024。

2026 人才、组织与能力

1. UC Berkeley EECS, Changing of the Guard: Welcoming Ana Arias as EECS Department Chair, 2026-07-02。
2. SFGATE, UC Berkeley AI expert leaves for Anthropic, 2026-07。
3. TechCrunch, Andrej Karpathy joins Anthropic’s pre-training team, 2026-05-19。
4. TechCrunch, John Jumper leaves DeepMind for Anthropic, 2026-06-20。
5. Reuters, Google Gemini co-lead Noam Shazeer leaves for OpenAI, 2026-06。
6. The Atlantic, Where Did All the Computer-Science Professors Go?, 2026-07-21。
7. OpenAI, Built to benefit everyone: our plan, 2026-06-08。
8. Anthropic, Introducing our Science Blog, 2026-03-23；Claude’s Constitution。
9. Stanford HAI, 2026 AI Index — Research and Development, 2026。

10. UChicago BFI, Attention and Money Is All You Need? Why Universities Are Struggling to Keep AI Talent。
11. Harvard Crimson, AI Wrote a Harvard Physicist's Most Recent Paper, 2026-04-24。
12. OpenAI Academy, Alex Lupsasca: black hole physics and hidden symmetries；Terence Tao: AI is ready for primetime in math and theoretical physics。
13. Leiden Declaration, Leiden Declaration on Artificial Intelligence and Mathematics, 2026-06。
14. International AI Safety Report, International AI Safety Report 2026, 2026。
15. METR, Early-2025 AI experienced open-source developer study；Time Horizon 1.1。
16. The Information, Workday CTO Joins Anthropic, 2026-04。
17. Anthropic, Introducing The Anthropic Institute；AI, R&D, and the possibility of recursive self-improvement, 2026。

> 所有网页资料访问于 2026-07-24。对 2026 年新近任职，优先采用本人、学校、公司公告；无法取得一手确认时，采用两家可靠媒体交叉核对或明确标注不确定性。

方法论说明

本报告使用横纵分析法：

- **纵向分析**追踪 1943—2026 年智能研究的生产方式：从手写规则、可训练表示，到 scaling、AI for Science 与 AI 辅助 AI 研发，并识别每次范式转换后稀缺资源和研究中心的变化。
- **横向分析**比较 2026 年 Anthropic、OpenAI、Google DeepMind、Meta、xAI、创业实验室与大学的资源、使命、人才路径和制度代价。
- **交汇分析**检验两条线能否共同解释“为什么是这些人、为什么是现在”，并把“AI 等于无上权力”的命题拆成认识论、基础设施、经济、规范、议程与强制权力。

研究局限有三点。公开名单会漏掉未披露任职，也容易把休假误写成永久离职；公司对能力和使命的陈述带有招募、融资与政策沟通动机；2026 年的大量事件仍在发展，无法用长期结果验证。因此，文中的确定结论集中在已发生的人才流动与基础设施变化，对自动化科学、递归研发和 AGI 时间表只做条件性判断。