

2025年人工智能 指数报告



Stanford University
Human-Centred
Artificial Intelligence

介绍

2025年人工智能指数报告

欢迎阅读第八版人工智能指数报告。恰逢人工智能对社会、经济和全球治理的影响不断加强的关键时期，我们在这一重要时刻发布了 2025 年人工智能指数报告。它也是我们迄今为止发布过的最全面的指数。今年的报告新增了对人工智能硬件发展状况的深入分析、对推理成本的新估算，以及对人工智能论文发表和专利申请趋势的新分析。我们还首次披露了企业采用负责任的人工智能实践的最新数据，并扩展了人工智能在科学和医学领域日益重要的作用的分析维度。

发起于 2017 年，作为“人工智能百年研究 (One Hundred Year Study of Artificial Intelligence)” 项目分支，人工智能指数报告一直致力于为政策制定者、新闻工作者、高管、研究人员和公众提供准确、经过严格验证和全球来源的数据。我们的使命始终如一：帮助这些利益相关方就人工智能的发展和部署做出更明智的决策。在这个从会议室到厨房餐桌到处都在讨论人工智能的世界里，这一使命显得尤为重要。

从地缘政治格局的变化和底层技术的快速发展，到人工智能在商业、决策和公共生活中不断扩大的作用，人工智能指数持续引领行业关键趋势的追踪与解读。纵向追踪始终是我们的核心任务。在这一高速发展的领域，本报告提供了重要的背景信息：帮助我们了解人工智能的现状、发展历程以及未来走向。

作为全球公认的人工智能领域权威资源之一，人工智能指数报告被《纽约时报》、彭博社和《卫报》等主要媒体引用，成为数百篇学术论文的文献参考，并服务于世界各地的政策制定者和政府机构。我们已经向埃森哲、IBM、富国银行和富达等公司提供了人工智能现状的简报，并将继续为全球人工智能生态系统输送独立见解。

联合主任致辞

随着人工智能持续重塑人类生活、企业界和公共话语体系，人工智能指数报告始终跟踪其进展情况，通过独立的、数据驱动的视角，跨时间、跨地域地全方位观察人工智能的发展、应用和影响。

对于人工智能来说，2024年是多么美好的一年。诺贝尔物理学奖和化学奖，以及因强化学习方面的奠基性工作而获得的图灵奖，都体现了人们对人工智能在推动人类知识进步方面所起作用的认可。曾经令人望而生畏的图灵测试已不再被视为一个雄心勃勃的目标，今天的精尖系统已经超越了它。与此同时，人工智能的应用正以前所未有的速度渗透社会生活，数以百万计的人们在专业工作和休闲活动中高频使用人工智能。随着高性能、低成本和开源模型的普及，人工智能的可及性和影响力必将进一步扩大。在经历了短暂的放缓之后，企业对人工智能的投资出现反弹。生成式人工智能初创企业融资案例数量增加了近三倍。商业应用在经过多年的低迷后于2024年迎来显著增长。人工智能已从边缘领域成为业务价值的核心驱动力。

各国政府也在加大参与力度。政策制定者们不再停留于讨论人工智能，他们正在对其进行投资。一些国家启动了价值数十亿美元体量的国家人工智能基础设施计划，包括能源扩容以支持人工智能发展的重大努力。全球协作机制日益完善，地方性措施也同步成型。

然而，信任仍然是一项重大挑战。公众对人工智能企业数据保护能力的信任度持续走低，对算法公平性和偏见的担忧依然存在。虚假信息继续构成风险，深度伪造技术在选举等场景的滥用引发广泛担忧。对此，各国政府正在推进新的监管框架，旨在促进透明度、负责任和公平性。公众的态度也在转变。2024年的一项全球调查显示，尽管疑虑犹存，公众对人工智能带来广泛社会效益的潜力的乐观情绪明显上升。

人工智能不再只是一个关于可能发生什么的故事，而是一个关于正在发生什么以及我们如何共同塑造人类未来的故事。敬请品读本年度《人工智能指数报告》并亲自洞见这一切。

Yolanda Gil 和 Raymond Perrault
《人工智能指数报告》联合主任

核心要点

1. **人工智能在严苛比较基准测试中的性能持续提升。** 2023 年, 研究人员推出了 MMMU、GPQA 和 SWE-bench 等一系列新型比较基准, 旨在测试前沿人工智能系统的极限。仅一年后, 性能就大幅提升: MMMU、GPQA 和 SWE-bench 的得分分别提高了 18.8%、48.9% 和 67.3%。除这些比较基准, 人工智能系统在生成高质量视频方面也取得了重大进展, 在某些特定场景下, 基于语言模型的智能体在时间受限的编程任务中甚至表现优于人类。

2. **人工智能日益融入日常生活。** 从医疗到交通, 人工智能正迅速从实验室走向日常生活。2023 年, 美国食品和药物管理局 (FDA) 批准的人工智能医疗设备达 223 款, 较 2015 年的 6 款实现跨越式增长。在公共道路上, 自动驾驶汽车已走出试验阶段: 美国头部运营商之一 Waymo 每周提供超过 15 万次自动驾驶乘车服务, 而百度推出的经济型 Apollo Go 自动驾驶出租车所提供的服务目前已覆盖中国众多城市。

3. **随着不断的研究显示出人工智能对生产效率的强大影响, 企业界全面拥抱人工智能, 投资与应用双创新高。** 2024 年, 美国私人人工智能投资达 1091 亿美元, 约为中国 (93 亿美元) 的 12 倍、英国 (45 亿美元) 的 24 倍。生成式人工智能的发展势头尤为强劲, 吸引了全球 339 亿美元的私人投资——同比增长了 18.7%。人工智能的商业应用也在加速普及, 78% 的企业在 2024 年应用了人工智能技术, 较前一年的 55% 有所提升。同时, 越来越多的研究证实, 人工智能不仅可以提高生产效率, 在多数情况下还有助于缩小劳动力的技能差距。

4. **美国在开发顶级人工智能模型方面仍处于领先地位, 但中国正在缩小与美国的差距。** 2024 年, 美国机构共开发了 40 个标志性的人工智能模型, 而中国只有 15 个, 欧洲只有 3 个。虽然美国在数量上保持领先, 但中国的模型在质量上迅速缩小了差距: 在 MMLU 和 HumanEval 等主要比较基准上的性能差距从 2023 年的两位数缩小到 2024 年的接近持平。中国在人工智能论文和专利方面继续保持领先。模型开发日益全球化, 中东、拉美和东南亚都推出了引人注目的模型。

5. **负责任的人工智能生态系统发展不平衡。** 在人工智能相关的事故激增的同时, 主要的工业模型开发商采用标准化的负责任的人工智能 (Responsible AI, RAI) 评测仍然很少见。不过, HELM Safety、AIR-Bench 和 FACTS 等新比较基准为评估真实性和安全性提供了前景广阔的工具。企业层面, 对负责任的人工智能风险的认知与实质性行动之间仍存在差距。相较而言, 各国政府则表现出了更强的紧迫感: 2024 年, 全球人工智能合作显著深化, 经合组织、欧盟、联合国和非盟等组织相继发布了监管框架, 聚焦透明度、可信度等负责任的人工智能核心原则。

核心要点（续）

6. 全球对人工智能的乐观情绪正在攀升，但地区间仍存在较大差异。在中国（83%）、印度尼西亚（80%）和泰国（77%）等国家，绝大多数人认为人工智能产品和服务利大于弊。相比之下，加拿大（40%）、美国（39%）和荷兰（36%）等地的乐观程度仍然较低。不过，人们的情绪正在转变。自 2022 年以来，多个原持怀疑态度的国家的乐观情绪大幅增长，包括德国（+10%）、法国（+10%）、加拿大（+8%）、英国（+8%）和美国（+4%）。

7. 人工智能变得更加高效、经济和易用。依托小型模型能力跃升，执行 GPT-3.5 级别的系统的推理成本在 2022 年 11 月至 2024 年 10 月间骤降 280 多倍。硬件层面，年化成本降幅达 30%，能效年提升率达 40%。开源模型正在缩小与闭源模型的差距，在某些比较基准上，性能差距从 8% 缩小到仅 1.7%。这些趋势加在一起，正在迅速降低先进人工智能的应用门槛。

8. 各国政府正在加强对人工智能的监管和投资。2024 年，美国联邦机构出台了 59 项与人工智能相关的法规，是 2023 年的两倍多，发布法规的机构数量也是 2023 年的两倍。全球范围内，75 个国家的人工智能立法提及率自 2023 年以来增长 21.3%，较 2016 年累计增幅达 9 倍。加强监管的同时，各国政府展开大规模投资：加拿大承诺投资 24 亿美元，中国启动了 475 亿美元的半导体基金，法国承诺投资 1090 亿欧元，印度拨款 12.5 亿美元，沙特阿拉伯更推出规模达千亿美元的“超越计划（Project Transcendence）”。

9. 人工智能和计算机科学教育加速普及，但在资源获取和准备程度方面仍存落差。全球三分之二的国家现在已实施或规划基础教育阶段计算机科学教育，覆盖率是 2019 年的两倍，其中非洲和拉美地区进展最为显著。美国过去 10 年中计算机专业本科毕业生人数增长 22%。然而，在许多非洲国家，由于电力等基础设施的不足，获得计算机学位的机会仍然有限。在美国，81% 基础教育阶段的计算机教师认同将人工智能纳入基础课程，但只有不到一半的教师认为自己具备相关教学能力。

10. 产业依然在人工智能竞争中引领——但技术前沿竞争加剧。2024 年，全球近 90% 的标志性的人工智能模型来自产业界，高于 2023 年的 60%，而学术界仍然是高引用率研究的首要来源。模型规模持续快速扩张——训练计算每五个月翻一番，数据集每八个月倍增，能耗年增速 100%。然而模型性能差距正在缩小，一年内，榜首和第十名模型的 Elo 技能得分差距从 11.9% 降至 5.4%，现在前两名的差距仅为 0.7%。技术前沿领域的竞争日趋激烈，头部阵营也日益集聚。

核心要点（续）

11. **人工智能因其科学影响力获得最高学术荣誉。**人工智能日益增长的重要性在重大科学奖项中得到了体现：两项诺贝尔奖分别授予深度学习（物理学奖）及人工智能在蛋白质折叠中的应用（化学奖），图灵奖则颁给了对强化学习的突破性贡献。

12. **复杂推理仍是一项挑战。**人工智能模型在国际数学奥林匹克竞赛问题等任务中表现出色，但在 PlanBench 等复杂推理比较基准中仍然举步维艰。即使存在理论正确解法，它们也常常无法可靠地解决逻辑任务。在精度至关重要的领域里，人工智能的有效性依然存在局限。

指导委员会

主席

Raymond Perrault
SRI 国际研究院

当选主席

Yolanda Gil
南加州大学信息科学研究所

成员

Erik Brynjolfsson
斯坦福大学

Jack Clark
Anthropic, OECD

John Etchemendy
斯坦福大学

Katrina Ligett
希伯来大学

Terah Lyons
摩根大通

James Manyika
谷歌
牛津大学

Juan Carlos Niebles
斯坦福大学
Salesforce

Vanessa Parli
斯坦福大学

Yoav Shoham
斯坦福大学
AI21 实验室

Russell Wald
斯坦福大学

Tobi Walsh
悉尼新南威尔士大学

工作人员和研究人员

研究主管兼总编辑

Nestor Maslej, 斯坦福大学

研究助理

Loredana Fattorini, 斯坦福大学

附属研究员

Elif Kiesow Cortez, 斯坦福法学院研究员

Julia Betts Lotufo, 研究员

Anka Reuel, 斯坦福大学

Alexandra Rome, 研究员

Angelo Salatino, 英国开放大学知识媒体研究所

Lapo Santarasci, 卢卡高等研究学院

研究生研究员

Emily Capstick, 斯坦福大学

Malou van Draanen Glismann, 斯坦福大学

Njenga Kariuki, 斯坦福大学

本科生研究员

Armin Hamrah, 克莱蒙特·麦肯纳学院

Sukrut Oak, 斯坦福大学

Ngorli Fiifi Paintsil, 斯坦福大学

Andrew Shi, 斯坦福大学

如何引用本报告

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, Sukrut Oak. “The AI Index 2025 Annual Report,” AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2025.

The AI Index 2025 Annual Report by Stanford University is licensed under Attribution-NoDerivatives 4.0 International.

公共数据和工具

《2025 年人工智能指数报告》附有原始数据和互动工具。我们邀请每位读者根据自己的工作和兴趣使用这些数据和工具。

- 原始数据和图表：报告中所有图表：的公开数据和高分辨率图像可在 Google Drive 上获取。
- Global AI Vibrancy Tool：比较 30 多个国家的人工智能生态系统。Global AI Vibrancy Tool 将于 2025 年夏季更新。

人工智能指数（AI Index）和斯坦福大学 HAI

人工智能指数（AI Index）是斯坦福大学以人为本人工智能研究院（Stanford Institute for Human-Centered Artificial Intelligence, HAI）的独立研究项目。



人工智能指数（AI Index）最初源自人工智能百年研究（One Hundred Year Study on Artificial Intelligence）项目（AI100）



首个官方中文版由 AI Index 与其亚洲合作伙伴 Digital Civilization 合作组织与统筹，作为拓展 AI Index 在亚洲影响力并推动区域生态体系建设的一项重要举措。我们欢迎来自各界的个人与机构提供反馈并参与合作，共同推动构建一个更具包容性、以人为本的人工智能社区。

人工智能指数欢迎对明年报告的反馈和新想法。请通过 nmaslej@stanford.edu 联系我们。

人工智能指数承认，尽管该报告由一支人类研究人员团队撰写，但其写作过程得到了人工智能工具的辅助。具体而言，作者使用了 ChatGPT 和 Claude 帮助完善和校对初稿。工作流程包括作者撰写原始内容，并将在编辑过程中作为工作流程的一部分使用人工智能工具。

支持型合作伙伴



分析研究合作伙伴



贡献者

人工智能指数谨此感谢以下各位专家在各章节和部分中为《2025 年人工智能指数报告》提供的数据、分析、建议及专业评论：

介绍

Loredana Fattorini, Yolanda Gil, Nestor Maslej, Vanessa Parli, Ray Perrault

第一章 研究与开发

Nancy Amato, Andrea Brown, Ben Cottier, Lucía Ronchi Darré, Virginia Dignum, Meredith Ellison, Robin Evans, Loredana Fattorini, Yolanda Gil, Armin Hamrah, Katrina Ligett, Nestor Maslej, Maurice Pagnucco, Ngorli Fiifi Paintsil, Vanessa Parli, Ray Perrault, Robi Rahman, Christine Raval, Vesna Sabljakovic–Fritz, Angelo Salatino, Lapo Santarasci, Andrew Shi, Nathan Sturtevant, Daniel Weld, Kevin Xu, Meg Young

第二章 技术性能

Rishi Bommasani, Erik Brynjolfsson, Loredana Fattorini, Tobi Gertsenberg, Yolanda Gil, Noah Goodman, Nicholas Haber, Armin Hamrah, Sanmi Koyejo, Percy Liang, Katrina Ligett, Nestor Maslej, Juan Carlos Niebles, Sukrut Oak, Vanessa Parli, Marco Pavone, Ray Perrault, Anka Reuel, Andrew Shi, Yoav Shoham, Toby Walsh

第三章 负责任的人工智能

Medha Bankhwal, Emily Capstick, Dmytro Chumachenko, Patrick Connolly, Natalia Dorogi, Loredana Fattorini, Ann Fitz–Gerald, Yolanda Gil, Armin Hamrah, Ariel Lee, Katrina Ligett, Shayne Longpre, Nestor Maslej, Katherine Ottenbreit, Halyna Padalko, Vanessa Parli, Ray Perrault, Brittany Presten, Anka Reuel, Roger Roberts, Andrew Shi, Georgio Stoev, Shekhar Tewari, Dikshita Venkatesh, Cayla Volandes, Jakub Wiatrak

第四章 经济

Medha Bankhwal, Erik Brynjolfsson, Mar carpanelli, Cara Christopher, Michael Chui, Natalia Dorogi, Heather English, Murat Erer, Loredana Fattorini, Yolanda Gil, Heather Hanselman, Rosie Hood, Vishy Kamalapuram, Kory Kantenha, Njenga Kariuki, Akash Kaura, Elena Magrini, Nestor Maslej, Katherine Ottenbreit, Vanessa Parli, Ray Perrault, Brittany Presten, Roger Roberts, Cayla Volandes, Casey Weston, Hansen Yang

第五章 科学与医学

Russ Altman, Kameron Black, Jonathan Chen, Jean–Benoit Delbrouck, Joshua Edrich, Loredana Fattorini, Alejandro Lozano, Yolanda Gil, Ethan Goh, Armin Hamrah, Fateme Nateghi Haredasht, Tina Hernandez–Boussard, Yeon Mi Hwang, Rohan Koodli, Arman Koul, Curt Langlotz, Ashley Lewis, Chase Ludwig, Stephen P. Ma, Abdoul Jalil Djiberou Mahamadou, David Magnus, James Manyika, Nestor Maslej, Gowri Nayar, Madelena Ng, Sophie Ostmeier, Vanessa Parli, Ray Perrault, Malkiva Pillai, Ossian Karl–Johan Ferdinand Rabow, Sean Riordan, Brennan Geti Simon, Kotoha Togami, Artem Trotsuk, Maya Varma, Quinn Waeiss, Betty Xiong

第六章 政策

Elif Kiesow Cortez, Loredana Fattorini, Yolanda Gil, Julia Betts Lotufo, Vanessa Parli, Ray Perrault, Alexandra Rome, Lapo Santarasci, Georgio Stoev, Russell Wald, Daniel Zhang

贡献者（续）

第七章 教育

John Etchemendy, Loredana Fattorini, Lili Gangas, Yolanda Gil, Rachel Goins, Laura Hinton, Sonia Koshy, Kirsten Lundgren, Nestor Maslej, Lisa Cruz 11ohatski, Vanessa Parli, Ray Perrault, Allison Scott, Andreen Soley, Bryan Twarek, Laurens Vehmeijer

第八章 公共观点

Emily Capstick, John Etchemendy, Loredana Fattorini, Yolanda Gil, Njenga Kariuki, Nestor Maslej, Vanessa Parli, Ray Perrault

人工智能指数谨此感谢以下各位专家在各章节和部分中为《2025 年人工智能指数报告》提供的数据、分析、建议及专业评论：

组织机构

Accenture

Arnab Chakraborty, Patrick Connolly, Shekhar Tewari, Dikshita Venkatesh, Jakub Wiatrak

Epoch AI

Ben Cottier, Robi Rahman

GitHub

Lucía Ronchi Darré, Kevin Xu

Lightcast

Cara Christopher, Elena Magrini

LinkedIn

03 Carpanelli, Akash Kaura Kory Kantenga, Rosie Hood, Casey Weston

McKinsey & Company

Medha Bankhwal, Natalia Dorogi, Katherine Ottenbreit, Brittany Presten, Roger Roberts, Cayla Volandes

Quid

Heather English, Hansen Yang

人工智能指数还感谢以下人士在准备本报告过程中提供的帮助：Jeanina Matias、Nancy King、Carolyn Lehman、Shana Lynch、Jonathan Mindes 和 Michi Turner；感谢 Christopher Ellis 在维护人工智能指数网站方面提供的帮助；以及感谢 Annie Benisch、Stacey Sickels Boyce、Marc Gough、Caroline Meinhardt、Drew Spence、Casey Weston、Madeleine Wright 和 Daniel Zhang 在帮助推广本报告方面所做的工作。

我们感谢 Jason Liu 在推动 AI Index 中文版过程中所发挥的重要作用。展望未来，我们将持续致力于提升 AI Index 的全球可及性，并加强在亚洲区域的生态协同建设。

目录

报告核心要点		12
第一章	研究与开发	24
第二章	技术性能	81
第三章	负责任的人工智能	160
第四章	经济	214
第五章	科学与医学	280
第六章	政策与治理	323
第七章	教育	364
第八章	公共观点	394
附录		414

获取公共数据

报告核心要点

第一章： 研究与开发

1. 产业界持续加大人工智能投入并主导标志性人工智能模型研发，而学术界则引领高影响力研究。产业界在标志性人工智能模型研发中的主导优势持续强化，这一趋势在既往两期人工智能指数报告中已有明确阐述。2024 年，近 90% 的前沿模型源自产业界（较2023 年的 60% 提升 30 个百分点）。学术界在过去三年中始终是高被引（前 100 名）论文的首要机构贡献者。

2. 中国在人工智能论文发表总量上领先，而美国在具有高影响力的研究方面占据优势。2023 年，中国在人工智能领域的论文发表量（23.2%）和引用量（22.6%）均位居全球首位。而在过去三年中，美国机构贡献了最多被引用次数排名前 100 的人工智能论文。

3. 人工智能论文发表总量持续增长，并在计算机科学领域占据日益重要的主导地位。从 2013 年到 2023 年，在计算机科学和其他科学学科领域发表的人工智能论文发表总数几乎增加了两倍，从大约 102,000 篇增加到超过 242,000 篇。从比例上看，人工智能在计算机科学论文中所占的份额已从 2013 年的 21.6% 上升到 2023 年的 41.8%。

4. 美国仍然是标志性人工智能模型的主要来源地。2024 年，美国机构开发了 40 个标志性人工智能模型，大大超过中国的 15 个和欧洲的 3 个。在过去十年，源自美国的标志性机器学习模型比任何其他国家都多。2024 年，美国机构共开发出 40 个前沿人工智能模型，显著超过中国的 15 个和欧洲地区的合计的 3 个。过去十年间，美国在前沿机器学习模型的研发数量上持续保持全球领先地位。

5. 人工智能模型正变得规模更大、算力需求更高、能耗更强。最新研究表明，标志性人工智能模型的训练算力需求约每 5 个月翻一番，大语言模型训练数据集规模每 8 个月翻一番，而训练所需功耗每年递增。大规模的产业投资持续推动着模型规模扩大和性能提升。

6. 人工智能模型的使用成本持续下降。以语言模型性能评估常用基准——MMLU 为例，达到 GPT-3.5 水平（64.8 分）的人工智能模型查询成本已从 2022 年 11 月的每百万词元 20 美元，降至 2024 年 10 月的仅每百万词元 0.07 美元（Gemini-1.5-Flash-8B 模型），18 个月内降幅超 280 倍。根据任务类型不同，大语言模型的推理价格年降幅已达 9 至 900 倍不等。

报告核心要点

第一章：

研究与开发(续)

7. 人工智能专利申请持续上升。从 2010 年到 2023 年，人工智能专利数量稳步大幅增长，从 3833 项激增到 122511 项。仅在去年，人工智能专利数量就增长了 29.6%。截至 2023 年，中国在人工智能专利总数上遥遥领先，占所有授权的 69.7%，而按人均计算，韩国和卢森堡则是人工智能专利的主要生产国。

8. 人工智能硬件变得更快、更便宜、更节能。最新研究表明，以 16 位浮点运算能力衡量的机器学习硬件性能正以每年 43% 的速度增长，每 1.9 年实现翻倍。其性价比显著提升——成本每年下降 30%，而能效则以每年 40% 的幅度持续改善。

9. 人工智能训练产生的碳排放量正在稳步上升。早期训练的人工智能模型，如 AlexNet 网络(2012 年)，碳排放量不大，仅为 0.01 吨。最新人工智能模型的训练碳排放量显著增加：2020 年 GPT-3 为 588 吨，2023 年 GPT-4 达 5,184 吨，而 2024 年 Llama 3.1 405B 更是高达 8,930 吨。作为对比，美国人年均碳排放量仅为 18 吨。

第二章：

技术性能

1. 人工智能达成新比较基准的速度比以往任何时候都快。2023 年，研究人员推出了 MMMU、GPQA 和 SWE-bench 等多个具有挑战性的新比较基准，旨在测试日益强大的人工智能系统极限。到 2024 年，人工智能在这些比较基准上的表现取得显著突破：MMMU 和 GPQA 测试成绩分别提升 18.8 和 48.9 个百分点；更引人注目的是，在 SWE-bench 编程测试中，人工智能系统的解题能力从 2023 年仅能解决 4.4% 的问题，跃升至 2024 年的 71.7%。

2. 开源模型迎头赶上。根据去年发布的人工智能指数报告，领先的开源模型曾大幅落后于闭源模型。而到 2024 年，这一差距已基本消失。具体来看，2024 年 1 月初，在 Chatbot Arena Leaderboard 上，顶尖闭源模型的性能优势为 8.0%；而到 2025 年 2 月，这一差距已缩小至 1.7%。

报告核心要点

第二章：

技术性能(续)

3. 中美人工智能模型能力差距缩小。2023 年, 美国顶尖人工智能模型性能曾大幅领先中国同类产品, 但这一态势现已改变。数据显示: 截至 2023 年底, 在 MMLU、MMMU、MATH 和 HumanEval 等比较基准中, 中美模型的性能差距分别为 17.5、13.5、24.3 和 31.6 个百分点; 而到 2024 年末, 这些差距已大幅收窄至 0.3、8.1、1.6 和 3.7 个百分点。

4. 前沿的人工智能模型的性能趋于收敛。根据去年的人工智能指数, Chatbot Arena Leaderboard 上排名第一与第十的模型间 Elo 分数差已从去年的 11.9% 收窄至 2025 年初的 5.4%。同样, 排名前两位的模型之间的差距也从 2023 年的 4.9% 缩小到 2024 年的 0.7%。人工智能领域的竞争日趋激烈, 如今有越来越多的开发者推出了高质量的模型。

5. 新型推理范式, 如测试时计算(test-time compute), 显著提升模型性能。2024 年, OpenAI 推出的 o1、o3 等模型采用迭代式输出推理架构。这种测试时计算方法极大地提高了模型的性能, o1 在国际数学奥林匹克资格考试中获得了 74.4% 的高分, GPT-4o 只有 9.3%。但该技术存在代价, o1 的运算成本激增至 GPT-4o 的 6 倍, 推理速度则降低 30 倍。

6. 更具挑战性的比较基准不断被提出。MMLU、GSM8K 和 HumanEval 等传统人工智能比较基准已趋近饱和, 加上 MMMU 和 GPQA 等更新的更具挑战性的比较基准表现持续提升, 促使研究人员探索更多针对领先人工智能系统的评价方法。其中值得注意的有人类最后的考试(Humanity's Last Exam), 这是一项严格的学术测试, 最先进的人工智能系统的得分率仅为 8.80%; 前沿数学(Frontier Math), 这是一项复杂的数学比较基准, 人工智能系统仅解决了 2% 的问题; “BigCodeBench”这是一项编码比较基准, 人工智能系统的成功率仅为 35.5%, 远低于人类 97% 的水平。

7. 高质量人工智能视频生成模型实现重大突破。2024 年, 多款能够根据文本输入生成高清视频的先进人工智能模型相继问世, 其中包括 OpenAI 的 SORA、Stable Video Diffusion 3D 和 4D、Meta 的 Movie Gen, 以及谷歌 DeepMind 的 Veo 2。与 2023 年的视频生成模型相比, 这些新一代模型在画质表现上取得显著提升。

报告核心要点

第二章： 技术性能(续)

8. **小型模型展现更强性能。**2022 年，在 MMLU 比较基准中达到 60% 以上分数的最小模型是拥有 5400 亿参数的 PaLM；而到 2024 年，微软推出的 Phi-3-mini 仅用 38 亿参数就实现了相同水平——相当于两年间参数规模缩减了 142 倍。
9. **复杂推理仍是人工智能面对的难题。**尽管通过思维链(Chain-of-Thought)等推理机制的引入显著提升了大语言模型的性能，这些系统仍无法可靠解决本可通过逻辑推理获得确定性答案的问题——包括数学运算和任务规划等，尤其当问题规模超出其训练范围时。这一缺陷严重影响了人工智能系统的可信度，使其难以胜任高风险场景的应用需求。
10. **人工智能智能体展现初步潜力。**2024 年推出的 RE-Bench 基准为评估 AI 智能体的复杂任务能力建立了严格标准。在短时任务(2 小时时限)中，顶级人工智能系统的得分可达人类专家的 4 倍；但随着时间延长至 32 小时，人类表现反超人工智能系统，得分达到 2:1 的优势。AI 智能体已在特定领域，如编写 Triton Kernels，达到人类专业水平，且能更快、更低成本地得出结果。

第三章： 负责任的人工智能

1. **目前，依据负责任的人工智能(RAI)准则对人工智能系统进行评估的做法尚未普及，但新的比较基准体系正在逐步形成。**去年的人工智能指数曾着重指出，目前缺乏针对大语言模型的标准化 RAI 比较基准。虽然这一问题依然存在，但 HELM Safety 和 AIR-Bench 等新比较基准的出现有助于填补这一空白。
2. **人工智能事件报告数量持续增加。**根据人工智能事件数据库(AI Incidents Database)统计，2024 年报告的人工智能相关事件增至 233 起，创历史新高，比 2023 年增加 56.4%。

报告核心要点

第三章： 负责任的人工智能(续)

3. 各类机构虽意识到负责任的人工智能风险，但风险缓解措施滞后。麦肯锡一项关于企业 RAI 实施情况的调查显示，尽管多数机构能识别关键 RAI 风险，但并非所有机构都采取了积极应对措施。在领导者最关注的风险维度中，模型准确性问题(64% 受访者提及)、合规性风险(63%)以及网络安全威胁(60%)位列前三，但值得注意的是，将这些风险列为核心关注点的受访者比例均未超过 65%。

4. 在全球范围内，政策制定者对负责任的人工智能表现出浓厚兴趣。2024 年，全球加强了人工智能治理方面的合作，重点是协商确定负责任的人工智能的原则。多个国际组织，包括经济合作与发展组织(OECD)、欧盟、联合国及非洲联盟，相继发布规范性框架，阐释了透明度与可解释性、可信度等 RAI 重点。

5. 公共数据资源正在迅速萎缩。人工智能模型的训练依赖于海量公开网络数据，但最新研究表明，2023 至 2024 年间数据使用限制显著增加，因为众多网站实施了新协议以限制人工智能训练的数据爬取。在 C4 通用爬取数据集持续维护的域名中，受限制文本数据的比例已从 5–7% 骤升至 20–33%。这种下降态势将影响数据多样性、模型对齐性和系统可扩展性，并可能催生数据约束条件下的新型学习范式。

6. 基础模型研究透明度提高，但仍任重道远。最新发布的基础模型透明度指数(Foundation Model Transparency Index)——一个跟踪基础模型生态系统透明度的项目——显示，主要模型开发者的平均透明度得分从 2023 年 10 月的 37% 提高到了 2024 年 5 月的 58%。虽然进展显著，但仍有相当大的改进空间。

7. 对事实性与真实性评估的比较基准正不断完善。早期比较基准，如 HalluEval 和 TruthfulQA，虽旨在评估人工智能模型的事实性与真实性，但未能在人工智能领域获得广泛应用。为此，更新、更全面的评估方案出现，包括升级版的 Hughes 幻觉评估模型排行榜(Hughes Hallucination Evaluation Model leaderboard)、FACTS 评估框架以及 SimpleQA 测试集。

8. 与人工智能相关的选举虚假信息在全球蔓延，但其影响仍不明确。2024 年，在十多个国家和超过十个社交媒体平台上出现了大量与人工智能相关的选举虚假信息，包括在美国总统大选期间。然而，人们对这一问题的可衡量影响仍存在诸多疑问，许多人认为虚假信息活动对选举的影响比实际情况更为深远。

报告核心要点

第三章：

负责任的人工智能(续)

9. 接受过显性无偏见训练的大语言模型仍会表现出隐性偏见。许多先进的大语言模型，包括 GPT-4 和 Claude 3 Sonnet，在设计时都采取了抑制显性偏见的措施，但它们仍然表现出隐性偏见。这些模型过度地将负面词汇与黑人群体关联，更多将女性与人文学科而不是理工科(STEM)领域联系在一起，并偏爱男性担任有领导力的角色，从而加剧了决策中的种族与性别偏见。虽然偏见评价结果在标准比较基准上有所改善，但人工智能模型偏见仍是一个普遍存在的问题。

10. 负责任的人工智能获得了学术研究人员的关注。2024 年，全球顶级人工智能会议收录的负责任的人工智能论文数量达到 1,278 篇，较 2023 年的 992 篇增长 28.8%，自 2019 年以来持续保持稳定的年增长率。这一上升趋势凸显了负责任的人工智能在人工智能研究界日益增长的重要性。

第四章：

经济

1. 全球私人人工智能投资创历史新高，增长 26%。2024 年，全球企业人工智能投资规模达到 2523 亿美元，其中私人投资同比增长 44.5%，并购交易规模较上年增长 12.1%。过去十年间，该领域经历显著扩张，自 2014 年以来，总投资规模增长逾十三倍。

2. 生成式人工智能投资规模激增。2024 年，生成式人工智能领域的私人投资达到 339 亿美元，比 2023 年增长 18.7%，是 2022 年水平的 8.5 倍以上。该领域目前占所有人工智能相关私人投资总额的 20% 以上。

3. 美国扩大了其在全球人工智能私人投资中的领先优势。2024 年，美国人工智能私人投资规模达到 1091 亿美元，相当于中国的近 12 倍(93 亿美元)，英国的 24 倍(45 亿美元)。在生成式人工智能领域，美国投资总额较中国、欧盟及英国投资总和还多 254 亿美元，较 2023 年 218 亿美元的差额继续扩大。

4. 人工智能的使用达到前所未有的水平。2024 年，受访企业报告采用人工智能技术的比例从 2023 年的 55% 跃升至 78%。同样，在至少一项业务职能中使用生成式人工智能的受访者数量增长逾一倍——从 2023 年的 33% 上升至 2024 年的 71%。

报告核心要点

第四章：

经济(续)

5. 人工智能已开始在多个业务职能领域产生财务效益，但大多数企业仍处于应用初期阶段。报告显示，在单一业务职能内应用人工智能并取得财务效益的企业中，多数反馈效益水平仍处于较低区间。在成本节约方面，在客户服务运营中使用人工智能的企业中有 49% 的受访者报告实现降本，供应链管理领域这一比例为 43%，软件工程领域为 41%。不过，这些企业报告的成本降幅大多不足 10%。在收入增长方面，营销与销售部门应用人工智能的企业中有 71% 的受访者报告收入提升，供应链管理领域为 63%，服务运营领域为 57%。但需要指出的是，这些收入增幅普遍低于 5%。

6. 人工智能的应用呈现出明显的区域差异，其中大中华区正迅速崛起。尽管北美地区仍保持着企业人工智能应用率的领先地位，但大中华区是同比增长率最高的地区之一，其企业人工智能采用率提升了 27%。欧洲紧随其后，实现了 23% 的增长，这表明全球人工智能格局正在快速演变，各国在人工智能应用领域的国际竞争也日趋激烈。

7. 中国在工业机器人领域仍占据主导地位，虽略有放缓。2023 年，中国安装 27.63 万台工业机器人，是日本的 6 倍，美国的 7.3 倍。自 2013 年超过日本以来，中国在全球工业机器人安装量中的份额从 20.8% 上升至 51.1%。虽然中国的机器人安装量继续超过世界其他国家的总和，但这一差距在 2023 年略有缩小，标志着其急剧扩张的势头略有放缓。

8. 协作型和交互式机器人使用日趋普及。2017 年，协作型机器人仅占所有新安装工业机器人的 2.8%，到 2023 年，这一数字攀升至 10.5%。同样，2023 年，除医疗机器人外，所有应用领域的服务机器人安装量均呈现增长趋势。这一趋势不仅表明机器人安装量的总体增长，还表明人们越来越重视将机器人部署在面向人类的岗位上。

9. 人工智能正在推动能源结构的重大变革，并引发对核能的新一轮关注。微软宣布以 16 亿美元重启三哩岛核反应堆，为人工智能提供动力，而谷歌和亚马逊也签订了核能协议，为人工智能业务提供支持。

10. 人工智能提高生产效率并缩小技能差距。去年的人工智能指数报告是首批强调人工智能对生产效率产生积极影响的研究报告之一。今年更多研究进一步验证了这些发现，证实人工智能不仅能提升生产效率，在多数情况下还有助于缩小高技能与低技能劳动者之间的能力差距。

报告核心要点

第五章： 科学与医学

- 1. 更先进的大规模蛋白质测序模型问世。**包括 ESM3 和 AlphaFold 3 在内的多个高性能大规模蛋白质测序模型相继推出。随着时间的推移，这些模型的规模显著扩大，使得蛋白质预测准确率不断提高。
- 2. 人工智能持续推动科学发现的飞速发展。**人工智能在科学进步中的作用不断扩大。2022 至 2023 年仅是人工智能驱动科研突破的初始阶段，而 2024 年出现更具突破性的进展，包括训练大语言模型智能体执行生物任务的 Aviary，以及显著增强野火预测能力的 FireSat。
- 3. 主流大语言模型的临床知识水平持续提升。**OpenAI 最近发布的 o1 在 MedQA 比较基准中创下了 96.0% 的新纪录，比 2023 年公布的最佳成绩提高了 5.8%。自 2022 年末以来，该测试性能已累计提升 28.4%。作为评估临床知识的重要比较基准，MedQA 可能正接近性能饱和，预示着需要更具挑战性的评估。这表明需要建立更具挑战性的评估体系。
- 4. 人工智能在关键临床任务中的表现优于医生。**一项新的研究发现，在诊断复杂的临床病例时，无论是有人工智能还是没有人工智能，仅 GPT-4 就能胜过医生。最近的其他研究表明，人工智能在癌症检测和识别高死亡率风险患者方面超过了医生。不过，一些初步研究表明，人工智能与临床医生的协同诊疗可产生最优结果，这一发现值得作为重点领域开展深入研究。
- 5. 美国食品及药物管理局（FDA）批准的人工智能医疗设备数量激增。**美国食品药品监督管理局于 1995 年批准了首款人工智能医疗设备。截至 2015 年，仅有 6 款此类设备获批，但这一数字到 2023 年激增至 223 款。
- 6. 合成数据在医学领域展现出巨大潜力。**2024 年发布的研究表明，人工智能生成的合成数据可以帮助模型更好地识别健康的社会决定因素，加强保护隐私的临床风险预测，并促进新药化合物的发现。2024 年最新研究表明，人工智能生成的合成数据可有效提升模型对健康社会决定因素的识别能力，优化隐私保护型临床风险预测，并促进新药化合物的发现。
- 7. 医学人工智能伦理研究文献逐年增加。**从 2020 年到 2024 年，医学人工智能伦理方面的论文数量几乎翻了两番，从 2020 年的 288 篇增加到 2024 年的 1031 篇。

报告核心要点

第五章： 科学与医学(续)

8. 基础模型进入医学领域。2024 年, 一大波大型医学基础模型发布, 涵盖从 Med-Gemini 等通用多模态模型, 到面向特定专科的 EchoCLIP(超声心动图)、视觉 FM(眼科) 及 ChexAgent(放射科)等专用模型。

9. 公共蛋白质数据库规模不断扩大。自 2021 年以来, 主要公共蛋白质科学数据库的条目数量显著增长, 其中包括 UniProt (增长 31%)、PDB(增长 23%) 及 AlphaFold(激增 585%)。这一扩展对科研发现具有重要影响。

10. 人工智能研究获得两项诺贝尔奖。2024 年, 人工智能驱动的研究获得了最高荣誉, 两项与人工智能相关的突破获得了诺贝尔奖。谷歌 DeepMind 的德米斯 · 哈萨比斯 (Demis Hassabis) 和约翰 · 朱珀 (John Jumper) 凭借 AlphaFold 在蛋白质折叠方面的开创性工作获得了诺贝尔化学奖。与此同时, 约翰 · 霍普菲尔德 (John Hopfield) 和杰弗里 · 辛顿 (Geoffrey Hinton) 因其在神经网络方面的奠基性贡献获得了诺贝尔物理学奖。

第六章： 政策

1. 美国各州正引领人工智能立法进程, 而联邦层面进展相对迟缓。2016 年, 只有一项州级人工智能相关法律获得通过, 而到 2023 年, 增加到 49 项。仅在去年, 这一数字就翻了一番多, 达到 131 项。虽然联邦层面的人工智能法案提案也有所增加, 但通过的数量仍然很少。

2. 世界各国政府都加大人工智能基础设施投资力度。加拿大宣布了一项 24 亿美元的人工智能基础设施一揽子计划, 而中国设立 475 亿美元半导体产业基金。法国承诺为人工智能基础设施投资 1170 亿美元, 印度承诺投资 12.5 亿美元, 而沙特阿拉伯的“超越计划”则包括对人工智能的 1000 亿美元投资。

3. 在全球范围内, 人工智能在立法程序中的提及率不断上升。在 75 个国家中, 2024 年立法程序中提及人工智能的次数增加了 21.3%, 从 2023 年的 1557 次增加到 1889 次。自 2016 年以来, 人工智能被提及的总数增长了 9 倍多。

报告核心要点

第六章：

政策(续)

4. 在全球范围内，人工智能安全研究机构加速扩张与协同合作。2024 年，各国相继成立国际人工智能安全研究机构。首批机构由美国和英国于 2023 年 11 月首届人工智能安全峰会闭幕后率先设立。随着 2024 年 5 月首尔人工智能峰会的召开，日本、法国、德国、意大利、新加坡、韩国、澳大利亚、加拿大及欧盟等国家和地区也相继承诺成立相关机构。

5. 美国人工智能相关联邦法规数量激增。2024 年，美国出台了 59 项人工智能相关法规，是 2023 年 25 项的两倍多。这些法规来自 42 个机构，是 2023 年出台法规的 21 个机构的两倍。

6. 美国多州加强深度伪造监管立法。2024 年之前，只有加利福尼亚、密歇根、华盛顿、得克萨斯和明尼苏达五个州颁布了法律，对选举中的深度伪造行为进行监管。2024 年，俄勒冈州、新墨西哥州和纽约州等 15 个州出台了类似措施。此外，到 2024 年，已有 24 个州通过了针对深度伪造的法规。

第七章：

教育

1. 美国高中计算机科学(Computer Science, CS)课程的普及率与选修人数较上一学年略有提升，但教育差距依然存在。学生的参与情况因州、种族和民族、学校规模、地理位置、收入、性别和残疾而异。

2. 美国的计算机科学教师希望传授人工智能，但认为自己不具备这样的能力。尽管 81% 的计算机科学教师认同应将人工智能应用及人工智能基础知识纳入计算机科学基础课程体系，但仅有不足半数的高中计算机科学教师认为自己具备开展人工智能教学的专业能力。

3. 全世界有三分之二的国家提供或计划提供 K-12 阶段的计算机科学教育。自 2019 年以来，这一比例翻了一番，其中非洲和拉丁美洲国家进展最为显著。然而，由于学校缺乏电力供应，非洲国家学生获得计算机科学教育的机会最少。

报告核心要点

第七章：

教育(续)

4. 2022 年至 2023 年期间，美国获得人工智能硕士学位的毕业生人数几乎翻了一番。尽管人工智能在学士学位和博士学位中的关注度提升将较为缓慢，但硕士学位的激增可能预示着所有学位层次都将出现这一发展趋势。

5. 在培养信息、技术和通信技术(ICT)毕业生方面，美国持续处于全球领先地位。西班牙、巴西和英国紧随美国之后，成为各层次毕业生最多的国家，而土耳其则是男女比例最均衡的国家。

第八章：

公众观点

1. 全球对人工智能产品和服务的态度持谨慎乐观的态度。在益普索(Ipsos)2022–2024年持续追踪的26个国家中，18国的受访者对人工智能产品与服务“利大于弊”的认同比例呈现上升趋势。在全球范围内，认为人工智能产品和服务利大于弊的个人比例从2022年的52%上升到2024年的55%。

2. 人工智能对日常生活影响的预期认知度持续攀升。在世界各地，三分之二的人现在认为，人工智能驱动的产品与服务将在未来3至5年内显著改变日常生活——这一比例较2022年上升6个百分点。除马来西亚、波兰和印度外，其余国家自2022年以来该认知度均有所提升，其中加拿大(增长17%)与德国(增长15%)涨幅最为显著。

3. 对人工智能公司伦理行为的怀疑正在增加，同时对人工智能公平性的信任正在下降。在全球范围内，人们对人工智能公司保护个人数据的信心从2023年的50%下降到2024年的47%。同样，与相比，如今相信人工智能系统不偏不倚、不受歧视的人也越来越少。

4. 人工智能乐观程度的地区差异依然存在。2023年人工智能指数首次指出，人工智能乐观程度的地区差异依然存在。在中国(83%)、印度尼西亚(80%)和泰国(77%)等国家，绝大多数人认为人工智能驱动的产品和服务利大于弊，而在加拿大(40%)、美国(39%)和荷兰(36%)，只有少数人持这种观点。

报告核心要点

第八章： 公众观点(续)

5. 美国人自动驾驶汽车仍持不信任态度。根据美国汽车协会(American Automobile Association, AAA)最新调查数据显示，61%的美国民众对自动驾驶汽车存在恐惧心理，仅有13%的受访者表示信任该技术。尽管这一恐惧比例较2023年68%的峰值有所下降，但仍高于2021年54%的水平。

6. 美国地方政策制定者普遍支持对人工智能实施监管。2023年，美国73.7%的地方政策制定者(涵盖镇、市、县三级政府)支持对人工智能实施监管，较2022年的55.7%显著提升。民主党人士的支持率(79.2%)明显高于共和党人士(55.5%)，但两党支持率相较2022年均呈现显著增长。

7. 此前对人工智能持最强烈怀疑态度的国家中，对人工智能的乐观态度呈现急剧上升趋势。在全球范围内，人们对人工智能产品和服务的乐观态度有所提高，其中以前最持怀疑态度的国家乐观态度增幅最大。2022年，英国(38%)、德国(37%)、美国(35%)、加拿大(32%)和法国(31%)是最不倾向于认为人工智能利大于弊的国家。此后，这些国家对人工智能的乐观态度分别提升了8%、10%、4%、8%和10%。

8. 劳动者预期人工智能将重塑就业结构，但其对岗位被替代的担忧程度相对较低。在全球范围内，60%的受访者认为人工智能将在未来五年内改变个人的工作方式。然而，有一小部分受访者(36%)认为，人工智能将在未来五年内取代他们的工作。

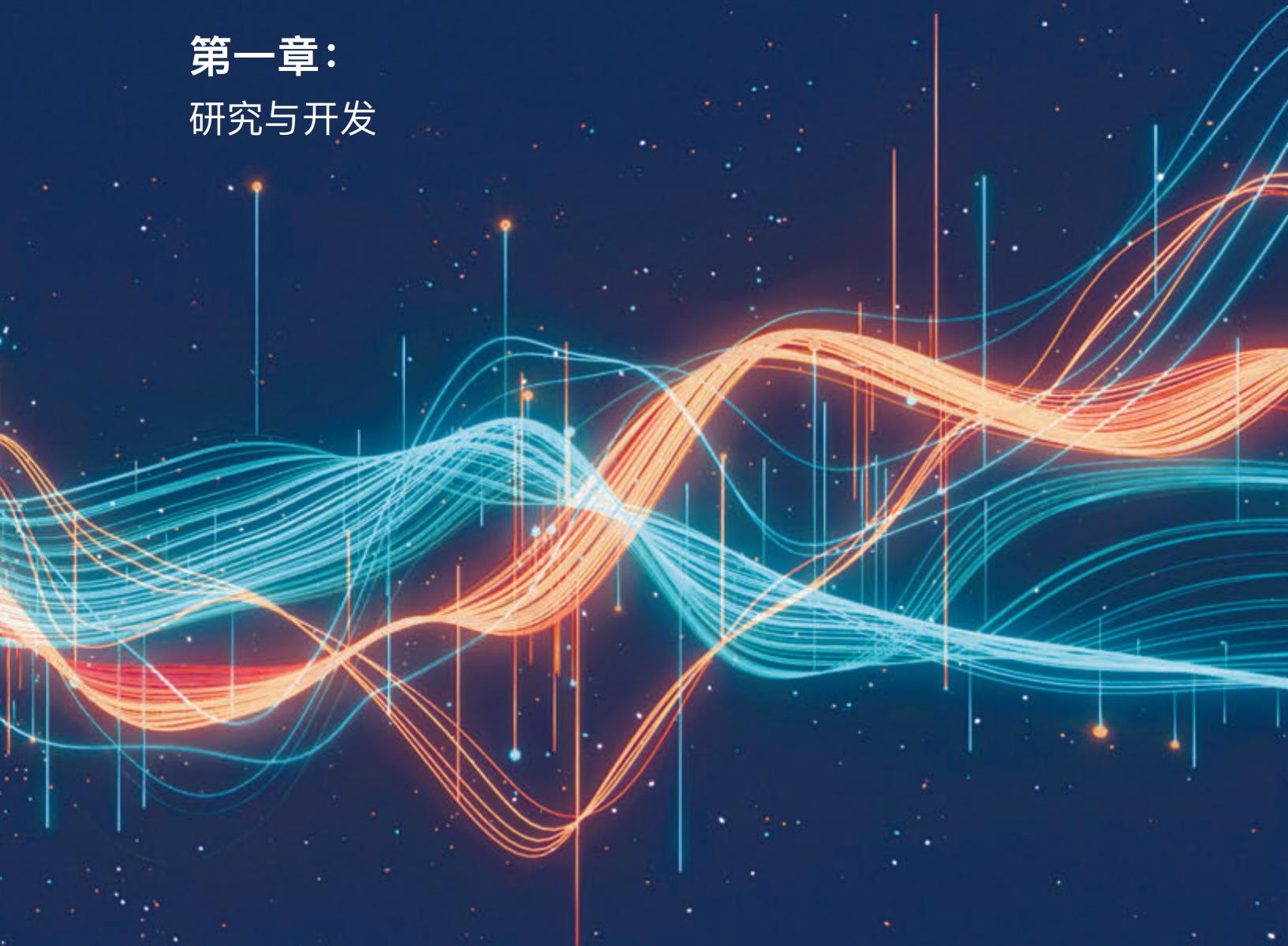
9. 美国地方政策制定者在人工智能政策优先事项上存在明显分歧。美国地方政府决策者虽普遍支持人工智能监管，但在具体政策优先事项上存在显著差异。支持率最高的政策包括更严格的数据隐私法规(80.4%)、失业人员再培训计划(76.2%)以及人工智能应用监管条例(72.5%)。然而，针对执法部门人脸识别禁令(34.2%)、工资下降补贴(32.9%)和全民基本收入(24.6%)等政策的支持率则显著降低。

10. 人工智能被视为提升效率的工具和娱乐体验的助推器，但其经济影响仍存疑虑。全球对人工智能影响的看法各不相同。55%的人认为人工智能将节省时间，51%的人预计它将提供更好的娱乐选择，但对其健康或经济效益有信心的人较少。只有38%的人认为人工智能将提升医疗健康水平，36%的人认为人工智能将改善国民经济，31%的人认为人工智能将对就业市场产生积极影响，37%的人认为人工智能将提升自己的工作效率。



第一章：

研究与开发



第一章：研究与开发

概述	26
章节核心要点	27
1.1 论文发表	29
概述	29
人工智能论文发表总量	29
按发表平台 (Venue) 分类	31
按国家 / 地区分类	32
按行业分类	36
按研究主题分类	38
发表的百强论文	39
按国家 / 地区分类	39
按行业类型	40
按机构类型	41
1.2 专利	42
概述	42
按国家 / 地区分类	43
1.3 标志性人工智能模型	46
按国家 / 地区分类	46
按行业分类	47
按研发主体分类	49
模型发布	50
参数发展趋势	52
算力发展趋势	56
重点：模型训练会面临数据枯竭吗？	59
推理成本	64
训练成本	65
1.4 硬件	68
概览	68
重点：能源效率和环境影响	71
1.5 人工智能会议	75
参会规模	75
1.6 开源人工智能软件	77
开源人工智能软件项目	77
星标	79

获取公共数据

第一章： 研究与开发

概述

本章探讨了人工智能研究与发展的最新趋势，首先系统分析人工智能论文发表、专利及标志性的人工智能系统，并基于国家和地区、研发机构与行业领域三维度对上述成果的开发方进行解析。本章同时涵盖了对人工智能模型训练成本、学术会议参与度及开源人工智能软件的分析。今年新增的内容包括人工智能硬件生态演进图谱、人工智能训练能耗与环境影响评估及模型推理成本时序分析。

第一章： 研究与开发

章节核心要点

1. 产业界持续加大人工智能投入并主导标志性人工智能模型研发，而学术界则引领高影响力研究。产业界在标志性人工智能模型研发中的主导优势持续强化，这一趋势在既往两期人工智能指数报告中已有明确阐述。2024 年，近 90% 的前沿模型源自产业界（较2023 年的 60% 提升 30 个百分点）。学术界在过去三年中始终是高被引（前 100 名）论文的首要机构贡献者。

2. 中国在人工智能论文发表总量上领先，而美国在具有高影响力的研究方面占据优势。2023 年，中国在人工智能领域的论文发表量（23.2%）和引用量（22.6%）均位居全球首位。而在过去三年中，美国机构贡献了最多被引用次数排名前 100 的人工智能论文。

3. 人工智能论文发表总量持续增长，并在计算机科学领域占据日益重要的主导地位。从 2013 年到 2023 年，在计算机科学和其他科学学科领域发表的人工智能论文发表总数几乎增加了两倍，从大约 102,000 篇增加到超过 242,000 篇。从比例上看，人工智能在计算机科学论文中所占的份额已从 2013 年的 21.6% 上升到 2023 年的 41.8%。

4. 美国仍然是标志性人工智能模型的主要来源地。2024 年，美国机构开发了 40 个标志性人工智能模型，大大超过中国的 15 个和欧洲的 3 个。在过去十年，源自美国的标志性机器学习模型比任何其他国家都多。2024 年，美国机构共开发出 40 个前沿人工智能模型，显著超过中国的 15 个和欧洲地区的合计的 3 个。过去十年间，美国在前沿机器学习模型的研发数量上持续保持全球领先地位。

5. 人工智能模型正变得规模更大、算力需求更高、能耗更强。最新研究表明，标志性人工智能模型的训练算力需求约每 5 个月翻一番，大语言模型训练数据集规模每 8 个月翻一番，而训练所需功耗每年递增。大规模的产业投资持续推动着模型规模扩大和性能提升。

6. 人工智能模型的使用成本持续下降。以语言模型性能评估常用基准——MMLU 为例，达到 GPT-3.5 水平（64.8 分）的人工智能模型查询成本已从 2022 年 11 月的每百万词元 20 美元，降至 2024 年 10 月的仅每百万词元 0.07 美元（Gemini-1.5-Flash-8B 模型），18 个月内降幅超 280 倍。根据任务类型不同，大语言模型的推理价格年降幅已达 9 至 900 倍不等。

第一章： 研究与开发

章节核心要点（续）

7. 人工智能专利申请持续上升。从 2010 年到 2023 年，人工智能专利数量稳步大幅增长，从 3833 项激增到 122511 项。仅在去年，人工智能专利数量就增长了 29.6%。截至 2023 年，中国在人工智能专利总数上遥遥领先，占所有授权的 69.7%，而按人均计算，韩国和卢森堡则是人工智能专利的主要生产国。

8. 人工智能硬件变得更快、更便宜、更节能。最新研究表明，以 16 位浮点运算能力衡量的机器学习硬件性能正以每年 43% 的速度增长，每 1.9 年实现翻倍。其性价比显著提升——成本每年下降 30%，而能效则以每年 40% 的幅度持续改善。

9. 人工智能训练产生的碳排放量正在稳步上升。早期训练的人工智能模型，如 AlexNet 网络（2012 年），碳排放量不大，仅为 0.01 吨。最新人工智能模型的训练碳排放量显著增加：2020 年 GPT-3 为 588 吨，2023 年 GPT-4 达 5,184 吨，而 2024 年 Llama 3.1 405B 更是高达 8,930 吨。作为对比，美国人年均碳排放量仅为 18 吨。

第一章：研究与开发

1.1 论文发表

下图展示了 2010 至 2023 年全球人工智能英文论文发表的梳理，按机构类型、成果类别及地域分布三维度分类统计。在今年报告中，人工智能指数新增一节内容，分析 100 篇被引用次数最多的人工智能论文发表的趋势，可为特别具有影响力的研究提供洞见。今年，人工智能指数利用 OpenAlex 数据库分析了人工智能研究成果的趋势。因此，今年报告中的数字与往年略有不同。¹ 鉴于出版物元数据的收集存在显著滞后，在某些情况下，需要等到一年中的中期才能完全收集到上一年的发表物，因此，在今年的报告中，人工智能指数团队决定只考察到 2023 年的出版趋势。

2013–2023 年全球范围内在 CS 领域人工智能论文发表数量

资料来源 2025 年人工智能指数图表：2025 年人工智能指数报告

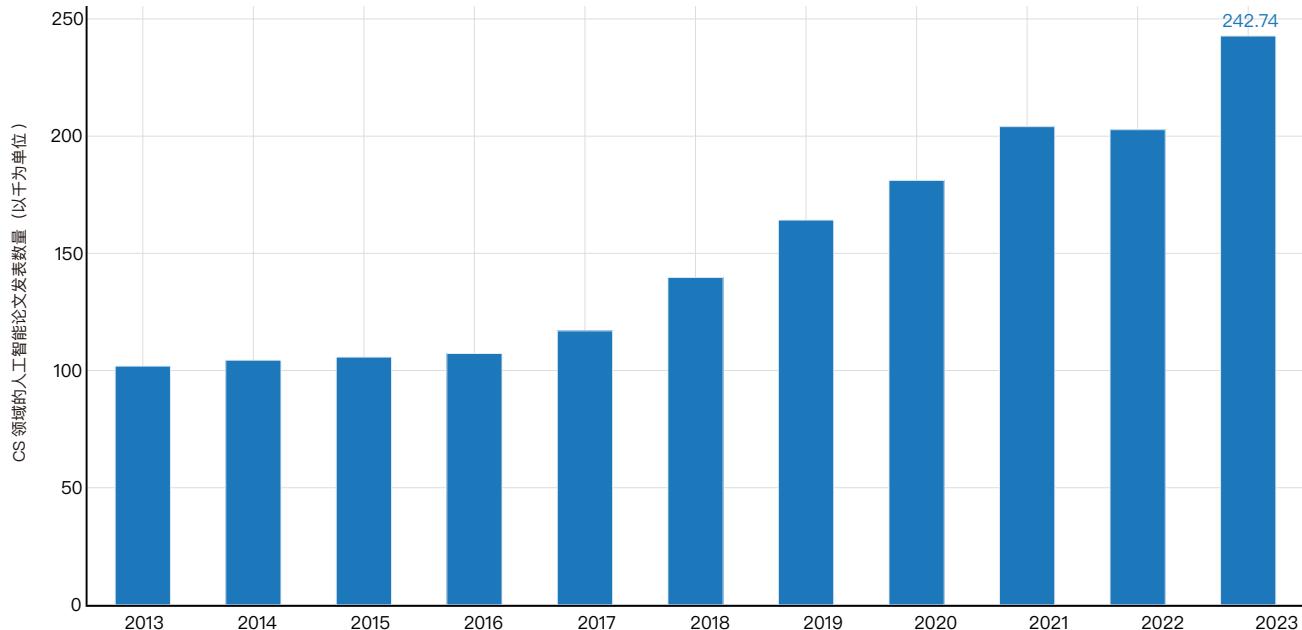


图 1.1.1

1. OpenAlex 是一个完全开放的学术元数据目录，包括科学论文、作者、机构等。人工智能指数使用 OpenAlex 作为书目数据库，并使用最新版本的 CSO 分类器对人工智能相关研究进行自动分类。前几年，该指数依赖于第三方提供商，其基础数据源和分类方法各不相同。因此，今年报告的研究结果与以往报告中的结果略有不同。此外，人工智能指数仅将分类器应用于 OpenAlex 归类计算机科学大领域的论文，这种方法可能导致人工智能相关论文的数量不足，因为它排除了社会科学等领域的研究，这些领域采用了人工智能方法，但不属于计算机科学指定的分类范围。

2. CSO 分类器 (v3.3) 是一个自动文本分类系统，旨在使用包含 15,000 个主题和 166,000 种关系的综合本体对计算机科学领域的研究论文进行分类，包括 GenAI、大语言模型和提示工程等新兴领域。它通过三个模块处理元数据（如标题和摘要）：语法模块用于准确匹配主题，语义模块利用词嵌入推断相关主题，后处理模块通过过滤异常值和添加相关的更高层次领域来完善结果。

第一章：研究与开发

1.1 论文发表

2013 年至 2023 年间，人工智能相关论文发表的总数翻了一番多，从 2013 年的约 10.2 万篇增至 2023 年的超过 24.2 万篇。过去一年间的增长率达 19.7%，这一增幅具有重要意义。

计算机科学的众多领域——从硬件与软件工程到人机交互——如今均对人工智能的发展都有所贡献。因此，观测到的增长现象反映出该人工智能更广泛且日益增强的关注度。

2013–2023 年全球范围在 CS 领域人工智能论文发表的统计（占总数百分比）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

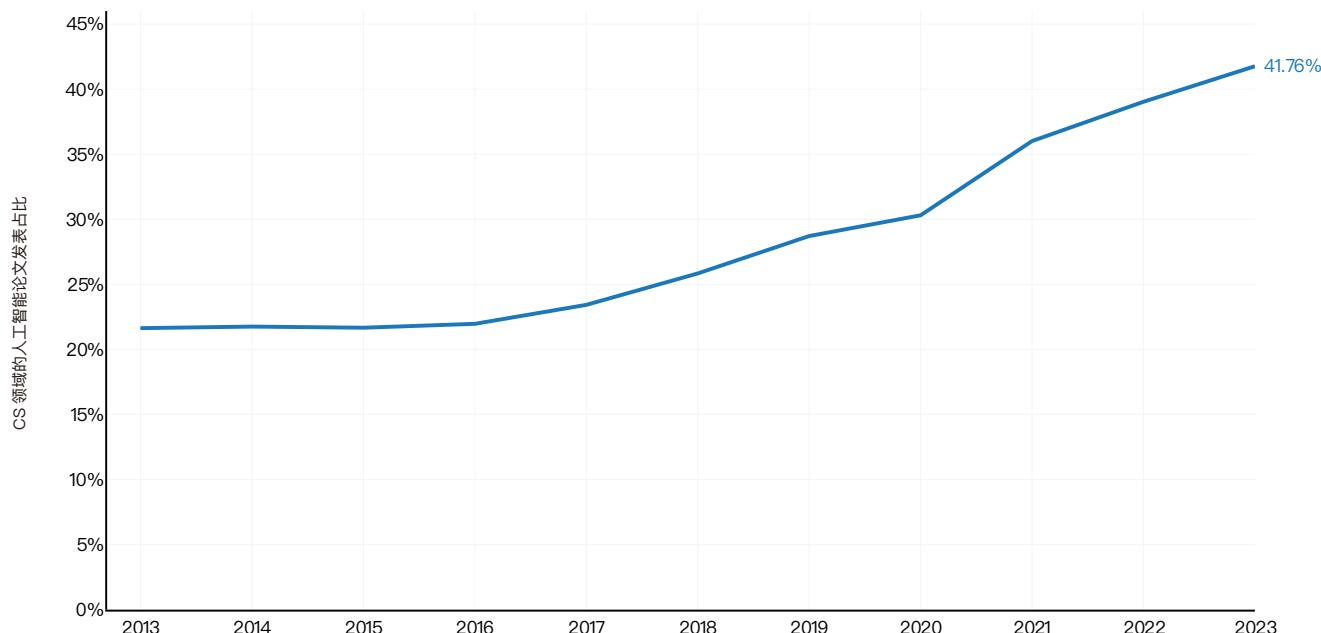


图 1.1.2

第一章：研究与开发

1.1 论文发表

图 1.1.2 显示了 OpenAlex 数据库中被归类为人工智能相关的计算机科学论文发表的发布占比。图 1.1.2 采用了与图 1.1.1 相同的数据，但以比例形式显示。人工智能论文发表的比例大幅增长，从 2013 年到 2023 年几乎翻了一番。

按发表平台(Venue)分类

人工智能研究人员在多种学术平台发表研究成果。图 1.1.3 按平台类型展示了人工智能领域论文总量的分布情况。2023

年，期刊论文占人工智能论文发表的最大份额(41.8%)，其次是会议论文发表(34.3%)。尽管自 2013 年以来，期刊和会议论文发表总量持续增长，但其在人工智能领域的占比却呈稳步下降趋势——期刊论文占比从 2013 年的 52.6% 降至 2023 年的 41.8%，会议论文发表占比同期从 36.4% 下降至 34.3%。与此相反，类似 arXiv 存储库的人工智能论文发表占比显著提升。

2013–2023 年按平台分类CS 领域人工智能论文发表的统计（占总数百分比）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

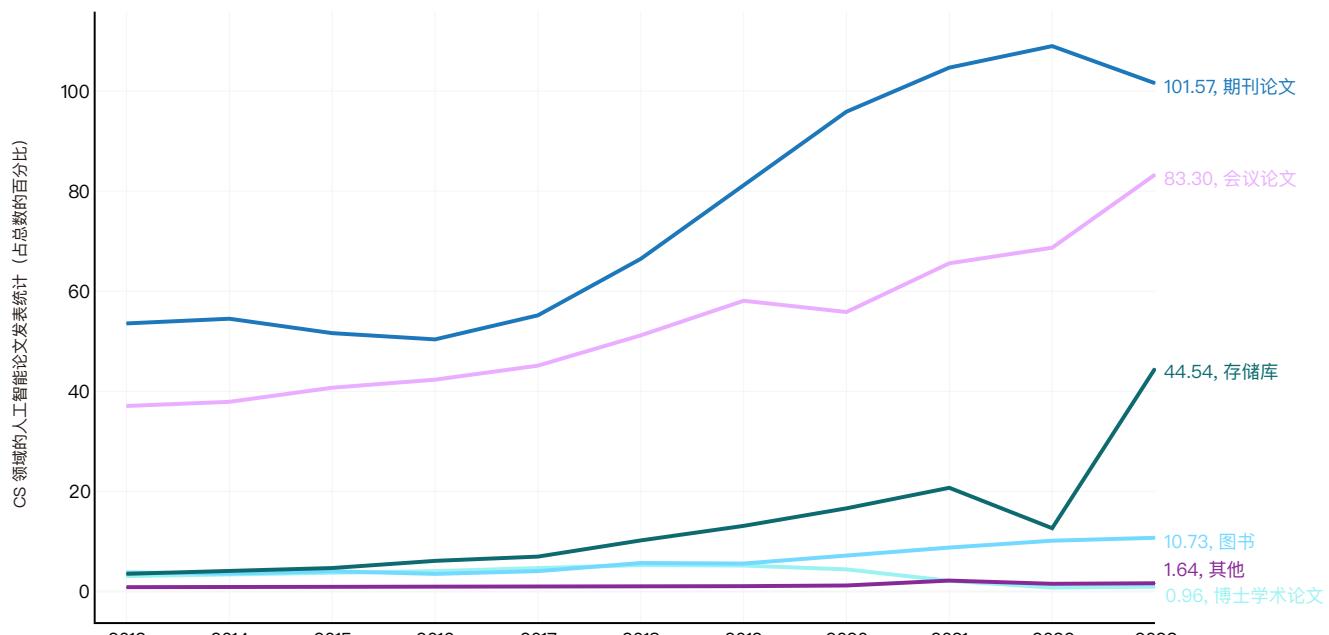


图 1.1.3

第一章：研究与开发

1.1 论文发表

按国家 / 地区分类

图 1.1.4 展示了不同地区人工智能论文发表情况随时间的变化。³ 在 2023 年，东亚和太平洋地区在人工智能研究产出方面领先，占所有人工智能论文发表的 34.4%，其次是欧洲和中亚 (18.2%) 以及北美 (10.3%)。⁴

图 1.1.4 分析了人工智能领域论文发表的地域分布情况，揭示了研究成果数量最多的地区；而图 1.1.5 则聚焦于引用情

况，统计了各区域研究成果在总引用量中的占比。截至 2023 年，东亚与太平洋地区的人工智能论文发表引用量占比最高，达到 37.1% (图 1.1.5)。回溯至 2017 年，该地区与北美地区的引用份额基本持平；此后，北美和欧洲地区的占比呈下降趋势，而东亚与太平洋地区的引用份额则显著攀升。

2013—2023 年按地区划分在 CS 领域人工智能论文发表的统计（占总数百分比）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

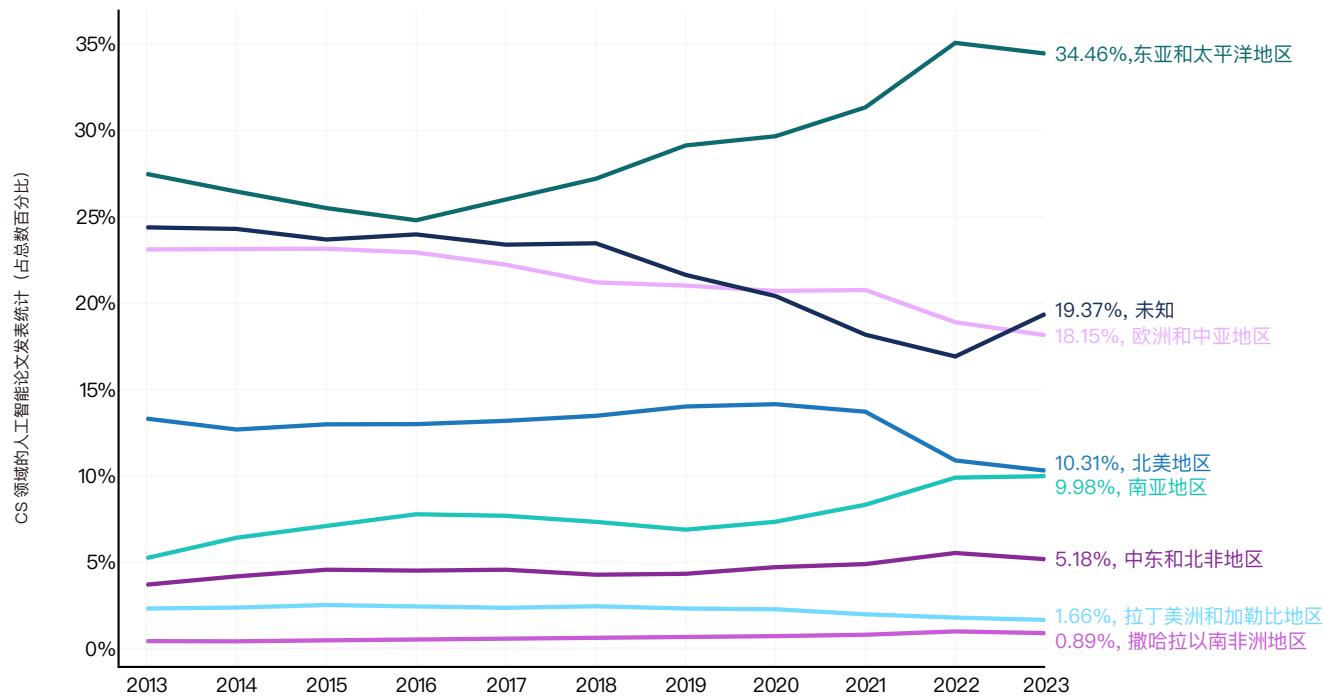


图 1.1.4

3. 本章中的地区是根据世界银行的分析分组分类的。人工智能指数使用作者身份数据中的“国家”字段来确定作者所属的国家。这个字段列出了根据机构隶属关系从OpenAlex中检索到的作者隶属的所有国家。这些隶属关系可以在论文中明确说明，也可以从作者最近发表的论文中推断。在按国家统计研究成果时，人工智能指数会对研究成果所关联的每个国家分配一个计数。例如，如果一篇论文有三位作者，其中两位隶属于美国的机构，一位隶属于中国的机构，那么这篇论文在美国和中国各被计算一次。

4. 当作者的机构归属缺失或不完整时，论文发表的国家归属可能为“未知”。造成这一问题的有很多，包括机构名称不规范或遗漏、平台功能缺陷、作者群体实践、隶属关系标签不规范、文档类型不一致或作者发表记录有限等。本文讨论的是与OpenAlex有关的问题；不过，遗漏机构的问题也与其他书目数据库有关。

第一章：研究与开发

1.1 论文发表

2013–2023 年按地区划分在 CS 领域人工智能领域论文发表的引用量（占总引用量百分比）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

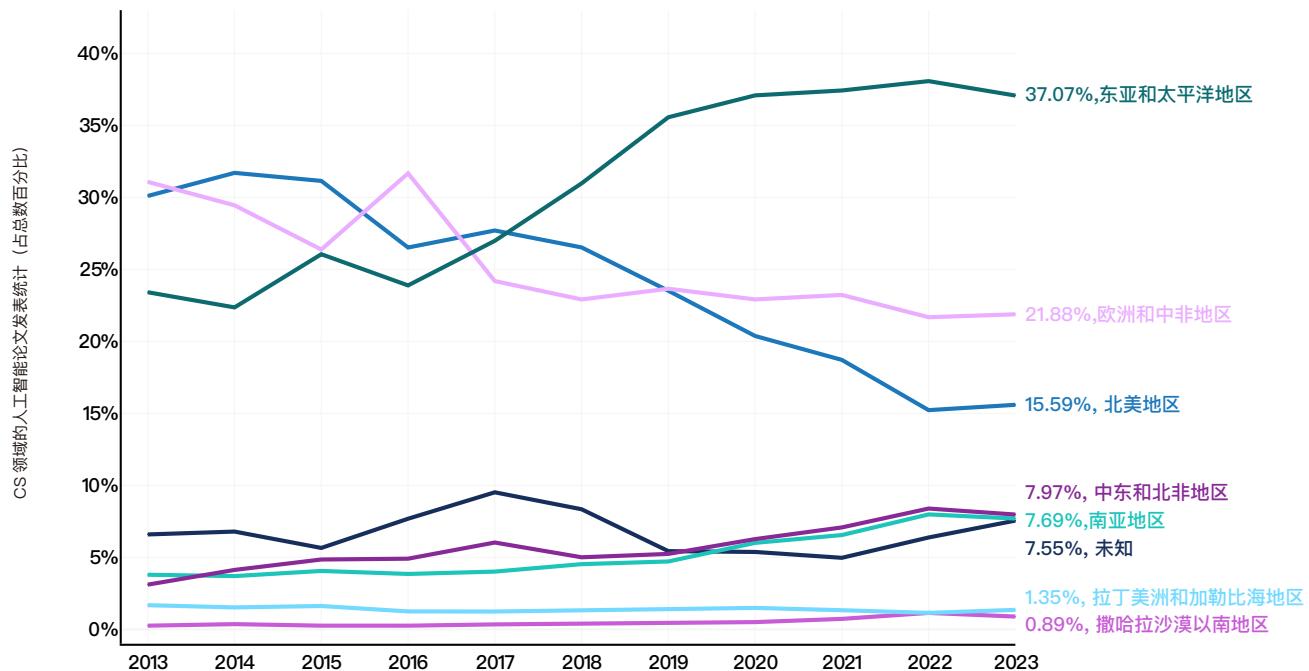


图 1.1.5

第一章：研究与开发

1.1 论文发表

2023 年, 中国在人工智能论文发表数量上居全球首位, 占比达 23.2%, 而欧洲和印度的占比分别为 15.2% 和 9.2% (见图 1.1.6)。⁵自 2016 年以来, 中国的份额持续稳步增长, 而欧

洲的占比则呈现下降趋势。美国在人工智能领域的论文发表比例在 2021 年前保持相对稳定, 但此后略有下降。

2013–2023 年在 CS 领域人工智能论文发表的统计（占总数百分比）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

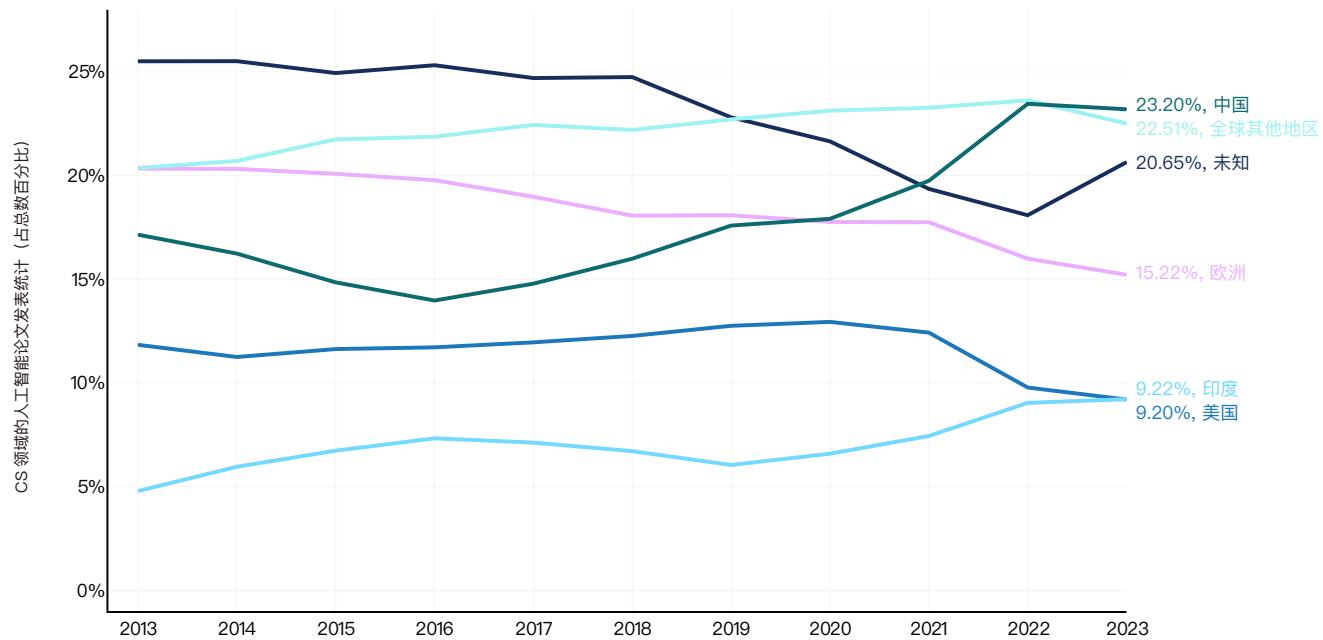


图 1.1.6⁶

5. 在本报告及其他章节中, 人工智能指数对“欧洲”的定义遵循联合国统计司所界定的国家列表。

6. 保持简明扼要, 人工智能指数对部分国家的结果进行了可视化。不过, 所有国家的完整结果都将在 AI Index's Global Vibrancy Tool 上公布, 该工具计划于 2025 年夏季更新。如需立即获取特定国家的研发数据, 请联系人工智能指数团队。

第一章：研究与开发

1.1 论文发表

2023年，中国人工智能论文的引用量占比达22.6%，位居全球首位，欧洲和美国分别以20.9%和13.0%的占比紧随其后(图1.1.7)。与论文发表总量趋势一致，2010年代末成为关键转折点——中国在这一时期超越欧美，成为人工智能领域被引用文献的主要来源地。

2013–2023年按选定地理区域分类在CS领域人工智能论文发表的引用量（占总引用量百分比）

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

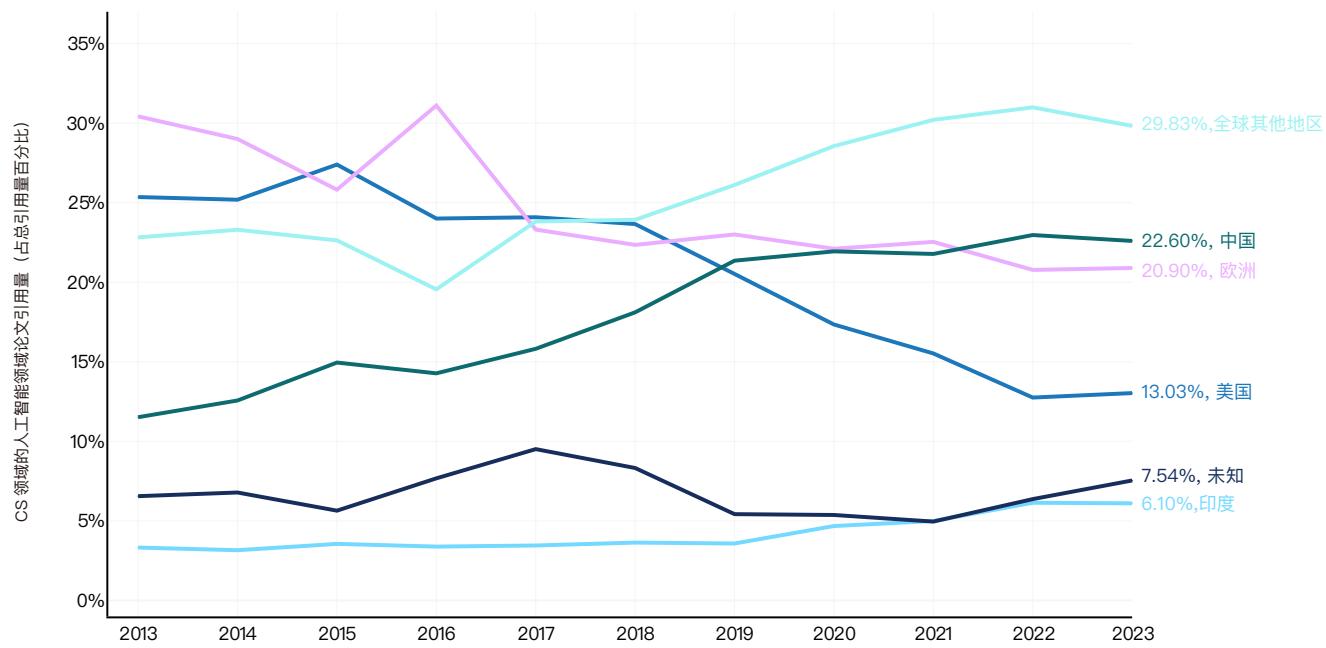


图 1.1.7

第一章：研究与开发

1.1 论文发表

按行业分类

学术机构仍是全球范围内人工智能论文发表的主要来源（图 1.1.8）。2013 年，学术机构贡献了 85.9% 的人工智能论文；到 2023 年，这一比例仍保持在 84.9% 的高位。2023 年，产业

界贡献了 7.1% 的人工智能论文，政府机构和非营利组织的占比分别为 4.9% 和 1.7%。

2013—2023 年按行业分类在 CS 领域人工智能发表物的统计

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

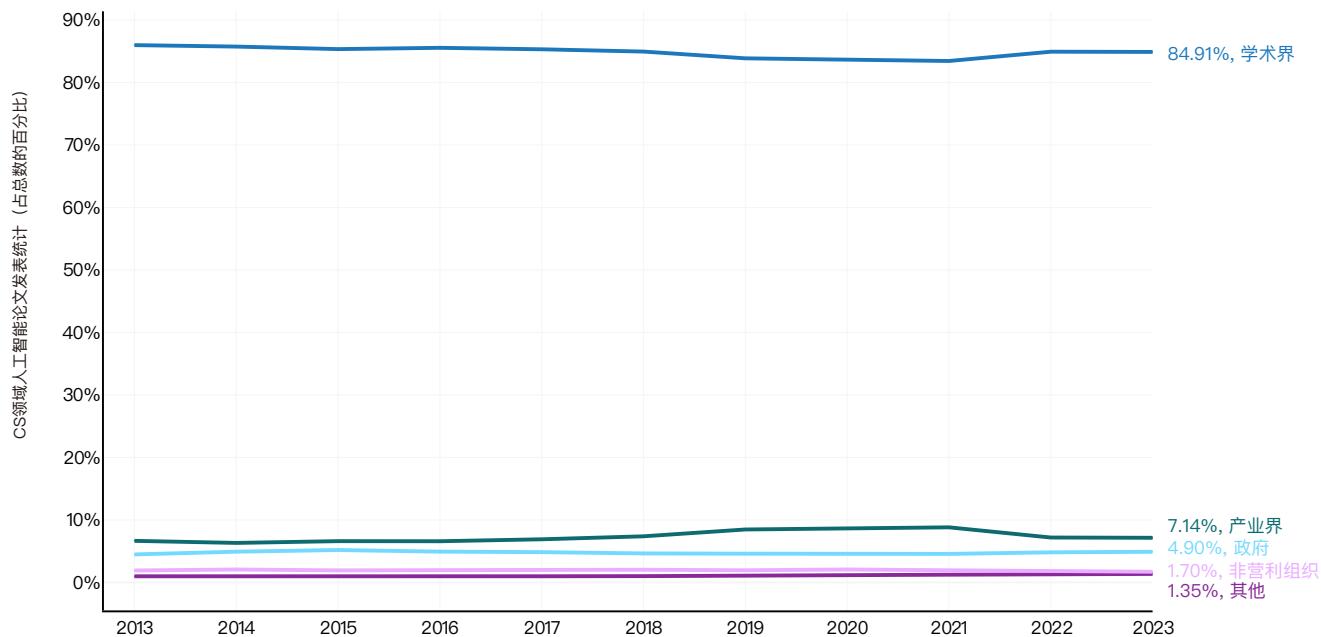


图 1.1.8⁷

7、图 1.1.8 和图 1.1.9 未标注隶属机构的论文未纳入最终可视化统计。

第一章：研究与开发

1.1 论文发表

人工智能论文发表的行业来源在不同地区呈现显著差异（图 1.1.9）。美国产业界贡献的人工智能论文占比达 16.49%，明显高于中国的 8.02%。在主要地区中，中国教育机构产出的人工智能论文占比最高，达到 84.45%。

2023 年按行业和选定地理区域分类在 CS 领域人工智能论文发表的统计（占总量的百分比）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

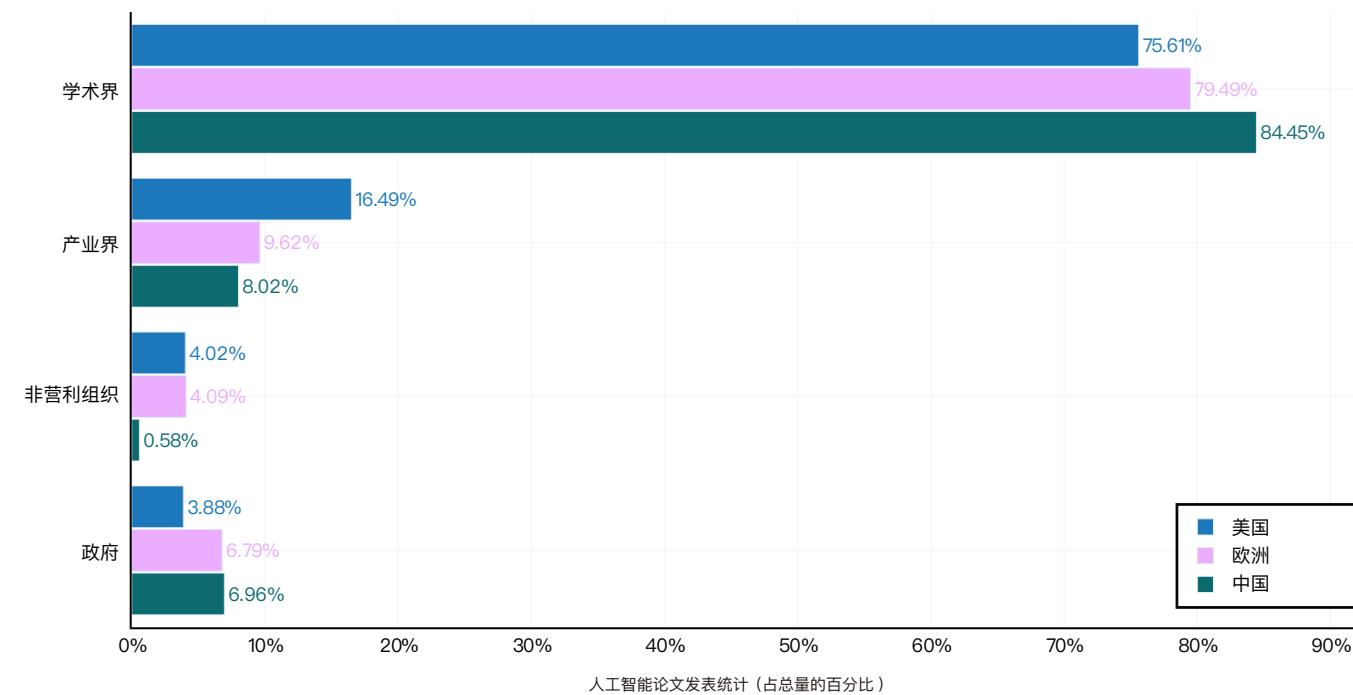


图 1.1.9

第一章：研究与开发

1.1 论文发表

按研究主题分类

机器学习是 2023 年人工智能领域最热门的研究主题，占总论文数量的 75.7%，其次是计算机视觉 (47.2%)、模式识别 (25.9%) 和自然语言处理 (17.1%) (图 1.1.10)。过去一年中，关于生成式人工智能的论文数量出现了显著增长。

2013–2023 年按精选核心主题分类人工智能论文发表的统计

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

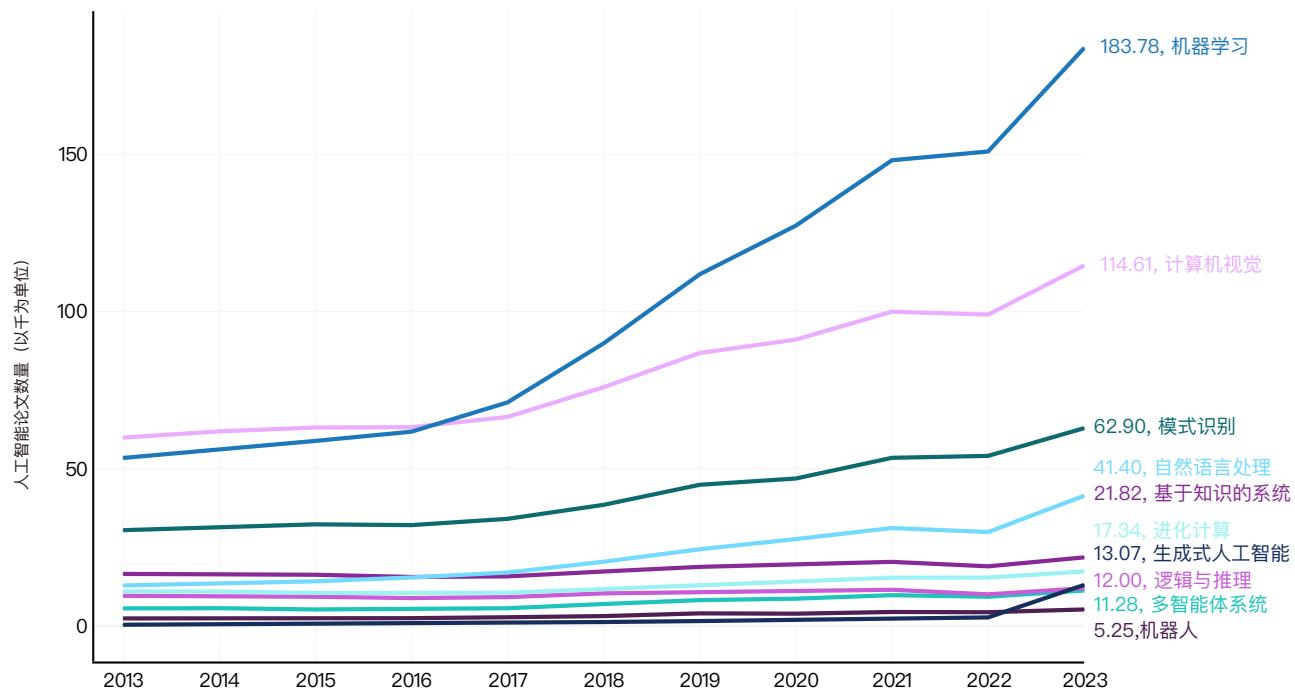


图 1.1.10 8

8、人工智能指数使用自己的主题分类器对论文进行分类。一份论文有可能被赋予多个主题标签。

第一章：研究与开发

1.1 论文发表

发表的百强论文

虽然追踪人工智能论文总量能提供一个对人工智能研究活动的宏观视角，但聚焦于被引用次数最多的论文则能揭示该领域最具影响力的研究成果。这项分析揭示了一些最具开创性和影响力的人工智能研究正在哪里兴起。今年，人工智能指数通过 OpenAlex 的引文数据，确定了 2021 年、2022 年和 2023 年被引用次数最多的 100 篇人工智能论文。⁹ 2023 年被引用次数最多的人工智能论文包括 OpenAI 的 GPT-4 技术报告、Meta 的 Llama 2 技术报告和谷歌的 PaLM-E 的技术报

告。值得注意的是，由于引用的滞后性，今年报告中被引用次数最多的论文在今后的版本中可能会有所变化。

按国家 / 地区分类

图 1.1.11 显示了各年被引用次数最多的 100 篇人工智能论文的地理分布。从 2021 年到 2023 年，美国一直是被引用最多的国家，2021 年为 64 篇，2022 年为 59 篇，2023 年为 50 篇。¹⁰ 自 2021 年以来美国在顶级人工智能论文中所占的比例逐渐下降。

2021–2023 年按选定地理区域划分的百强高被引论文统计

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

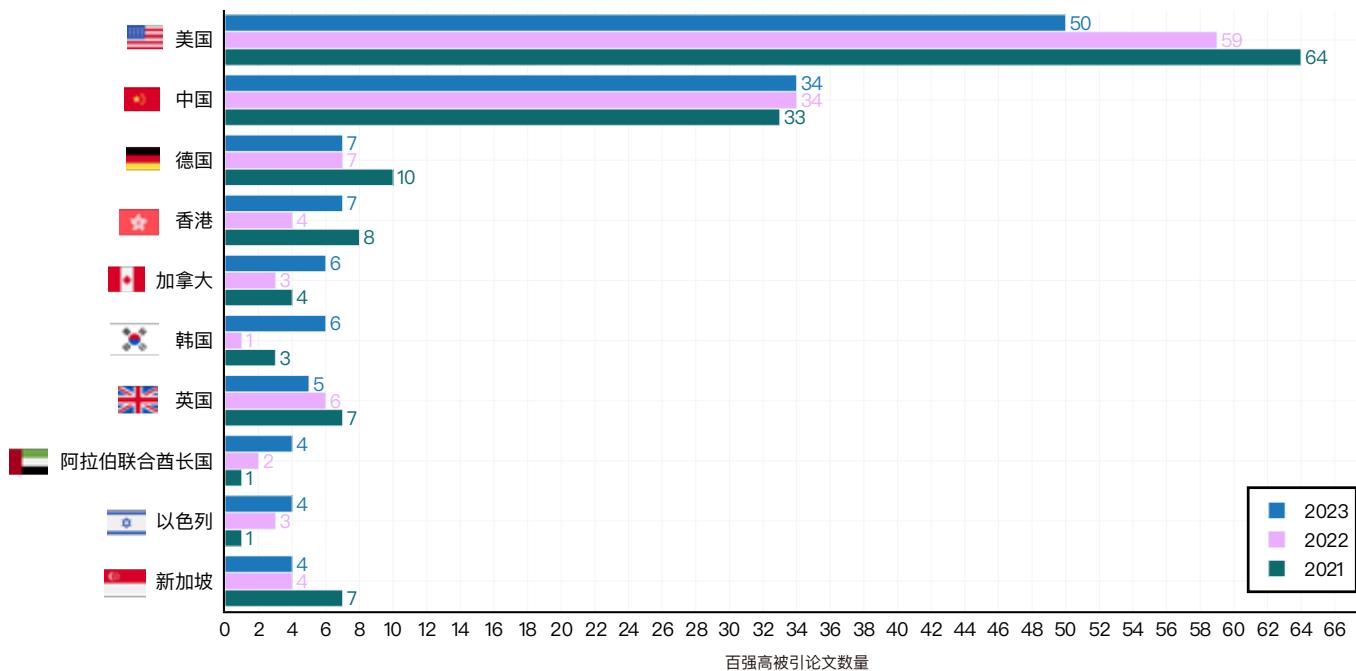


图 1.1.11

9、完整研究方法指南及百强论文名录详见附录部分。

10、一篇出版物可以有多个来自不同国家或组织的作者。例如，如果一篇论文包括来自多个国家的作者，则每个国家只计一次。因此，本部分数字的总和超过 100。

第一章：研究与开发

1.1 论文发表

按行业分类

学术界持续产出最多被引的人工智能论文,2023年为42篇,2022年为27篇,2021年为34篇(图1.1.12)。值得注意的是,产业界贡献出现了显著下降,百强论文的数量从2021

年的17篇和2022年的19篇骤降至2023年的仅7篇。随着人工智能研究竞争日益激烈,许多行业人工智能实验室降低发表论文频率或减少披露研究细节。

2021–2023年按选定行业划分的百强高被引论文的统计

资料来源:2025年人工智能指数 | 图表:2025年人工智能指数报告

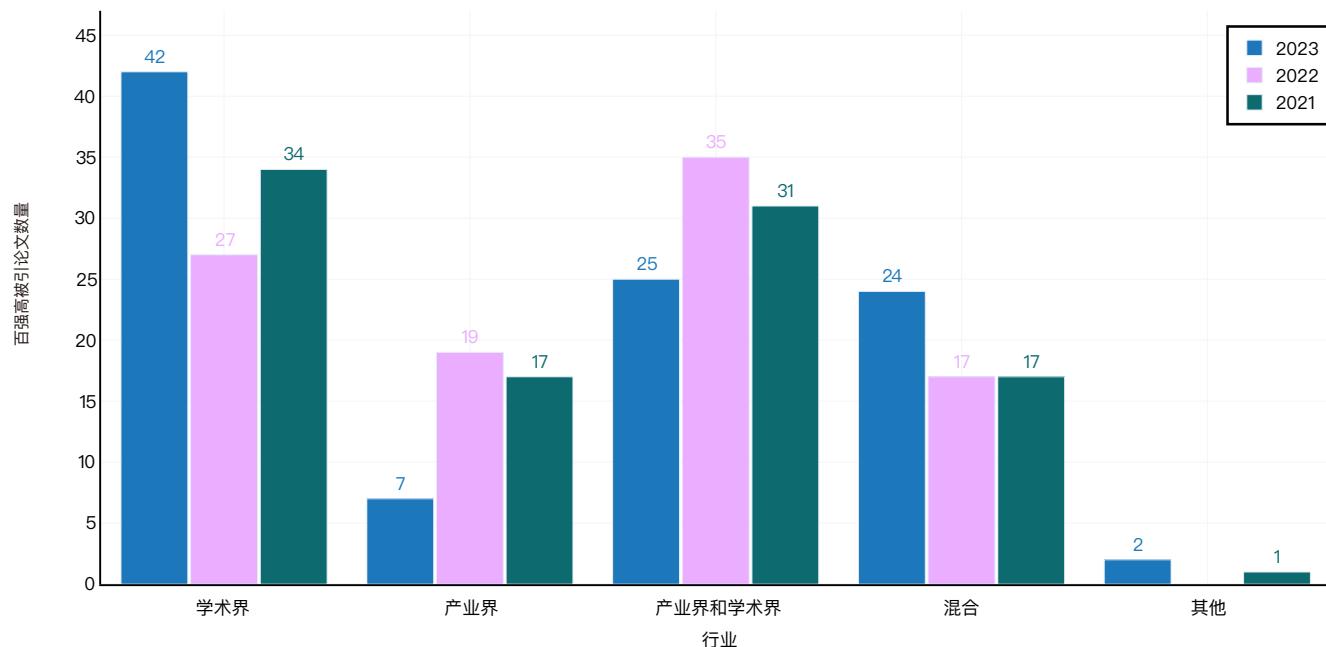


图 1.1.12¹¹

¹¹“混合”名称包括所有非产业界和学术界的跨部门合作(如产业界和政府、学术界和非营利组织)。有些机构缺少2021年的数据,因为它们当年没有论文入选前100名。由于论文的多个作者可能来自不同机构,因此图1.1.12中的机构标签总可能超过100个。另外,由于有两篇论文的作者所属部门不明,因此图1.1.12中的论文总数为98篇。

第一章：研究与开发

1.1 论文发表

按机构类型

图 1.1.13 展示了 2021 至 2023 年间全球人工智能领域高被引百强论文来源机构分布情况。部分机构在图表中可能出现空白柱，这表明该机构在某年未发表百强论文。此外，图 1.1.13 仅列出了排名前 10 的机构，尽管许多其他机构也做出了重要

贡献。谷歌每年均位居榜首，但在 2023 年与清华大学并列第一，两者均有 8 篇论文入选百强。2023 年，卡内基梅隆大学是排名最高的美国学术机构。

2021—2023 年按机构类型划分的百强高被引论文的统计

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

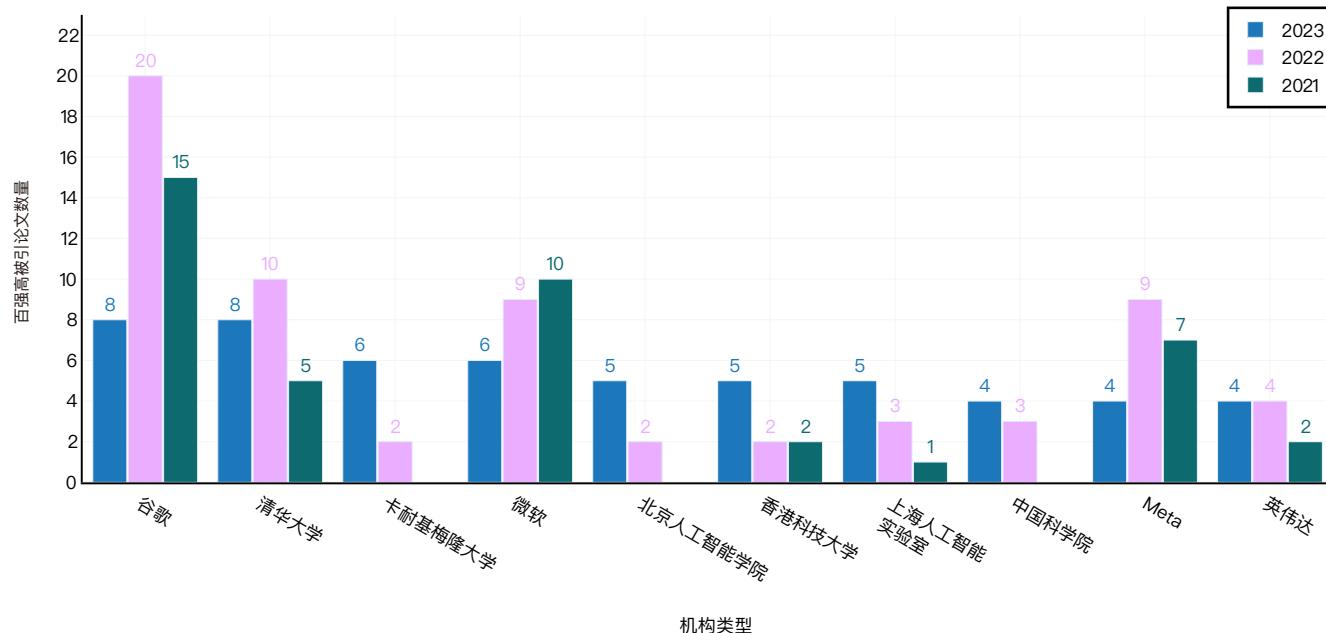


图 1.1.13

第一章：研究与开发

1.2 专利

本节通过分析全球人工智能专利的时序演变趋势，揭示人工智能领域技术创新、研究进展与产业发展的关键动态。此外，分析人工智能专利可以揭示这些技术进步如何在全球范围内分布。与论文发表数据类似，人工智能专利数据的可获得性也存在明显的延迟，2023年是可获得数据的最近年份。本节中的数据来自欧洲专利局（EPO）提供的综合数据库 PATSTAT Global 中的专利级目录记录。¹²

1.2 专利 概述

图 1.2.1 显示了 2010 年至 2023 年全球人工智能专利的增长情况。在过去十几年中，人工智能专利数量稳步大幅增长，从 2010 年的 3833 项增至 2023 年的 122511 项。去年，人工智能专利总量增长了 29.6%。

2010—2023 年全球人工智能专利授权数量

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

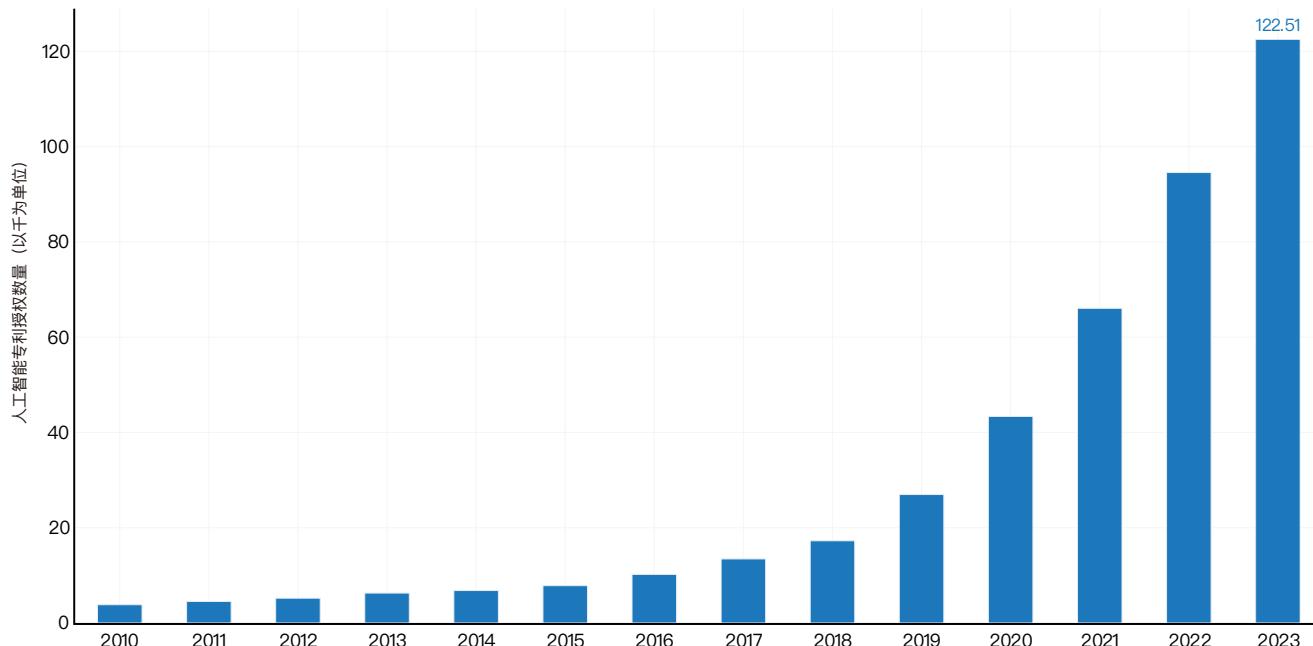


图 1.2.1

12、有关本节专利分析方法的更多详情，请参阅附录。

第一章：研究与开发

1.2 专利

按国家 / 地区分类

图 1.2.2 展示了授予的人工智能专利的区域分布，即全球不同地区提交的专利数量。截至 2023 年，截至 2023 年，全球获授权的人工智能专利中，绝大多数(82.4%)来自东亚和太平

洋地区，北美地区以 14.2% 的占比位列第二。自 2010 年以来，东亚和太平洋地区与北美在人工智能专利授权方面的差距不断扩大。

2010—2023 年按地区划分的 AI 专利授权量（占全球总量的百分比）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

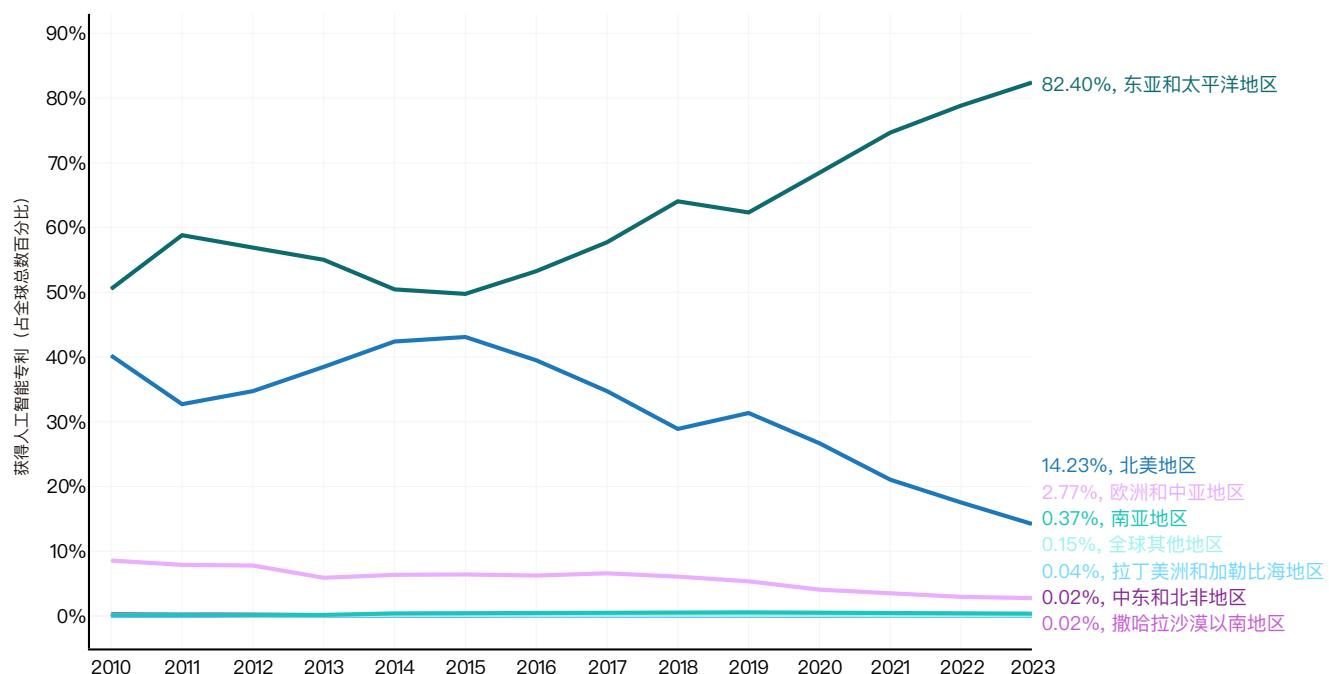


图 1.2.2¹³

13、不同国家和地区的专利标准和法律各不相同，因此在解释这些图表时应谨慎。更详细的国家级专利信息将在 AI Index's Global Vibrancy Tool 的后续版本中发布。

第一章：研究与开发

1.2 专利

按地理区域细分，全球获批的人工智能专利中，绝大多数来自中国（69.7%）和美国（14.2%）（图 1.2.3）。来自美国的人工智能专利占比已从 2015 年的峰值（42.8%）有所下降。

图 1.2.3 和图 1.2.4 记录了哪些国家在人均人工智能专利方面处于领先地位。2023 年，每 10 万居民中人工智能专利授

权最多的国家是韩国（17.3 项），其次是卢森堡（15.3 项）和中国（6.1 项）（图 1.2.3）。图 1.2.5 显示了 2013 年至 2023 年人均人工智能专利授权量的变化。在此期间，卢森堡、中国和瑞典的人均人工智能专利增幅最大。

2010—2023 年按地区划分的人工智能专利授权量（占全球总量的百分比）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

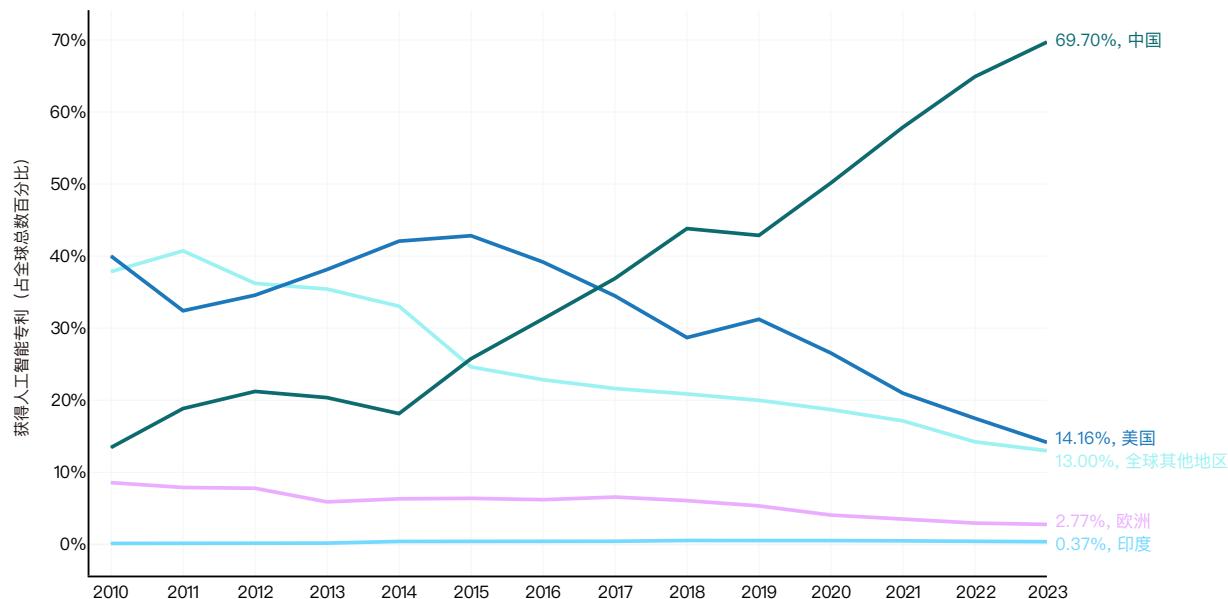


图 1.2.3

第一章：研究与开发

1.2 专利

2023 年按国家分类每 10 万居民授权的人工智能专利数量

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

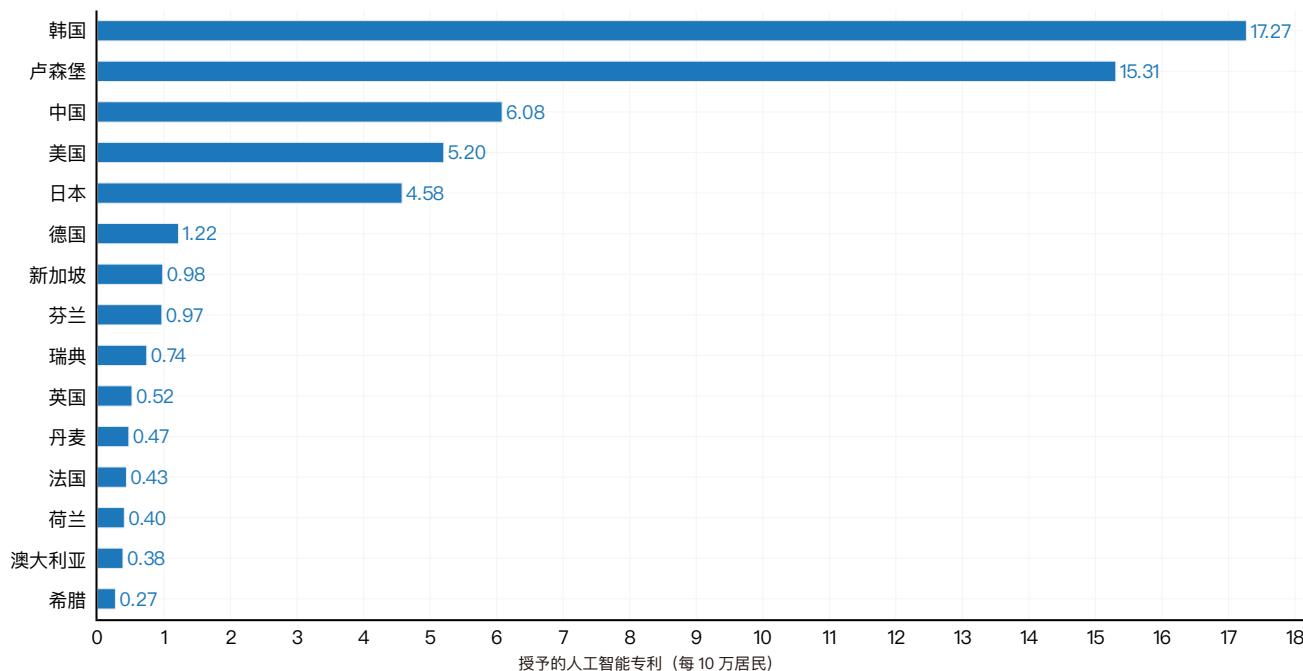


图 1.2.4

2013 年与 2023 年按国家划分的每 10 万居民授予的人工智能专利数量的百分比变化对比

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

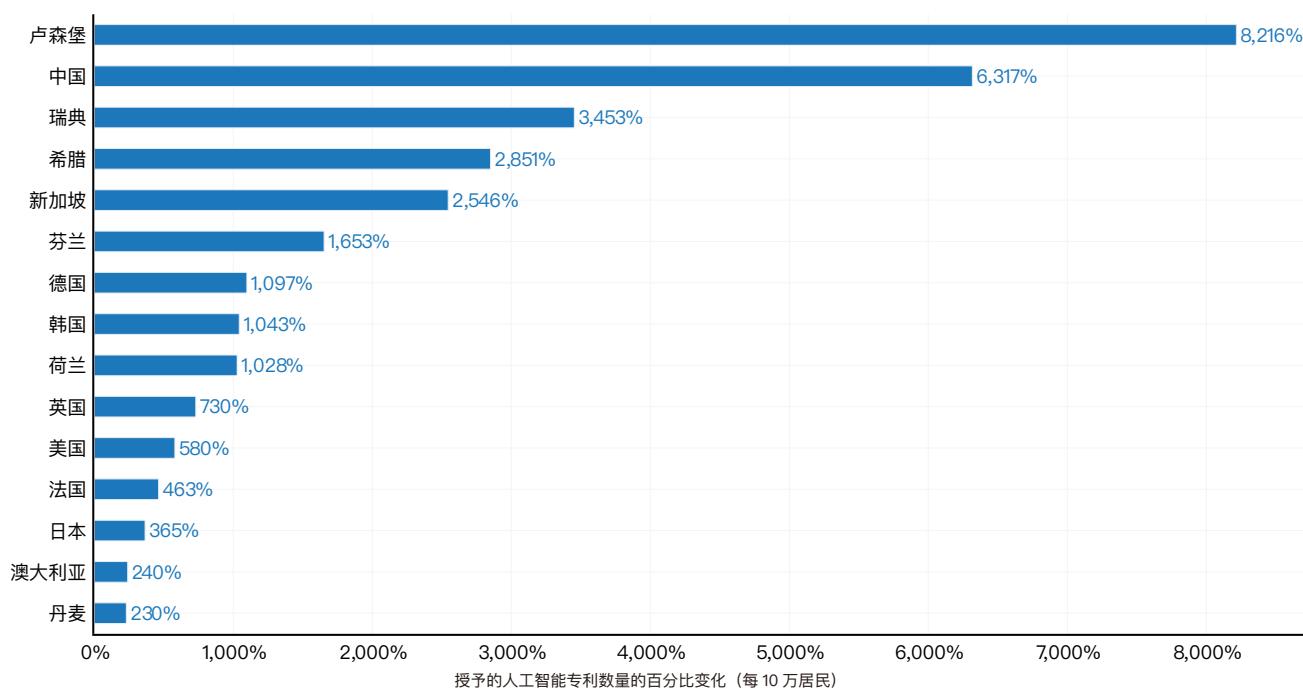


图 1.2.5

第一章：研究与开发

1.3 标志性人工智能模型

本节探讨标志性人工智能模型。人工智能指
数数据提供商 Epoch AI 使用 “标志性机器学习
模型 (notable machine learning) ”一词来指代
人工智能 / 机器学习生态系统中特别有影响力的
模型。Epoch 维护着一个涵盖 1950 年代以来发
布的 900 个人工智能模型的数据库，其收录标准
包括：技术突破性、历史里程碑意义或高被引率
等核心指标。由于 Epoch 对数据进行了人工整
理，因此一些被部分人视为标志性的模型可能未
被收录。通过分析这些模型，我们可以全面了解机
器学习领域近年来和过去几十年的发展变化。数
据集中可能缺少某些模型，但数据集可以揭示相
对趋势。标志性人工智能模型包括 GPT-4o、
Claude 3.5 和 AlphaGeometry。

在本节中，人工智能指数从不同角度探讨了
标志模型的发展趋势，包括起源国、起源组织、模
型发布梯度、参数数量和计算使用情况。最后，分
别对机器学习的训练成本以及推理成本进行了探
讨与分析。

1.3 标志性人工智能模型

按国家 / 地区分类

为了展示人工智能领域不断演变的地缘政治格局，人工智能指数展示了标志性
模型所属的国家。图 1.3.1 展示了归因于研究人员所属机构所在地的标志性人工智能
模型总数。¹⁶ 2024 年，美国以 40 个标志性人工智能模型遥遥领先，中国以 15 个
紧随其后，法国则有 3 个。2024 年全球主要经济体包括美国、中国和欧盟均报告
说，2024 年发布的标志性模型少于上一年(图 1.3.2)。自 2003 年以来，美国开
发的模型数量超过了英国、中国和加拿大等其他主要国家(图 1.3.3)。

模型发布总量下降确切原因难以确定，但这可能源于多种因素的综合作用：
训练数据规模的不断扩大、人工智能技术的日益复杂化，以及开发新建模方法所面临
的挑战日益严峻。Epoch AI 当前收录的标志性模型可能遗漏了部分受关注度较低国
家的发布成果。人工智能指数与 Epoch 合作致力于提高人工智能模型生态系统中的
全球代表性。如果读者认为缺少了某些国家的模型，欢迎联系人工智能指数团队，我
们将努力解决这个问题。

2024 年按选定地理区域划分的标志性人工智能模型数量

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

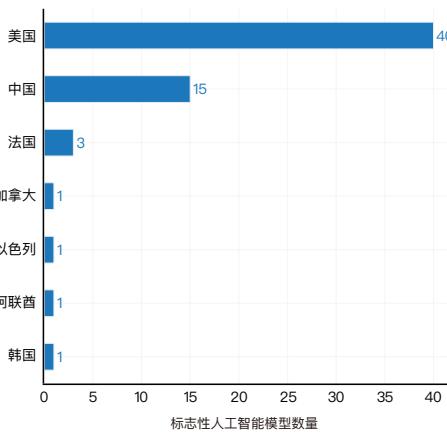


图 1.3.1¹⁷

2003–2024 年按选定地理区域划分的标志性人工智能模型数量

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

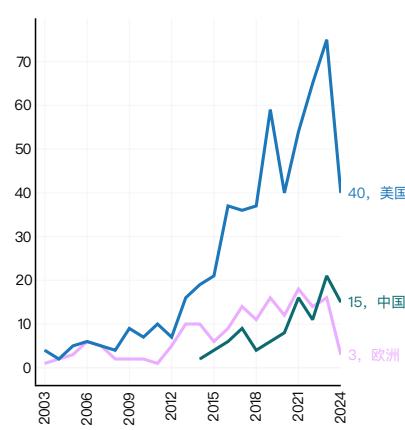


图 1.3.2

14、“人工智能系统 (AI system)”是指基于人工智能的计算机程序或产品，如 ChatGPT。“人工智能模型 (AI Model)”包括一组在训练过程中学到的参数值，例如 GPT-4。

15、Epoch AI 数据库会不断添加新的和历史性的模型，因此今年人工智能指数中包含的模型逐年总数可能与去年报告中公布的数据不完全一致。数据统计截取于 2025 年 3 月 17 日。

16、如果介绍机器学习模型的论文中至少有一位作者与某个国家的机构有关联，那么该模型就与该国家有关。如果模型的作者来自多个国家，则可能出现重复计算的情况。

17、该图表：展示了所选择的部分国家 / 地区的模型发布情况。有关各国模型发布情况的更全面数据，将于即将发布的 AI Index's Global Vibrancy Tool 中提供。

第一章：研究与开发

1.3 标志性人工智能模型

2003–2024 年按地理区域划分的标志性人工智能模型数量（总量）

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

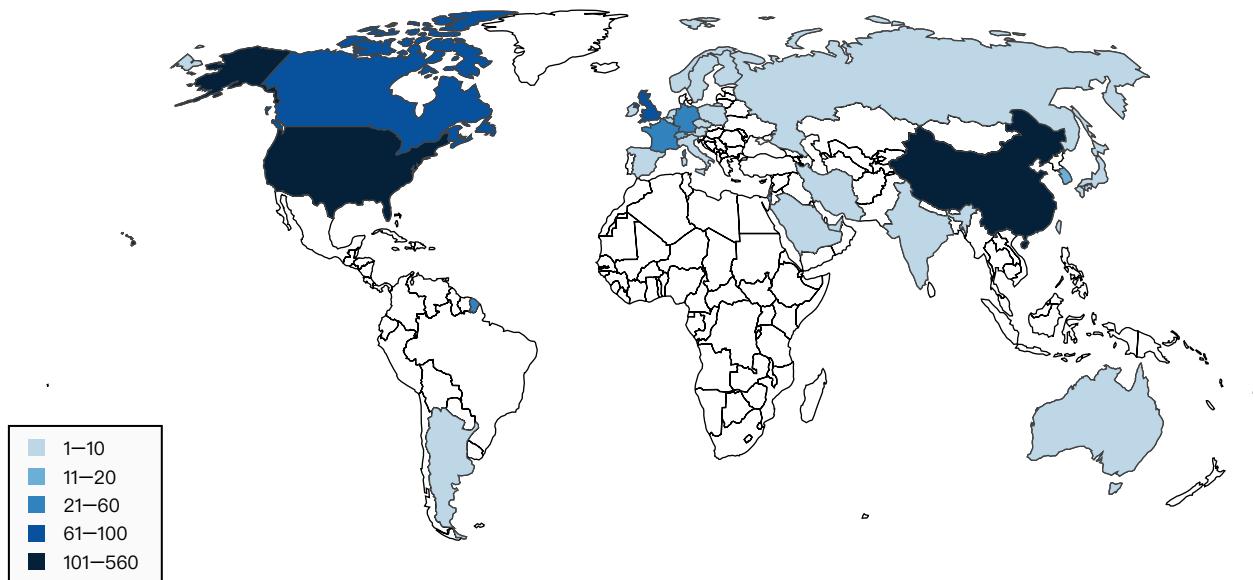


图 1.3.3

按行业分类

图 1.3.4 展示了按模型发布年份划分的标志性人工智能在各领域的来源分布。Epoch 根据来源对模型进行了分类：产业界包括谷歌、Meta 和 OpenAI 等公司；学术界包括清华大学、麻省理工学院和牛津大学等大学；政府指国家附属研究机构，如英国的 Alan Turing Institute for AI 和阿布扎比的 Technology Innovation Institute；研究集体包括非营利性人工智能研究组织 Allen Institute for AI 和 Fraunhofer Institute。

2014 年之前，学术界在发布机器学习模型方面一直处于领先地位。自那以后，工业界开始领跑。根据 Epoch AI 的数据，在 2024 年，工业界将产生 55 个标志性人工智能模型。¹⁸ 随着时间的推移，产学研合作推动的模型数量持续增长。过去十年间，来自产业界的知名人工智能模型占比稳步上升，至 2024 年已达到 90.2%。

¹⁸ 在解释这一数字时应谨慎。学术模型数量为零并不意味着 2023 年学术机构没有产生任何标志性模型，而是意味着 Epoch AI 没有发现任何标志性模型。此外，学术研究成果往往需要更长时间才能获得认可，即便是那些引入重要架构的高被引论文，也可能需要数年时间才能产生广泛影响。

第一章：研究与开发

1.3 标志性人工智能模型

2003—2024 年按行业划分的标志性人工智能模型数量

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

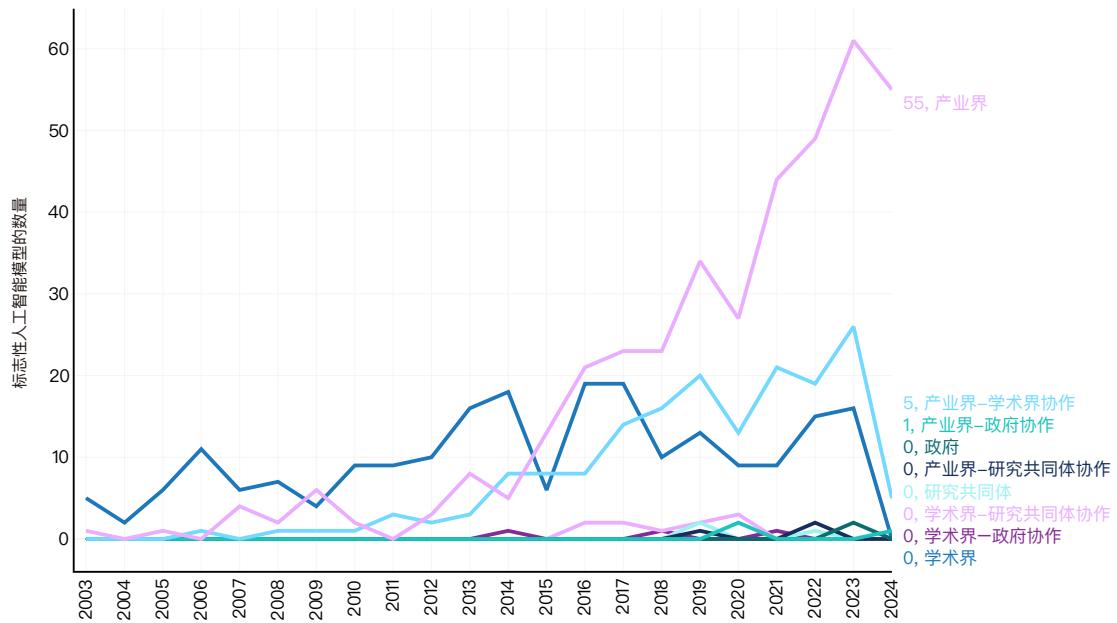


图 1.3.4

2003—2024 年按行业划分的标志性人工智能模型（占总数百分比）

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

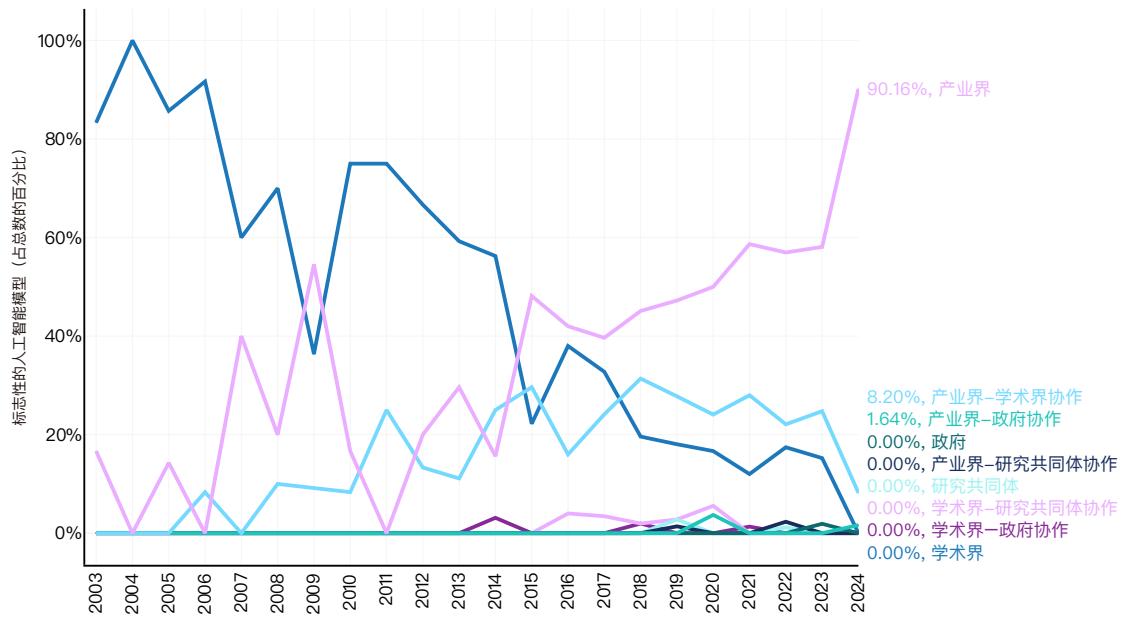


图 1.3.5

第一章：研究与开发

1.3 标志性人工智能模型

按研发主体分类

图 1.3.6 与图 1.3.7 分别呈现了 2024 年度及过去十年间，机器学习领域标志性模型研发的主导机构分布情况。2024 年，贡献最多的是谷歌（7 个）、OpenAI（7 个模型）和阿里巴巴（4 个）。自 2014 年以来，谷歌以 187 个标志性模型遥遥领先，其

次是 Meta（82 个）和微软（39 个）。在学术机构中，卡内基梅隆大学（25 个）、斯坦福大学（25 个）和清华大学（22 个）自 2014 年以来在标志性模型研发方面成果最多。

2024 年按组织划分的标志性人工智能模型数量

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告



图 1.3.6¹⁹

2014—2024 年按组织划分的标志性人工智能模型数量（总计）

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

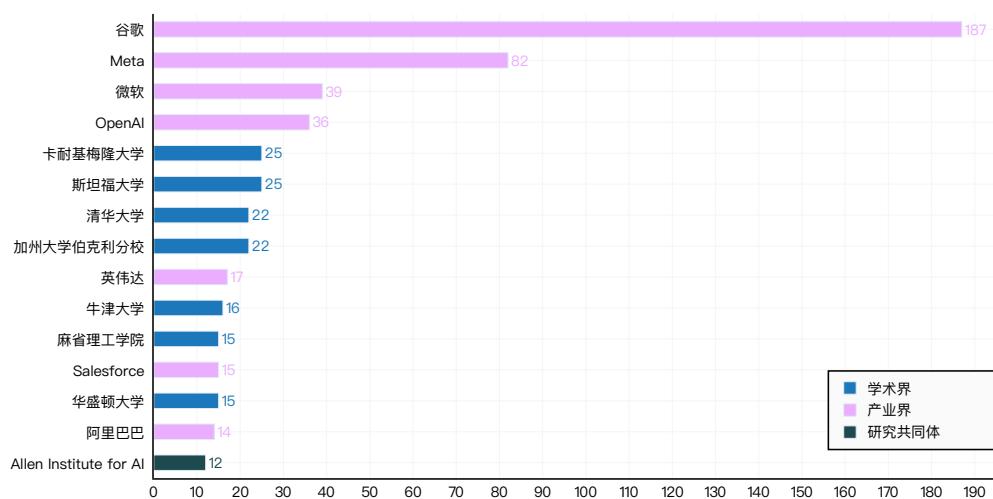


图 1.3.7

19、在组织统计数据中，DeepMind 发布的研究被归入谷歌。

模型发布

机器学习模型按照开放程度和使用权限可分为多种发布类型。API 访问模型，如 OpenAI 的 o1，允许用户通过查询与模型进行交互，而无需直接访问其底层权重。限制条件下的开源权重模型，如 DeepSeek 的 V3，提供对其权重的访问，但施加了一些限制，如禁止商业使用或二次分发。托管访问非 API 类模型，如 Gemini 2.0 Pro，指仅通过平台界面可用，不提供程序化调用接口的模型。无限制开源权重模型，如 AlphaGeometry，是完全开放的，允许自由使用、修改和再分发。非商业开源权重模型，如 Mistral Large 2，共享权重，但仅限于研究或非商业目的使用。最后，未发布模型，如 ESM3 98B，依然专有，只有其开

发人员或选定的合作伙伴才能访问。未知指的是访问类型不明确或未公开的型号。

图 1.3.8 展示了各类模型发布时所采用的不同访问权限类型。²⁰ 2024 年，API 访问是最常见的发布类型，61 个模型中有 20 个以这种方式提供，其次是限制使用的开源权重和未发布模型。

图 1.3.9 从比例维度呈现了机器学习模型访问权限类型的历时演变情况。2024 年，大多数人工智能模型是通过 API 访问发布的(32.8%)，这一比例自 2020 年以来稳步上升。

2014—2024 年按访问类型划分的标志性人工智能模型数量

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

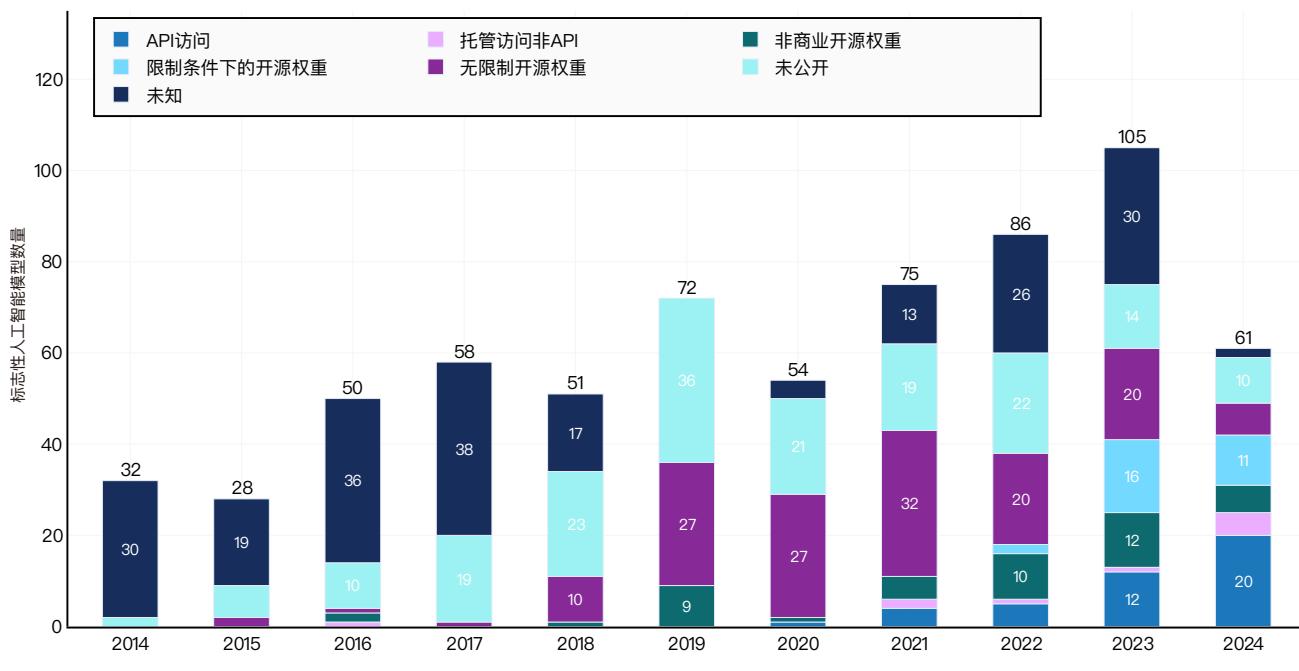


图 1.3.8 21

²⁰ 托管访问是指使用第三方远程提供的计算资源或服务（如软件、硬件或存储），而不是亲自拥有或管理这些资源或服务。托管访问不是在本地运行软件或基础设施，而是通过云或其他远程服务（通常是互联网）访问这些资源。例如，通过 AWS、谷歌 Cloud 或微软 Azure 等平台使用 GPU，而不是在自己的硬件上运行 GPU，就属于托管访问。

21、Epoch 数据库中的所有模型并非都按访问类型分类，因此图 1.3.8 至 1.3.10 中的总数可能与本章其他地方报告的总数不完全对齐。

第一章：研究与开发

1.3 标志性人工智能模型

2014—2024 年按访问类型划分标志性人工智能模型（占总数百分比）

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

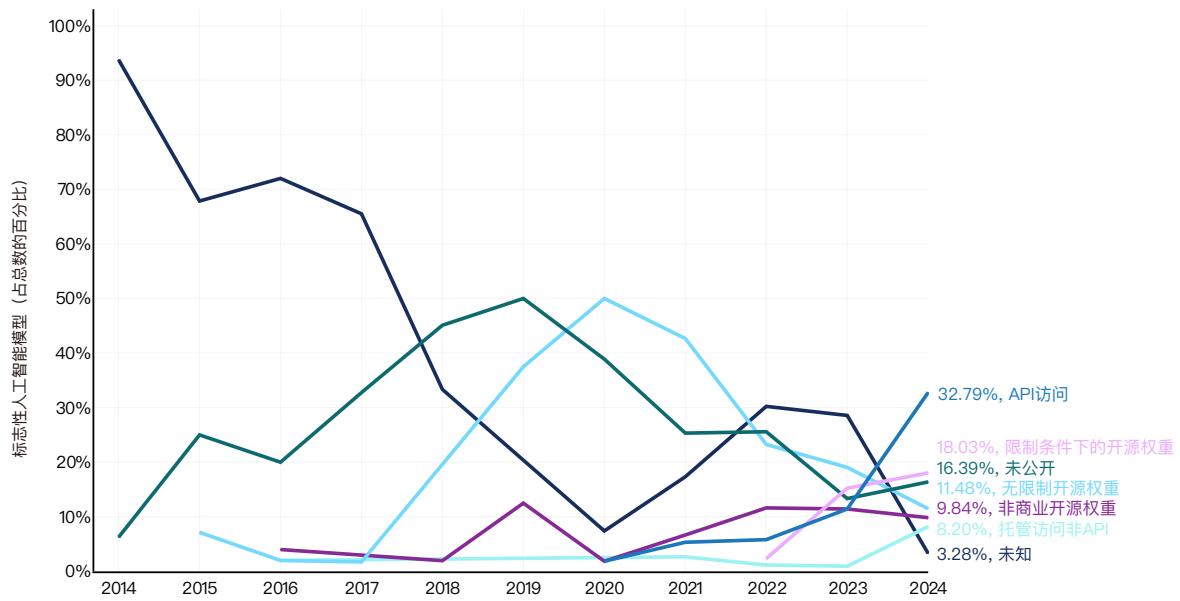


图 1.3.9

在传统的开源软件发布中，所有组件，包括培训代码，通常都会公开。然而，人工智能技术却往往并非如此，即使发布模型权重的开发人员也可能保留培训代码。如图 1.3.10 所示，标志

性人工智能模型可按代码开放程度进行分类。2024 年，其中 60.7% 的模型在发布时未同步公开训练代码。

2014—2024 年按训练代码访问类型划分的标志性人工智能模型数量

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

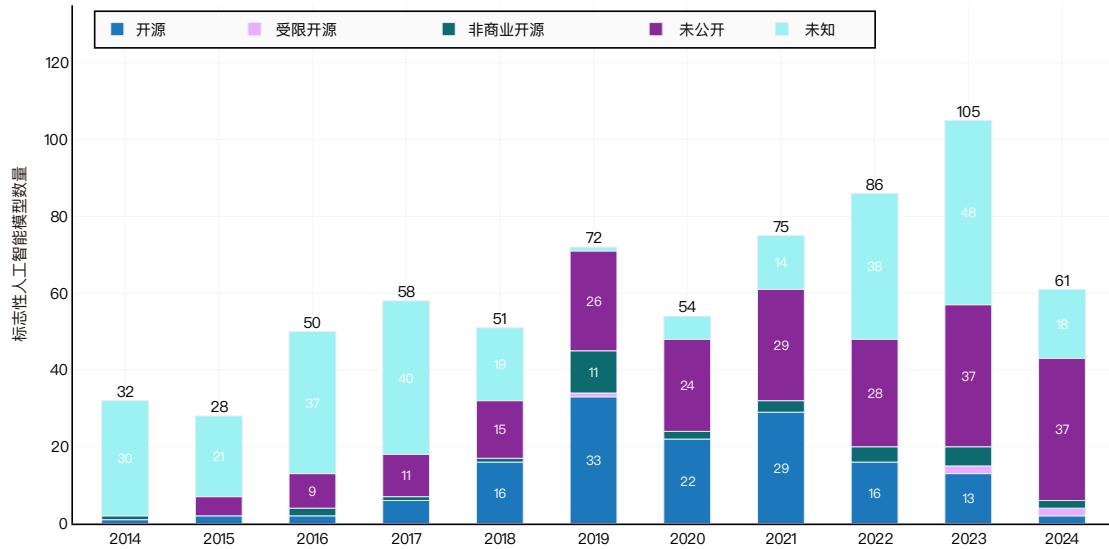


图 1.3.10

参数发展趋势

机器学习模型中的参数是在训练过程中学习到的数值，决定了模型如何解释输入数据和进行预测。参数较多的模型需要更多的数据来训练，但它们可以承担更多的任务，通常优于参数较少的模型。

图 1.3.11 展示了 Epoch 数据库中机器学习模型的参数数量，并按模型来源的行业进行了分类。图 1.3.12 展示了相同的数据，但选取了较少的标志性模型。自 2010 年代初以来，模型

参数量呈现急剧增长态势，这反映了以下关键因素：架构复杂度持续提升、训练数据日益丰富、硬件设施不断改进，以及大模型效能已获验证。高参数量模型在产业界表现尤为突出，这表现出企业机构具备雄厚资金实力，足以支撑海量数据训练所需的巨额计算成本。下列部分图表：采用对数刻度，以准确反映近年来人工智能模型参数量及计算需求的指数级增长态势。

2003—2024 年按行业划分的标志性人工智能模型参数数量

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

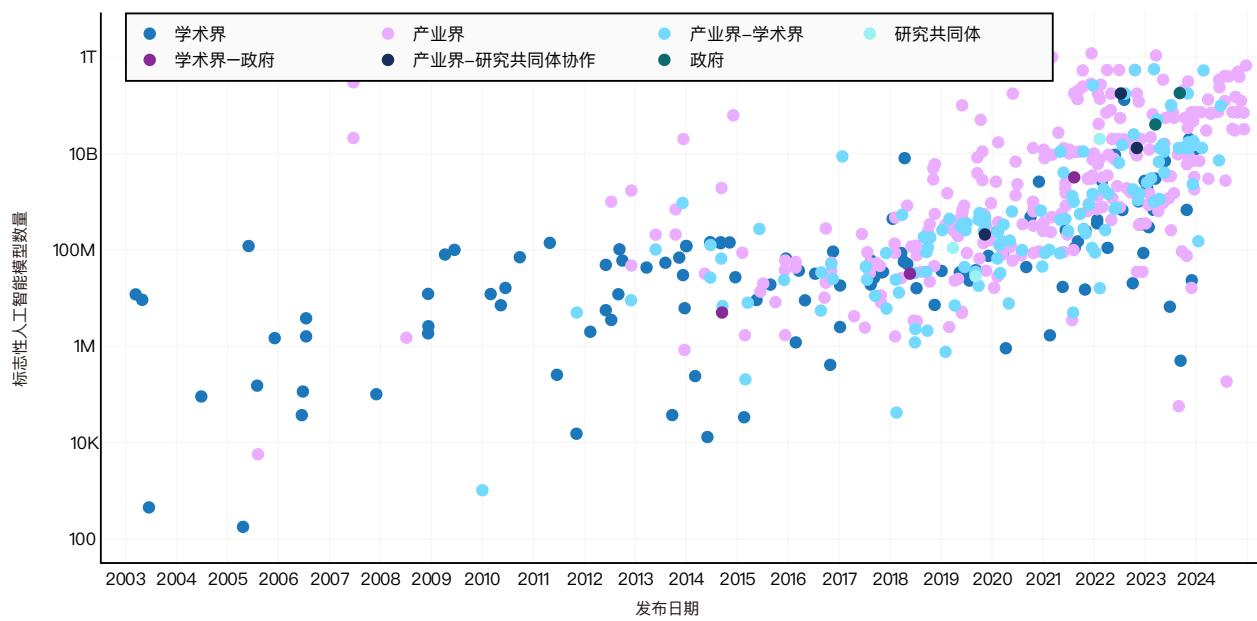


图 1.3.11

第一章：研究与开发

1.3 标志性人工智能模型

2012–2024 年按行业划分的标志性人工智能模型参数数量

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

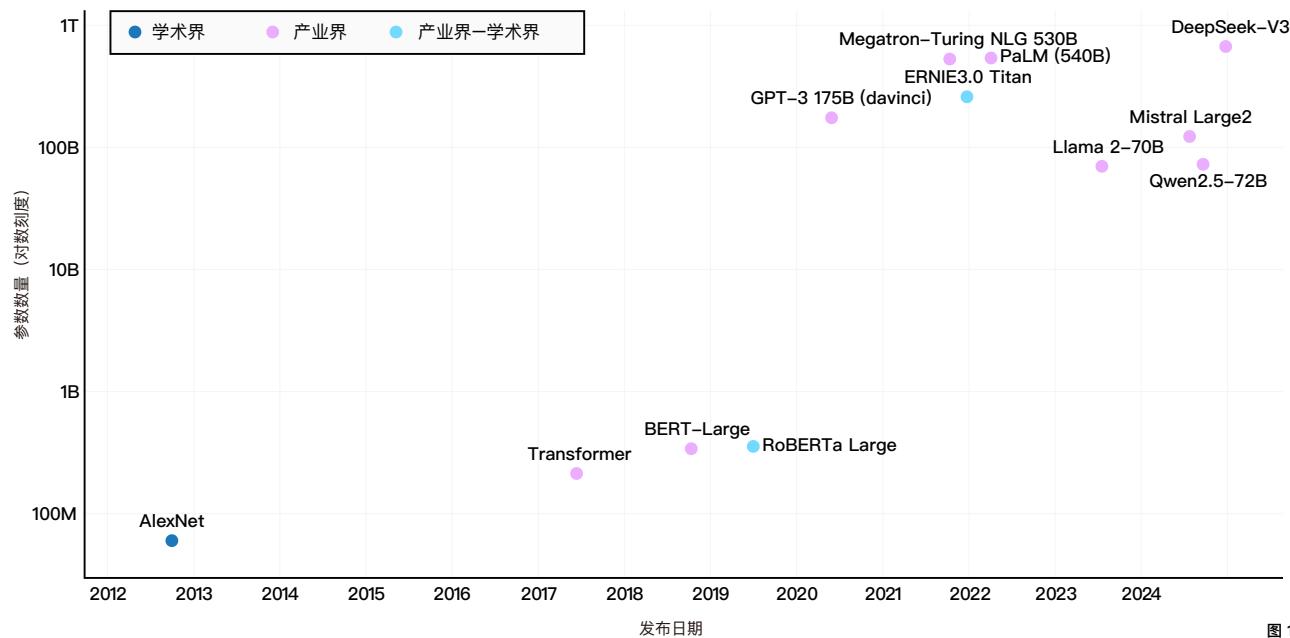


图 1.3.12

第一章：研究与开发

1.3 标志性人工智能模型

随着模型参数数量的增加，用于训练人工智能系统的数据量也在增加。图 1.3.13 展示了用于训练标志性机器学习模型的数据集规模的增长。2017 年发布并被广泛认为引发了大语言模型革命的 Transformer 模型，是在大约 20 亿个词元的基础上训练出来的。到 2020 年 GPT-3 175B——最初的 ChatGPT

的基础模型之一——估计是在 3740 亿个词元上训练出来的。相比之下，Meta 的旗舰大语言模型，即 2024 年夏天发布的 Llama 3.3，则是在大约 15 万亿个词元上训练出来的。根据 Epoch AI 的数据，大语言模型训练数据集的规模大约每八个月翻一番。

2010–2024 年标志性人工智能模型人工智能模型训练数据集规模

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

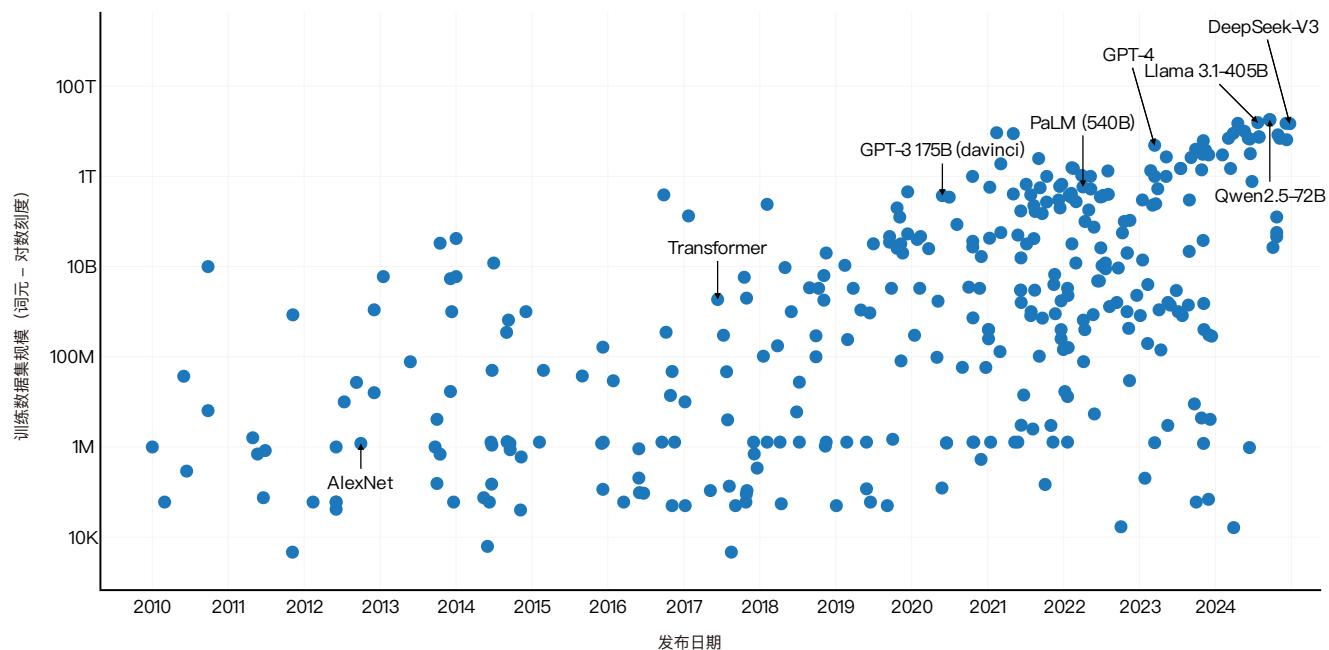


图 1.3.13

第一章：研究与开发

1.3 标志性人工智能模型

在越来越大的数据集上训练模型导致训练时间显著延长（图 1.3.14）。一些最先进的模型，如 Llama 3.1–405B，需要大约 90 天的时间来训练——这在当今标准下是一个典型的训练周期。谷歌于 2023 年底发布的 Gemini 1.0 Ultra 耗时约 100

天。这与 AlexNet 形成了鲜明对比，AlexNet 是首批利用 GPU 提高性能的模型之一，在 2012 年仅用五到六天就完成了训练。值得注意的是，AlexNet 的训练硬件远不及后者先进。

2010–2024 年标志性人工智能模型训练时长

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

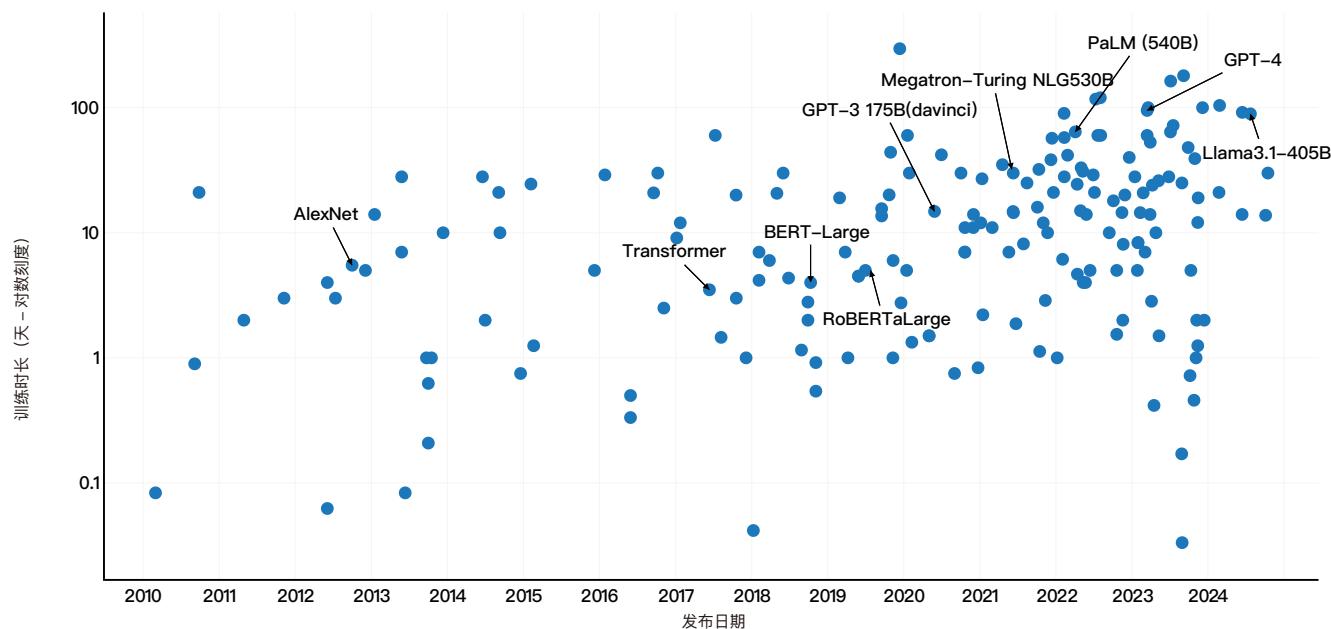


图 1.3.14

算力发展趋势

在人工智能模型领域，“compute”（计算资源）特指训练和运行机器学习模型所需的基础算力资源。通常而言，模型复杂度与训练数据集规模将直接影响所需算力资源的多少。模型复杂度越高、训练数据量越大，其训练过程所需的算力规模就越大。在最终训练运行之前，研究人员会在整个研发阶段进行多次测试运行。虽然单个模型的训练成本相对较低，但多次研发迭代所需的累计费用，以及必要数据集费用，将快速攀升至可观规模。需注意，当前数据仅反映最终训练阶段的成本，而非完整研发流程的总投入。

图 1.3.15 展示了近 22 年间标志性机器学习模型所需的训练算力变化情况。值得注意的是，近年来重要人工智能模型的算力消耗已呈现指数级增长态势。²² 据 Epoch 估算，标志性人工智能模型的训练算力大约每五个月翻一番。这一趋势在过去五年中尤为明显。算力需求的快速增长具有重要影响。以计算密集型模型为例，其往往会产生更大的环境足迹（environmental footprints），而企业机构通常比学术组织拥有更丰富的计算资源。作为参考，人工智能指数第 2 章分析了计算资源提升与模型性能改进之间关联性。

2003—2024 年按行业划分的标志性人工智能模型训练计算量

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

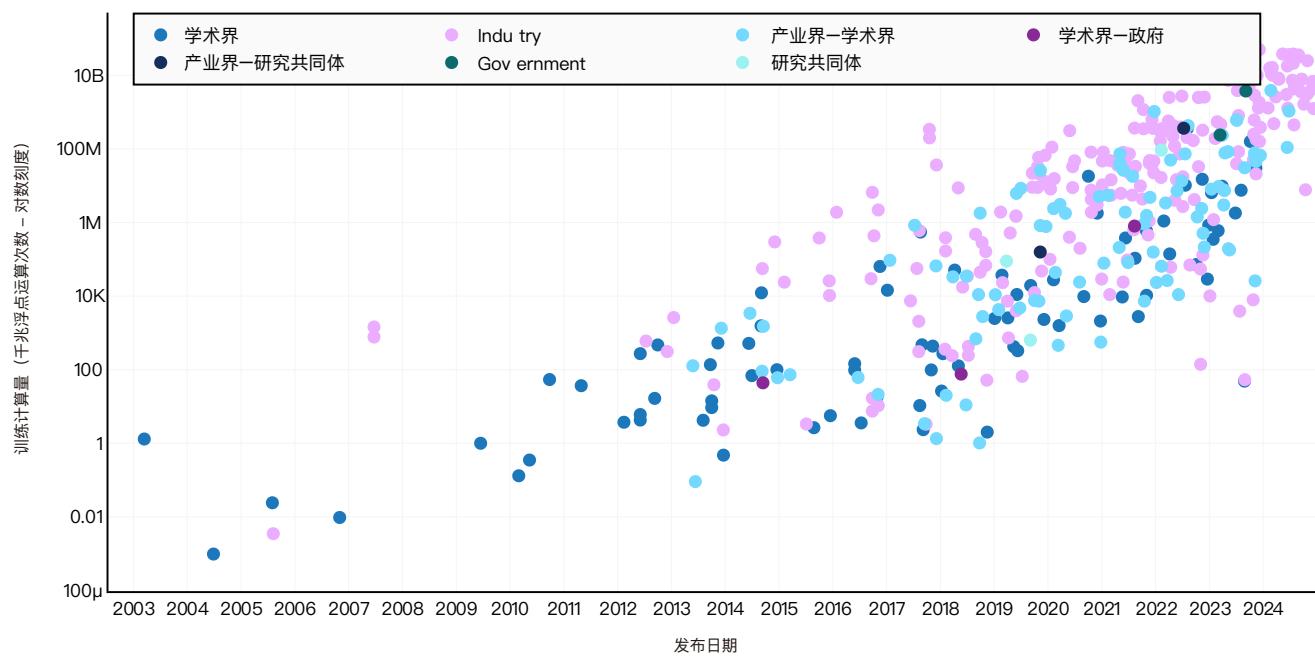


图 1.3.15²³

22. FLOP (floating-point operation) 指“浮点运算”。浮点运算是涉及浮点数的单一算术运算，如加法、减法、乘法或除法。处理器或计算机每秒可执行的 FLOP 数量是衡量其计算能力的指标。FLOP 率越高，计算机的计算能力就越强。用于训练人工智能模型浮点运算次数反映了该模型在开发过程中对算力的需求。

23. 训练算力估算 是人工智能模型分析的一个重要方面，但往往需要间接测量。在无法获得直接报告的情况下，Epoch 通过使用硬件规格和使用模式或根据模型架构和训练数据计算算术运算来估算计算量。在这两种方法都不可行的情况下，比较基准性能可作为代理，通过比较模型与已知计算值来推断训练算力。有关 Epoch 方法的全部详情，请参阅其网站的文档部分。

第一章：研究与开发

1.3 标志性人工智能模型

图 1.3.16 显示了自 2012 年以来标志性机器学习模型的训练算力需求变化情况。例如，AlexNet 网络是推广使用 GPU 改进人工智能模型的标准做法的模型之一，其训练估计需要 470 petaFLOP。²⁴ 2017 年发布的原 Transformer 需要约 7,400 petaFLOP。OpenAI 的 GPT-4o，当前最先进的基础模

型之一，需要 380 亿 petaFLOP。现在，开发尖端人工智能模型需要海量数据、巨额算力及雄厚的资金支持，而这些都是学术界无法获得的。大多数领先的人工智能模型都来自产业界，去年的人工智能指数首次强调了这一趋势。虽然今年的差距略有缩小，但这一趋势依然存在。

2012–2024 年按领域划分的标志性人工智能模型训练计算量

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

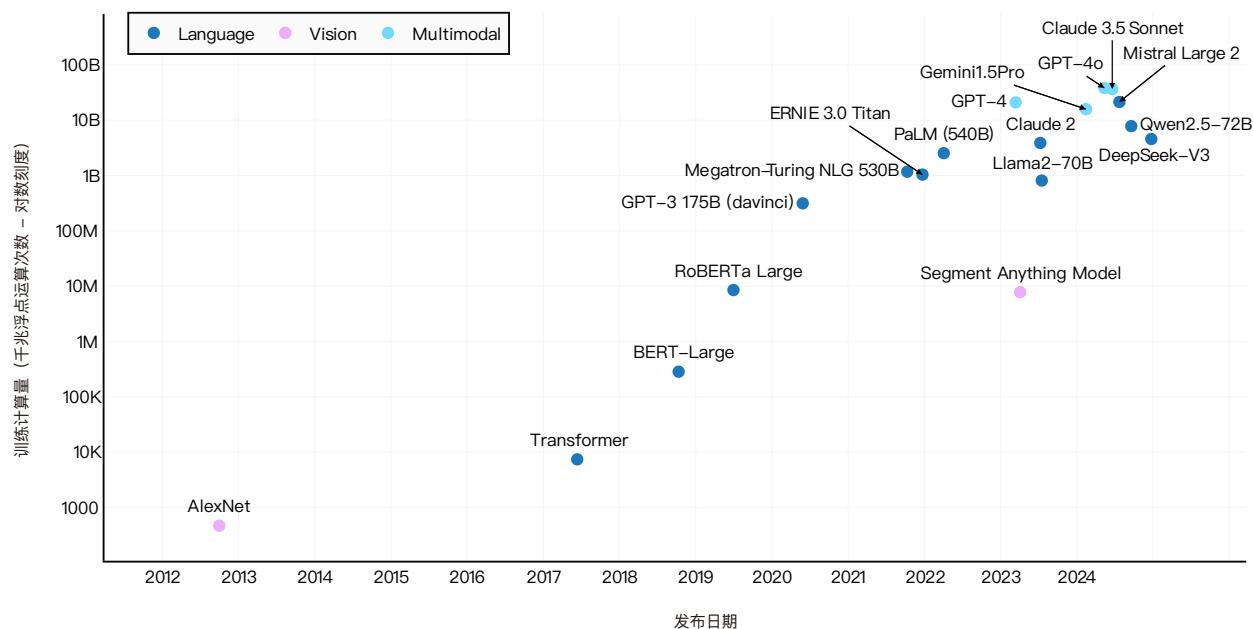


图 1.3.16

²⁴ petaFLOP (PFLOP) 是一个衡量计算性能的单位，1 PFLOP 相当于每秒四千万亿 (10^{15}) 次浮点运算

第一章：研究与开发

1.3 标志性人工智能模型

2024年12月，Deep Seek推出了V3模型，引起了广泛关注，尤其是因为该模型在计算资源需求远低于许多领先的大语言模型的情况下，实现了卓越的性能。图1.3.17比较了美国和中国一些标志性机器学习模型的训练计算量，凸显了一个关键趋势：

美国顶级人工智能模型的计算量通常远高于中国模型。根据Epoch AI的研究数据，自2021年底以来，中文前十大语言模型的训练算力年均增长约3倍，显著低于2018年以来全球其他地区5倍的年均增速。

2018—2024年美国与中国部分标志性人工智能模型训练算力分析

资料来源：Epoch AI, 2025 | 图表：2025年人工智能指数报告

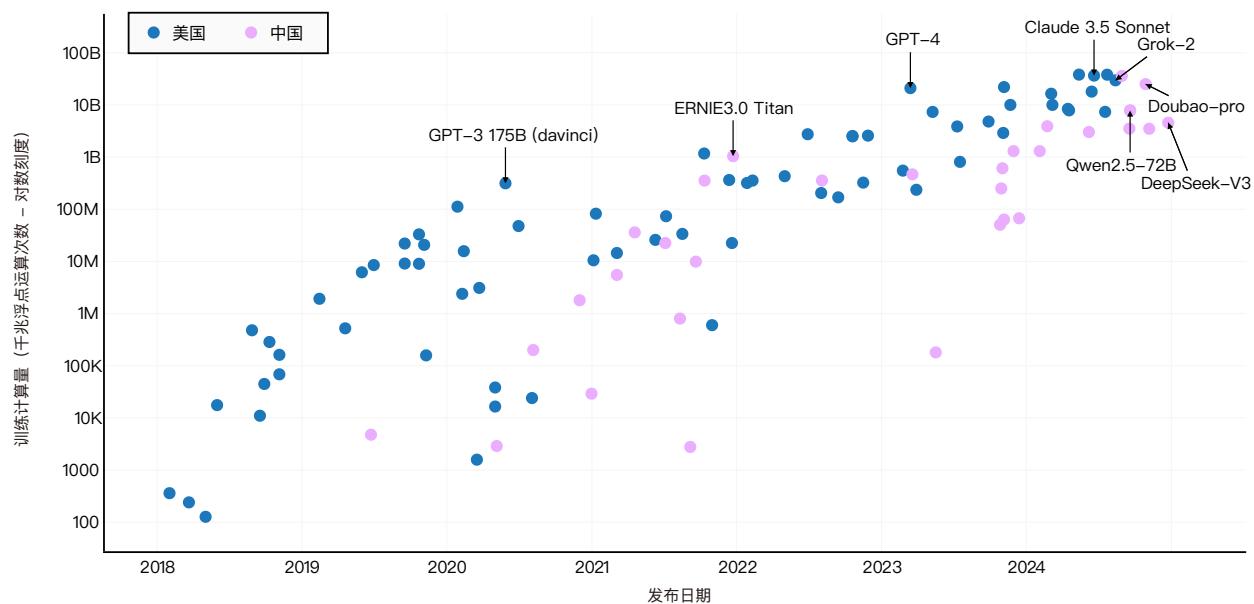


图1.3.17

重点：

模型训练会面临数据枯竭吗？

人工智能系统在算法上取得实质性改进的主要驱动力之一，是在越来越大的数据集上扩展模型及其训练。然而，随着互联网训练数据的日益枯竭，人们越来越担心这种扩展方法的可持续性以及数据瓶颈的可能性，因为在这种情况下，规模收益会逐渐减少。去年的人工智能指数探讨了这场辩论中的各种因素，包括现有互联网数据的可用性以及在合成数据上训练模型的潜力。今年的新研究表明，现有数据存量的持续时间可能比之前预期的要长。

Epoch AI 更新了之前对人工智能研究人员何时可能耗

尽数据的估计。在最新研究中，该团队根据词元计数估算了可用于训练模型的有效数据总存量（图 1.3.18）。Common Crawl，一个常用于人工智能训练的开放式网络爬虫数据库，是人工智能训练中经常使用的网络抓取数据开放存储库，据估计，它包含的词元中位数为 130 万亿个。索引网络包含约 510 万亿个词元，而整个网络包含约 3100 万亿个词元。此外，图片总存量估计为 300 万亿，视频为 1350 万亿。

数据存量中位数估计值

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数

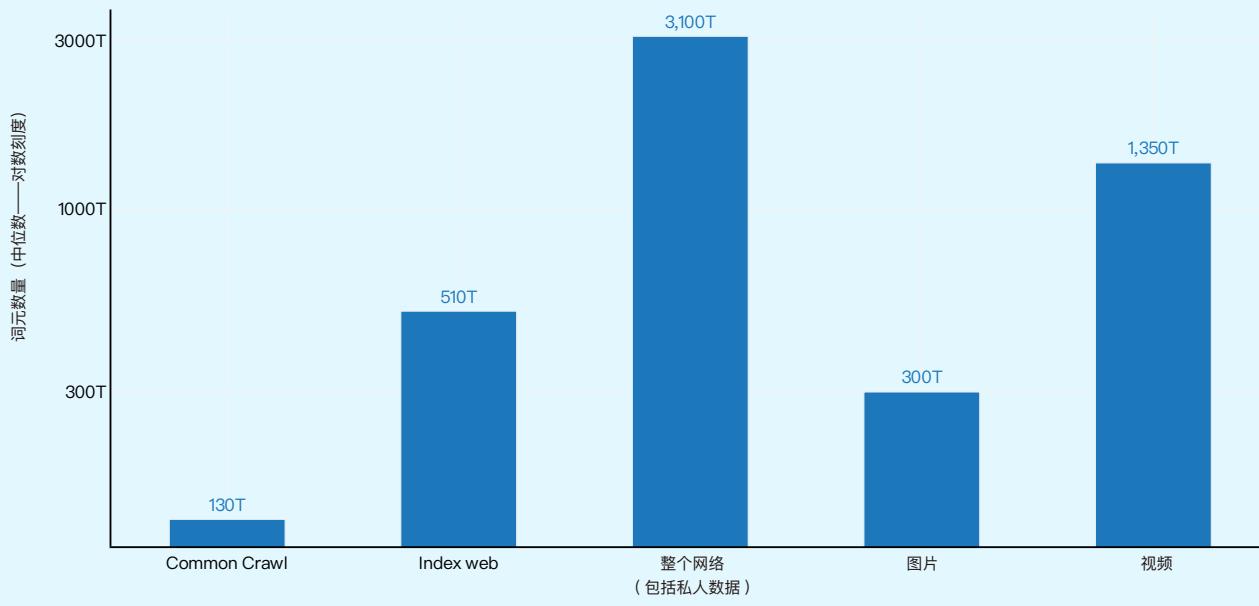


图 1.3.18

重点：

模型训练会面临数据枯竭（续）

Epoch AI 研究团队预计，在 80% 的置信区间内，当前的训练数据存量将在 2026 年至 2032 年期间全部用完（图 1.3.19）。数据耗尽的具体时间受多种因素影响。关键因素之一是数据集规模的历史增长，这取决于互联网用户生成和贡献内容的数量。另一关键因素在于算力使用效率——若采用最优算力配置方案进行模型训练，现有数据存量可支撑更长时间。但是，如果为提升高效推理计算能力

而对模型过度训练，则数据存量可能更快耗尽。当人工智能模型被过度训练时，即它们被训练的时间超过了典型的收益递减点，它们可能会实现更高的推理计算效率，也就是说，它们可以使用更少的计算能力来处理提示（进行预测、生成文本等）。然而，代价是数据存量（即可用于训练模型的数据）的加速消耗。

公共文本和数据使用存量的预测

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数

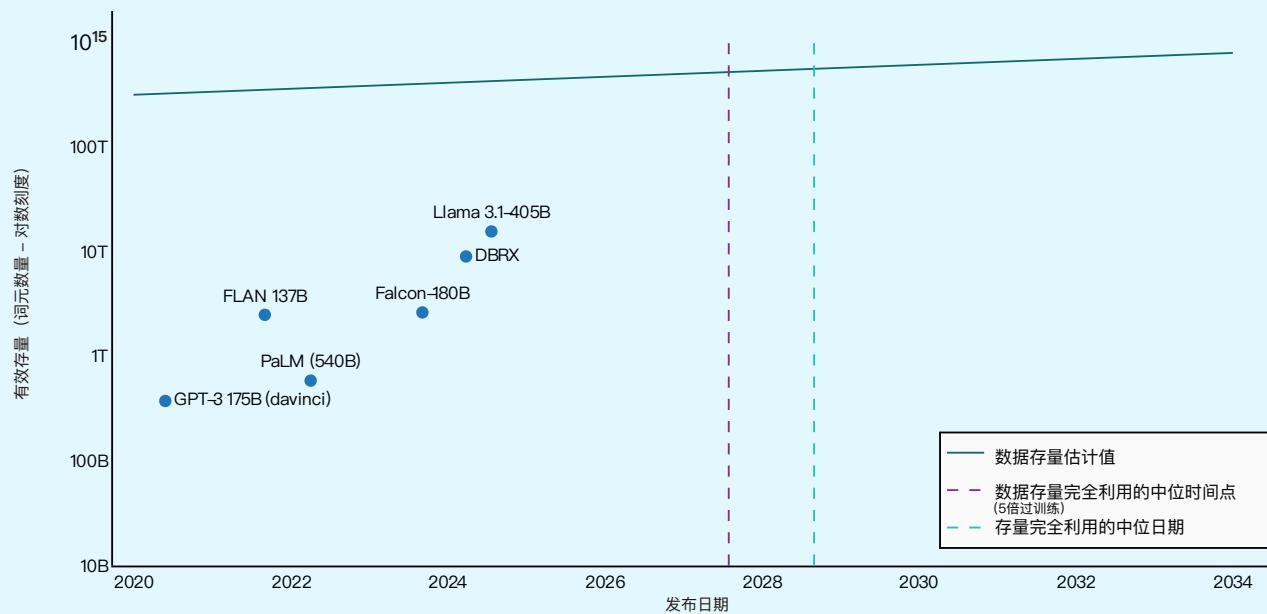


图 1.3.19

重点：

模型训练会面临数据枯竭（续）

这些预测与 Epoch 早期的估计略有不同，后者曾预测高质量文本数据将在 2024 年耗尽。修订后的预测反映了一种更新后的方法论，该方法结合了新研究，表明网络数据比精选语料库表现更好，且模型可以多次在同一数据集上训练。研究发现，经过精心过滤的网络数据是有效的，且重复训练同一数据集是可行的，这扩展了对可用数据量的估计。因此，Epoch 研究人员推迟了数据耗尽可能发生的时间预测。

使用合成数据（即由人工智能模型自身生成的数据）

来训练模型也被认为是解决潜在数据短缺的一种方案。2024 年人工智能指数报告指出，这种方法存在局限性，即模型在多次使用合成数据训练后，可能会丢失分布尾部的表征，从而导致模型输出质量下降。这一现象在不同模型架构中均被观察到，包括变分自编码器（VAEs）、高斯混合模型（GMMs）和大语言模型（LLMs）。然而，最新研究表明，当合成数据与真实数据叠加使用而非完全替代时，模型崩溃现象不会发生。尽管这种叠加不一定会提升性能或降低测试损失（测试损失越低，模型性能越好），但也不会导致像完全替换数据时那样的性能退化（图 1.3.20）。

数据积累对在 TinyStories 上预训练的语言模型的影响

资料来源：Gerstgrasser 等，2024 | 图表：2025 年人工智能指数报告

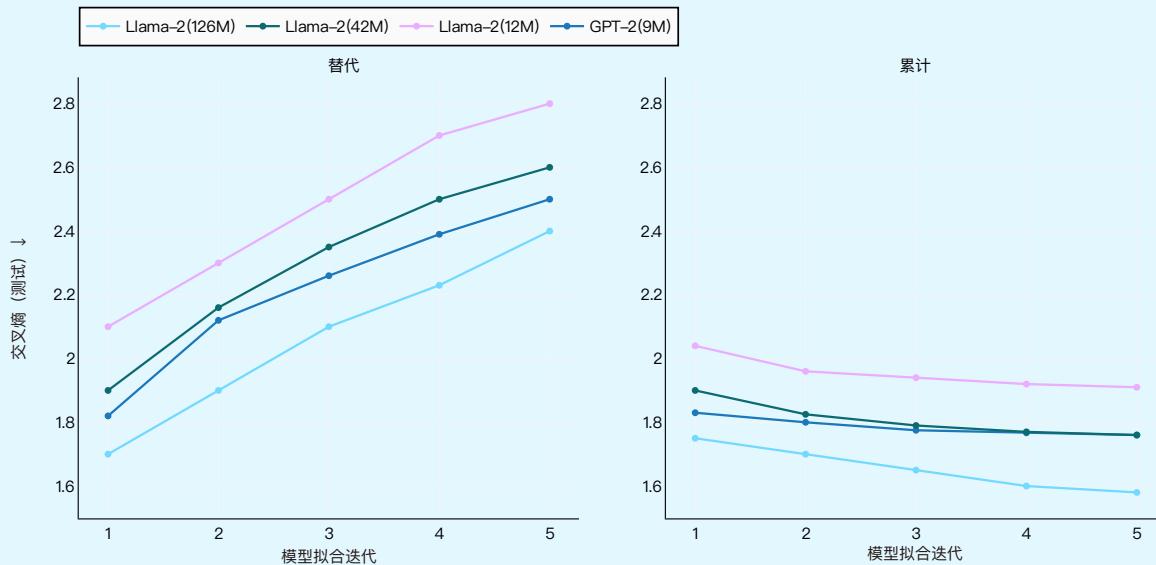


图 1.3.20

重点：

模型训练会面临数据枯竭（续）

本年度，高保真合成数据 (high-fidelity synthetic data) 的生成技术取得了进展。然而，合成数据与真实数据总体上仍有区别，与真实数据相比，目前没有可扩展的方法能使大语言模型在合成数据上的训练表现与真实数据相当。斯洛文尼亚研究团队通过多种架构和数据集对比了合成数据与真实数据的训练效果，评估了合成关系数据在保留原始数据的关键特征 ("保真度") 和对下游任务的有用性 ("实用性") 方面的表现。他们发现，大多数方法生成的合成数据可被系统性识别，尤其在涉及关系信息时更为明显。此外，与真实数据训练的模型相比，其高效注意力通常会下降，但有些方法仍能获得中等偏上的预测分数。在少数实验中，合成数据表现更优，例如使用 Synthetic Data Vault (SDV) 合成数据训练 XGBoost 分类器的效果优于沃尔玛数据，平均平方误差 (MSE) 更低。另有证据表明，合成数据在医疗领域潜力显著：某些模型架构通过合成增强数据集训练，可使分类和预测任务的 F1 分数或 AUROC (接收者操作特征曲线下面积) 在少数类上提升 5%–10%。²⁵

众所周知，大语言模型会产生幻觉并提供与事实不符的输出结果，因此人们对合成生成数据质量和保真度表示担忧。在对数据集中的幻觉内容进行训练时，模型的输出质量可能加速恶化。为解决这一问题，人们开发了新技术。例如，斯坦福大学和北卡罗来纳大学教堂山分校的研究人员利用自动事实检查和置信度评分来对模型响应对的事实性评分进行排序。这些研究人员推出 FactTune-FS 方法在事实性改进方面往往优于其他基于 RLHF 和解码的方法（图 1.3.21）。此外，Human-in-the-loop 标记首选回答的方法也被用于对齐语言模型，虽然有效，但成本较高。最后，训练前可通过事后过滤和去偏方法剔除合成数据中的异常值。

25、AUROC (接收者操作特征曲线下面积) 是评估人工智能模型性能的常用指标，尤其适用于分类任务。

第一章：研究与开发

1.3 标志性人工智能模型

重点：

模型训练会面临数据枯竭（续）

事实准确性（传记类问题的答案正确率）

资料来源：Tian 等，2023 | 图表：2025 年人工智能指数报告

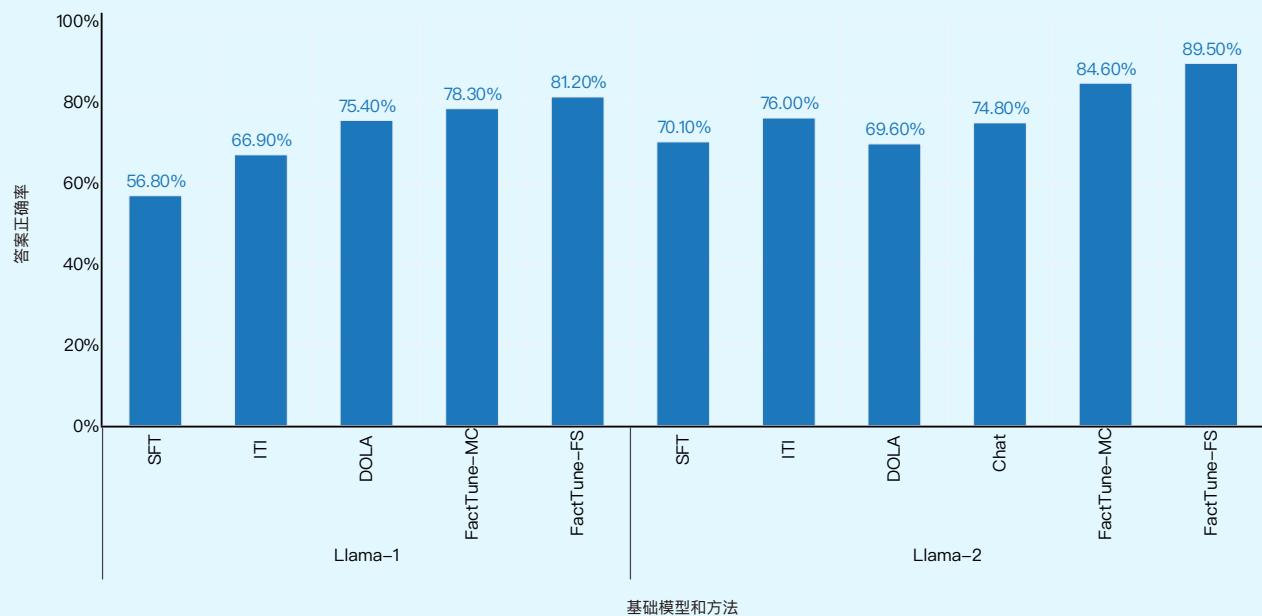


图 1.3.21

随着合成数据的普及，特别是人工智能生成的网络内容所占比例越来越大，未来的模型将不可避免地在非人类生成的材料上进行训练。虽然合成数据具有近乎无限供应的优势，但要有效地利用合成数据进行模型训练，还需要深入了解其对学习动态和学习效果的影响。数据集扩展的

一种方法是数据增强，即通过修改真实数据（如图像倾斜或混合）在保留关键特征的前提下创造新的变化。合成数据生成和数据增强都为增强人工智能模型提供了机会，但其有效运用还需要进一步研究。

第一章：研究与开发

1.3 标志性人工智能模型

推理成本

去年的人工智能指数强调了前沿大语言模型系统训练成本的快速上升。今年，除了更新对训练成本的分析外，该指数还研究了前沿系统的推理成本是如何随时间推移而变化的。推理成本指查询已训练模型所需的费用，通常以每百万词元的美元价格衡量。人工智能词元定价数据来自 Artificial Analysis 和 Epoch AI 的 API 定价专有数据库，报告价格为输入与输出词元价格的 3:1 加权平均值为分析推理成本。

人工智能指数与 Epoch 合作测量了在固定人工智能性能阈值下成本的下降情况。这种标准化方法有助于更精确的比较。虽然新模型价格可能更高，但其性能也显著提升——若直接与性能较低的老模型比较，可能掩盖真实趋势：单位美元获得的人工智能性能已大幅提高。

例如，在 MMLU 测试（评估语言模型性能的常用比较基准）中达到 GPT-3.5 水平（64.8 分）的模型，其推理成本从 2022 年 11 月的每百万词元 20 美元降至 2024 年 10 月的 0.07 美元（Gemini-1.5-Flash-8B），约 1.5 年内下降超 280 倍。在 GPQA（比 MMLU 更具挑战性的比较基准）上得分超过 50% 的模型的成本也呈现出类似的趋势。在这方面，推理成本从 2024 年 5 月的每百万词元 15 美元降至同年 12 月的 0.12 美元（Phi 4）。Epoch AI 估计，根据任务不同，大语言模型的推理成本正以每年 9 至 900 倍的速度下降。

2022–2024 年选定的比较基准的推理成本

资料来源：Epoch AI, 2025; Artificial Analysis, 2025 | 图表：2025 年人工智能指数报告

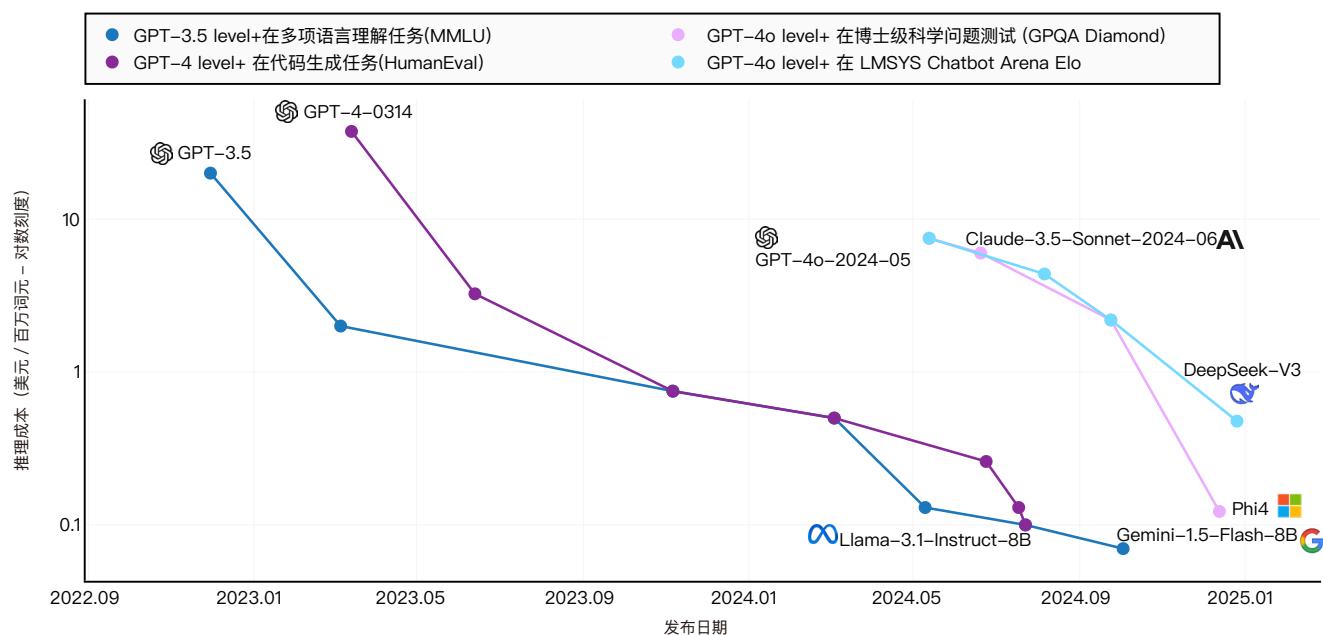


图 1.3.22

第一章：研究与开发

1.3 标志性人工智能模型

实现特定性能水平的推理成本已显著下降，但最先进模型仍比前文提及的其它模型更昂贵。图 1.3.23 展示了 OpenAI、Meta 和 Anthropic 等开发商的领先模型每百万词元的成本。

本。²⁶ 这些顶级模型的定价通常高于同公司的较小模型，反映了尖端性能所需的溢价。

选定的模型的输出价格（每百万词元）

资料来源：Artificial Analysis, 2025 | 图表：2025 年人工智能指数报告

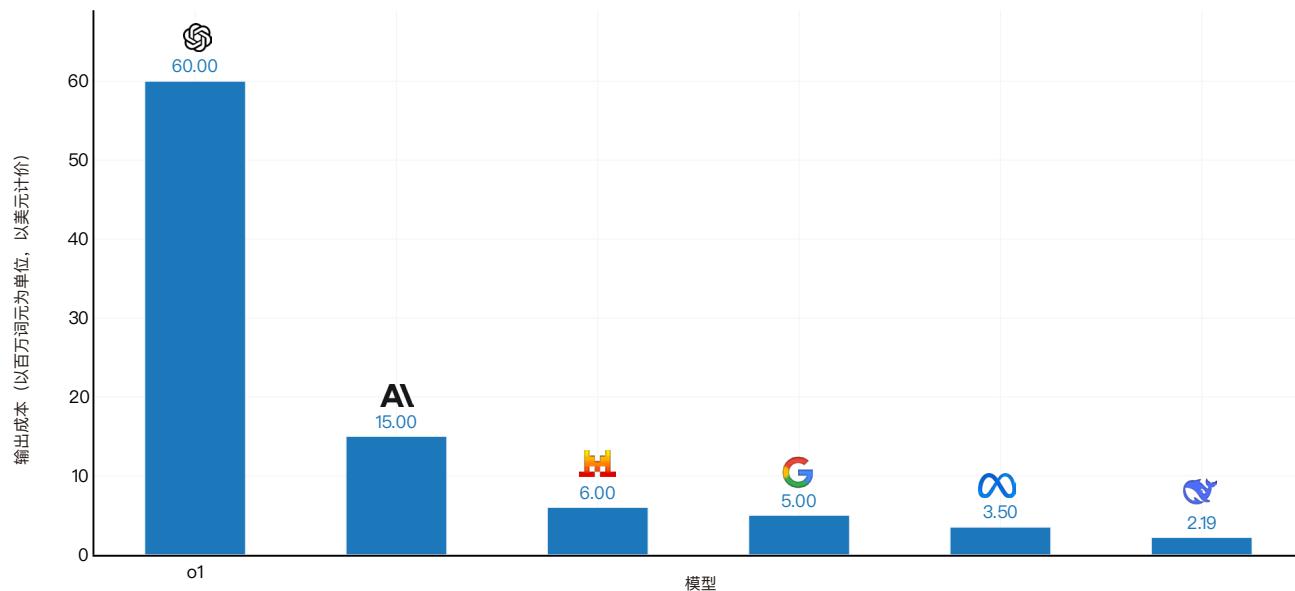


图 1.3.23

训练成本

围绕基础模型经常讨论的是其高昂的训练成本。虽然人工智能公司很少披露确切数字，但据广泛估计，成本高达数以百万美元计，而且还在持续上升。但普遍估计成本已达数百万美元计且持续攀升。例如，OpenAI CEO Sam Altman 透露 GPT-4 训练费用超 1 亿美元；2024 年 7 月 Anthropic CEO Dario Amodei 指出，已投入约 10 亿美元规模的训练成本。较新的 DeepSeek-V3 据报成本较低(约 600 万美元)，但总体

而言训练仍极其昂贵。²⁷

了解与训练人工智能模型相关的成本仍然很重要，但详细的成本信息仍然很少。去年，人工智能指数发布了对基础模型训练成本的初步估算，人工智能指数再次与 Epoch AI 合作，更新并完善了这些估算。为了计算尖端模型的成本，Epoch 团队根据论文、新闻稿和技术报告中的信息，分析了培训时间、硬件类型、数量和使用率等因素。²⁸

26. 该指数直观显示了截至 2025 年 2 月公开定价的部分先进机型。自发布以来，更新的机型可能已经发布，定价也可能发生变化。

27. 一些报道对 DeepSeek-V3 的既定成本提出了质疑，认为如果将员工工资、资本支出和研究费用考虑在内，实际开发成本要高得多。

28. 有关 Epoch 研究方法的详细报告见本文。

第一章：研究与开发

1.3 标志性人工智能模型

图 1.3.24 显示了根据云计算租赁价格估算的部分人工智能模型的相关训练成本。图 1.3.25 则呈现了人工智能指数掌握的所有模型训练成本估算。

人工智能指数的估算验证了人们的猜测，即近年来模型训练成本大幅增加。例如，2017 年提出的 Transformer 模型，该架构支撑了几乎所有现代大语言模型，训练成本约为 670 美元；2019 年发布的 RoBERTa Large，在 SQuAD 和 GLUE 等经典理解比较基准中达到当时最优水平，训练成本约 16 万美元；至 2023 年，OpenAI 的 GPT-4 训练成本估算已达 7900 万美元。

2024 年 Epoch 在少数可估算成本的模型中，发现 Llama 3.1-405B 的训练成本约 1.7 亿美元。随着人工智能领域的竞争日趋激烈，企业对其训练过程的披露越来越少，使得估算计算成本变得越来越困难。正如之前的人工智能指数报告所指出的，人工智能模型的训练成本与其计算需求之间存在直接的关联。如图 1.3.26 所示，计算需求更高的模型训练成本显著增加。

2019—2024 年选定的人工智能模型估计训练成本

资料来源：Epoch AI, 2024 | 图表：2025 年人工智能指数报告

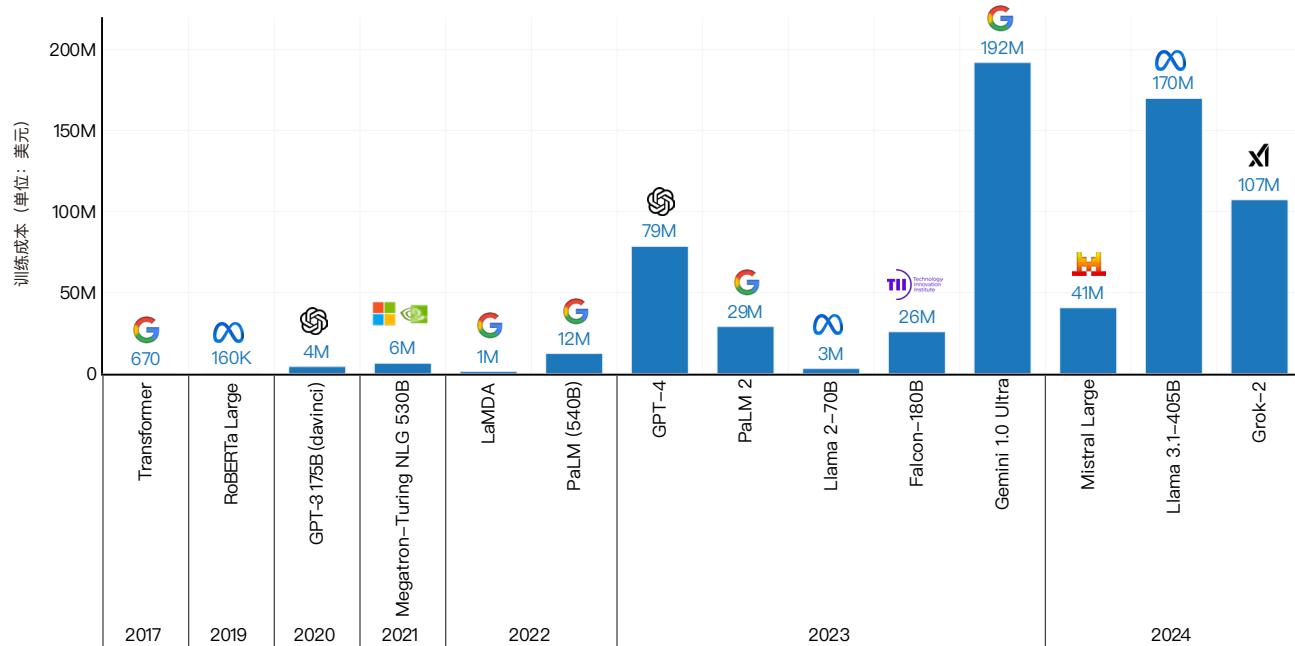


图 1.3.24 29

29、本节中报告的费用数字是经过通货膨胀调整的。

第一章：研究与开发

1.3 标志性人工智能模型

2016–2024 年选定的人工智能模型训练成本估算

资料来源：Epoch AI, 2024 | 图表：2025 年人工智能指数报告

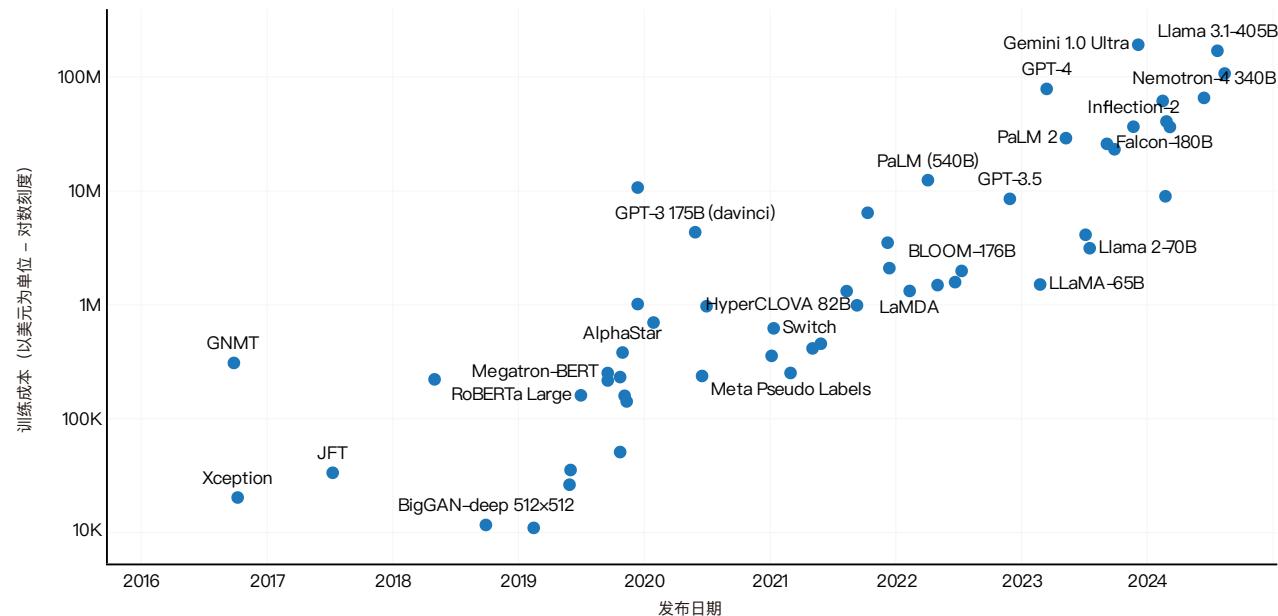


图 1.3.25

选定的人工智能模型训练成本估算及算力

资料来源：Epoch AI, 2024 | 图表：2025 年人工智能指数报告

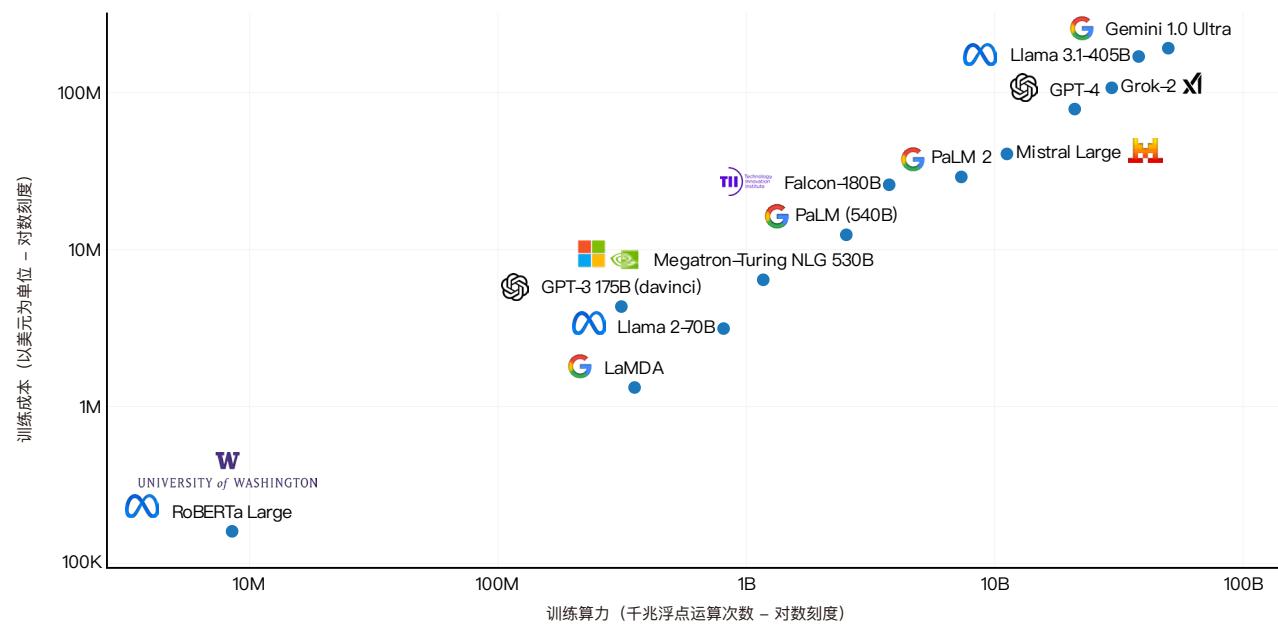


图 1.3.26

第一章：研究与开发

1.4 硬件

硬件进步对推动人工智能发展具有关键作用。虽然扩大模型规模和使用更大数据集训练带来了显著性能提升，但这些进展主要得益于硬件的改进——尤其是更强大、更高效的 GPU（图形处理器）的发展。GPU 加快了复杂计算的速度，使模型能够并行处理海量数据并显著缩短训练时间。本节利用 Epoch AI 的数据分析机器学习硬件的主要趋势及其对人工智能发展的影响。

虽然本节目目前强调的是计算性能(FLOP/s)，但网络带宽(GPU 的通信速度)同样至关重要。虽然有关数据中心网络带宽的数据有限，但未来版本的人工智能指数将致力于纳入这方面的信息。

1.4 硬件

概览

图 1.4.1 展示了不同精度类型的机器学习硬件的峰值计算性能，其中精度是指计算中用于表示数值（尤其是浮点数）的比特数。精度的选择取决于具体目标。例如，低精度硬件需要的比特数更少，内存带宽更低，是优化计算速度和能效的理想选择。这尤其有利于边缘 / 移动设备的人工智能模型或推理速度优先的场景。另一方面，精度更高的硬件可以保留更高的数值准确率，因此对于科学计算和对精度误差敏感的应用至关重要。在下图可视化的精度中，FP32 精度最高，TF32 为中高精度，Tensor-FP16/BF16 和 FP16 则是为速度与效率优化的低精度格式。

Epoch 估计以 16 位浮点运算为单位，机器学习硬件的运算能力在 2008–2024 年间的年增长率约为 43%，每 1.9 年翻一番。据 Epoch 分析，这一进步源于晶体管数量增加、半导体制造工艺改进以及人工智能专用硬件的发展。

2008–2024 年不同精度下机器学习硬件的峰值计算性能

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

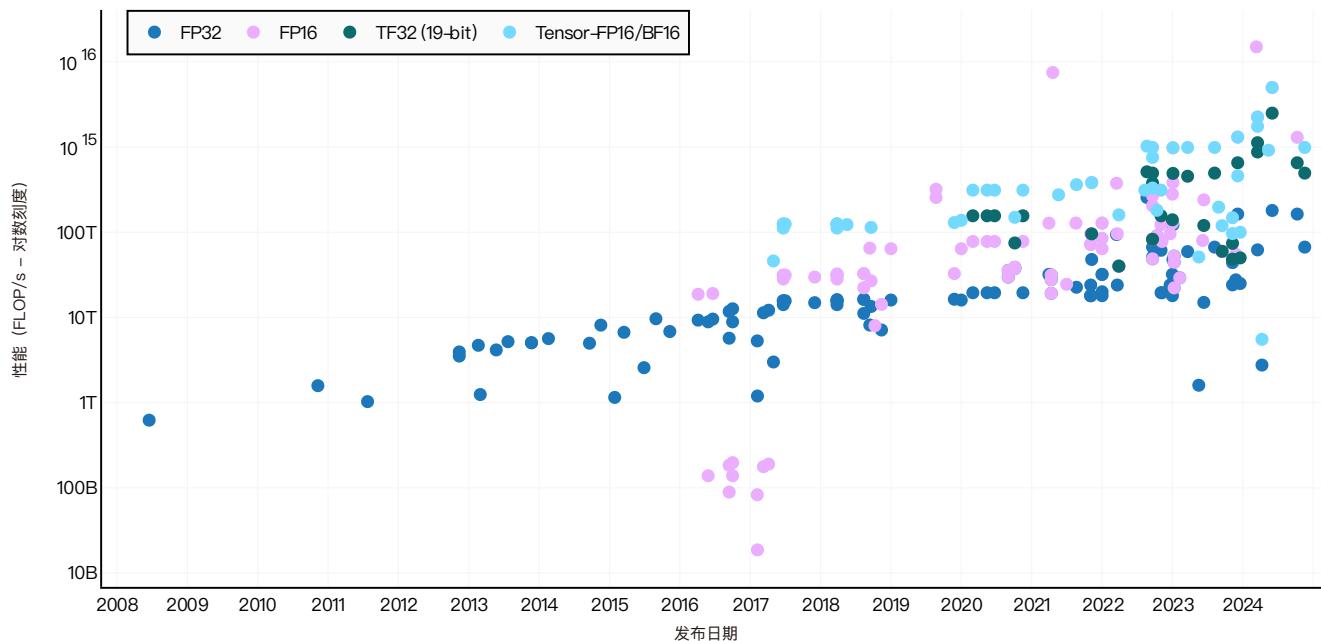


图 1.4.1

第一章：研究与开发

1.4 硬件

主流机器学习硬件的性价比持续提升。图 1.4.2 展示了部分英伟达数据中心 GPU 的性能，这些都是人工智能训练最常用的，单位为每秒 FLOP。图 1.4.3 显示了这些 GPU 的性价比，单位为每美元每秒 FLOP。例如，2022 年 3 月发布的 H100 图形处理器、达到每美元每秒 220 亿 FLOP，性价比约为 A100

(2020 年 6 月推出)的 1.7 倍，P100(2016 年 4 月发布)的 16.9 倍。据 Epoch 估算，性能水平固定的硬件成本每年下降 30%，这使得人工智能训练越来越经济实惠、可扩展，并有利于模型改进。

英伟达数据中心 GPU 在机器学习方面的领先性能

资料来源：Epoch AI, 2024 | 图表：2025 年人工智能指数报告

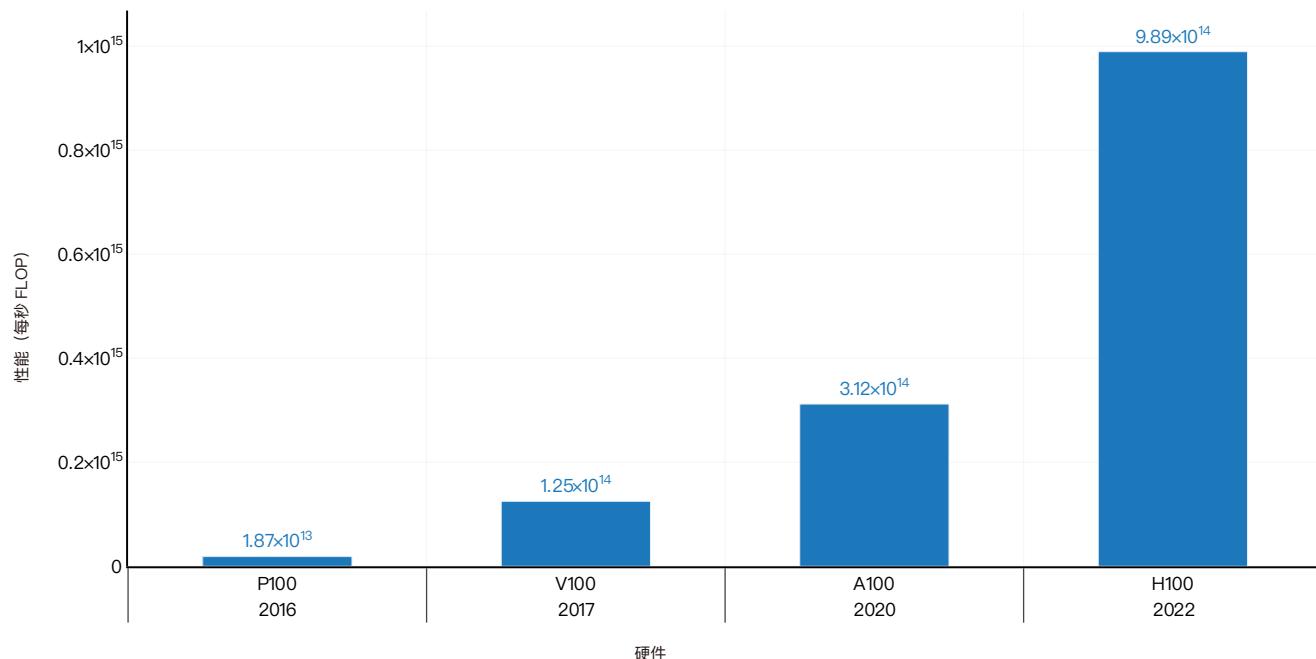


图 1.4.2

第一章：研究与开发

1.4 硬件

图 1.4.4 基于 Epoch AI 标志性机器学习模型数据集，统计了训练这些模型所使用的硬件。截至2024年，最常用的硬

件是 A100（6个模型使用），其次是 V100。使用 H100 训练的模型数量正在快速增长，2024年底已达 15 个。

机器学习领域领先的英伟达数据中心 GPU 的性价比

资料来源：Epoch AI, 2024 | 图表：2025 年人工智能指数报告

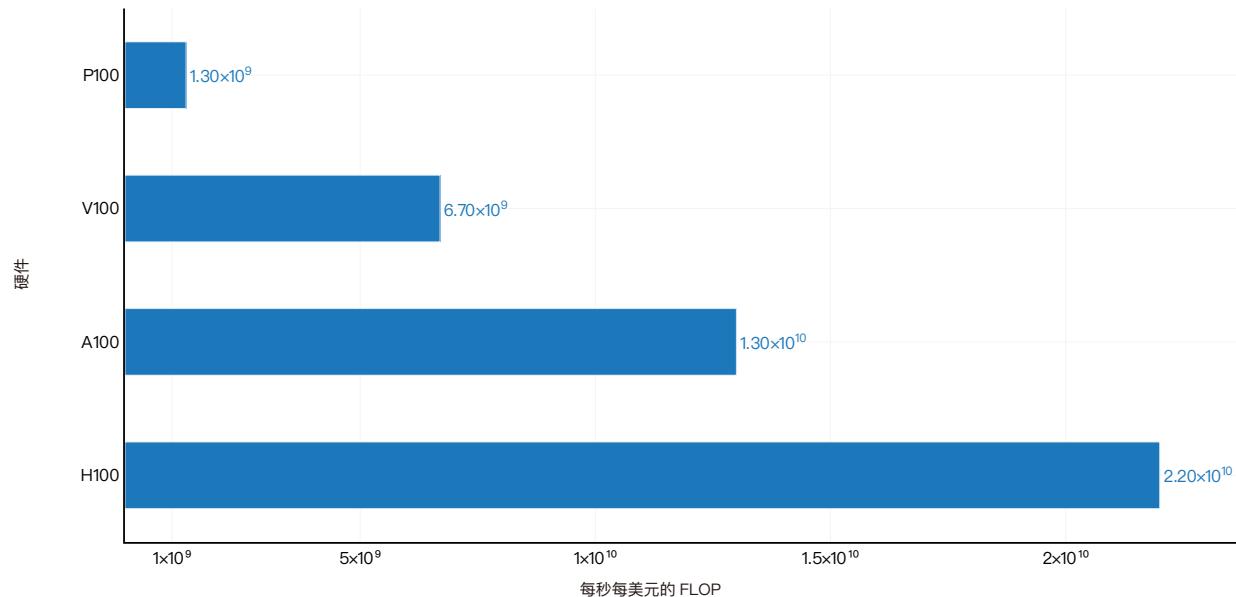


图 1.4.3

机器学习领域领先的英伟达数据中心 GPU 的性价比

资料来源：Epoch AI, 2024 | 图表：2025 年人工智能指数报告

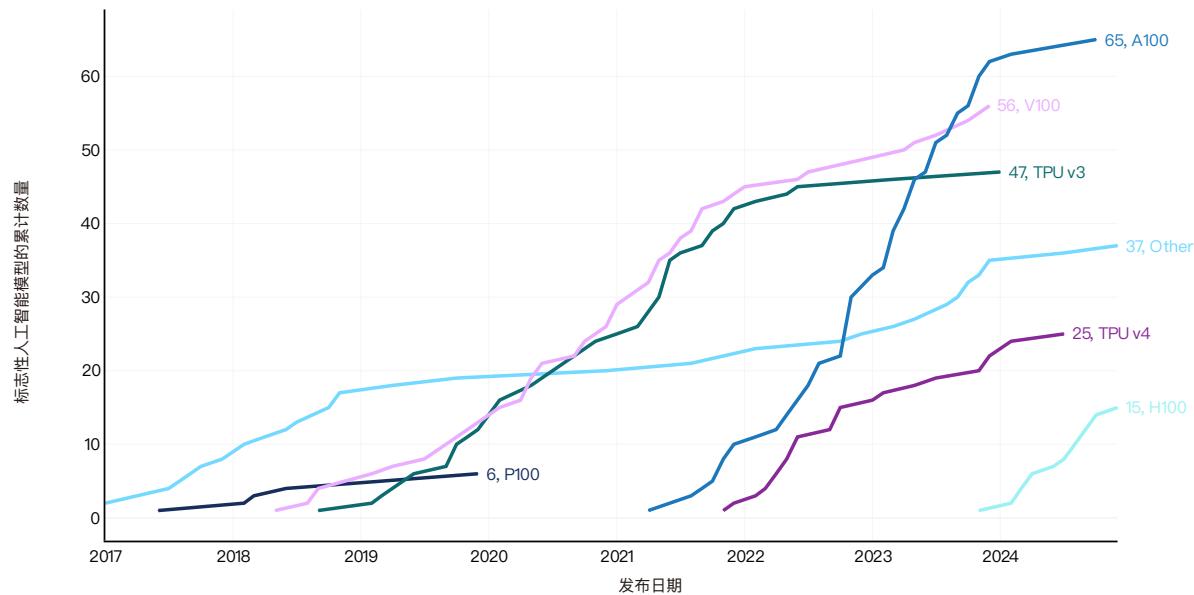


图 1.4.4

重点：

能源效率和环境影响

训练人工智能系统需要大量能源，因此机器学习硬件的能效是一个关键因素。Epoch AI 报告称，随着时间的推移，机器学习硬件的能效越来越高，提高约 40%。图 1.4.5 展示了 Tensor-FP16 精度硬件的能效(以每瓦 FLOP 计量)。

例如，2024 年 3 月发布的英伟达 B100 能效达 2.5 万亿 FLOP/ 瓦，而 2016 年 4 月发布的 P100 仅为 740 亿 FLOP/ 瓦，这意味着 B100 的能效是 P100 的 33.8 倍。

2016–2024 年领先机器学习硬件的能效

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

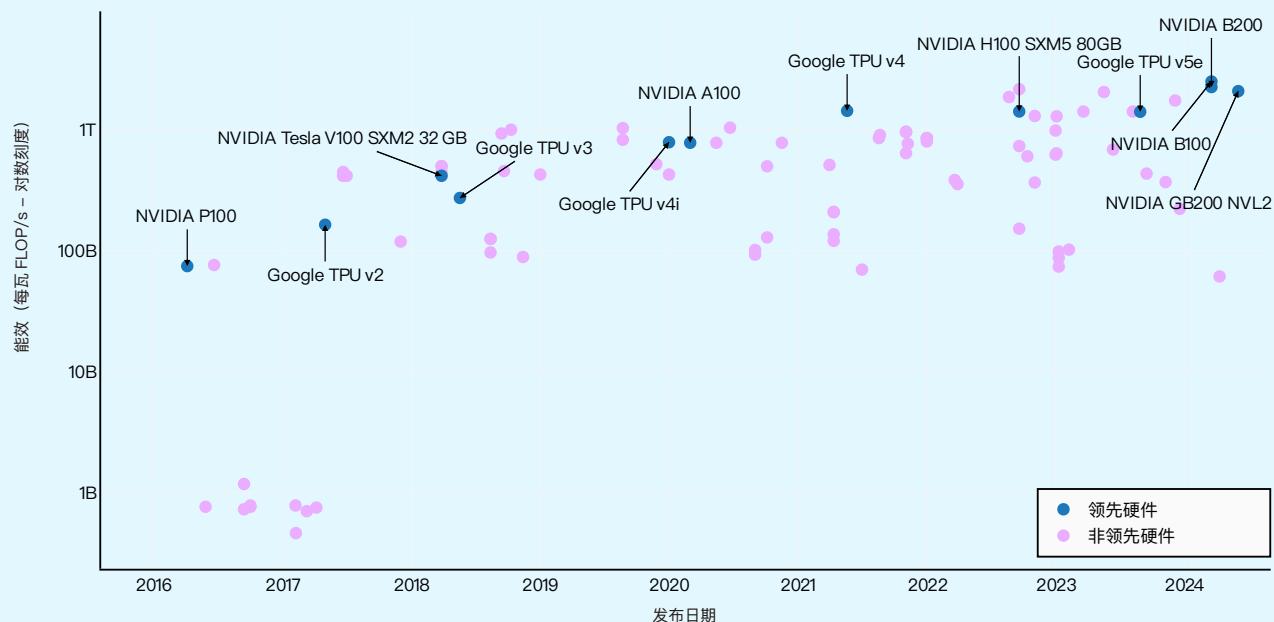


图 1.4.5

重点：

能源效率和环境影响

尽管人工智能硬件的能效有了很大提高，但训练人工智能系统所需的总功耗仍在快速上升。图 1.4.6 展示了训练各种最先进人工智能模型的总功耗（以瓦为单位）。例如，2017 年提出的原 Transformer 模型功耗约为 4,500 瓦，而谷歌早期旗舰大语言模型 PaLM 功耗达 260 万瓦，是 Transformer 的近 600 倍。2024 年夏季发布的 Llama 3.1-405B 功耗达 2,530 万瓦，较原 Transformer 增长超 5,000 倍。根据 Epoch AI，训练标志性人工智能模型所需

的功耗每年翻一番。人工智能模型能耗的持续增长反映了其训练过程中对越来越大数据集的依赖趋势。

不难理解，随着时间的推移，用于训练人工智能系统的总电量在增加，模型排放的碳量也在增加。决定人工智能系统碳排放量的因素很多，包括模型中的参数数量、数据中心用电效率 (PUE) 以及电网碳强度。³⁰

2011–2024 年训练前沿模型所需的总功率消耗

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告

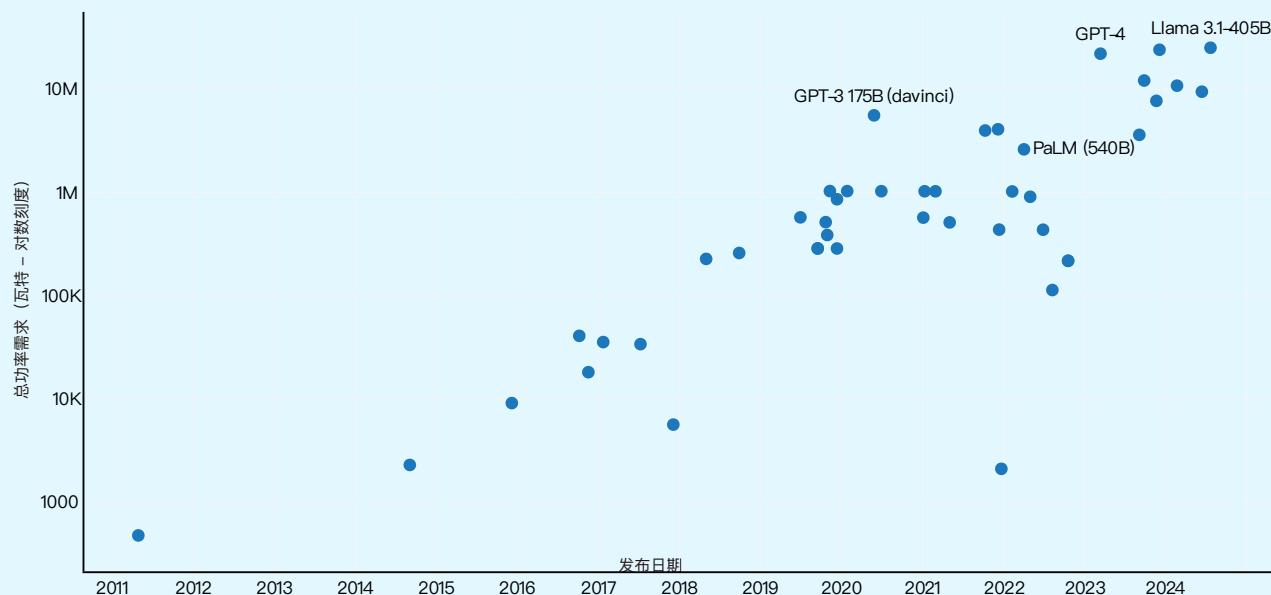


图 1.4.6

³⁰ 电力使用效率 (PUE) 是用于评估数据中心能效的一个指标。计算公式为数据中心总能耗 (含制冷) 与 IT 设备能耗之比，PUE 值越高，数据中心的效率越低。

重点：

能源效率和环境影响(续)

图 1.4.7 展示了按发布年份排序的部分人工智能模型的碳排放量。为了估算这些排放量，人工智能指数使用了模型开发者公布的碳数据，并辅以广泛使用的在线人工智能训练排放量计算器的测算结果。由于多数开发者未披露模型碳足迹，这一补充计算十分必要。该计算器根据训练所用硬件类型、总训练时间、云服务商和训练地区来估算排放量。³¹

随着时间的推移，训练标志性人工智能模型所产生的碳排放量稳步上升。虽然 AlexNet 的排放量可以忽略不计，但据报道，GPT-3(2020 年发布)在训练期间的碳排放量约为 588 吨，GPT-4(2023 年)的排放量为 5184 吨，Llama 3.1 405B(2024 年)的排放量为 8930 吨。2024 年发布的 DeepSeekV3，其性能与 OpenAI 的 o1 相当，估计其排放量与五年前发布的 GPT-3 相当。作为参照，美国人平均每人每年的碳排放量为 18.08 吨。

2012–2024 年特定人工智能模型及实际活动产生的碳排放量估算

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告



图 1.4.7

31、人工智能指数从各种在线来源获取了排放计算器的输入数据，如训练硬件和持续时间。为了验证计算器的准确率，我们将计算器的估计值与开发人员报告的实际排放量进行了比较，发现结果基本一致。完整的估算方法详见附录。

重点：
能源效率和环境影响(续)
选定的人工智能模型估算的碳排放量及参数数量

资料来源：Epoch AI, 2025 | 图表：2025 年人工智能指数报告



图 1.4.8

第一章：研究与开发

1.5 人工智能会议

人工智能会议是研究人员展示研究成果、与同行和合作者建立联系的重要平台。在过去二十年，这些会议的规模、数量和影响力都在不断扩大。本节将探讨主要人工智能会议的参会趋势。

1.5 人工智能会议

参会规模

图 1.5.1 展示了 2010 年以来部分人工智能会议的参会人数。2020 年，新冠疫情迫使会议在线举行，参会人数显著增加。随后，可能由于会议恢复到线下形式，参会人数有所下降，2022 年参会人数恢复到疫情前的水平。此后，参会人数稳步增长，2023 至 2024 年增幅达 21.7%。³² 自 2014 年起，年参会人数

增长超 6 万，既反映人工智能研究热度上升，也体现新会议涌现。神经信息处理系统大会 (NeurIPS) 仍是最受欢迎的人工智能会议，2024 年吸引近 2 万名参会者(图 1.5.2–1.5.3)。在主要的人工智能会议中，NeurIPS、CVPR、ICML、ICRA、ICLR、IROS 和 AAAI 去年的参会人数都有所增加。

2010–2024 年期间部分人工智能会议参会情况

资料来源：Artificial Analysis, 2025 | 图表：2025 年人工智能指数报告

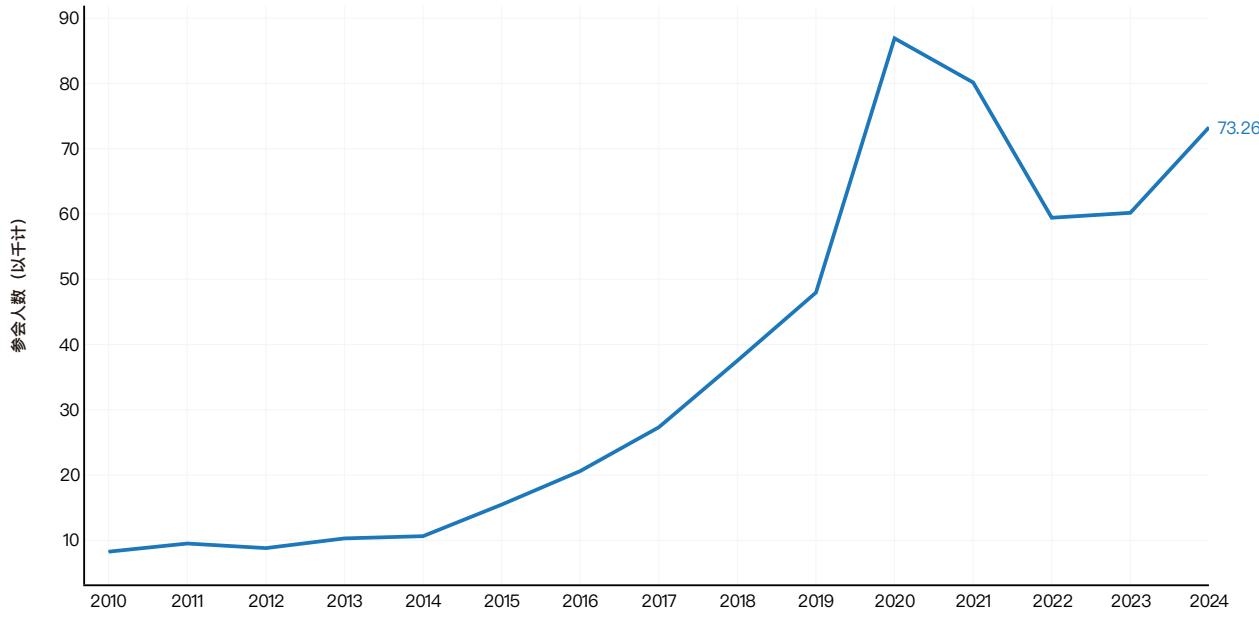


图 1.5.1

32、对于近年来许多会议采用虚拟或混合形式举办的情况，这些数据应谨慎解读。会议组织方指出，由于虚拟会议使得全球研究者的参与更为便利，准确统计线上会议的出席人数存在困难。AI Index 报告中的总出席人数涵盖了虚拟参会、混合参会和线下参会三种形式。本次统计覆盖的会议包括：AAAI、AAMAS、CVPR、EMNLP、FAccT、ICAPS、ICCV、ICLR、ICML、ICRA、IJCAI、IROS、KR、NeurIPS 和 UAI 等人工智能领域重要会议。

第一章：研究与开发

1.5 人工智能会议

2010—2024 年大型会议的参会人数

资料来源：AI Index, 2024 | 图表：2025 年人工智能指数报告

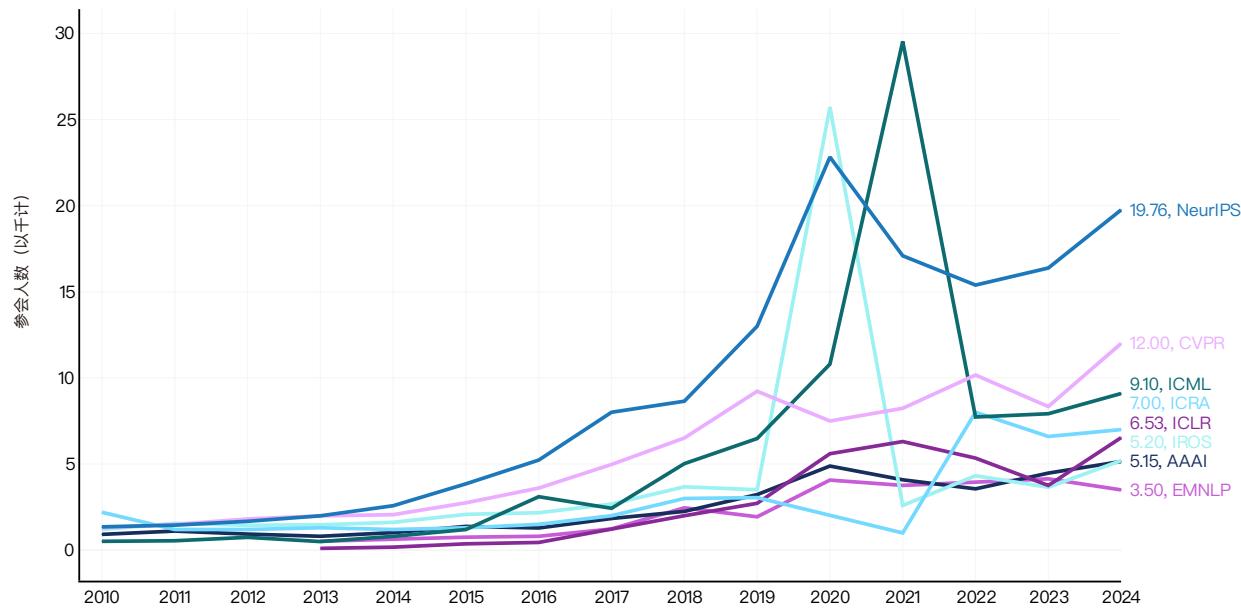


图 1.5.2³³

2010—2024 年小型会议的参会人数

资料来源：AI Index, 2024 | 图表：2025 年人工智能指数报告

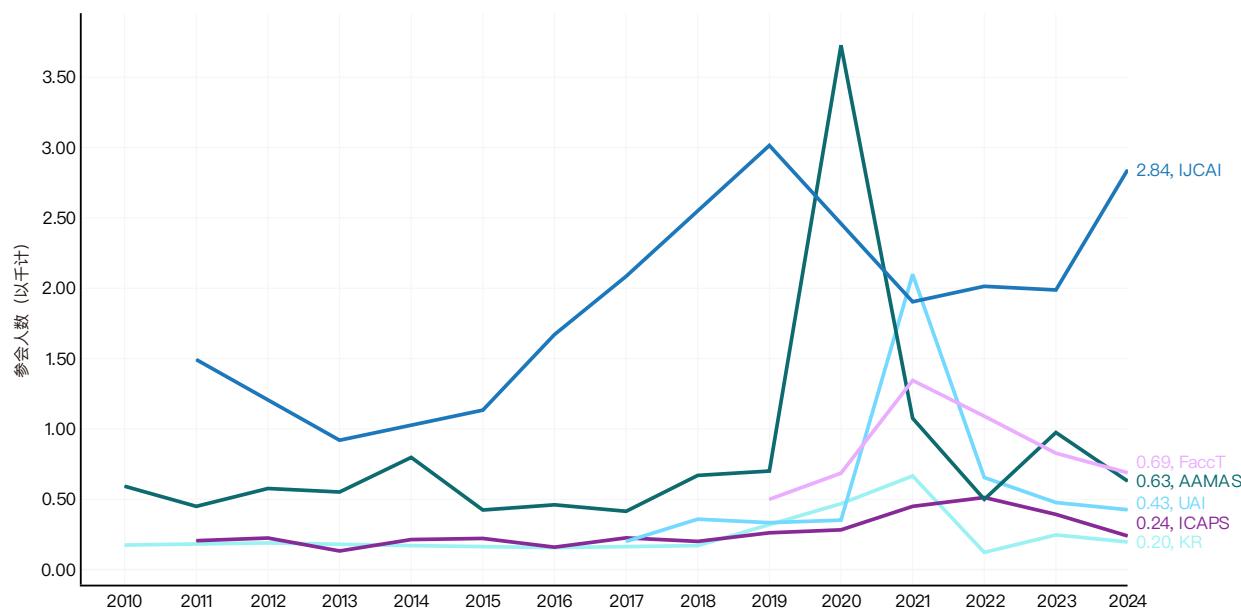


图 1.5.3

33、2021 年 ICML 参会人数大幅飙升，很可能是由于该年的会议是线上举行的。

第一章：研究与开发

1.6 开源人工智能软件

GitHub 是一个基于 Web 的平台，使个人和团队能够托管、审查和协作代码库。作为软件开发者广泛使用的工具，GitHub 提供代码管理、项目协作和开源软件支持功能。本节基于来自 GitHub 的数据，深入分析论文数据中未反映的开源人工智能软件开发的广泛趋势。³⁴

1.6 开源人工智能软件

开源人工智能软件项目

GitHub 项目由一系列文件组成，包括源代码、文档、配置文件和图像，这些文件共同构成了一个软件项目。图 1.6.1 显示了随着时间推移 GitHub 人工智能项目的总数的变化。³⁵ 自 2011 年以来，与人工智能相关的 GitHub 项目数量持续增长，从 2011 年的 1,549 个增至 2024 年的约 430 万个。值得注意的是，仅去年一年，GitHub 人工智能项目总数激增了 40.3%。

2011–2024 年 GitHub 上的人工智能项目数量

资料来源：GitHub, 2024 | 图表：2025 年人工智能指数报告

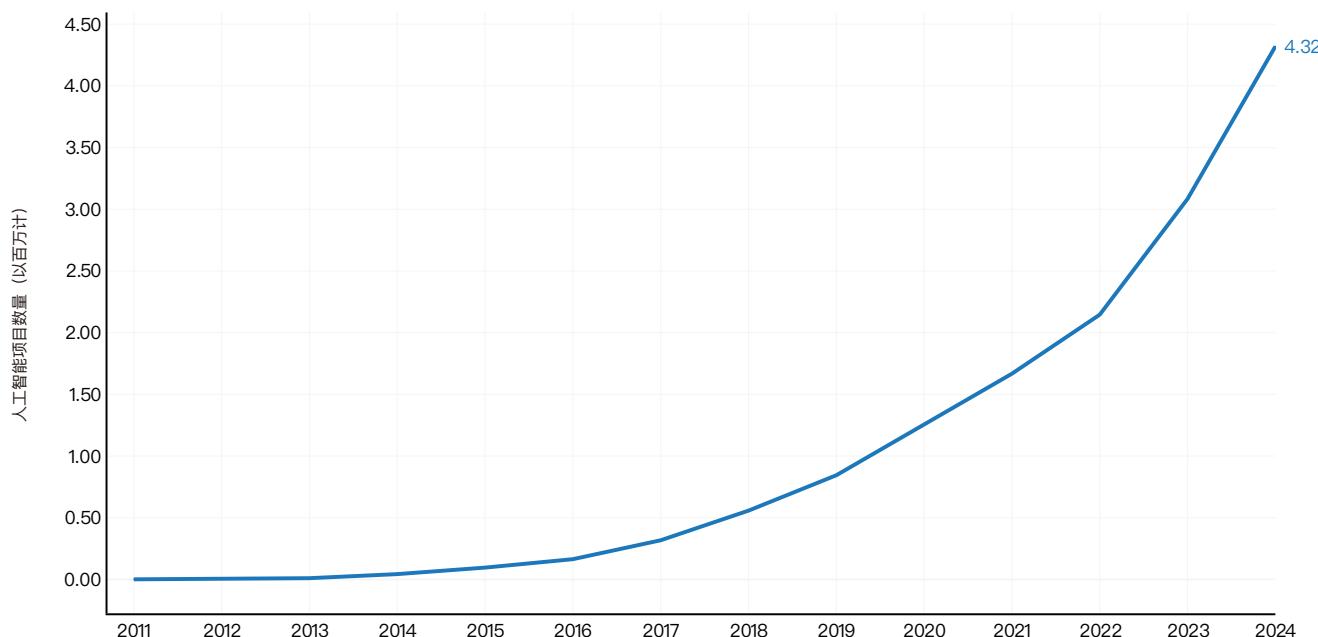


图 1.6.1

34、今年 GitHub 更新了其方法，以捕捉更广泛的人工智能相关主题，包括更多最新发展。因此，今年的人工智能指数中的数据可能与以往版本的数据不对齐。中国研究人员经常使用 GitHub 以外的其他网站共享代码，如 Gitee 和 GitCode，但本报告不包括这些网站的数据。完整的方法说明见附录。

35、GitHub 使用人工智能主题分类方法来识别与人工智能相关的知识库。有关该方法的详细信息，请参阅附录。

第一章：研究与开发

1.6 开源人工智能软件

图 1.6.2 展示了 2011 年以来 GitHub 人工智能项目的地理分布。截至 2024 年，美国贡献了 23.4% 的 GitHub 人工智能项目，占比最高；印度以 19.9% 位居第二，欧洲以 19.5% 紧

随其后。值得注意的是，自 2016 年起，美国开发者在 GitHub 开源人工智能项目中的占比持续下降，近年趋于稳定。

2011–2024 年按地理区域划分的 GitHub 人工智能项目占比

资料来源：GitHub, 2024 | 图表：2025 年人工智能指数报告

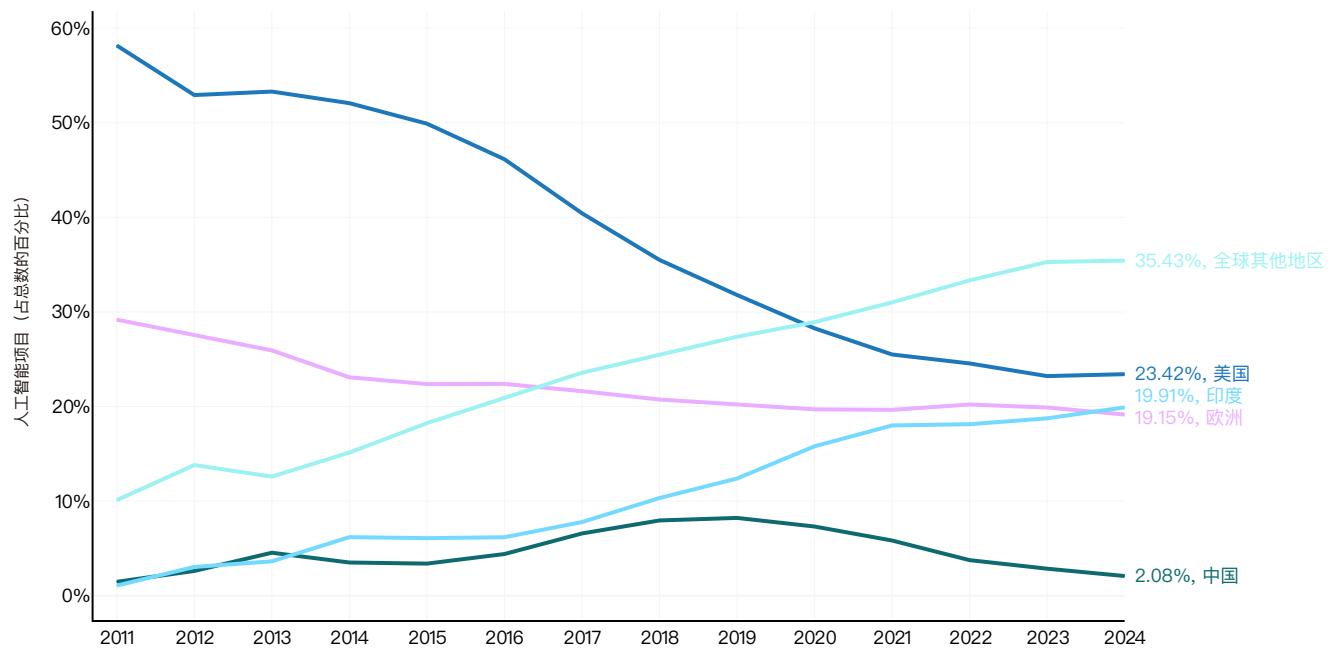


图 1.6.2

第一章：研究与开发

1.6 开源人工智能软件

星标

GitHub 用户可通过“加星 (starring)”功能表达对代码仓库的关注，类似于社交媒体点赞，代表对开源项目的支持。最受关注的仓库包括 TensorFlow、OpenCV、Keras 和 PyTorch 等库，这些项目不仅在人工智能领域，更在整个开发者社区中广受欢迎。TensorFlow、Keras 和 PyTorch 是构建和部署机器学习模型的常用库，而 OpenCV 则提供计算机视觉相关工具，

如目标检测和特征提取。

GitHub 上人工智能相关项目的星标总数持续增长，从 2023 年的 1,400 万增至 2024 年的 1,770 万（图 1.6.3）。³⁶ 此前 2022 至 2023 年间星标数已实现翻倍激增。

2011–2024 年人工智能项目在 GitHub 上的星标数量

资料来源：GitHub, 2024 | 图表：2025 年人工智能指数报告

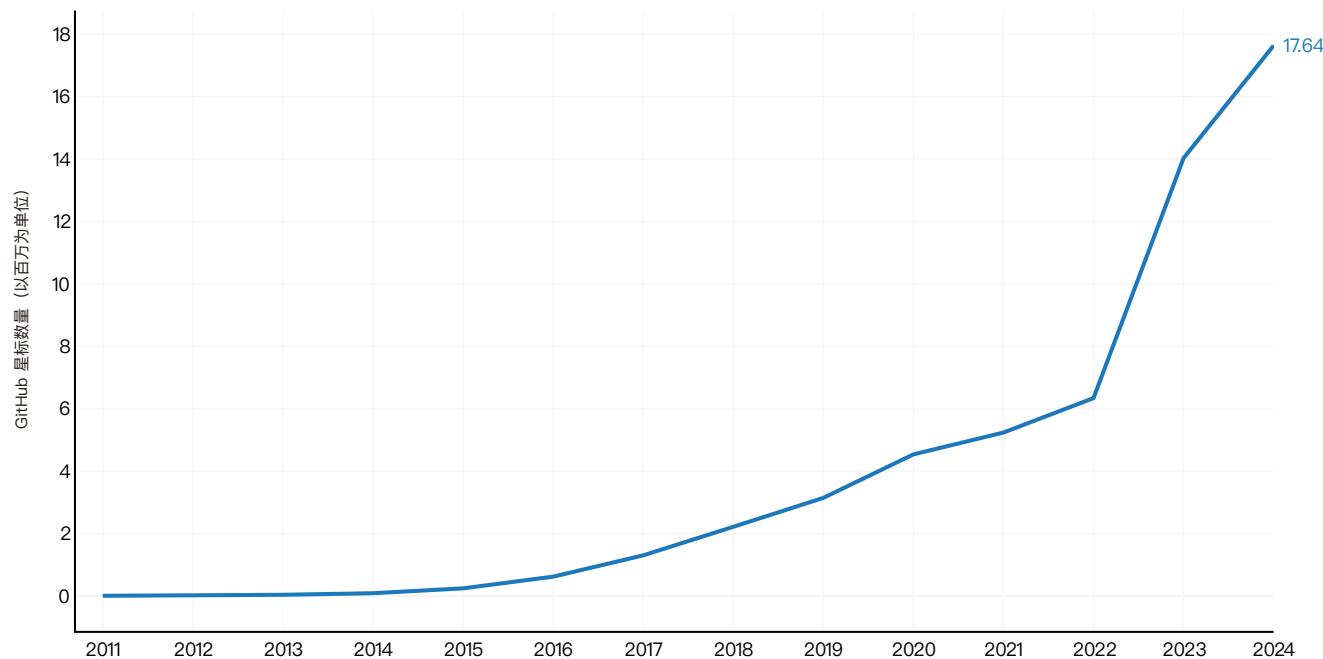


图 1.6.3

³⁶、图 1.6.3 显示年度新增星标数，而非历史累计总量。

第一章：研究与开发

1.6 开源人工智能软件

2024 年，美国在 GitHub 星标数统计中位居全球首位，总计达 2110 万（图 1.6.4）。所有被抽样调查的主要地理区域，包

括欧洲、中国和印度，其本国项目在 GitHub 上获得的星标数量均较上年同期有所增长。

2011—2024 年按地理区域划分的 GitHub 星标数量

资料来源：GitHub, 2024 | 图表：2025 年人工智能指数报告

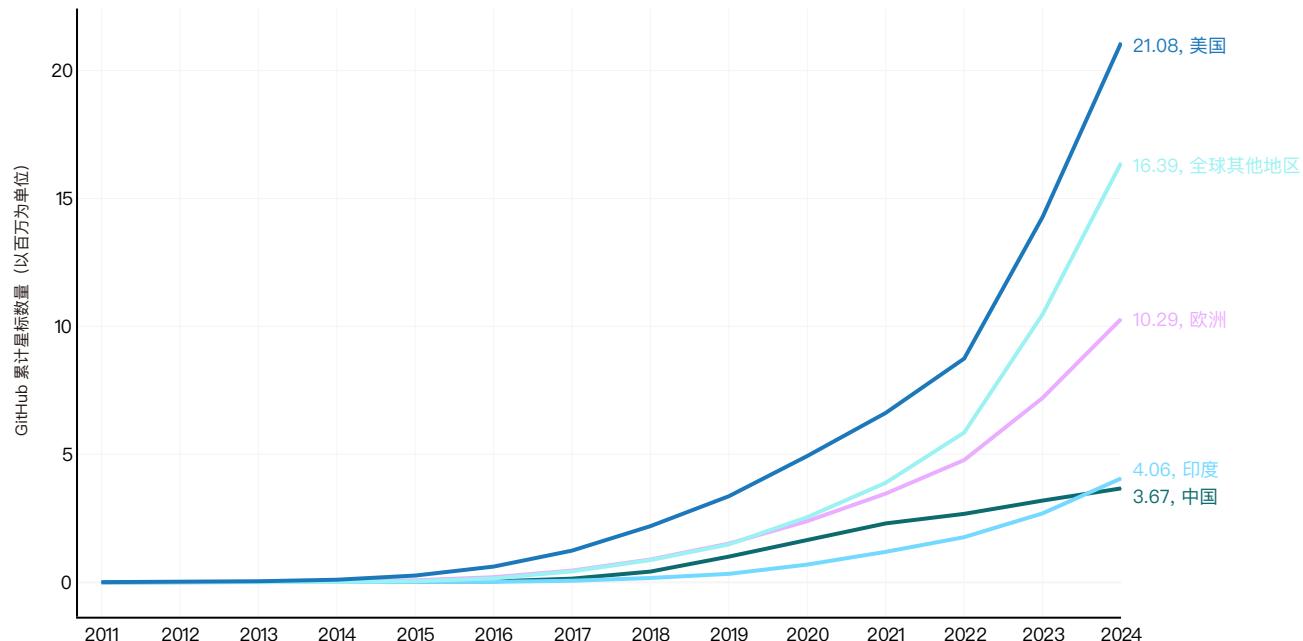


图 1.6.4

第二章： 技术性能



第二章：技术性能

概述	84
章节要点	85
2.1 2024 年人工智能技术发展概览	87
时间表：重要模型和数据集发布	87
人工智能性能状况	93
总体回顾	93
闭源权重模型与开源权重模型的对比	94
美国与中国技术性能对比	96
小型模型性能提升	98
前沿模型性能趋同	99
人工智能基准比较	100
2.2 语言	103
语言理解	104
MMLU：大规模多任务语言理解	104
生成任务	105
Chatbot Arena Leaderboard	105
Arena-Hard-Auto	107
WildBench	108
重点：o1, o3, 与推理时间计算	110
MixEval	112
RAG：检索增强生成	113
Berkeley Function Calling Leaderboard	113
MTEB：大规模文本嵌入式比较基准	115
重点：长上下文检索评估	117
2.3 图像与视频	119
理解能力	119
VCR: 视觉常识推理	119
MVBF	120
生成能力	122
Chatbot Arena: 视觉	123
重点：视频生成的崛起	124
2.4 语音	126
语音识别	126
LSR2: Lip Reading Sentences 2	126
2.5 编程	128
HumanEval	128
SWE-bench	129
BigCodeBench	130
Chatbot Arena: 编程能力测评	131
2.6 数学	132
GSM8K	132
MATH	133
Chatbot Arena: 数学能力测评	134
FrontierMath	134
重点：学习与定理证明	136
2.7 推理	137
通用推理	137
MMMU：面向专家级 AGI 的大规模多学科 多模态理解与推理基准	137
GPQA: A Graduate-Level Google-Proof Q&A Benchmark	138
ARC-AGI	139
Humanity's Last Exam(人类终极考试,HLE)	141
规划	143
PlanBench	143

第二章：技术性能（续）

2.8 AI 智能体	144
VisualAgentBench	144
RE-Bench	145
GAIA	147
2.9 机器人与自主运动	148
机器人	148
RLBench	148
重点：人形机器人	150
重点：DeepMind 的进展	151
重点：机器人基础模型	154
自动驾驶汽车	155
发展	155
技术创新与新比较基准	156
安全标准	157

获取公共数据

第二章： 技术性能

概述

本年度的人工指数报告技术性能章节全面概述了 2024 年人工智能领域的的主要进展。开篇从宏观层面总结了人工智能技术发展动态，涵盖重大人工智能技术发布、人工智能能力现状以及关键趋势——包括开源权重模型性能提升、前沿模型性能趋同，以及中国大语言模型的质量改进。随后，本章节详细分析了各类人工智能能力的当前发展水平，涉及语言理解与生成、检索增强生成、编程、数学、推理、计算机视觉、语音以及代理型人工智能。今年新增了对机器人与自动驾驶汽车性能趋势的扩展分析。

第二章： 技术性能

章节要点

- 人工智能达成新比较基准的速度比以往任何时候都快。**2023 年，研究人员推出了 MMMU、GPQA 和 SWE-bench 等多个具有挑战性的新比较基准，旨在测试日益强大的人工智能系统极限。到 2024 年，人工智能在这些比较基准上的表现取得显著突破：MMMU 和 GPQA 测试成绩分别提升 18.8 和 48.9 个百分点；更引人注目的是，在 SWE-bench 编程测试中，人工智能系统的解题能力从 2023 年仅能解决 4.4% 的问题，跃升至 2024 年的 71.7%。
- 开源模型迎头赶上。**根据去年发布的人工智能指数报告，领先的开源模型曾大幅落后于闭源模型。而到 2024 年，这一差距已基本消失。具体来看，2024 年 1 月初，在 Chatbot Arena Leaderboard 上，顶尖闭源模型的性能优势为 8.0%；而到 2025 年 2 月，这一差距已缩小至 1.7%。
- 中美人工智能模型能力差距缩小。**2023 年，美国顶尖人工智能模型性能曾大幅领先中国同类产品，但这一态势现已改变。数据显示：截至 2023 年底，在 MMLU、MMMU、MATH 和 HumanEval 等比较基准中，中美模型的性能差距分别为 17.5、13.5、24.3 和 31.6 个百分点；而到 2024 年末，这些差距已大幅收窄至 0.3、8.1、1.6 和 3.7 个百分点。
- 前沿的人工智能模型的性能趋于收敛。**根据去年的人工智能指数，Chatbot Arena Leaderboard 上排名第一与第十的模型间 Elo 分数差已从去年的 11.9% 收窄至 2025 年初的 5.4%。同样，排名前两位的模型之间的差距也从 2023 年的 4.9% 缩小到 2024 年的 0.7%。人工智能领域的竞争日趋激烈，如今有越来越多的开发者推出了高质量的模型。
- 新型推理范式，如测试时计算 (test-time compute)，显著提升模型性能。**2024 年，OpenAI 推出的 o1、o3 等模型采用迭代式输出推理架构。这种测试时计算方法极大地提高了模型的性能，o1 在国际数学奥林匹克资格考试中获得了 74.4% 的高分，GPT-4o 只有 9.3%。但该技术存在代价，o1 的运算成本激增至 GPT-4o 的 6 倍，推理速度则降低 30 倍。

第二章： 技术性能

章节要点（续）

6. 更具挑战性的比较基准不断被提出。MMLU、GSM8K 和 HumanEval 等传统人工智能比较基准已趋近饱和，加上 MMMU 和 GPQA 等更新的更具挑战性的比较基准表现持续提升，促使研究人员探索更多针对领先人工智能系统的评价方法。其中值得注意的有人类最后的考试(Humanity's Last Exam)，这是一项严格的学术测试，最先进的人工智能系统的得分率仅为 8.80%；前沿数学(Frontier Math)，这是一项复杂的数学比较基准，人工智能系统仅解决了 2% 的问题；“BigCodeBench”这是一项编码比较基准，人工智能系统的成功率仅为 35.5%，远低于人类 97% 的水平。

7. 高质量人工智能视频生成模型实现重大突破。2024 年，多款能够根据文本输入生成高清视频的先进人工智能模型相继问世，其中包括 OpenAI 的 SORA、Stable Video Diffusion 3D 和 4D、Meta 的 Movie Gen，以及谷歌 DeepMind 的 Veo 2。与 2023 年的视频生成模型相比，这些新一代模型在画质表现上取得显著提升。

8. 小型模型展现更强性能。2022 年，在 MMLU 比较基准中达到 60% 以上分数的最小模型是拥有 5400 亿参数的 PaLM；而到 2024 年，微软推出的 Phi-3-mini 仅用 38 亿参数就实现了相同水平——相当于两年间参数规模缩减了 142 倍。

9. 复杂推理仍是人工智能面对的难题。尽管通过思维链(Chain-of-Thought)等推理机制的引入显著提升了大语言模型的性能，这些系统仍无法可靠解决本可通过逻辑推理获得确定性答案的问题——包括数学运算和任务规划等，尤其当问题规模超出其训练范围时。这一缺陷严重影响了人工智能系统的可信度，使其难以胜任高风险场景的应用需求。

10. 人工智能智能体展现初步潜力。2024 年推出的 RE-Bench 基准为评估 AI 智能体的复杂任务能力建立了严格标准。在短时任务(2 小时时限)中，顶级人工智能系统的得分可达人类专家的 4 倍；但随着时间延长至 32 小时，人类表现反超人工智能系统，得分达到 2:1 的优势。AI 智能体已在特定领域，如编写 Triton Kernels，达到人类专业水平，且能更快、更低成本地产出结果。

第二章：技术性能

2.1 2024年人工智能技术发展概览

本章节首先对 2024 年发布的重要模型进行了高度概述，并回顾了当前人工智能技术性能的现状。

2.1 2024年人工智能技术发展概览

时间表：重要模型和数据集发布

根据人工智能指数指导委员会的评选，以下是 2024 年最标志性的模型与数据集发布

日期	名称	类别	创建者	意义	图片
2024年1月19日	Stable LM 2	大语言模型	Stability AI	Stability AI 的最新语言模型基于 Stable LM 改进，性能显著提升。该模型仅16亿参数，专为笔记本电脑和智能手机等便携设备高效运行而设计。	 图 2.1.1 资料来源: Wikipedia, 2025
2024年2月8日	Aya 数据集	数据集	Cohere for AI、北京智源研究院、Cohere、宾汉姆顿大学	作为 Cohere 的 Aya 计划的一部分，发布了一个包含 114 种语言 5.13 亿条提示完整比对的数据集。该论文及配套数据集标志着多语言指令微调领域的重大突破。	 图 2.1.2 资料来源: Cohere, 2025
2024年2月15日	Gemini 1.5 Pro	大语言模型	谷歌 DeepMind	Gemini 模型凭借 100 万词元的上下文窗口刷新行业基准，远超 GPT-4 Turbo 的 12.8 万词元限制。	 图 2.1.3 资料来源: 谷歌, 2024
2024年2月20日	SDXL -Lightning	文生图	字节跳动	由 TikTok 开发方字节跳动推出，是当时最快的文生图系统之一，可在 1 秒内生成高质量合成图像。其速度通过渐进对抗蒸馏技术实现，而非传统的基于扩散的方法。	 图 2.1.4 资料来源: Hugging Face, 2025
2024年3月4日	Claude 3	大语言模型	Anthropic	Anthropic 最新的 大语言模型在几乎所有行业的基准比较中都优于 GPT-4 和 Gemini，显著减少错误拒绝率并提高准确性。	 图 2.1.5 资料来源: Anthropic, 2025

第二章：技术性能

2.1 2024年人工智能技术发展概览

2024年3月17日	Inflection-2.5	大语言模型	Inflection AI	Inflection 旗舰产品 "PI" 搭载该模型，仅用 GPT-4 40% 的计算资源即实现同等性能。发布两周后，微软以 6.5 亿美元收购 Inflection。	 图 2.1.6 资料来源: Inflection, 2025
2024年3月19日	Moirai与LOTSA	模型/数据集	Salesforce	Salesforce 发布通用预测基础模型 Moirai，及跨 9 大领域、包含 270 亿观测值的时序数据集 LOTSA。	 图 2.1.7 资料来源: Salesforce, 2025
2024年3月27日	DBRX	大语言模型	Databricks	Databricks 开源的专家混合模型 (MoE)，性能超越 Mixtral 和 Grok 等同类小型 MoE 模型。该仅含解码器的 Transformer 模型拥有 1320 亿参数（每输入激活 368 亿），训练数据达 12 万亿词元。	 图 2.1.8 资料来源: Databricks, 2025
2024年4月2日	Stable Audio 2	文生曲/曲生曲	Stability AI	最新版本的 Stable Audio 是 Stability 的 AI 歌曲生成器，新增支持音频到音频功能。用户可以上传歌曲，并使用自然语言提示进行操作，实现定制歌曲。	 图 2.1.9 资料来源: Stability AI, 2025
2024年4月17日	Llama 3	大语言模型	Meta	Llama 3 系列首发 80 亿及 700 亿参数文本模型，成为同规模性能最优模型之一。	 图 2.1.10 资料来源: Meta, 2025
2024年5月13日	GPT-4o	多模态	OpenAI	GPT-4o 是一种新型多模态模型，支持文本、音频、图像和视频任意组合的输入与输出，它对音频的响应时间短至 320 毫秒，与人类的反应时间相当。	 图 2.1.11 资料来源: OpenAI, 2024

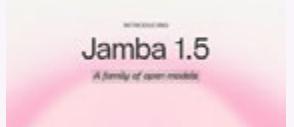
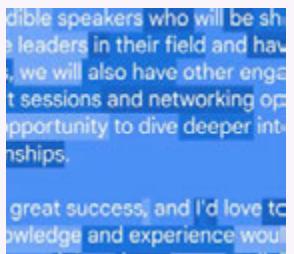
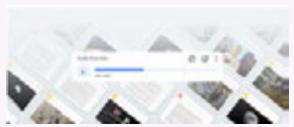
第二章：技术性能

2.1 2024 年人工智能技术发展概览

2024年6月7日	Qwen2	大语言模型	阿里巴巴	阿里巴巴开发的 Qwen2 系列包含基础模型和指令微调模型，在多项比较基准中性能媲美 Llama 3-70B 和 Mixtral-8x22B 等竞品。	 图 2.1.12 资料来源: Qwen, 2024
2024年6月17日	Runway Gen-3	文生视频/图生视频	跑道Runway	Runway 升级版视频生成模型为行业树立新标杆，特别擅长生成具有生动表情的逼真人像。	 图 2.1.13 资料来源: Runway, 2024
2024年7月23日	Llama 3.1405B	大语言模型	梅塔Meta	Meta发布迄今最大模型 Llama3.1 系列最终版，4050亿参数，成为当时公开可用的最强基础模型，性能比肩众多闭源权重模型。	 图 2.1.14 资料来源: Meta, 2024
2024年8月12日	Falcon Mamba	大语言模型	阿布扎比技术创新研究所	基于 Mamba 状态空间语言模型 (State Space 语言模型, SSLM) 架构开发的 70 亿参数 Falcon 模型，作为少数由政府机构研发的人工智能模型之一，通过动态参数调整机制和输入信息过滤功能，在运算效率上显著超越传统基于 Transformer 架构的同类模型。	 图 2.1.15 资料来源: Hugging Face, 2025
2024年8月13日	Grok-2	文生文/文生图	xAI	Grok 由 xAI 开发，一个高级文本和图像生成模型，在图像创建、高级推理和问题解决方面表现突出。它的发布引人注目，尽管 xAI 在2023年3月才成立，但它的技术性能很快就能与领先机型相媲美。	 图 2.1.16 资料来源: xAI, 2025

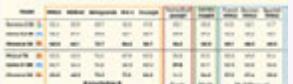
第二章：技术性能

2.1 2024 年人工智能技术发展概览

2024年8月15日	Imagen 3	文生图	谷歌 Labs	谷歌新一代人工智能图像生成器在 GenAI-Bench 图像比较基准中创下最高 Elo 评分，为人工智能生成视觉效果设定了新的质量标准。	
2024年8月22日	Jamba 1.5	大语言模型	AI21 Labs	首个融合了状态空间与 Transformer 架构的大语言模型，为基于文本的应用提供高质量的结果。这种混合方法在文本应用中实现高速响应与高质量输出的平衡。	
2024年8月29日	SynthID v2	工具	谷歌	SynthID v2 是谷歌水印和识别软件 SynthID 的升级版本。新增支持人工智能生成的图像、视频、音频和文本内容，并提供增强的跟踪和验证功能。	
2024年9月11日	NotebookLM 播客工具	文生播客	谷歌 Labs	继 Synthpod 之后，第二款端到端人工智能博客生成器问世并迅速走红。该工具因其便捷性，深受利用 NotebookLM 进行学习的学生群体以及使用人工智能生成的摘要进行工作听读的科技从业者欢迎。	
2024年9月12日	o1-preview	语言/数学 /生物	OpenAI	OpenAI 的“o”系列首款模型，专为高级推理和处理复杂任务而设计。在数学、科学和编程等复杂任务上的推理能力显著超越 GPT。	
2024年9月17日	NVLM (D,H,X)	视觉-语言	英伟达	英伟达发布了三个用于视觉语言任务的开放存取模型，在 OCRBench（光学字符识别）和 VQAv2（自然语言理解）上取得了最高分。	

第二章：技术性能

2.1 2024 年人工智能技术发展概览

2024年9月19日	Qwen2.5	大语言模型	阿里巴巴	Qwen2.5 是中国电子商务巨头阿里巴巴推出的最新系列基础模型，包含高效小型模型及专为编程和数学优化的专用模型。	 图 2.1.23 资料来源: Qwen, 2025
2024年10月16日	Minstral	大语言模型	Mistral	Minstral 是包含 30 亿和 80 亿参数两款紧凑模型，在所有主流行业比较基准中超越同规模的 Gemma 和 Llama 模型。	 图 2.1.24 资料来源: Mistral, 2025
2024年10月22日	Anthropic 计算机控制	代理能力	Anthropic	Anthropic Computer Use 是 Claude 3.5 Sonnet 用户的一项突破性计算机控制功能，允许 Claude 在用户的计算机上实时移动光标、输入文本和自主完成任务。	 图 2.1.25 资料来源: Anthropic, 2025
2024年10月28日	苹果智能系统	iPhone 功能	苹果	苹果整合 Image Playground(图像生成)、Genmoji(表情定制)、Siri 与 ChatGPT 联动等 AI 功能套件。	 图 2.1.26 资料来源: 苹果, 2025
2024年12月3日	Nova Pro	多模态	亚马逊	Nova Pro 是亚马逊网络服务 Nova 系列最强模型，擅长处理视觉与文本信息，在金融文档分析领域表现尤为突出。	 图 2.1.27 资料来源: Amazon, 2025
2024年12月11日	Gemini 2	大语言模型	谷歌 DeepMind	Gemini 升级版，新增计算机控制功能及图像 / 音频生成能力，速度较 1.5 Pro 提升 2 倍，编程和图像分析性能显著增强。	 图 2.1.28 资料来源: 谷歌, 2025

第二章：技术性能

2.1 2024 年人工智能技术发展概览

2024年12月12日	Sora	文生视频	OpenAI	OpenAI 备受期待的视频生成模型，可为 ChatGPT Pro 用户生成 1080p/20 秒视频（Plus 用户 720p/5 秒）。自 2024 年初技术圈已流传演示版本，官方为提升模型安全性而延迟发布。	 图 2.1.29 资料来源: OpenAI, 2025
2024年12月13日	Global MMLU	数据集	Cohere	一个多语言评估集，包含 42 种语言的专业翻译 MMLU 问题，旨在作为提供更全球化的人工智能比较基准。它评估 AI 在多种语言中的表现，同时解决了原始 MMLU 数据集中存在的西方偏见，据估计，该数据集中 28% 的问题依赖于西方文化知识。	 图 2.1.30 资料来源: Singh 等, 2025
2024年12月20日	o3 (beta)	多模态	OpenAI	OpenAI 最新前沿模型，面向人工智能研究人员安全测试发布，在 SWE 编程、竞赛数学、博士级科学和研究数学等比较基准中超越所有前代模型，并以 87.5% 成绩创下 ARC-AGI 基准新纪录。	 图 2.1.31 资料来源: VentureBeat, 2025
2024年12月27日	DeepSeek-V3	大语言模型	DeepSeek	DeepSeek V3，该开源模型以显著少于顶尖模型的算力资源开发，在 MMLU 和 GPQA 等比较基准中性能超越领先模型。	 图 2.1.32 资料来源: Dirox, 2025

第二章：技术性能

2.1 2024 年人工智能技术发展概览

人工智能性能状况

本章节节人工智能指数呈现 2024 年人工智能领域的主要发展趋势和全方位高层视角览。

总体回顾

去年人工智能指数报告指出，人工智能已在多数任务中超越人类水平，仅剩竞赛级数学和视觉常识推理等少数例外。过去一年间，人工智能系统持续改进，在多个原属挑战性的比较基准中已实现对人类表现的超越。

选定的人工智能指数技术性能比较基准与人类表现对比

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

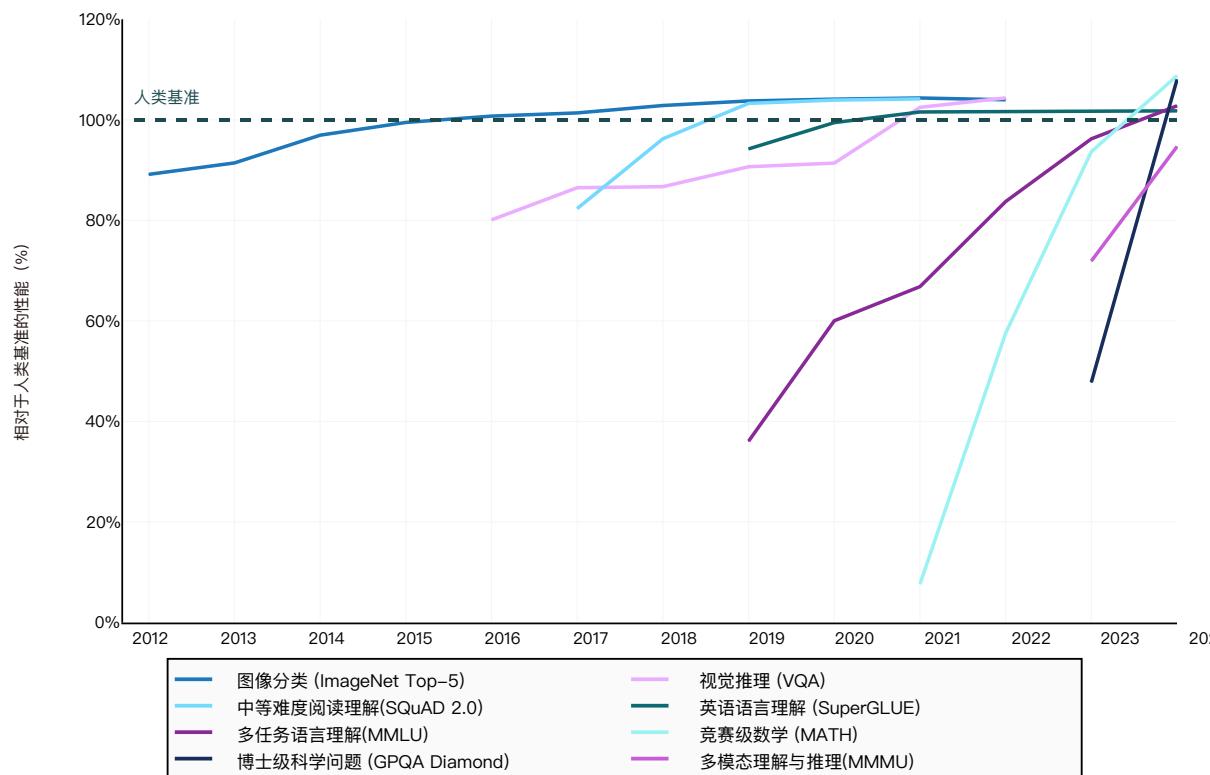


图 2.1.33²

1. 人工智能比较基准是一种用于评估人工智能系统在特定任务中性能的标准化测试。例如 ImageNet 作为经典基准，包含大量标注图像，人工智能系统的任务就是对这些图像进行准确分类。追踪比较基准进展是人工智能领域衡量系统发展的标准方法。
2. 在图 2.1.33 中，这些值经过缩放，以建立一个比较不同基准的标准指标体系。缩放函数经过校准，以将每年最佳模型的性能作为特定任务的人类基准的百分比来衡量。例如，值为 105% 表示该模型的性能比人类基准高 5%。

第二章：技术性能

2.1 2024 年人工智能技术发展概览

截至 2024 年，人类能力仍领先人工智能的任务领域已所剩无几。即便在这些领域，人工智能与人类之间的性能差距也在快速缩小。例如在竞赛级数学基准 MATH 上，最先进的人工智能系统目前领先人类表现 7.9 个百分点（2024 年初差距仅为 0.3 分）。³ 同样，针对复杂跨学科专家级问题的 MMMU 比较基准中，2024 年最佳模型 o3 得分 78.2%，仅比人类基准 82.6% 低 4.4 分。而 2023 年末谷歌 Gemini 在该测试中仅获 59.4%，充分展示了 AI 在复杂认知任务上的快速进步。

闭源权重模型与开源权重模型的对比

人工智能模型可以以不同的开放程度发布。如谷歌的 Med-Gemini 等模型完全闭源，仅限开发者使用；OpenAI 的 GPT-4o 和 Anthropic 的 Claude 3.5 等通过 API 提供有限公共访问，但未公开权重，因而无法独立修改或全面审查。相比之下，Meta 的 Llama 3.3 和 Stable Video 4D 等模型完全公开权重，允许任何人自由修改和使用。⁴

关于模型开放性的争论呈现两极分化。开源权重支持者强调其打破市场垄断、促进创新、提升安全性与透明度等优势。例如 Meta 的 Llama 模型已衍生出 Meditron 医疗工具、军事应用及全球众多开源项目。反对者则警告开源权重可能助长虚假信息传播和生物武器研发等安全风险，因此需要采取更加谨慎和可控的方法。

去年的人工智能指数报告指出闭源与开源大语言模型存在显著性能差距。图 2.1.34 展示了顶尖闭源权重和开源权重大语言模型在 Chatbot Arena Leaderboard 上的性能趋势，该平台是用于大语言模型性能比较基准的公共平台。2024 年 1 月顶尖闭源权重模型领先开源权重模型 8.0%，至 2025 年 2 月该差距已缩小至 1.7%。

这一趋势在其他问答比较基准中同样明显。2023 年闭源权重模型在 MMLU、HumanEval、MMMU 和 MATH 等主要基准上全面领先，持续优于开源权重模型，但到 2024 年差距大幅收窄（图 2.1.35）。例如，2023 年底，闭源权重模型在 MMLU 上领先开源权重模型 5.9 个百分点，但到 2024 年底，这一差距缩小到只有 0.1 个百分点。这一快速提升主要得益于 Meta 在夏季发布的 Llama 3.1，以及随后推出的其他高性能开源权重模型，如 DeepSeek 的 V3。

3、本图中的比较基准数据以及本章节其他章节中的比较基准数据是在 2025 年 1 月初收集的。自人工智能指数发布以来，个别比较基准得分可能有所提高。

4、在软件界，“开放源代码”指的是根据许可证发布的软件许可证授予用户自由使用、研究、修改和分发软件及其源代码的权利。但开源权重模型未必完全开源，因其底层代码或训练数据通常未公开。

第二章：技术性能

2.1 2024 年人工智能技术发展概览

LMSYS Chatbot Arena 中顶级闭源模型与开源模型的性能对比

资料来源：LMSYS, 2025 | 图表：2025 年人工智能指数报告

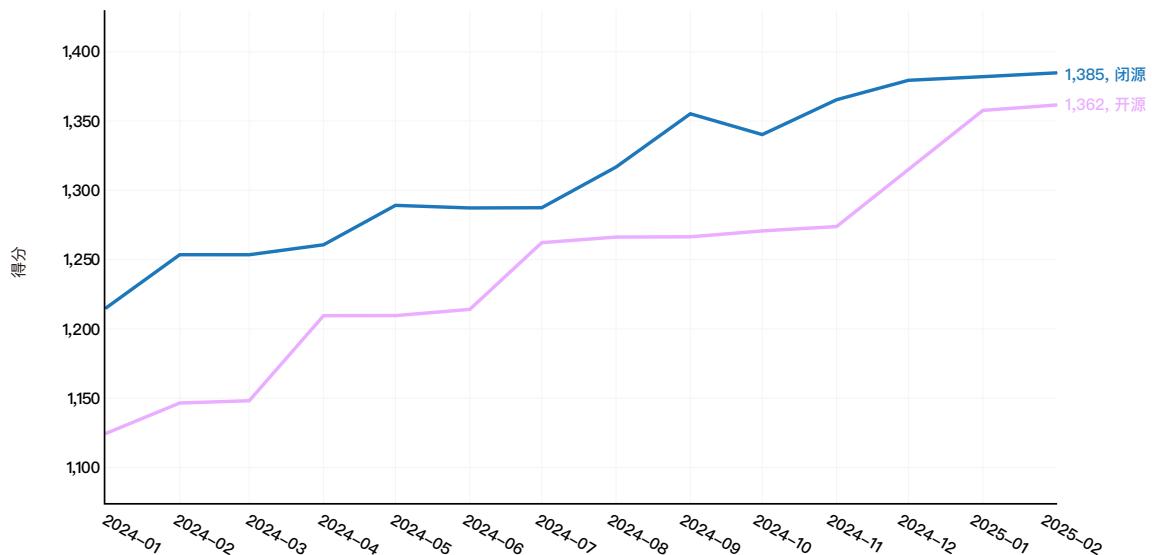


图 2.1.34

在选定的比较基准中，顶级闭源模型与开源模型的性能对比

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

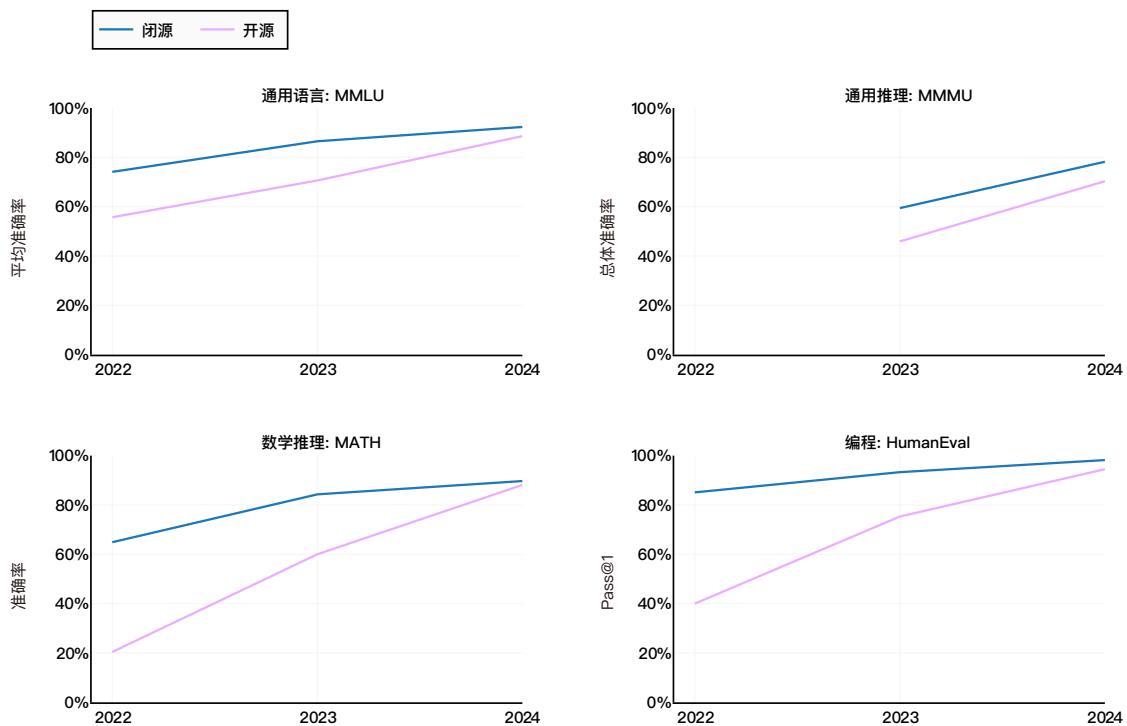


图 2.1.35

第二章：技术性能

2.1 2024 年人工智能技术发展概览

美国与中国技术性能对比

美国在人工智能研究和模型开发领域长期占据主导地位，中国则稳居第二。然而，最新证据表明，这一格局正在快速变化，中国开发的模型正逐步赶超美国同行。

2023 年，美国领先模型的性能显著优于中国模型。在 LMSYS Chatbot Arena 平台上，2024 年 1 月，美国顶尖模型的表现比中国最佳模型高出 9.3%。但到 2025 年 2 月，这一差距已缩小至仅 1.70%（图 2.1.36）。2023 年底，在 MMLU、MMMU、MATH 和 HumanEval 等比较基准中，中美模型的性能差距分别为 17.5、13.5、24.3 和 31.6 个百分点（图 2.1.37）。

而到 2024 年底，这些差距已大幅缩小至 0.3、8.1、1.6 和 3.7 个百分点。DeepSeek-R1 的发布引发了广泛关注，除此之外，另一原因在于该公司称其成果仅需通常训练此类模型所需硬件资源的一小部分即可实现。除了对美国股市造成影响外，DeepSeek-R1 的发布还引发了对美国半导体出口管制有效性的质疑。

在 LMSYS Chatbot Arena 美国和中国的模型的性能对比

资料来源：LMSYS, 2025 | 图表：2025 年人工智能指数报告

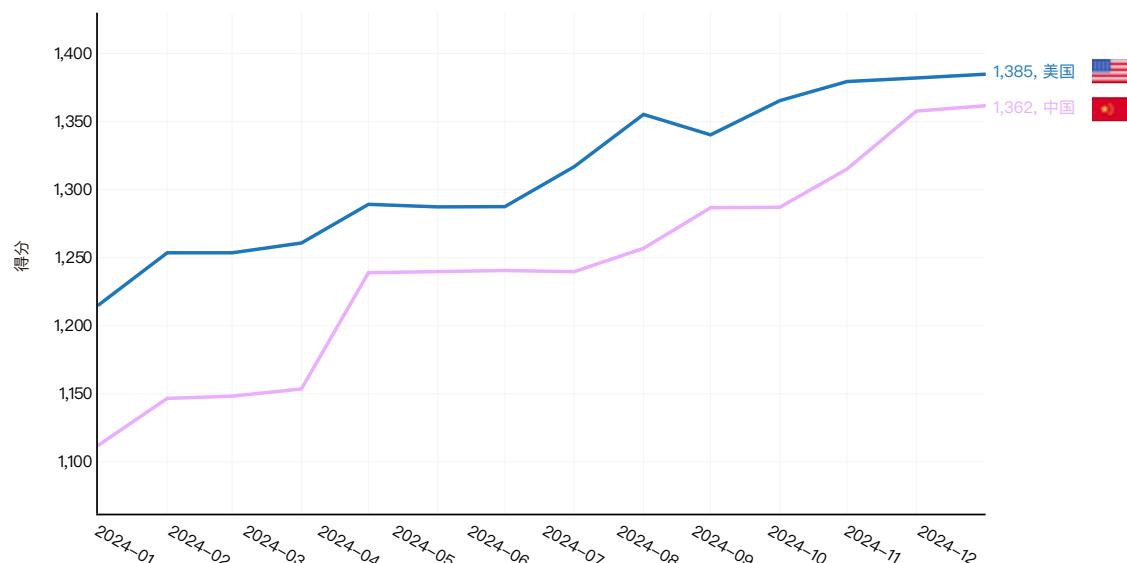


图 2.1.36

第二章：技术性能

2.1 2024 年人工智能技术发展概览

美国与中国的顶级模型在选定比较基准上的对比

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

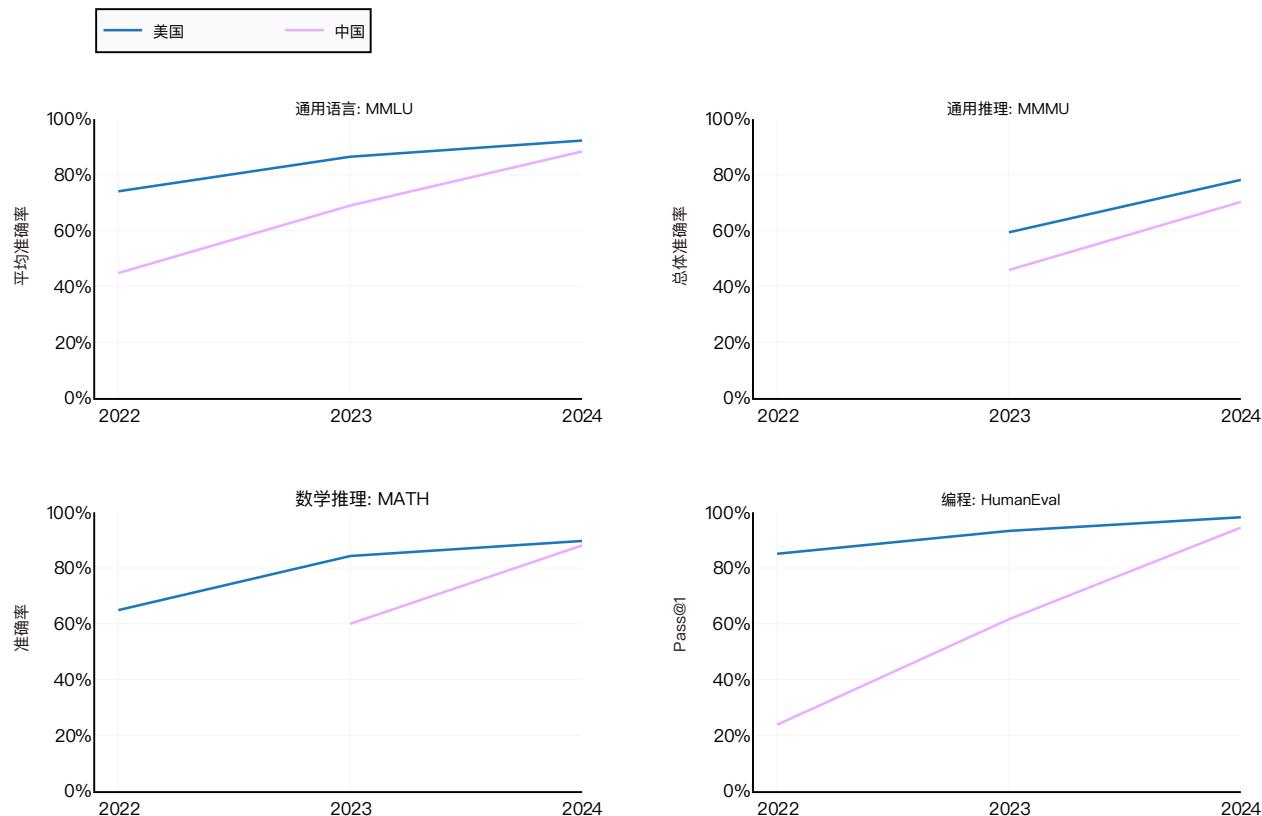


图 2.1.37

第二章：技术性能

2.1 2024 年人工智能技术发展概览

小型模型性能提升

近年来，人工智能的进步主要依赖于规模化(scaling)——即通过增加模型规模和训练数据来提升性能。尽管规模化显著增强了人工智能能力，但近期一个显著趋势是高性能小型模型的涌现。图 2.1.38 展示了在 MMLU (一种广泛使用的语言模型比较基准) 中得分超过 60% 的最小模型规模变化。作为背景参考，早期支持 ChatGPT 的模型(如 GPT-3.6 Turbo)在 MMLU 上的得分约为 70%。2022 年，达到 MMLU 60% 分数的最小模型是拥有 5400 亿参数的 PaLM；而到 2024 年，微软的 Phi-3 Mini 仅以 38 亿参数就达到了相同阈值，标志着两年间模型规模缩小了 142 倍。

2024 年是小型人工智能模型的突破之年。几乎所有主流人工智能开发商都发布了高性能紧凑模型，包括 GPT-40 mini、o1-mini、Gemini 2.0 Flash、Llama 3.1 8B 和 Mistral Small 3.5。⁵ 小型模型的崛起具有重要意义，其原因有以下几点：它体现了算法效率的提升，使开发者能够以更少的数据和更低的训练成本实现更高性能。这些效率提升与日益增长的数据集结合，可能催生性能更优的模型。此外，小型模型的推理速度通常更快、成本更低，其出现也降低了企业和开发者将 AI 整合到业务中的门槛。

2022–2024 年在 MMLU 的评估中得分超过 60% 的最小型人工智能模型

资料来源：Abdin 等，2024 | 图表：2025 年人工智能指数报告

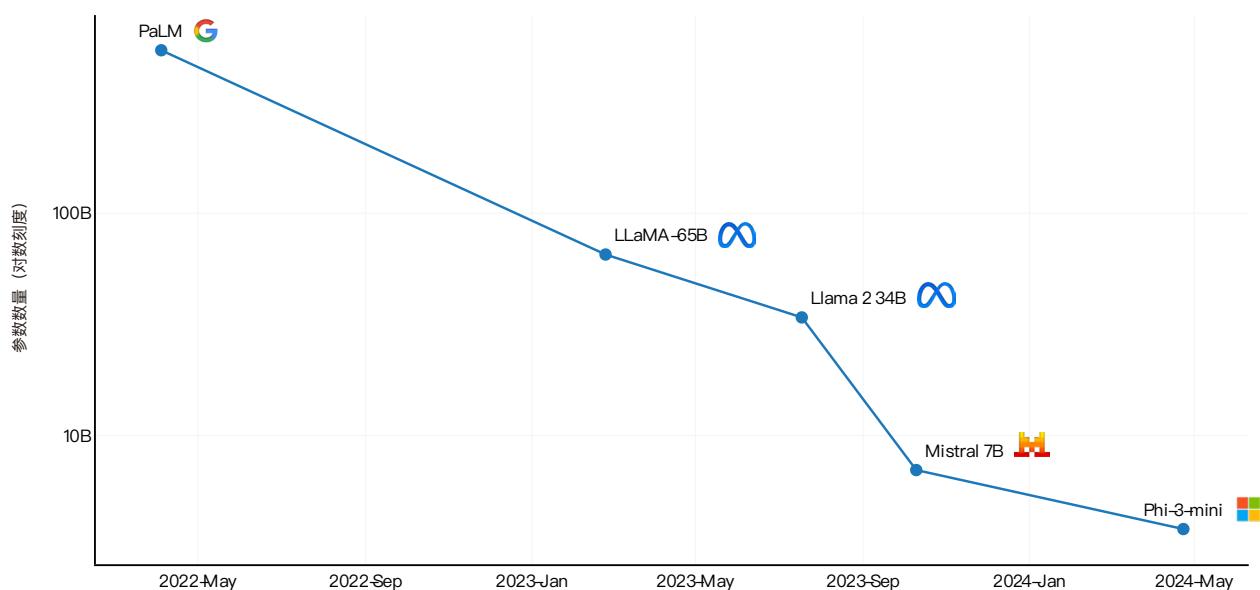


图 2.1.38

5. 这些仅是 2024 年发布的少量小型模型。

第二章：技术性能

2.1 2024 年人工智能技术发展概览

前沿模型性能趋同

近年来，人工智能前沿模型的性能逐渐趋同，多家供应商现已能提供高性能模型。这一现象标志着自 2022 年底以来的转变——当时 ChatGPT 的发布（被广泛视为 AI 进入公众视野的突破性事件）正值 OpenAI 和谷歌两大巨头主导市场的时期。OpenAI（成立于 2015 年）于 2020 年发布 GPT-3，而谷歌则在 2022 年推出了 PaLM 和 Chinchilla 等模型。

此后，新竞争者陆续入场，包括 Meta 的 Llama 系列、Anthropic 的 Claude、High-Flyer 的 DeepSeek、Mistral 的 Le Chat 以及 xAI 的 Grok。随着竞争加剧，模型性能差距日益

缩小（图 2.1.39）。根据去年的人工智能指数报告，在广泛使用的人工智能排名平台 Chatbot Arena Leaderboard 上，第一名与第十名模型的性能差距为 11.9%；而到 2025 年初，这一差距已缩小至 5.4%。同样，前两名模型的差异从 2023 年的 4.9% 降至 2024 年的仅 0.7%。人工智能领域竞争日趋激烈，印证了 2023 年的预测：人工智能企业缺乏抵御竞争对手的技术护城河。

选定的供应商在 LMSYS Chatbot Arena 的顶级模型表现

资料来源：LMSYS, 2025 | 图表：2025 年人工智能指数报告

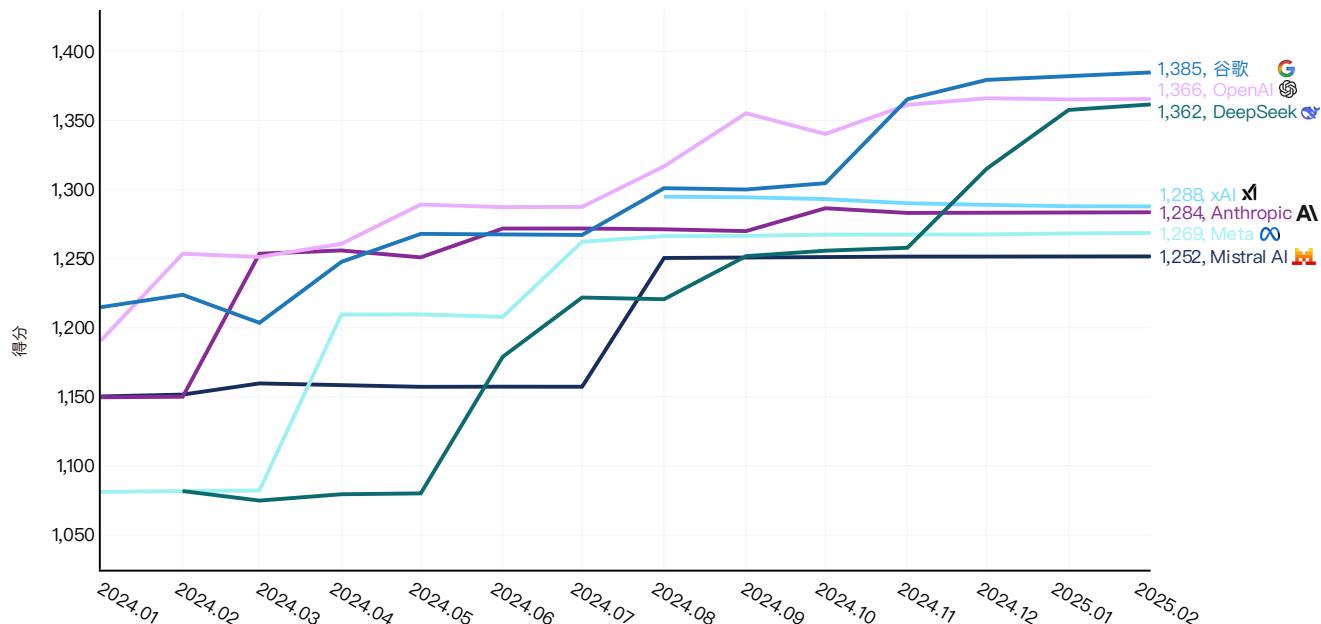


图 2.1.39

第二章：技术性能

2.1 2024 年人工智能技术发展概览

人工智能基准比较

多年来，人工智能指数报告一直通过比较基准来追踪人工智能系统的技术进展。尽管比较基准仍是关键工具，但必须认识到其局限性，并引导社区采用更有效的基准比较实践。

如去年人工智能报告所述，许多主流 AI 比较基准正趋于饱和。随着人工智能系统快速发展，即使新设计的更具挑战性的测试，其适用周期往往仅能维持数年。部分专家认为，学术比较基准的新时代可能即将结束。要真正评估人工智能系统的能力，需要更严格、更全面的评估方法。

此外，当模型开发者发布新模型时，他们通常会报告比较基准分数，而这些分数通常会被更广泛的社区所接受。然而，这种方法也有缺陷。在某些情况下，公司会使用非标准的提示技术，从而使模型间的比较变得不可靠。例如，谷歌在推出 Gemini Ultra 时，报告的 MMLU 比较基准分数使用了思维链提示技术，而其他开发人员并不使用。第三方研究也发现，部分模型在独立测试中的表现低于开发者最初报告的结果。

某些关键智能维度难以通过比较基准衡量。比较基准对于评估某些智能能力（如视觉和语言）是有效的，因为任务是离散的——例如正确地对图像进行分类或回答选择题。然而，但在多智能体系统、人机交互等领域则面临挑战，主因包括人类行为的多变性和答案的多样性。制定比较基准更具挑战性。

此外，人工智能的发展通常是在旨在衡量人类表现的竞赛中进行评估的，例如游戏以及对人类或机器提出的其他公开挑战。国际象棋和扑克等游戏需要高度的智力，几十年来，人工智能系统不断改进，已经能够在越来越复杂的游戏中击败最优秀的人类。具有物理组件或团队能力的游戏也是衡量人工智能进步的好方法，机器人界已经开展了各种挑战性的游戏竞赛，例如机器人足球赛（RoboCup）。人工智能的另一个竞赛领域涉及

协调和团队合作，多智能体系统在分布式推理方面表现出了进步。

人工智能界长期以来一直在开发比较基准。人工智能之所以能够取得重大进展，是因为不同方法和手段都可以根据比较基准所代表的同一黄金标准进行评估。在机器学习领域，不同领域、不同类型数据的比较基准推动了重大进展。许多比较基准由第三方自动评估，不会向人工智能开发人员公开测试数据，这使得评估结果更加可靠。最近一个有趣的趋势是，各种比较基准任务都由同一个模型来处理。例如，自然语言多年来一直是作为一系列独立任务（如理解、生成、问题解答）来处理的，每个任务都有自己的模型和比较基准。同样，语音任务也与语言理解或生成任务分开进行比较基准。如今，同一个模型可以处理所有语言任务，在某些情况下，一个模型可以处理语言、图像和多模态任务。这是人工智能在整合原本独立的智能任务和能力方面取得的一项非常重要的进步。

人工智能系统在比较基准中表现出持续超越力，其快速进步或许最能从人工智能长期面临的著名挑战——图灵测试——的重要性日益下降中得到体现。该测试最初由艾伦·图灵在 1950 年的论文《计算机器与智能》（“Computing Machinery and Intelligence”）中提出，用于评估机器表现出类似人类智能的能力。在测试中，人类评判者与机器和人类进行基于文本的对话；如果评判者无法可靠地区分它们，则认为机器通过了图灵测试。最新证据表明，大语言模型的进步已使人们难以区分顶尖语言模型与人类的差异，这标志着现代人工智能模型能够通过图灵测试。尽管该测试的优缺点长期存在争议，它仍是衡量机器智能的重要历史文化基准。对其相关性的质疑，凸显了近年来大语言模型的巨大进步以及人们对有效计算机科学

第二章：技术性能

2.1 2024 年人工智能技术发展概览

比较基准和人工智能测量不断发展的看法。

在机器人领域，出现了许多应对与物理世界互动和推理自然规律的模型。许多机器人比较基准（如 ARMBench）侧重于感知任务。然而，其他比较基准（如 VIMA-Bench）则评估机器人在模拟环境中的性能，这些环境同时融合了感知、通信和深度学习。

比较基准也会受到污染，即大语言模型遇到其训练数据中出现过的测试题。Scale 最近的一项研究发现，许多大语言模型在广泛使用的数学比较基准 GSM8K 上的表现存在严重污染。一些研究人员试图通过引入 LiveBench 等比较基准来应对这些污染问题，这些比较基准会定期更新来自陌生来源的新问题，这些问题不太可能出现在大语言模型的训练数据中。

最后，研究表明，许多比较基准的构建存在缺陷。在 BetterBench 中，研究人员系统地分析了 24 个著名的比较比

较基准，并发现了系统性缺陷：14 个未报告统计学意义，17 个缺乏结果复制脚本，大多数文档不完善，限制了其可重复性和评估模型的有效性。尽管广泛使用，但 MMLU 等比较基准对质量标准的遵守程度较低，而 GPQA 等比较基准的表现则明显更好。为了解决这些问题，该论文提出了一个涵盖比较基准开发所有阶段（设计、实施、文档和维护）的 46 项标准框架（图 2.1.40）。该论文还引入了一个公开可访问的存储库，以实现持续更新并提高比较基准的可比性。图 2.1.41 来自 BetterBench，评估了许多著名比较基准的可用性和设计。这些发现强调了标准化比较基准的必要性，以确保人工智能评估的可靠性，并防止对模型性能做出误导性的结论。比较基准有可能影响政策决策和组织内的采购决策，凸显了评估的一致性和严谨性的重要性。

比较基准生命周期的五个阶段

资料来源：Reuel 等，2024

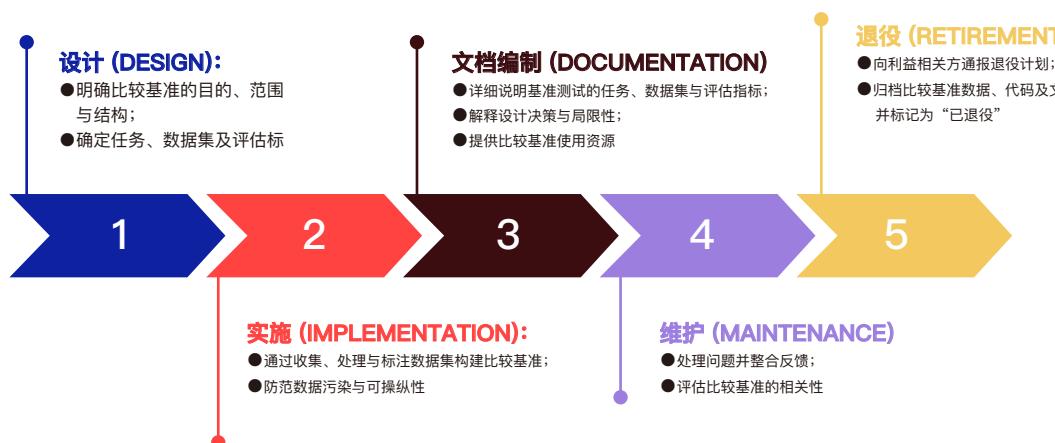


图 2.1.40

第二章：技术性能

2.1 2024 年人工智能技术发展概览

在选定的比较基准中，设计与可用性得分对比

资料来源：Reuel 等，2024 | 图表：2025 年人工智能指数报告

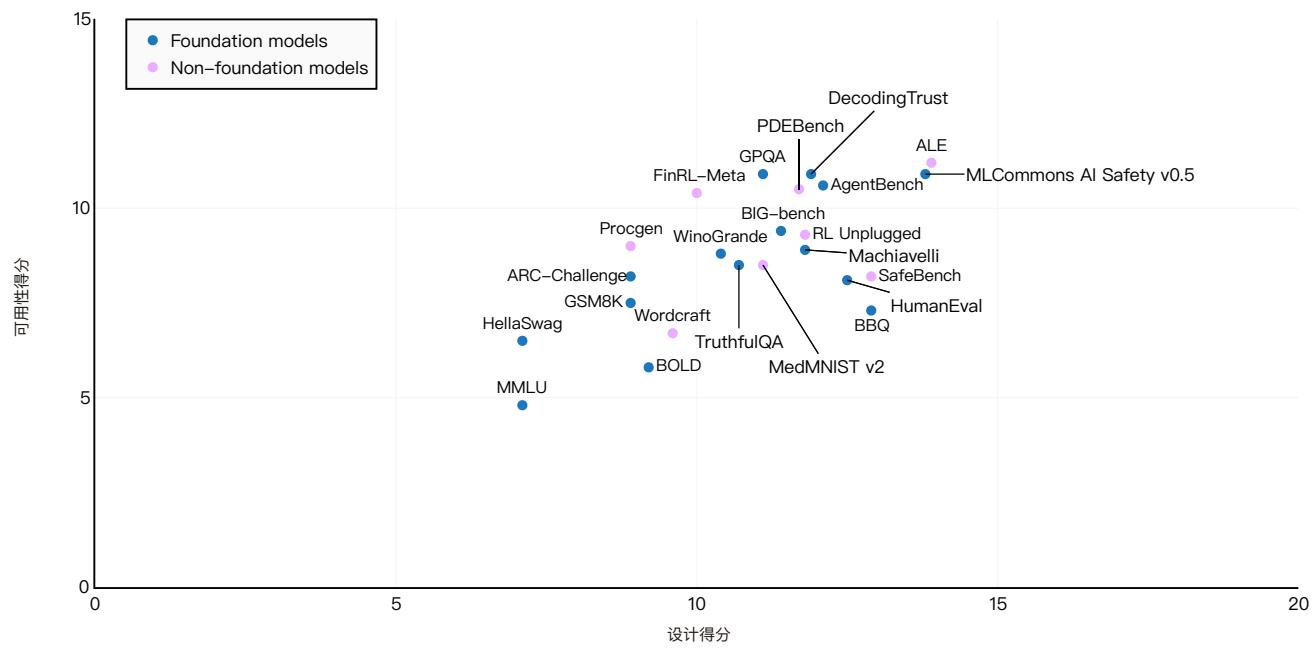


图 2.1.41

在本章节中，人工智能指数继续报告比较基准，并认可其在跟踪人工智能技术进步方面的重要性。按照惯例，该指数从 leaderboard、Papers With Code 和 RankedAGI 等公共存储库以及公司论文、博客文章和产品发布中获取比较基准分数。该指数基于以下假设：公司报告的分数准确且真实。本节中的

比较基准分数为 2025 年 2 月中旬的最新数据。但是，自人工智能指数发布以来，可能已经发布了超越当前最先进分数的新模型。

第二章：技术性能

2.2 语言

自然语言处理（NLP）使计算机能够理解、解释、生成和转换文本。当前最先进的模型，如 OpenAI 的 GPT-4o、Anthropic 的 Claude 3.5 和谷歌的 Gemini，能够生成流畅连贯的文本，并展现出高水平的语言理解能力（图 2.2.1）。与早期版本仅限于文本输入和输出不同，新一代语言模型现已能够在更广泛的输入和输出模态中进行推理，包括音频、图像和面向目标的任务（图 2.2.2）。

GPT-4o 的示例输出

资料来源：2025 年人工智能指数

问：斯坦福 HAI 是什么？

答：斯坦福 HAI（斯坦福以人本人工智能研究所）是斯坦福大学的一个研究机构，致力于推动人工智能发展以改善人类生活。该机构促进人工智能研究人员、社会科学家、政策制定者和行业领袖之间的跨学科合作，确保人工智能的开发和应用符合人类价值观和社会需求。

斯坦福 HAI 的核心领域包括：

研究：开展关于人工智能对社会、伦理及治理影响的课题。

教育：提供人工智能相关的课程、奖学金和培训项目。

政策与推广：与政策制定者和国际组织合作，推动负责任的人工智能法规。

倡议活动：举办研讨会、发布报告（如人工智能指数报告），并开发工具（如全球 AI 活力指数工具）。

该研究所在全球人工智能讨论中发挥着领导作用，旨在让人工智能广泛惠及人类，同时应对其潜在风险和挑战。更多信息可访问斯坦福 HAI 官网。

图 2.2.1

Gemini 2.0 在代理工作流中的应用

资料来源：2025 年人工智能指数

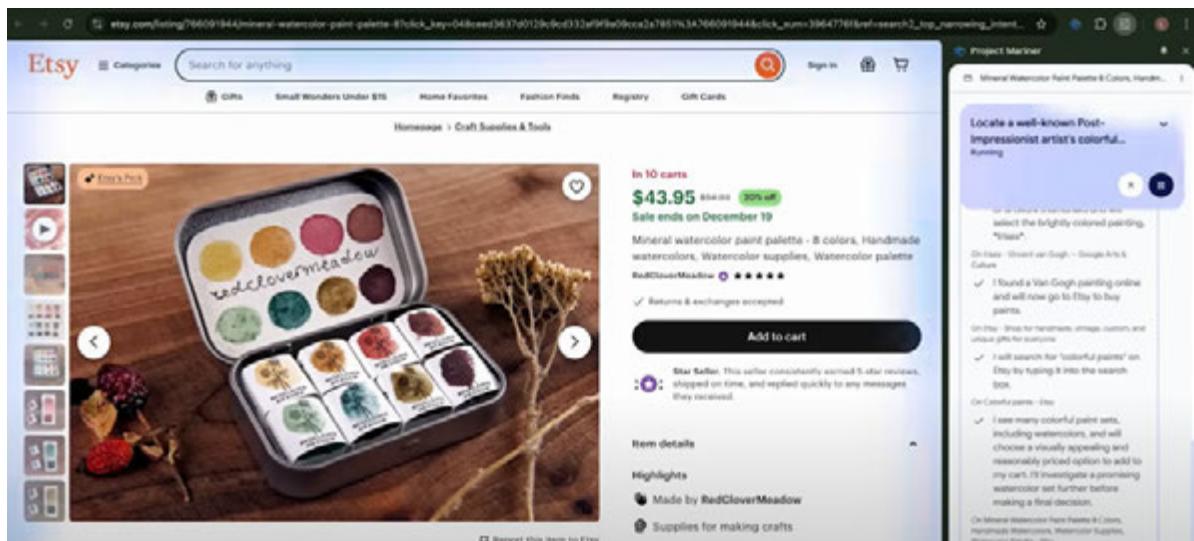


图 2.2.2

第二章：技术性能

2.2 语言

语言理解

英语语言理解对人工智能系统提出了多方面的挑战，包括阅读理解与逻辑推理等任务。

MMLU：大规模多任务语言理解

大规模多任务语言理解 (MMLU) 比较基准通过零样本或少样本场景评估模型在 57 个学科中的表现，涵盖人文、STEM (科学、技术、工程、数学) 及社会科学等领域(图 2.2.3)。MMLU 已成为评估大语言模型能力的核心比较基准：GPT-4o、Claude 3.5、Gemini 2.0 等前沿模型均基于此测试进行了评估。

MMLU 比较基准由加州大学伯克利分校、哥伦比亚大学、芝加哥大学和伊利诺伊大学厄巴纳 – 香槟分校的研究团队于 2020 年创建。

截至 2024 年 9 月，MMLU 最高分为 92.3%，由 OpenAI 的 o1-preview 模型取得。作为对比，2023 年 3 月发布的 GPT-4 得分为 86.4%。值得注意的是，早期测试模型 RoBERTa 在 2019 年仅获得 27.9% 的分数 (图 2.2.4)。这一最新成果标志着五年间性能提升了 64.4 个百分点。

MMLU 示例题目

来源 : Hendrycks 等 , 2021

- 微观经济学** 政府限制和监管垄断的原因之一是
- (A) 生产者剩余减少，消费者剩余增加。
 - (B) 垄断价格保证了生产效率，但社会失去了配置效率。
 - (C) 垄断企业不会进行重要的研发活动。
 - (D) 由于价格上升和产量下降，消费者剩余减少。



图 2.2.3

MMLU: 平均准确率

资料来源 : Papers With Code, 2025 | 图表: 2025 年人工智能指数报告

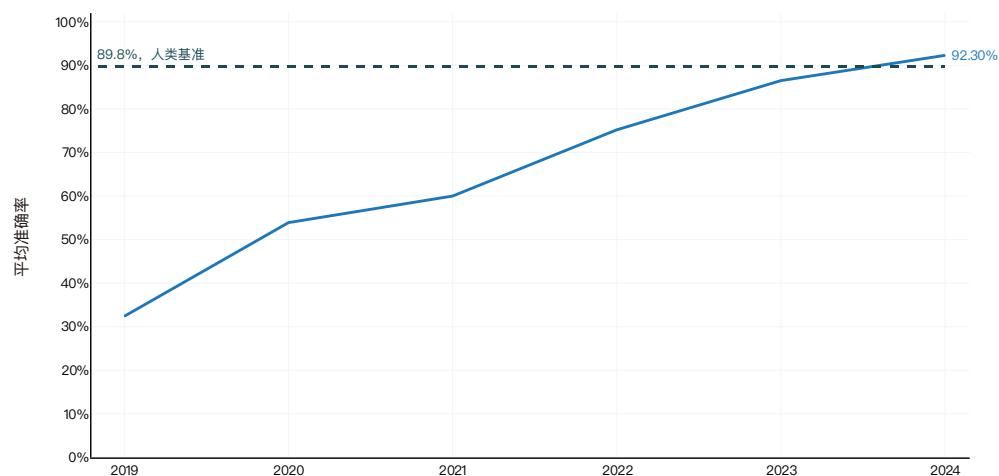


图 2.2.4

第二章：技术性能

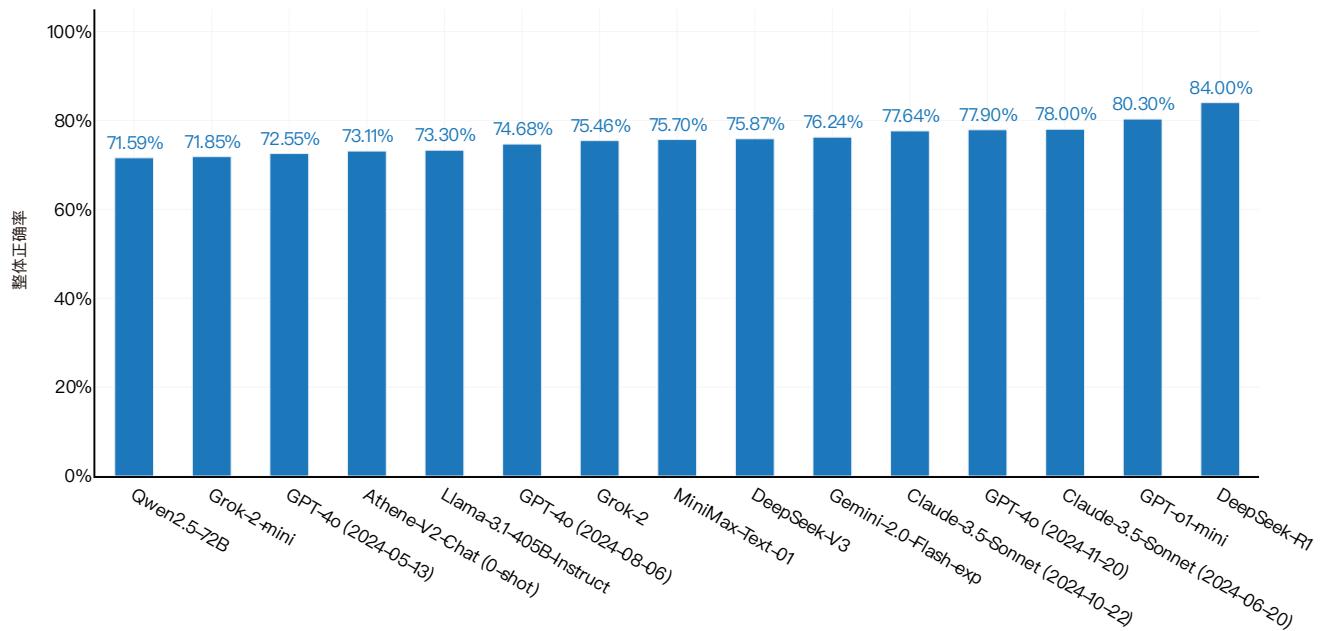
2.2 语言

尽管MMLU备受关注，但它也面临显著批评。有观点认为，该比较基准包含错误或过于简单的问题，可能无法有效挑战日益先进的系统。2024年，多伦多大学、滑铁卢大学和卡内基梅隆大学的研究团队推出了MMLU-Pro，这是MMLU的一个更具挑战性的版本。该版本剔除了噪声问题和琐碎问题，扩展了复杂题目，并增加了模型的选项数量。图2.2.5展示了MMLU-Pro的性能趋势，其中DeepSeek-R1以84.0%的得分位居榜首。

此外，测试环境也引发了担忧。开发者有时会使用非标准的提示技术报告MMLU分数，这些技术可能提升性能，但会导致误导性比较。此外，证据表明，开发者公开报告的分数与学术研究者后续评估的结果可能存在差异，有时甚至相差五个百分点。因此，MMLU的性能结果需谨慎解读。

MMLU-Pro: 整体正确率

资料来源：MMLU-Pro Leaderboard, 2025 | 图表：2025年人工智能指数报告



生成任务

在生成任务中，人工智能模型需测试其生成流畅且实用的语言回答的能力。

Chatbot Arena Leaderboard

随着高性能大语言模型的崛起，了解公众更偏好哪些模型变得愈发重要。LMSYS于2023年推出的Chatbot Arena Leaderboard，是首个全面评估公众对大语言模型偏好的平台之一。该leaderboard允许用户向两个匿名模型提问，并投票选出更优的回答（图2.2.6）。截至2025年初，该平台已累计超过100万次投票，用户将谷歌的Gemini系列模型评为社区最受欢迎的选择。

第二章：技术性能

2.2 语言

Chatbot Arena Leaderboard 上的模型回答示例

来源：Chatbot Arena Leaderboard, 2024

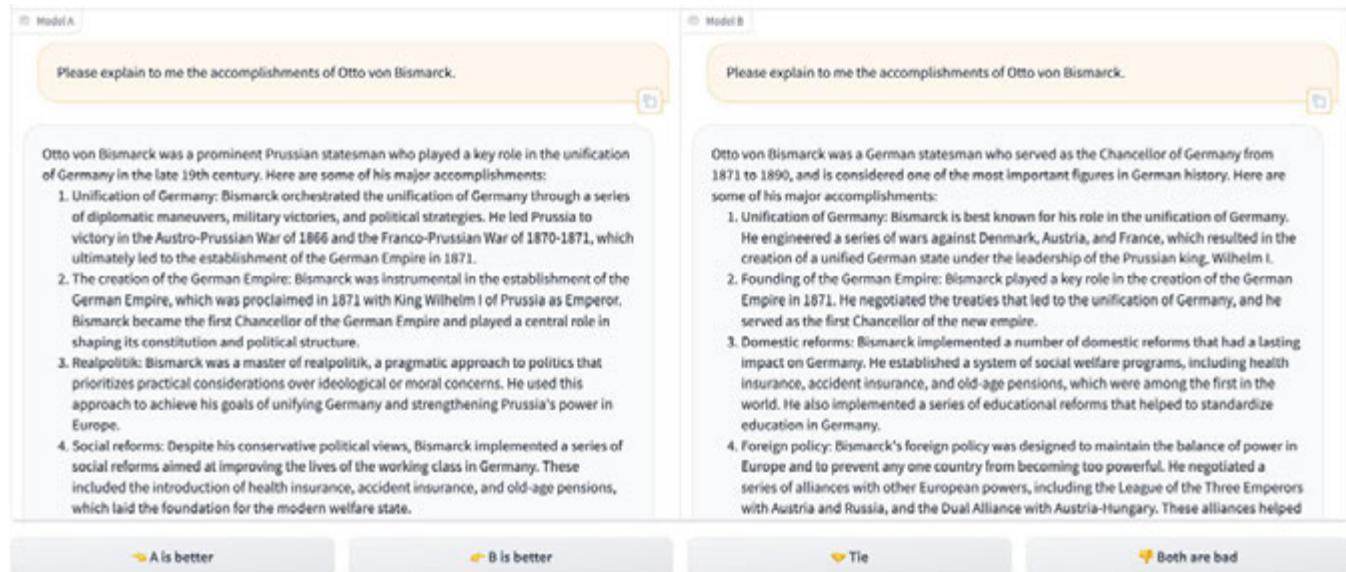


图 2.2.6

图 2.2.7 展示了截至 2025 年 1 月 Chatbot Arena Leaderboard 上前 10 名模型的概况。值得注意的是，顶尖模型之间的性能差距逐渐缩小。根据 2024 年人工智能指数，2023

年榜首与第 10 名模型的 Arena 评分差异为 11.9%，而到 2025 年，这一差距已降至 5.4%。这一趋同现象表明，近期大语言模型的质量正趋于均衡。

LMSYS Chatbot Arena 大语言模型 Elo 评分（总体）

来源：LMSYS, 2025 | 图表：2025 年人工智能指数报告

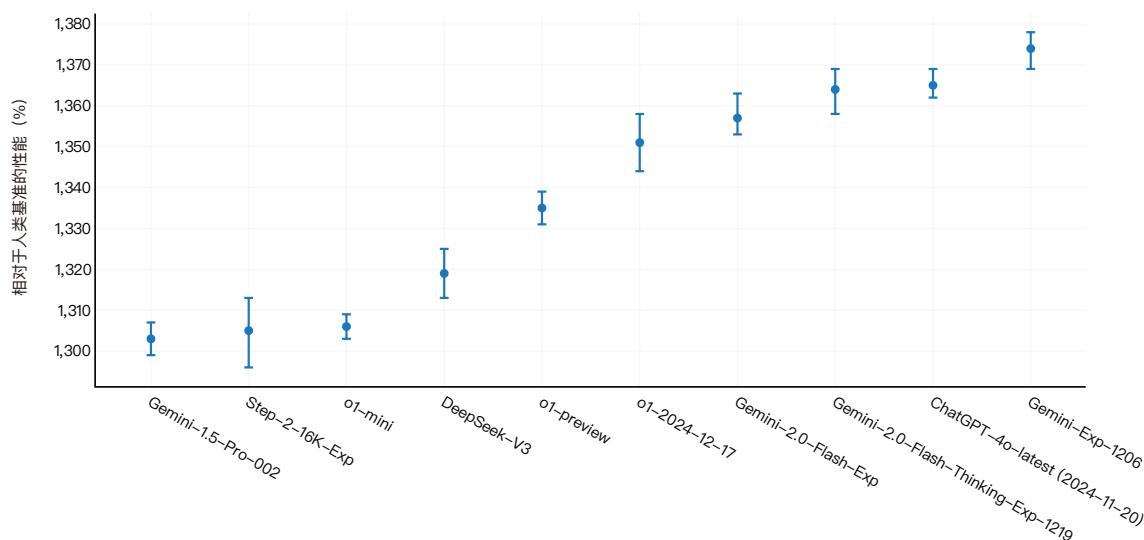


图 2.2.7

6、Arena 评分是 Arenaleaderboard 用于比较模型性能的相对排名系统。评分方法详见 Chatbot Arena Leaderboard 的文章。

第二章：技术性能

2.2 语言

Arena-Hard-Auto

随着人工智能能力快速提升，开发新比较基准面临的主要挑战在于，人工创建高质量基准成本高昂且耗时。为此，加州大学伯克利分校的研究团队今年推出了 BenchBuilder。该工具利用大语言模型构建自动化流程，从大规模众包数据集中筛选高质量开放式提示，可在无需大量人工干预的情况下更新或创建新基准。LMSYS 团队使用该工具开发了 Arena-Hard-Auto，这是一个专为评估指令调优大语言模型设计的基准（图 2.2.8）。Arena-Hard-Auto 包含 500 个来自 Chatbot Arena 的高难度用户查询，并以 GPT-4 Turbo 作为评判模型，将其响应与基线模型（GPT-4-0314）对比。

截至 2024 年 11 月，Arena-Hard-Autoleaderboard 上得分最高的模型为 o1-mini (92.0)、o1-preview (90.4) 和 Claude-3.5-Sonnet (85.2)（图 2.2.9）。该基准还设有风格控制 leaderboard，用于评估模型回答风格对用户偏好的潜在影响。风格 leaderboard 的首位是 Anthropic 公司 2024 年 11 月发布的 Claude Sonnet 3.5 变体（图 2.2.10）。然而，Arena-Hard-Auto 等自动化基准因问题分布不均而受到批评，例如超过 50% 的问题仅聚焦编程与调试，这限制了对大语言模型能力的全面评估。

Chatbot Arena Leaderboard 上的模型回答示例

来源：Chatbot Arena Leaderboard, 2024

	评估方式	是否支持开放式问题	提示语整理方式	提示语来源
Arena-Hard-Auto	自动评估	支持	自动评估	可配置
MMLU,MATH,GPOA	自动评估	不支持	人工手动整理	固定预设
MT-Bench, AlpacaEval	自动评估	支持	人工手动整理	固定预设
Live Bench, Live code Bench	自动评估	不支持	人工手动整理	固定预设
chatbot Arena	人工评估	支持	众包用户群体共同整理	用户群体

图 2.2.8

未修改的 Arena-Hard-Auto

资料来源：LMSYS, 2025 | 图表：2025 年人工智能指数报告

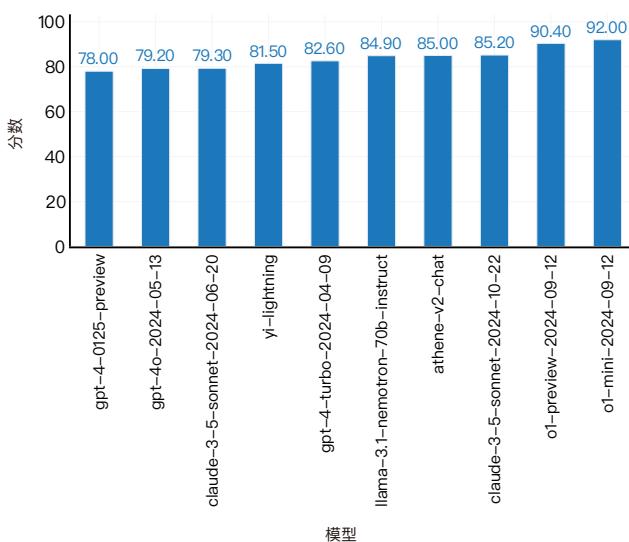


图 2.2.9

带风格控制的 Arena-Hard-Auto

资料来源：LMSYS, 2025 | 图表：2025 年人工智能指数报告

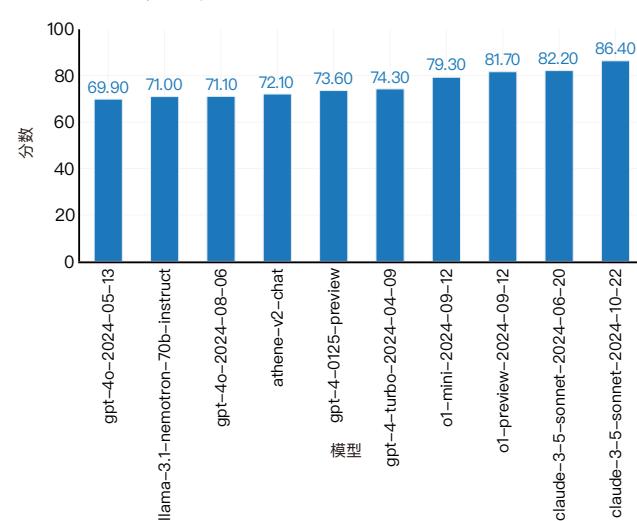


图 2.2.10

第二章：技术性能

2.2 语言

WildBench

WildBench 由 Allen Institute for AI 和华盛顿大学研究人员开发，是 2024 年推出的一个比较基准，用于在具有挑战性的现实世界查询中对大语言模型进行评估。创建者强调了现有大语言模型评估的几个局限性。例如，MMLU 仅关注学术

问题，未涵盖开放式现实场景；而 LMSYS 等基准虽涉及现实挑战，但过度依赖人工审核，且缺乏对全部模型使用统一数据集的评估一致性（图 2.2.11）

WildBench 的评估框架

来源：Lin 等，2024

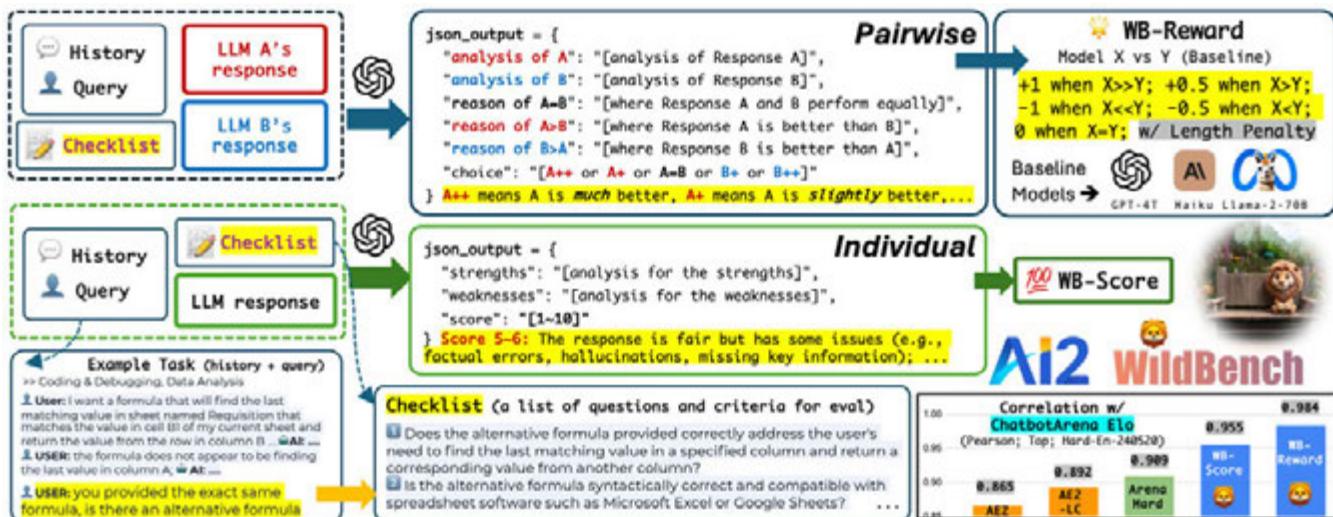


图 2.2.11

第二章：技术性能

2.2 语言

WildBench 通过自动化评估框架解决了现有基准的缺陷，其问题集涵盖语言模型可能遇到的多样化真实场景（“野外”问题）（图 2.2.11）。这些问题从超过 100 万条人机对话记录中精选而出，并定期更新以保证时效性。开发者还维护实

时 leaderboard 以追踪模型性能变化。目前，WildBench 表现最佳的模型为 GPT-4o，其 WB-Elo 评分为 1227.1，以微弱优势领先第二名 Claude 3.5 Sonnet (1215.4 分)（图 2.2.12）。

WildBench: WB-Elo (长度控制)

资料来源 : WildBench Leaderboard, 2025 | 图表: 2025 年人工智能指数报告

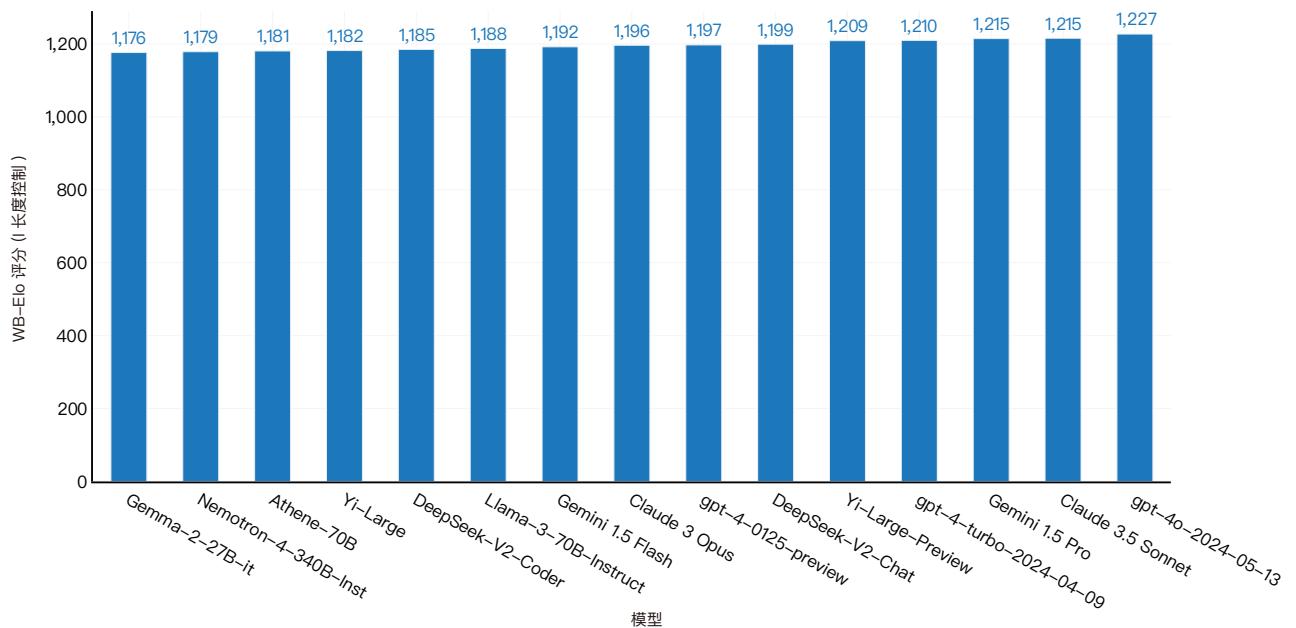


图 2.2.12

第二章：技术性能

2.2 语言

重点：

o1, o3, 与推理时间计算

OpenAI 最新发布的 o1 和 o3 模型标志着人工智能模型在“思考”能力和高级推理表现上的范式转变。o1 和 o3 在编程、量子物理和逻辑等多样化任务中展现出卓越性能，其高级推理能力归功于链式思考过程和答案迭代校验机制。这意味着模型会将复杂问题分解为更小的可执行步骤，

从而提升输出质量。例如，在解码乱序文本任务中，o1 会展示比 GPT-4 更详尽的思维推理过程（图 2.2.15）。这种人工智能系统在回答过程中迭代优化的机制被称为推理时间计算。

o1 的链式思考过程

资料来源：OpenAI, 2024

GPT-4o

用户输入：

oyfjdnisdr rtqwainr acxz mynzbhhx -> Think
step by step

使用上述方法解码以下密文：

oyekajjzdf aaptcg suaokybhai ouow aqht my-
nznvaatzacdfoulxxz

解码步骤说明：

配对字母：将密文中的字母两两分组（例如 oy fj dn is dr）。

字母转数字：将每个字母转换为字母表中的对应序号（A=1, B=2, ..., Z=26）。

求和：计算每对字母的数字之和。

求平均值：将和除以 2，得到平均值。

数字转字母：将平均值转换为对应的字母（1=A, 2=B, ..., 26=Z）。

应用示例：

以第一对字母 oy 为例：

$$\begin{aligned} o &= 15, \quad y = 25 \rightarrow \text{和} = 15 + 25 = 40 \rightarrow \text{平均值} = 40 / 2 \\ &= 20 \rightarrow \text{字母} = T \end{aligned}$$

图 2.2.13

第二章：技术性能

2.2 语言

重点：

o1, o3, 与推理时间计算(续)

图 2.2.14 对比了 GPT-4o 与 o1、o1-preview 在多项比较基准中的表现。⁷ 例如, o1 在 MMLU 上以 2.8 分优势超越 GPT-4o, 在 MATH 测试中领先 34.5 分, 在

GPQA Diamond 中高出 26.7 分, 在极难的 AIME 2024 数学竞赛中更是领先 65.1 分。而 o3 则展现出当前最复杂的推理能力, 在 ARC-AGI 机器智能比较基准中以 87.5% 的准确率刷新了此前 55.5% 的纪录。

GPT-4o 、o1-preview 、o1 在选定的比较基准上对比

资料来源 : OpenAI, 2024

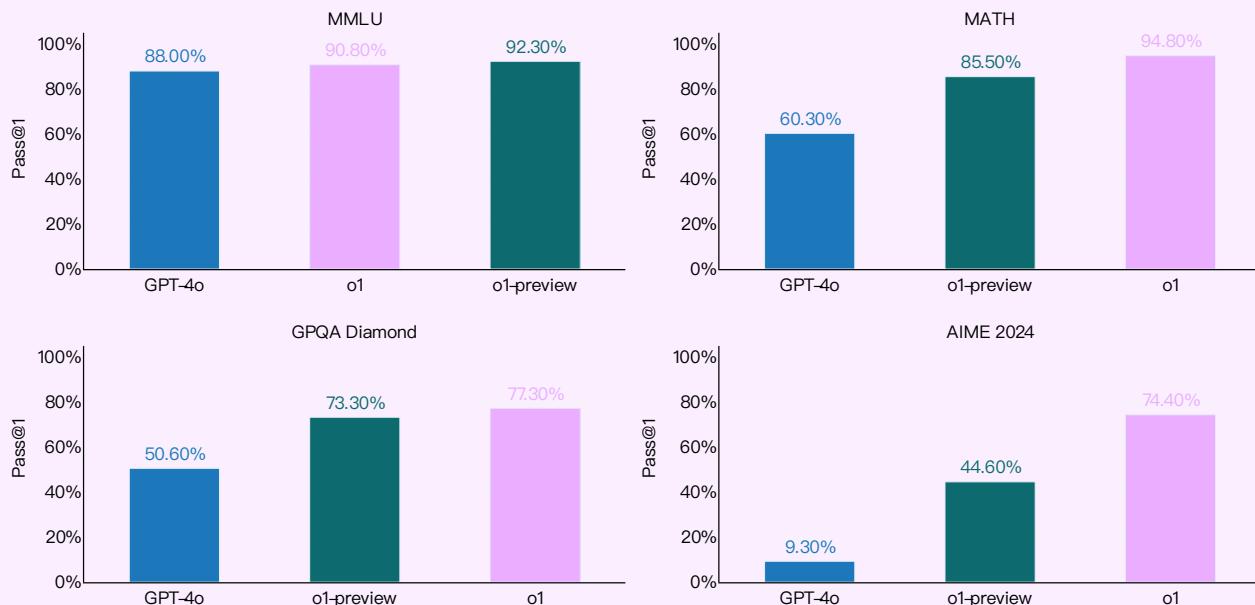


图 2.2.14

然而, 这些模型的强大推理能力伴随着显著的成本提升——包括财务成本和延迟成本。例如, GPT-4o 的输入 / 输出词元成本分别为每百万 2.5 美元和 10 美元, 而 o1 则高达 15 美元和 60 美元。⁸ 此外, o1 的首词元延迟达 29.7 秒, 是 GPT-4o (0.72 秒) 的 40 倍。o3 的延迟虽未公开, 但推测更高。o1 与 o3 的强劲能力将持续推动高级人工智能系统和智能体的发展。

OpenAI 于 2024 年 9 月 12 日向 ChatGPT Plus 和 Teams 用户发布 o1-preview, 并于 2024 年 12 月 5 日推出 o1 正式版 (同时发布月费 200 美元的 ChatGPT Pro 订阅服务以获取 o1 访问权限)。

7、o1-preview 是 o1 的早期预览版本, 在全面发布前提供有限访问。

8、o3 目前仅通过 Open 人工智能安全测试计划向特定研究人员和开发者开放。

第二章：技术性能

2.2 语言

MixEval

MixEval 由新加坡国立大学、卡内基梅隆大学和 Allen Institute for AI 的研究团队联合推出，是解决当前大语言模型评估领域局限性的一项新比较基准。MixEval 结合了 Chatbot Arena 中的综合性真实用户查询与 MMLU 基于标准答案的

题目（图 2.2.15），包含多个评估套件，其中 MixEval-Hard 是更具挑战性的版本，聚焦高难度查询，成为评估模型处理复杂问题的有效工具。

MixEval 的评估框架

资料来源：Ni 等, 2024

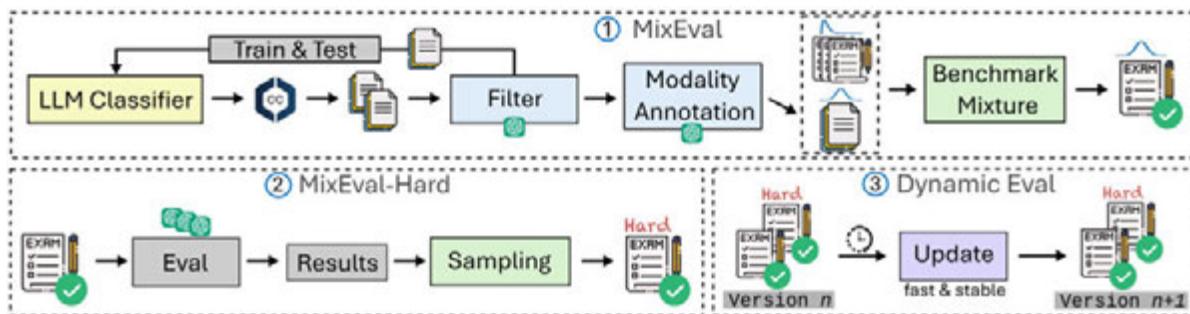


图 2.2.15

在 MixEval-Hard 比较基准中，得分最高的模型是 OpenAI 的 o1-preview (72.0 分)，其次是 Claude 3.5 Sonnet-0620 模

型 (68.1 分)，第三名为 Llama-3-405B-Instruct 模型 (66.2 分) (图 2.2.16)。这三款模型均发布于 2024 年。

聊天模型在 MixEval-Hard 的得分

资料来源：MixEval Leaderboard, 2025 | 图表：2025 年人工智能指数报告

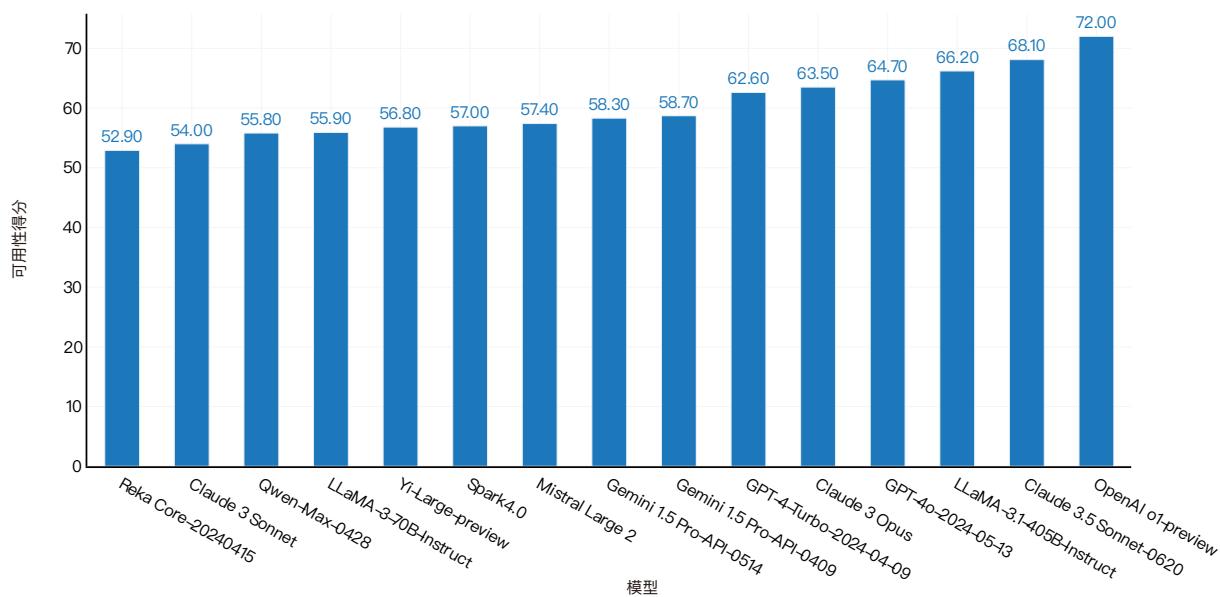


图 2.2.16

RAG: 检索增强生成

检索增强生成（RAG）是在大语言模型中测试的一种越来越常见的能力。这种方法将大语言模型与检索机制整合在一起，以增强其响应生成能力。模型首先从文件或文档中检索相关信息，然后根据检索到的内容生成适合用户查询的响应。RAG 的用例多种多样，包括从大型数据库中回答精确问题，以及利用公司文件信息解决客户查询。

近年来，RAG 备受研究机构和企业关注。例如，Anthropic 于 2024 年 9 月推出“上下文检索”技术，显著提升 RAG 模型的检索能力。2024 年还发布了多项 RAG 评估基准，如

Ragnarok (RAG 竞技场) 和 CRAG (综合 RAG 基准)。此外，针对特定场景的专项基准 (如金融问答基准 FinanceBench) 也已问世。

Berkeley Function Calling Leaderboard

Berkeley Function Calling Leaderboard 评估大语言模型准确调用函数或工具的能力。评估套件包括 2,000 多个问题 - 函数 - 答案对，涉及多种编程语言（如 Python、Java、JavaScript 和 REST API）和多个测试领域（图 2.2.17）。

Berkeley Function Calling Leaderboard 数据构成

资料来源：Yan 等，2024

Berkeley Function-Calling Leaderboard Evaluation Data Composition

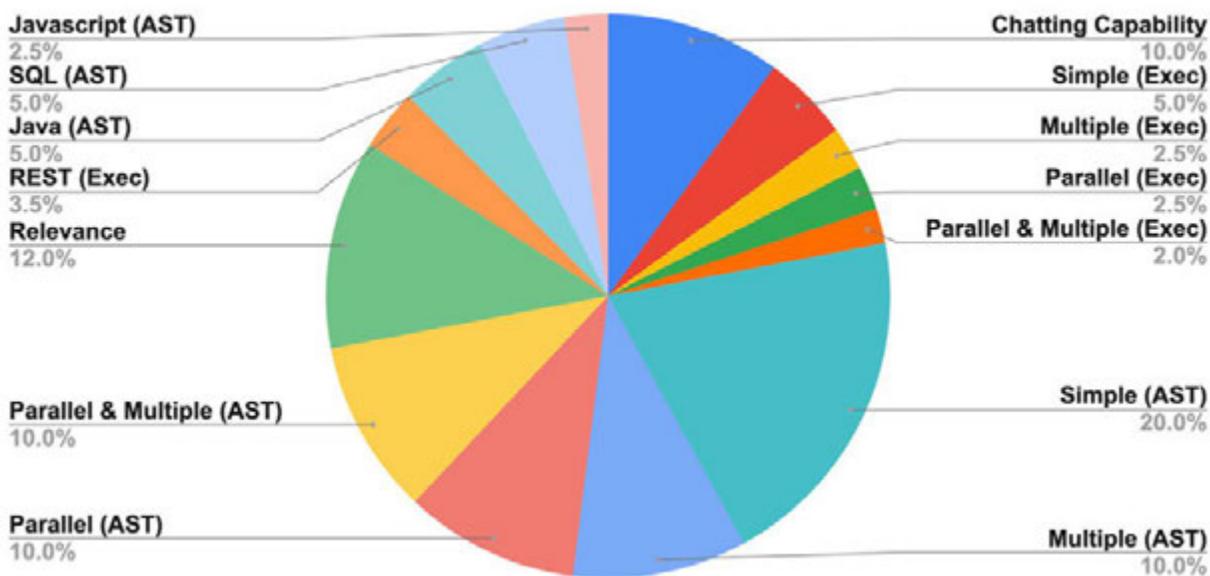


图 2.2.15⁹

9. 在这种情况下，AST (抽象语法树) 指的是涉及在结构层面分析或处理代码的任务，将代码解析为一棵语法元素树。标有 “AST” 的评估可能会测试人工智能模型以结构化方式理解、生成或处理代码的能力。Exec (基于执行) 表示需要实际执行函数调用以验证正确性的任务。标有 “Exec” 的评估可能会评估人工智能模型是否能够正确调用和执行函数，确保产生预期的输出结果。

第二章：技术性能

2.2 语言

在 Berkeley Function Calling Leaderboard 上，表现最佳的模型是 watt-tool-70b，这是基于 Llama-3.3-70B-Instruct 专门针对函数调用任务微调的变体，其整体准确

率达到 74.24% (图 2.2.18)。排名第二的是 GPT-4o 的 11 月版本，得分为 72.08。2024 年间，模型在该比较基准中的性能显著提升，年底顶尖模型的整体准确率较年初提高了 50 个百分点。

伯克利函数调用：整体准确率

资料来源：Berkeley Function-Calling Leaderboard, 2025 | 图表：2025 年人工智能指数报告

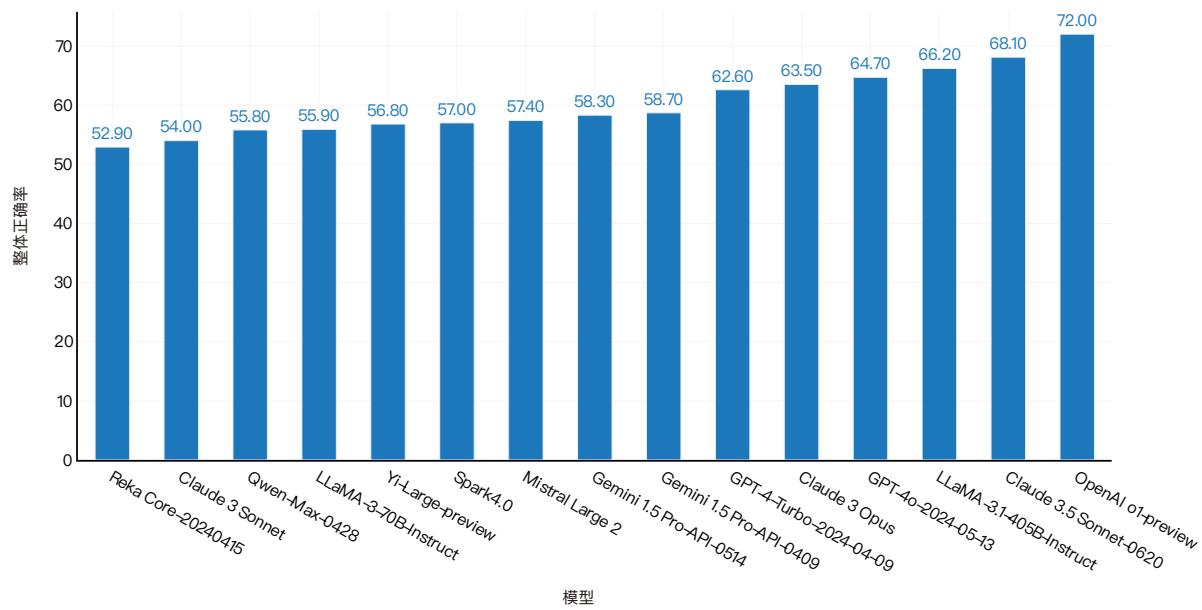


图 2.2.18

第二章：技术性能

2.2 语言

MTEB: 大规模文本嵌入式比较基准

大规模文本嵌入式比较基准（MTEB）平台由 Hugging Face 和 Cohere 公司的团队创建，于 2022 年底推出，旨在全面评估模型在各种嵌入任务中的技术性能。嵌入涉及将数据（如词、文本或文档）转换为数字向量，以捕捉大致的语义和向量之间的距离。嵌入是 RAG 的重要组成部分。在 RAG 任务中，当用户输入查询时，模型将其变换为嵌入向量。

这种 Transformers 可使模型搜索相关信息。MTEB 包括跨越 112 种语言的 58 个数据集和 8 个嵌入任务（图 2.2.19）。¹⁰ 例如，在 bitext 挖掘任务中，有两组来自两种不同语言的句子，对于第一组中的每个句子，模型的任务是在第二组中找到最佳匹配。

MTEB 比较基准的任务

资料来源：Muennighoff 等，2023

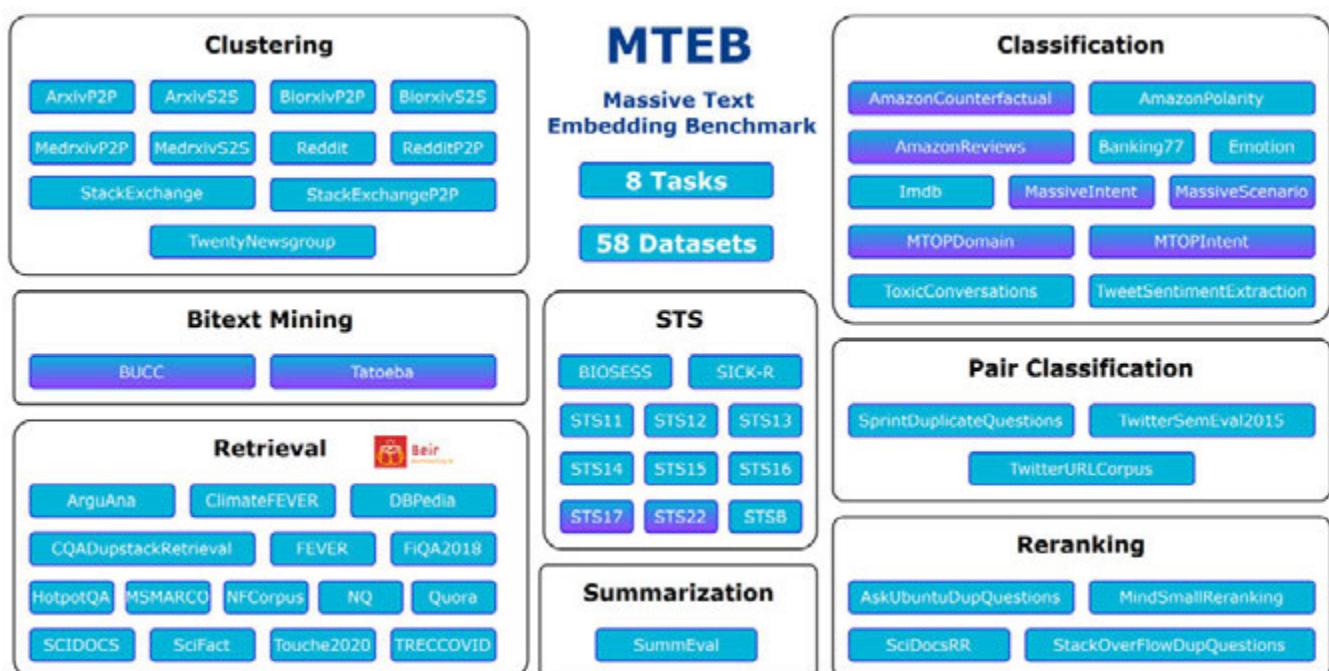


图 2.2.19

10. 基准覆盖的 8 类任务包括：双语文本挖掘、分类、聚类、配对分类、重排序、检索、语义文本相似度和摘要。各任务细节详见 MTEB 论文。

第二章：技术性能

2.2 语言

截至 2025 年初，MTEB 比较基准中表现最佳的嵌入模型是 Voyage AI 的 voyage-3-m-exp，得分为 74.03 分。Voyage AI 专注于打造高品质的 AI 嵌入模型。voyage-3-m-exp 是基于 voyage-3-large（专为嵌入任务设计的大规模基础模型）的变体，采用套娃表征学习（Matryoshka Representation

Learning）和量化感知训练等策略优化性能。该模型以微弱优势超越 NV-Embed-v2 (72.31 分)，后者在 2024 年大部分时间占据榜首（图 2.2.20）。MTEB 基准于 2022 年底首次推出时，领先模型的平均分仅为 59.5 分。因此，在过去两年中，比较基准的成绩显著提高。

MTEB 英语子集（56 个数据集）平均得分

资料来源：MTEB Leaderboard, 2025 | 图表：2025 年人工智能指数报告

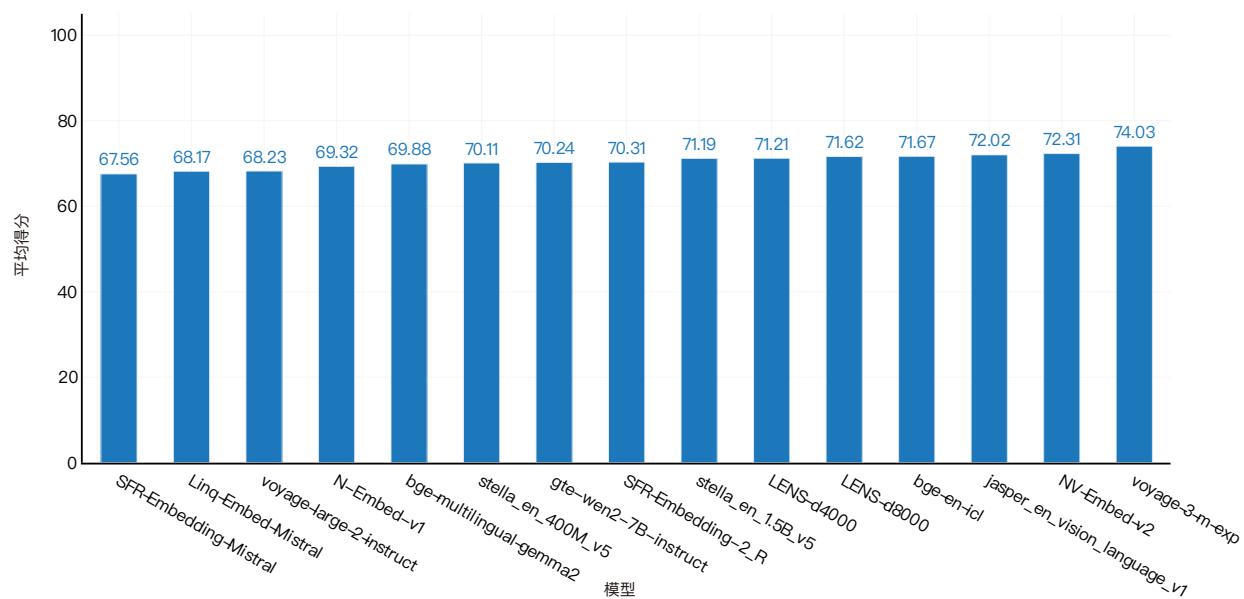


图 2.2.20

第二章：技术性能

2.2 语言

重点：

长上下文检索评估

随着人工智能模型的进步，它们处理更长上下文的能力也显著提高。例如，OpenAI 和 Meta 在 2023 年发布的 GPT-4 和 Llama 2 模型，其上下文窗口分别为 8,000 和 4,000 个词元。相比之下，更近期的模型如 GPT-4o（2024 年 5 月）和 Gemini 2.0 Pro Experimental（2025 年 2 月）的上下文窗口范围已扩展至 12.8 万至 200 万。这些扩展的上下文窗口使用户能够输入并处理越来越大的数据量，从而实现更复杂、更详细的交互。

随着大语言模型的上下文窗口不断扩大，评估其在长语境中的技术性能也变得越来越重要。然而，现有的长语境评估方法相对有限。通常情况下，这些评估侧重于“大海捞针”的场景，即要求模型从冗长的文本中检索特定的信息片段。这些评估虽然有用，但只能提供对模型的基础评估。

2024 年，为解决长上下文模型评估的局限性并改进其评估，推出了几款新的评估套件。其中一个比较基准是英伟达在 2024 年推出的 RULER，其通过检索性能、多跳推理、信息聚合和问答等任务综合评估长上下文性能。在 RULER 测试中，Gemini-1.5-Pro 以 95.5 分位居榜首，GPT-4 (89.0 分) 和 GLM4 (88.0 分) 次之（图 2.2.21）。研究还发现，多数宣称支持 32K 词元上下文的模型实际有效长度不足宣称值的一半（图 2.2.22）。事实上，RULER 团队证明，虽然大多数流行的大语言模型都宣称上下文大小为 32K 词元或更大，但其中只有一半能在 32K 长度时保持令人满意的技术性能。这意味着它们的实际操作上下文窗口比开发者宣称的要短（图 2.2.22）。

RULER 加权平均得分（递增）

资料来源：Hsieh 等，2024 | 图表：2025 年人工智能指数报告

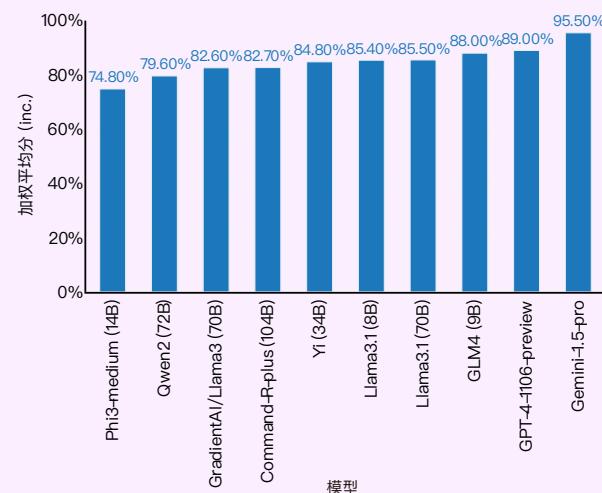


图 2.2.21

ULER 宣称和有效上下文长度的对比

资料来源：Hsieh 等，2024 | 图表：2025 年人工智能指数报告

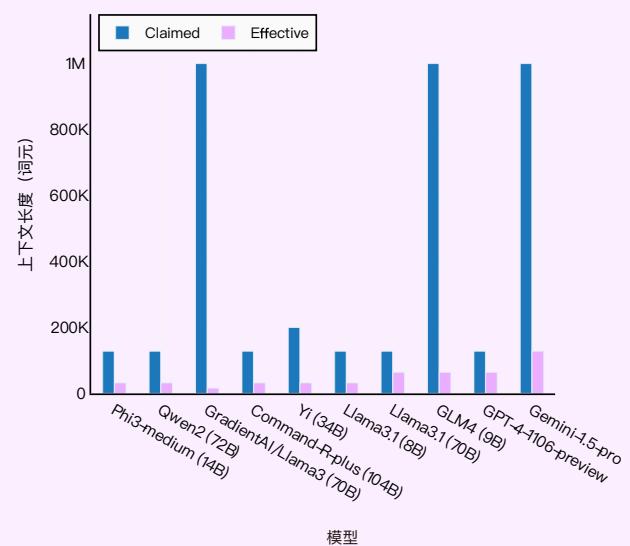


图 2.2.22

第二章：技术性能

2.2 语言

重点：

长上下文检索评估(续)

HELMET（长上下文模型高效全面评估框架）是英特尔与普林斯顿大学 2024 年联合推出的另一个长上下文评估基准。其开发动机源于现有基准的三大缺陷：下游任务覆盖不足、上下文长度过短无法测试进阶能力，以及评估指标不可靠（图 2.2.23）。相比 RULER，HELMET 更全面，包含合成召回、段落重排序、引用生成等 7 类长上下文评估任务。图 2.2.24 展示了

多款知名模型在 8K、32K 和 128K 上下文设置下的 HELMET 平均表现。虽然 GPT-4、Claude 3.5 Sonnet 和 Llama 3.1-70B 等模型在较长的上下文场景中性能下降，但其他模型，如 Gemini 1.5 Pro 和 GPT-4 八月版，则保持了技术性能。RULER 和 HELMET 等比较基准的推出显示了大语言模型的快速发展如何迫使研究人员重新思考和完善评估方法。

长上下文基准对比

资料来源：Yen 等，2024

图 2.2.23

	Type of tasks							Benchmark features		
	Cite	RAG	Re-rank	Long-QA	Summ	ICL	Synthetic Recall	Robust Eval.	$L \geq 128k$	Controllable L
ZeroSCROLLS	✗	✗	✗	✓	✓	✗	✗	✗	✗ [†]	✗
LongBench	✗	✓	✗	✓	✓	✓	✓	✗	✗ [†]	✗
L-Eval	✗	✓	✗	✓	✓	✗	✗	✓ [‡]	✗ [†]	✗
RULER	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓
oOBENCH	✗	✗	✗	✓	✓	✗	✓	✗	✓	✓
HELMET (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

HELMET：平均得分

资料来源：Yen 等，2024 | 图表：2025 年人工智能指数报告

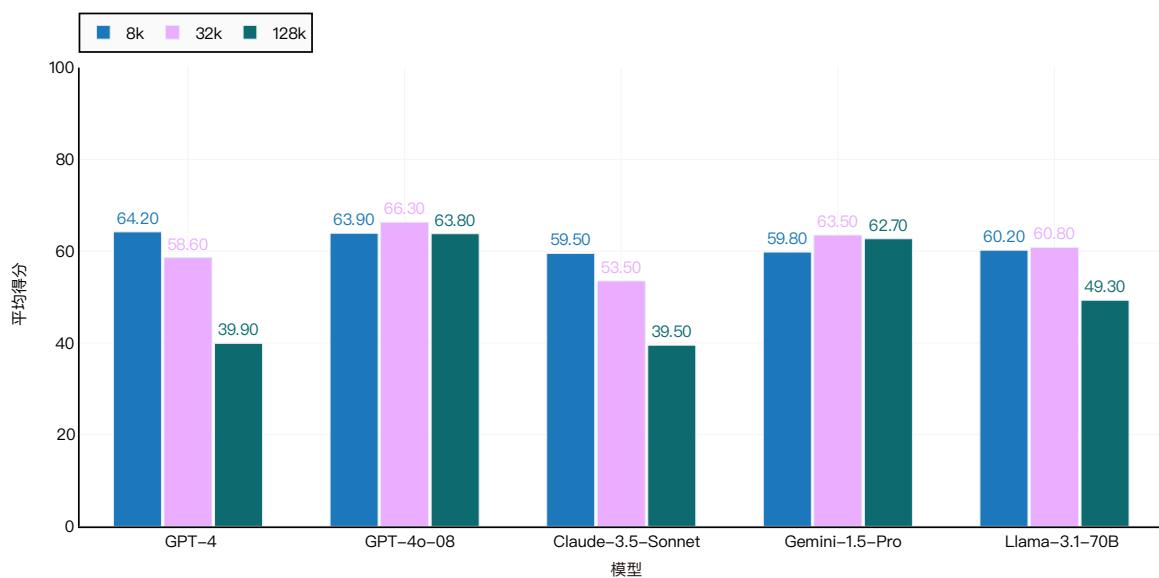


图 2.2.22

第二章：技术性能

2.3 图像与视频

计算机视觉使机器能理解图像 / 视频内容，并根据文本或其他提示生成逼真的视觉输出。该技术广泛应用于自动驾驶、医学影像和游戏开发等领域。

2.3 图像与视频

理解能力

视觉模型通过理解图像 / 视频内容并进行推理的能力接受评估。视觉理解是深度学习时代最早被广泛测试的人工智能能力之一，李飞飞创建的 ImageNet（在过往的人工智能指数报告中被深度阐述）曾是图像理解的基础基准。随着技术进步，研究人员将评估重点转向更复杂的任务，如视频理解或图像常识推理。

在 ImageNet 数据集时代，视觉算法的任务较为直接（例如，将图像归入预定义的类别）。而现代基准如 VCR 和 MVBench 则提出开放式挑战，其中不存在固定的分类或类别。在这种情况下，模型需处理自然语言问题，从开放图像集中识别对象，并基于图像内容或前期知识生成答案。

VCR 挑战赛示例问题

资料来源：Zellers 等，2018



How did [person2] get the money that's in front of her?

- a) [person2] is selling things on the street.
- b) [person2] earned this money playing music.
- c) She may work jobs for the mafia.
- d) She won money playing poker.

/ chose b
because...

- a) She is playing guitar for money.
- b) [person2] is a professional musician in an orchestra.
- c) [person2] and [person1] are both holding instruments, and were probably busking for that money.
- d) [person1] is putting money in [person2]'s tip jar, while she plays music.

第二章：技术性能

2.3 图像与视频

VCR比较基准是人工智能指数中为数不多的几个人工智能系统一直未达到人类基准水平的比较基准之一。然而,2024年是一个转折点,人工智能系统终于达到了这个基准。2024年7月,leaderboard上出现了一个模型,得分达到85.0,

与人类比较基准相匹配(图2.3.2)。这一里程碑标志着自2023年以来比较基准成绩提高了4.2%。甚至以前难以达到的比较基准现在也已超越。

视觉常识推理（VCR）任务：Q->AR 分数

资料来源：VCR Leaderboard, 2025 | 图表：2025年人工智能指数报告

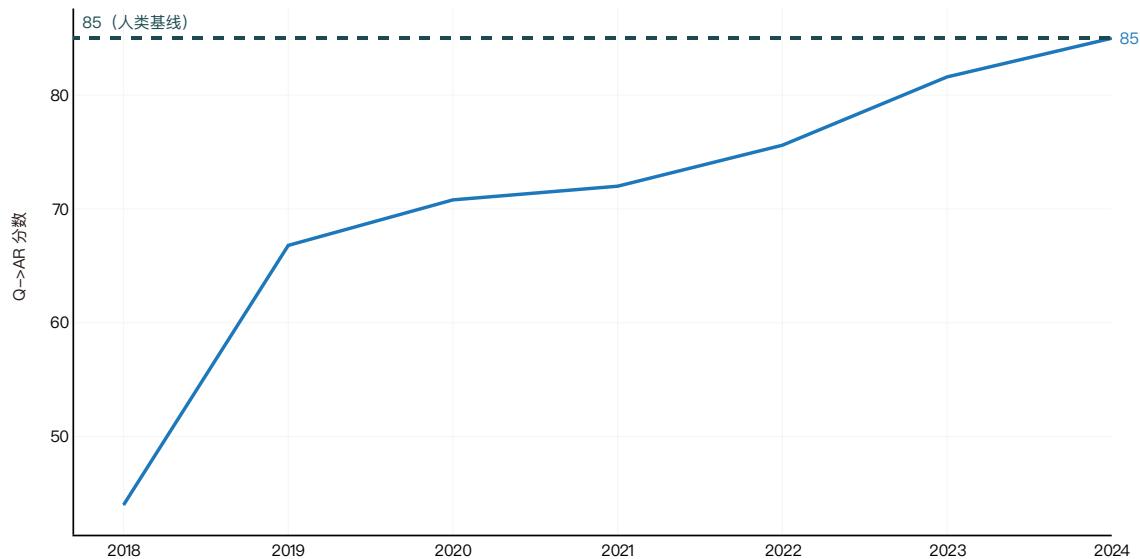


图 2.3.2

MVBench

由香港与中国大陆研究团队于2023年推出的MVBench是挑战性多模态视频理解基准。¹¹与早期仅测试空间理解的静态图像任务不同,MVBench包含需跨帧时序推理的复杂视频任务(图2.3.3)。

MVBench 任务示例

资料来源：Li等,2023

图 2.3.3



11、研究人员分别来自中国科学院、中国科学院大学、上海人工智能实验室、香港大学、复旦大学及南京大学

第二章：技术性能

2.3 图像与视频

截至2024年，MVBench leaderboard榜首为基于Qwen2.5-7B-Instruct语言模型构建的Video-CCAM-7B-v1.2，其得分为69.23，较2023年底引入该比较基准以来，提高了14.6%（图2.3.4）。这些结果凸显了人工智能模型在

动态视频理解能力方面取得的缓慢但稳步的进展。这些结果凸显了人工智能模型在动态视频理解能力方面取得的缓慢但稳步的进展。

MVBench: 平均准确率

资料来源：MVBench Leaderboard, 2025 | 图表：2025年人工智能指数报告

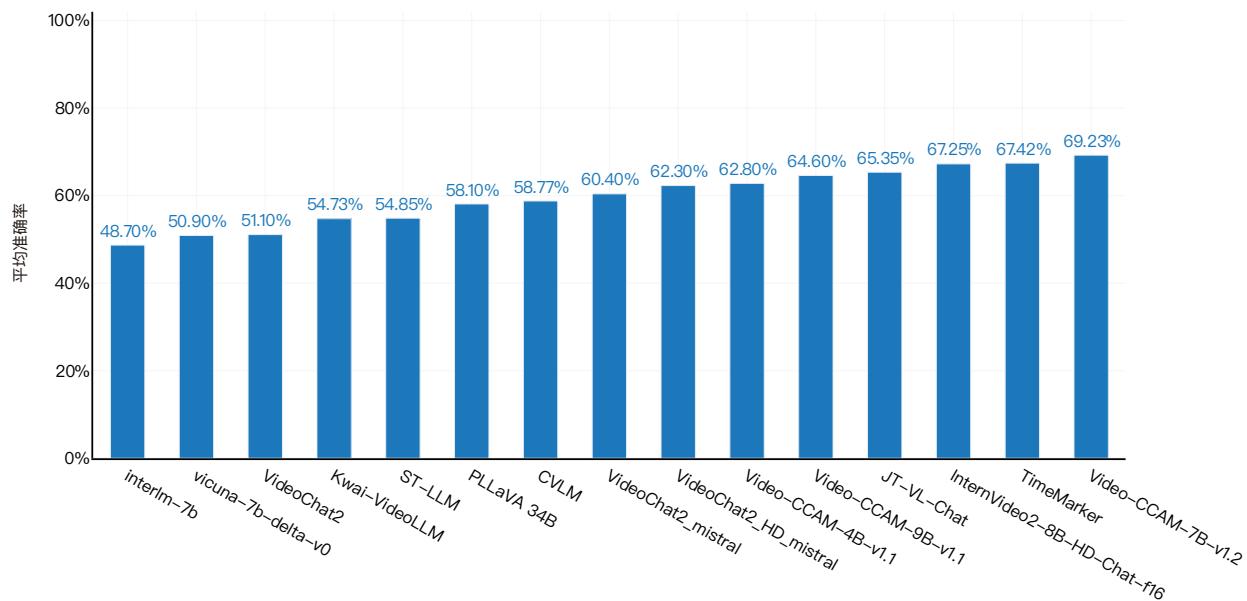


图 2.3.4

第二章：技术性能

2.3 图像与视频

生成能力

图像生成旨在创建与真实图像难以区分的合成内容。如去年人工智能指数所述，当前图像生成器已高度成熟，多数人难以分辨人工智能生成的人脸图像与真实照片（图 2.3.6）。图 2.3.6 展示了从 2022 年到 2025 年，针对“哈利·波特”的超逼真图像”这一提示，不同 Midjourney 模型版本生成的图像。这表明 Midjourney 在两年时间里生成超逼真图像的能力有了显著提高。在 2022 年，该模型生成的哈利·波特形象是卡通的、不准确的，但到了 2025 年，它可以生成呈现惊人真实的渲染效果。

哪张脸是真实的？

资料来源：Which Face Is Real, 2024·



图 2.3.5

Midjourney 版本迭代：“超写实哈利·波特图像”

资料来源：Midjourney, 2024



图 2.3.6

第二章：技术性能

2.3 图像与视频

Chatbot Arena: 视觉

人工智能社区越来越接受公共评估平台，例如 Chatbot Arena Leaderboard，以评估包括顶级人工智能图像生成器在内的领先人工智能系统的性能。该 leaderboard 还设有视觉 leaderboard，对 50 多个视觉模型的性能进行排名。用户可以提交文本到图像的提示，例如“蝙蝠侠在喝咖啡”，并为他们喜欢的生成结果投票（图 2.3.7）。截至目前，视觉竞技场已获得超过 15 万次投票。

截至 2025 年初，leaderboard 上排名第一的视觉模型是谷歌的 Gemini-2.0-Flash-Thinking-Exp-1219（图 2.3.8）。与其他 Chatbot Arena 类别（如整体、编码和数学）类似，领先模型的性能非常接近。例如，排名第一的模型与排名第四的模型 ChatGPT-4o-latest（2024-11-20）之间的差距仅为 3.4%。

Chatbot Arena 视觉板块示例

资料来源：Chatbot Arena Leaderboard, 2025

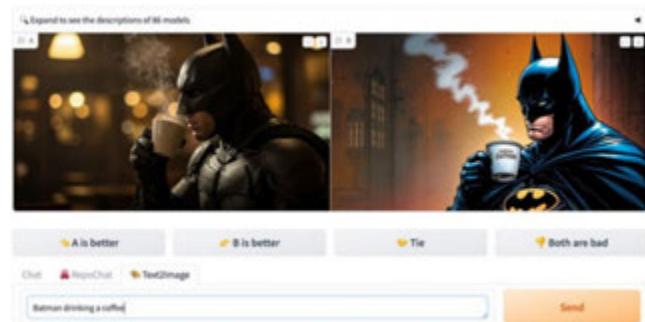


图 2.3.7

LMSYS Chatbot Arena 大语言模型 Elo 评分（视觉）

资料来源：LMSYS, 2025 | 图表：2025 年人工智能指数报告

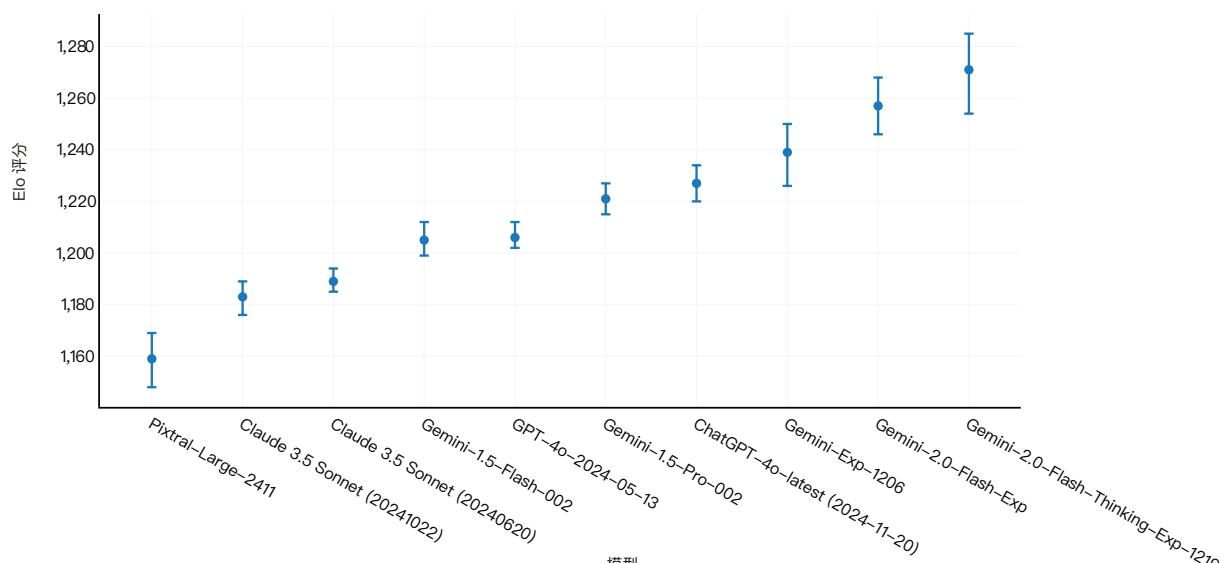


图 2.3.8

第二章：技术性能

2.3 图像与视频

重点：

视频生成的崛起

正如去年的人工智能报告所强调的，近年来，能够根据文本提示生成视频的模型逐渐兴起。早期的模型虽然展现出一定的潜力，但仍存在明显的局限性，例如生成的视频质量较低、缺少声音，或仅能生成非常短的片段。然而，2024 年标志着 AI 视频生成技术的重大飞跃，多家行业领先企业发布了先进的视频生成系统。

2023 年 11 月，Stability AI 推出了其 Stable Video Diffusion 模型，这是该公司首个能够生成高质量视频的基础模型（图 2.3.9）。该模型采用三步流程：文本到图像的预

训练、视频预训练以及高质量视频的微调。随后，2024 年 3 月，Stability AI 发布了 Stable Video 3D 模型，该模型能够从单张图像生成物体的多个 3D 视角和视频。2024 年 2 月，OpenAI 推出了其视频生成模型 Sora 的预览版，并于 2024 年 12 月正式向公众开放。Sora 能够生成长达 20 秒、分辨率高达 1080p 的视频（图 2.3.10）。作为一种扩散模型，它首先生成一个基础视频，随后通过多步去噪逐步优化，以提升视频质量。

Stable Video Diffusion 生成的静态帧

资料来源：Stability AI, 2025

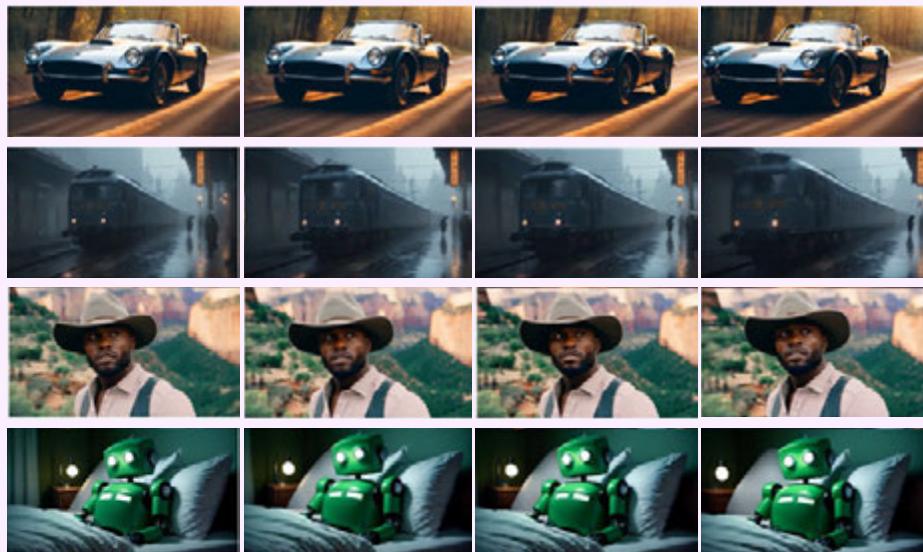


图 2.3.9

Sora 生成的静态帧

资料来源：OpenAI, 2024

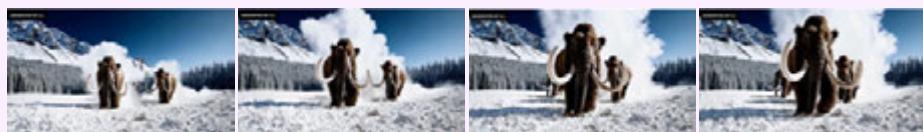


图 2.3.10

第二章：技术性能

2.3 图像与视频

重点：

视频生成的崛起(续)

其他科技巨头也纷纷进入视频生成领域。2024 年 10 月，Meta 发布了其 Movie Gen 模型的最新版本。与早期版本不同，新版 Movie Gen 具备基于指令的高级视频编辑功能、从图像生成个性化视频的能力，并支持为视频添加声音。Meta 最先进的 Movie Gen 模型能够生成 16 秒、每秒 16 帧、分辨率为 1080p 的视频。谷歌也在 2024 年取得了显著进展，推出了两款重要的视频生成模型：5 月发布的 Xeo 和 12 月发布的 Xeo_2。谷歌的内部比较基准显示，Veo 2 在性能上优于其他领先的视频生成模型，如 Meta 的 MovieGen、Kling v1.5 和 Sora Turbo。在用户对比中，Veo 2 生成的视频普遍更受青睐(图 2.3.11)。

Veo 2: 整体偏好

资料来源：DeepMind, 2024 | 图表：2025 年人工智能指数报告

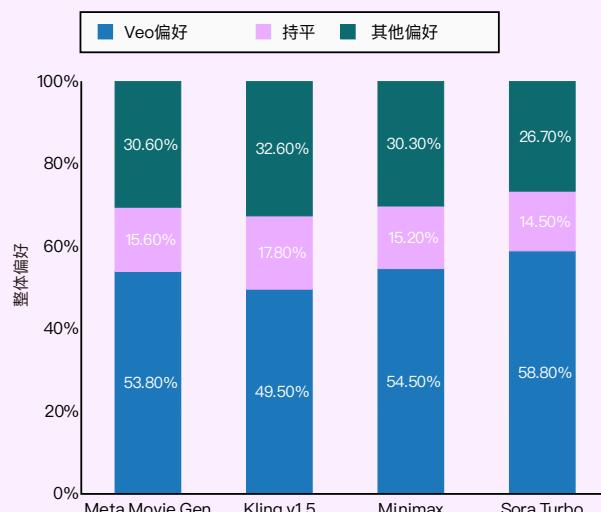


图 2.3.11

规模相对较小的玩家也在视频生成领域做出了显著贡献，例如 Runway 的 Gen-3 Alpha、Luma 的 Dream Machine 和快手的 Kling 1.5。通过对比 2023 年和 2024 年生成的视频，可以清晰地看到这一领域的显著进步。互联网上流行的提示词“威尔·史密斯吃意大利面”生动地展示了这一进步：2025 年由热门视频生成器 Pika 生成的视频，其质量较 2023 年的版本有了显著提升(图 2.3.12)。

威尔·史密斯吃意大利面，2023 年和 2025 年对比

资料来源：Pika, 2025



图 2.3.12

第二章：技术性能

2.4 语音

人工智能系统在人类语音处理方面表现卓越，其音频能力包括将口语转录为文本以及识别个体说话者。近年来人工智能在生成合成音频内容方面也取得了显著进展。

2.4 语音

语音识别

语音识别是指人工智能系统识别口语并将其转换为文本的能力。语音识别技术已取得长足进步，如今许多计算机程序和短信应用均配备了听写设备，能够可靠地将语音转录为文字。

LSR2: Lip Reading Sentences 2

牛津大学与 BBC 联合发布的 Lip Reading Sentences 2 (LRS2) 数据集于 2017 年推出，是目前最全面的公开唇读数据集之一，专注于真实场景下的应用（图 2.4.1）。该数据集包含来自各类脱口秀和新闻节目的视听片段。在自动语音识别 (ASR) 任务中，系统的语音转录能力通过词错误率 (WER) 进行评估，分数越低表示转录越精确。

BBC lip reading sentences 2 数据集中的静态图像

资料来源：Chung 等，2024



图 2.4.1

第二章：技术性能

2.4 语音

今年，Whisper–Flamingo 模型在 LRS2 比较基准中创下新标准，词错误率(WER)达到 1.3%，超越了 2023 年 1.5% 的最先进水平(图 2.4.2)。然而，由于当前 WER 已处于极低水平，

进一步显著提升的可能性较小，表明该比较基准可能已接近饱和。

LRS2: 词错误率 (WER)

资料来源：Papers With Code, 2025 | 图表：2025 年人工智能指数报告

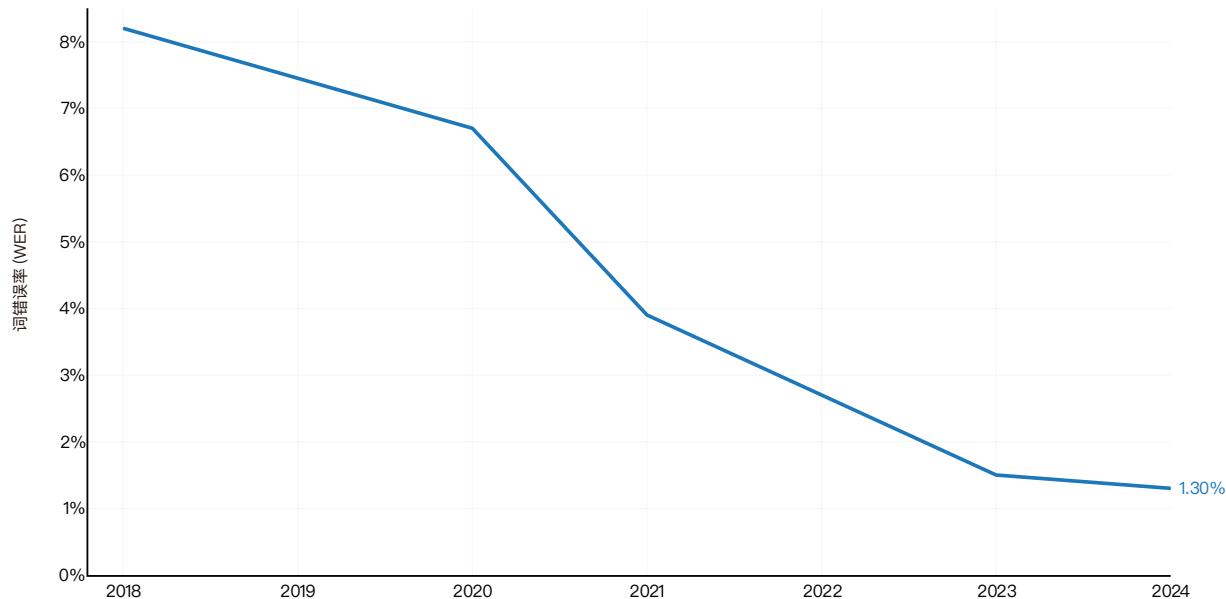


图 2.4.2

第二章：技术性能

2.5 编程

编程涉及生成计算机可执行的指令以完成任务。近年来，大语言模型已成为熟练的编程助手，为计算机科学家提供了重要支持。越来越多的证据表明，许多程序员认为人工智能编程助手非常实用。正如去年的人工智能指数报告所强调的，大语言模型的编程能力日益精进，以至于许多基础编程比较基准（如 HumanEval）正逐渐趋于饱和。为此，研究人员已将重点转向测试大语言模型在更复杂编程挑战中的表现。

2.5 编程

HumanEval

HumanEval 是 Open 人工智能研究人员于 2021 年推出的比较基准，通过 164 道手工编写的编程题目评估人工智能系统的编码能力（图 2.5.1）。目前 HumanEval 性能领先的模型是 Claude 3.5 Sonnet (HPT)，其得分达到了 100%（图 2.5.2）。

HumanEval 示例题目

资料来源：Chen 等，2023

```
def incr_list(l: list):
    """Return list with elements incremented by 1.
    >>> incr_list([1, 2, 3])
    [2, 3, 4]
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [6, 4, 6, 3, 4, 4, 10, 1, 124]
    """
    return [i + 1 for i in l]
```

图 2.5.1

HumanEval: Pass@1

资料来源：Papers With Code, 2025 | 图表：2025 年人工智能指数报告

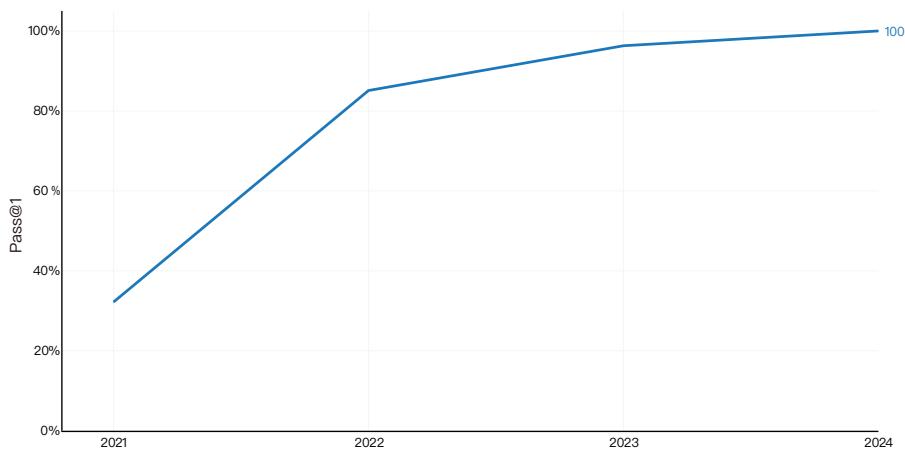


图 2.5.2

第二章：技术性能

2.5 编程

SWE-bench

2023年10月，普林斯顿大学和芝加哥大学的研究人员推出了SWE-bench数据集，该数据集包含2,294个源自真实GitHub问题和热门Python代码库的软件工程问题（图2.5.3）。SWE-bench为AI编程能力提供了更高难度的测试，要求系统跨多个函数协调修改、适应不同执行环境并进行复杂推理。SWE-bench包含一个经过筛选的Lite子集以简化评估，以及一个由人工标注的Verified子集。以下图表展示了Verified子集的得分。

SWE-bench突显了大语言模型在曾被视为极具挑战性任务上的快速进步。2023年底，SWE-bench上表现最佳的模型得分仅为4.4%。而到2025年初，排名第一的模型——OpenAI的o3模型——在Verified基准集上成功解决了71.7%的问题（图2.5.4）。这一显著性能提升表明，人工智能研究人员可能很快需要开发更具挑战性的编程比较基准来有效测试大语言模型。

SWE-bench示例模型输入

资料来源：Jimenez等，2023

```
Model Input
```

▼ Instructions • 1 line
You will be provided with a partial code base and an issue statement explaining a problem to resolve.

▼ Issue • 67 lines
napoleon_use_param should also affect "other parameters" section Subject: napoleon_use_param should also affect "other parameters" section

Problem
Currently, napoleon always renders the Other parameters section as if napoleon_use_param was False, see source

```
def _parse_other_parameters_section(self, section):
    # type: (unicode) -> List[unicode]
    fields = self._consume_fields()
    if self._config.napoleon_use_param: ...
```

def _parse_parameters_section(self, section):
 # type: (unicode) -> List[unicode]
 fields = self._consume_fields()
 if self._config.napoleon_use_param: ...

▼ Code • 1431 lines
► README.rst • 132 lines
► sphinx/ext/napoleon/docstring.py • 1295 lines
► Additional Instructions • 57 lines

图2.5.3

SWE-bench：问题解决率

资料来源：SWE-bench Leaderboard, 2025; OpenAI, 2024 | 图表：2025年人工智能指数报告

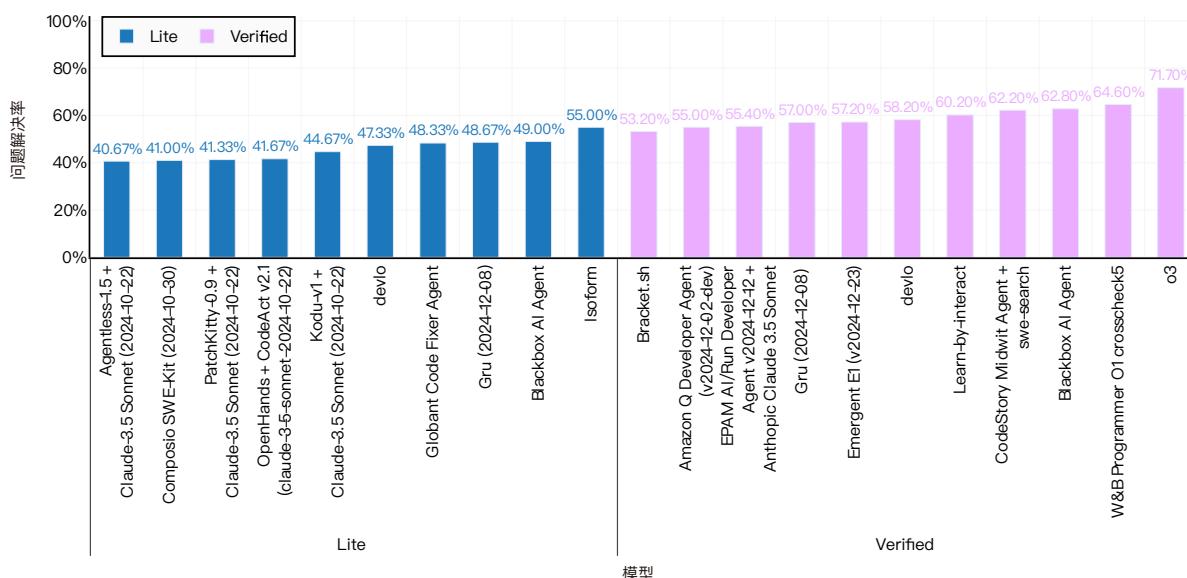


图2.5.4

第二章：技术性能

2.5 编程

BigCodeBench

现有编程比较基准的一个局限性在于，许多测试仅局限于短小、自包含的算法任务或独立函数调用。然而，解决复杂实际任务通常需要调用多样化函数的能力（如数据分析工具或网页开发工具）。高效的编程还要求模型能理解自然语言表达的编码指令——这一能力未被当前多数编程基准所测试。

为弥补现有编程基准的不足，一个国际团队于 2024 年发布了 BigCodeBench——一个全面、多样且极具挑战性的编程比较基准（图 2.5.5）。该基准要求大语言模型跨 139 个库和 7

大领域调用多重函数调用，涵盖 1,140 项细粒度任务。当前人工智能系统在该基准上表现欠佳：即使在基准的“完整”任务（基于结构化文档字符串的代码补全）和“指令”任务（基于自然语言指令的代码补全）的困难子集上，当前最优模型（OpenAI 的 o1）平均得分仅 35.5（图 2.5.6）。模型在基准全集上表现略优（图 2.5.7）。BigCodeBench 凸显了人工智能系统在实现人类级编程熟练度方面仍存在的差距。

BigCodeBench 中的编程任务

资料来源：Zhuo 等，2024

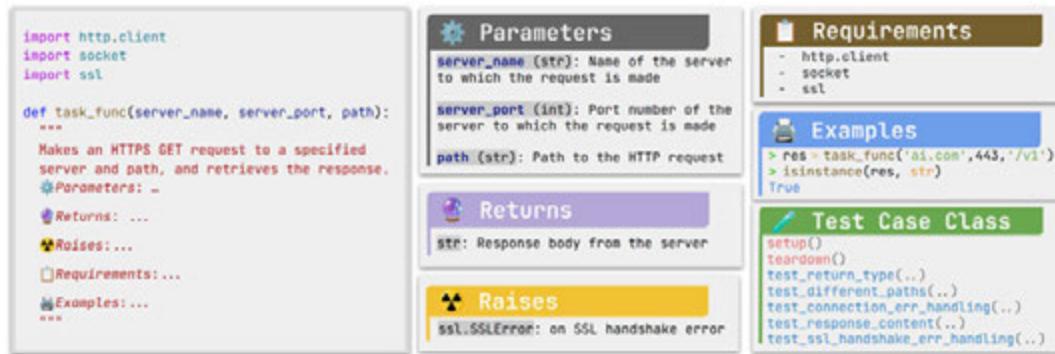


图 2.5.5

BigCodeBench 困难集测试：Pass@1 (平均得分)

资料来源：Hugging Face, 2025 | 图表：2025 年人工智能指数报告

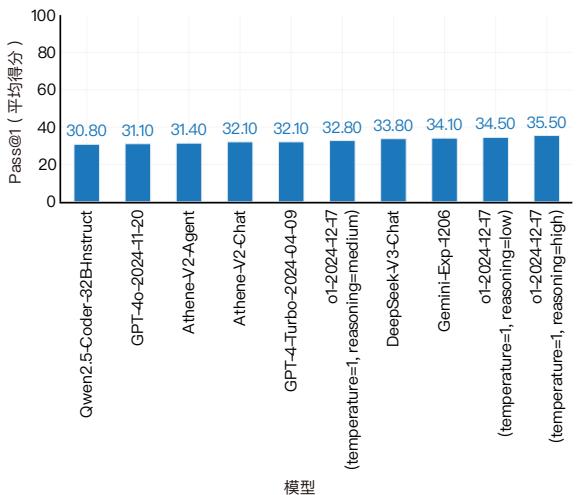


图 2.5.6

BigCodeBench 全集测试：Pass@1 (平均得分)

资料来源：Hugging Face, 2025 | 图表：2025 年人工智能指数报告

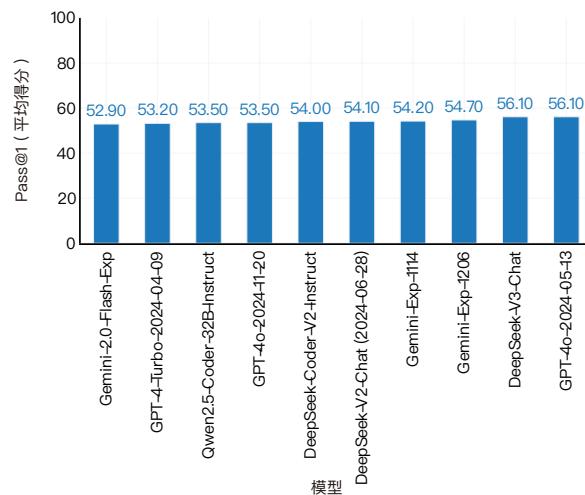


图 2.5.7

第二章：技术性能

2.5 编程

Chatbot Arena: 编程能力测评

Chatbot Arena 的大语言模型 leaderboard 新增了编程能力筛选功能，为开发者及社区评估不同模型的编程能力提供了宝贵参考。这种公开反馈为评估模型性能增添了新的维度。

目前评分最高的编程专用大语言模型是 Gemini-Exp-1206，其竞技场得分为 1,369 分，OpenAI 的最新 o1 模型以 1,361 分紧随其后。中国模型中，DeepSeek-V3 以 1,317 分领先，与榜首相差 3.8%（图 2.5.8）。

LMSYS Chatbot Arena 关于大语言模型的 Elo 评分 (编程)

资料来源：LMSYS, 2025 | 图表：2025 年人工智能指数报告

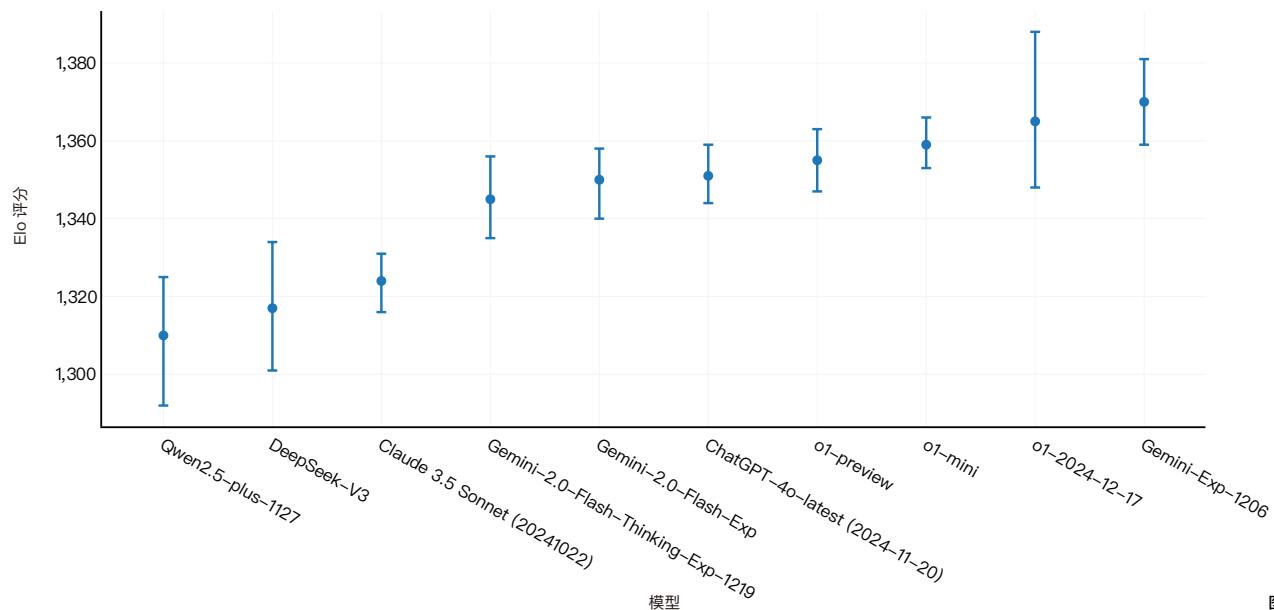


图 2.5.8

第二章：技术性能

2.6 数学

数学解题比较基准用于评估人工智能系统的数学推理能力，测试范围涵盖从小学水平到竞赛标准的各类数学问题。

2.6 数学

GSM8K

GSM8K 是 OpenAI 于 2021 年推出的数据集，包含约 8,000 道多样化的小学数学应用题，要求人工智能模型通过多步算术运算生成解决方案（图 2.6.1）。与 MMLU 一样，GSM8K 已成为评估先进的大语言模型的常用比较基准。但近期该基准可能面临数据污染和性能饱和的问题。

GSM8K 上表现最佳的模型是采用 HPT 提示策略优化的 Claude Sonnet 3.5 变体，其准确率达到 97.72%（图 2.6.2），较 2023 年的 91.00% 有显著提升。然而 2024 年，Mistral、Meta 和 Qwen 的多个模型得分均接近 96%，表明 GSM8K 基准可能已趋近饱和。

GSM8K 问题示例

资料来源：Cobbe 等，2023

Problem: Beth bakes 4 2-dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?
Solution: Beth bakes 4 2-dozen batches of cookies for a total of $4 \times 2 = 8$ dozen cookies.
 There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of $12 \times 8 = 96$ cookies.
 She splits the 96 cookies equally amongst 16 people so they each eat $96 \div 16 = 6$ cookies.
Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 82 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs \$3.50?
Solution: Mrs. Lim got 82 gallons - 18 gallons = $82 - 18 = 64$ gallons this morning.
 So she was able to get a total of 82 gallons + 82 gallons + 64 gallons = $82 + 82 + 64 = 228$ gallons.
 She was able to sell 200 gallons - 24 gallons = $200 - 24 = 176$ gallons.
 Thus, her total revenue for the milk is $\$3.50 \text{ per gallon} \times 176 \text{ gallons} = \441.00 .
Final Answer: 441

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each. 2 of the people have 4, and 1 person has 6. How many sodas are left over when the party is over?
Solution: Tina buys 3 12-packs of soda, for $3 \times 12 = 36$ sodas.
 If 6 people attend the party, so half of them is 3.
 $3 \times 3 = 9$ sodas
 Each of those people drinks 3 sodas, so they drink $3 \times 3 = 9$ sodas
 Two people drink 4 sodas, which means they drink $2 \times 4 = 8$ sodas
 With one person drinking 6, that brings the total drank to $9 + 8 + 6 = 23$ sodas.
 As Tina started off with 36 sodas, that means there are $36 - 23 = 13$ sodas left.
Final Answer: 13

图 2.6.1

GSM8K：准确率

资料来源：Papers With Code, 2024 | 图表：2025 年人工智能指数报告

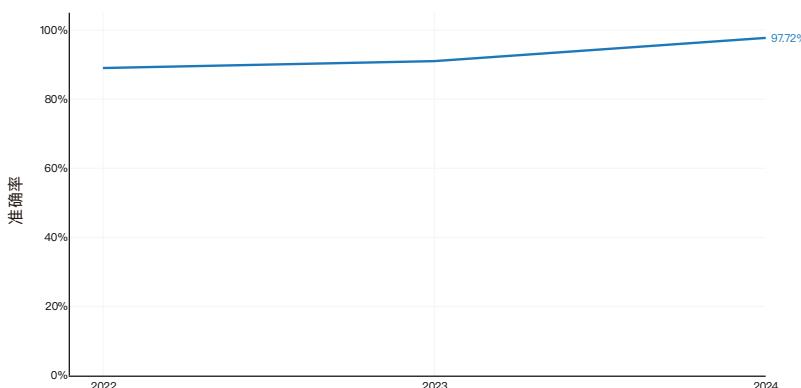


图 2.6.2

第二章：技术性能

2.6 数学

MATH

MATH 数据集是由加州大学伯克利分校和芝加哥大学的研究人员于 2021 年推出的数学问题集，包含 12,500 道具有挑战性的竞赛级题目(图 2.6.3)。该数据集发布之初，人工智能系统表现欠佳，仅能解决其中 6.9% 的问题。但此后性能显著提升：2025 年 1 月，OpenAI 发布的 o3-mini(高性能版)模型在 MATH 数据集上取得突破性进展，解题准确率高达 97.9% (图 2.6.4)。根据去年人工智能指数的评估，MATH 曾是少数几个人工智能系统尚未超越人类基线的基准之一，但这一现状已被改写。

来自 MATH 数据集的问题示例

资料来源：Hendrycks 等，Hendrycks 等，2023 年

Problem: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

Solution: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

Problem: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.

Solution: Complete the square by adding 1 to each side. Then $(x + 1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt[4]{2}$. The desired product is then $(-1 + \cos(\frac{\pi}{8})\sqrt[4]{2})(-1 - \cos(\frac{\pi}{8})\sqrt[4]{2}) = 1 - \cos^2(\frac{\pi}{8})\sqrt{2} = 1 - \frac{(1+\cos(\frac{\pi}{4}))}{2}\sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}$.

图 2.6.3

MATH 文字解题：准确率

资料来源：Papers With Code, 2024 年；OpenAI, 2025 年：《论文与代码》，2024 年；OpenAI, 2025 年 | 图表：2025 年人工智能指数报告

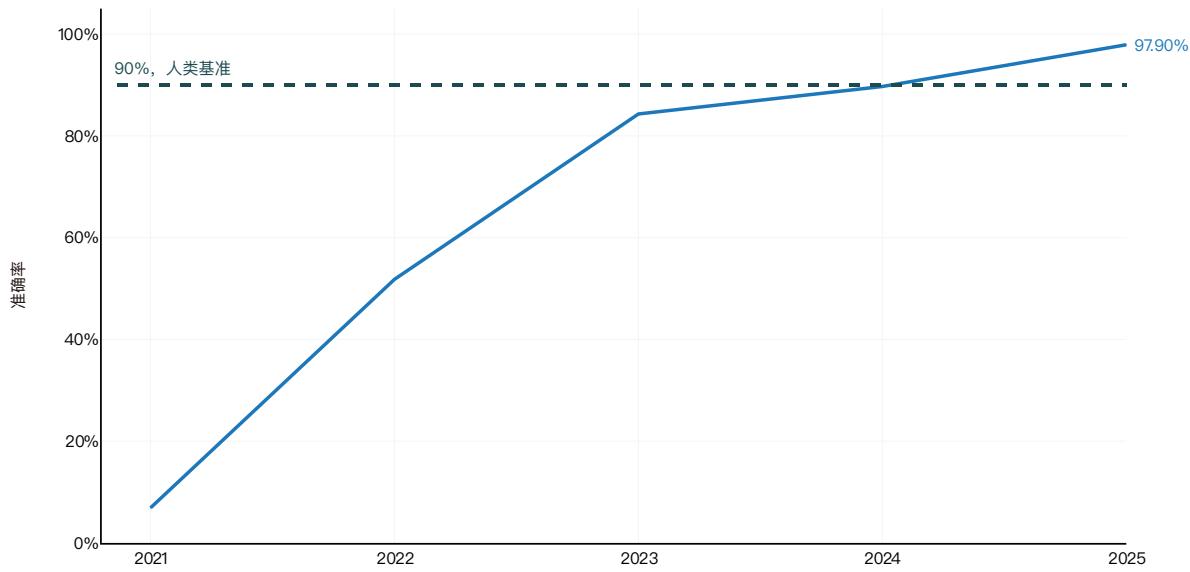


图 2.6.4

第二章：技术性能

2.6 数学

Chatbot Arena: 数学能力测评

Chatbot Arena 新增数学筛选功能，允许公众根据模型生成数学相关答案的表现进行排名。Math Arena 评估了超过 181 个模型，并收集了 34 万份公众投票。与通用和编程竞技

场中 Gemini 系列模型领先的情况不同，数学竞技场排名第一的是 OpenAI 于 2024 年 12 月发布的 o1 变体模型（图 2.6.5）。

LMSYS Chatbot Arena 大语言模型评分：Elo 评分（数学）

资料来源：LMSYS, 2025 | 图表：2025 年人工智能指数报告

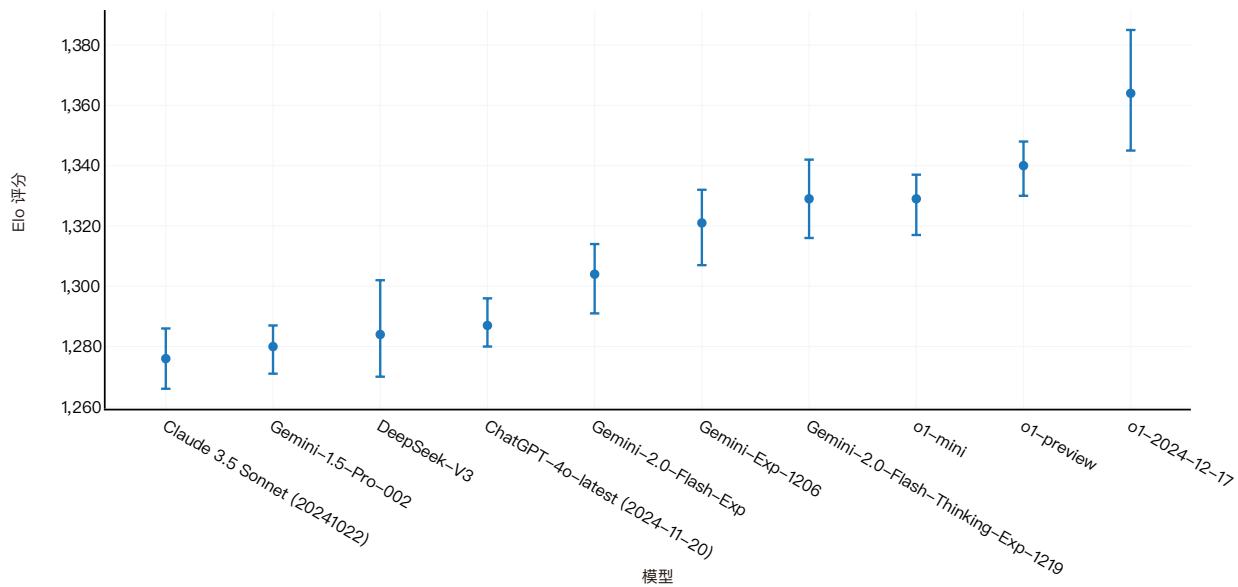


图 2.6.5

FrontierMath

数学界成员指出当前数学比较基准的局限性，呼吁开发新的比较基准以评估日益先进的人工智能系统。主要挑战在于饱和现象：人工智能系统在 GSM8K 和 MATH 等高中和大学水平数学测试中已接近完美表现。为进一步突破边界，研究人员提出需要测试真正高阶数学的基准，包括数论、实分析、代数几何和范畴论等问题。

FrontierMath 是 Epoch AI 推出的新基准，包含数百道原创高难度数学问题。这些问题由数学家团队审核，通常需要数

小时、数天甚至协作研究才能解决。图 2.6.6 展示了基准中的示例题目。Epoch 人工智能评估了六款领先大语言模型在 FrontierMath 上的表现：o1-preview、o1-mini、GPT-4o、Claude 3.5 Sonnet、Grok 2 Beta 和 Gemini 1.5 Pro 002。基准发布时，表现最佳的 Gemini 1.5 Pro 仅解决了 2.0% 的问题——远低于其在其他数学基准中的成绩（图 2.6.7）。而 OpenAI 的 o3 模型得分为 25.2%。FrontierMath 开发者希望该基准能在未来数年持续成为尖端人工智能系统的严峻挑战。

第二章：技术性能

2.6 数学

FrontierMath 示例问题

资料来源：Glazer 等，2024

Sample problem 1: Testing Artin's primitive root conjecture

Definitions. For a positive integer n , let $v_p(n)$ denote the largest integer v such that $p^v \mid n$.
 For p a prime and $a \not\equiv 0 \pmod{p}$, we let $\text{ord}_p(a)$ denote the smallest positive integer σ such that $a^\sigma \equiv 1 \pmod{p}$. For $x > 0$, we let

$$\text{ord}_{p,x}(a) = \prod_{\substack{q \leq x \\ q \text{ prime}}} q^{v_q(\text{ord}_p(a))} \prod_{\substack{q > x \\ q \text{ prime}}} q^{v_q(p-1)}.$$

Problem. Let S_x denote the set of primes p for which $\text{ord}_{p,x}(2) > \text{ord}_{p,x}(3)$, and let d_x denote the density

$$d_x = \frac{|S_x|}{|\{p \leq x : p \text{ is prime}\}|}$$

of S_x in the primes. Let

$$d_\infty = \lim_{x \rightarrow \infty} d_x.$$

Compute $\lfloor 10^6 d_\infty \rfloor$.

Answer: 367707

MSC classification: 11 Number theory

Sample problem 2: Find the degree 19 polynomial

Construct a degree 19 polynomial $p(x) \in \mathbb{C}[x]$ such that $X := \{(p(x) = p(y)) \subset \mathbb{P}^1 \times \mathbb{P}^1$ has at least 3 (but not all linear) irreducible components over \mathbb{C} . Choose $p(x)$ to be odd, monic, have real coefficients and linear coefficient -19 and calculate $p(19)$.

Answer: 1876572071974094803391179

MSC classification: 14 Algebraic geometry; 20 Group theory and generalizations; 11 Number theory generalizations

图 2.6.6

FrontierMath: 问题解决率

资料来源：Glazer 等，2024; OpenAI, 2025 | 图表：2025 年人工智能指数报告

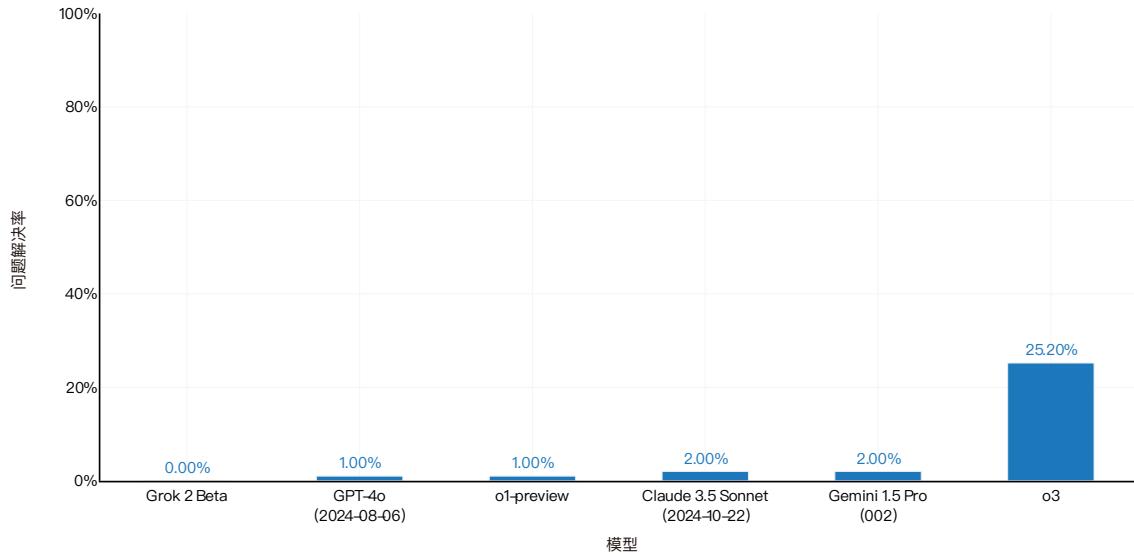


图 2.6.7

第二章：技术性能

2.6 数学

重点：

学习与定理证明

DeepMind 运用其系统 AlphaProof 和 AlphaGeometry 2，成功解决了 2024 年国际数学奥林匹克竞赛 (IMO) 6 道题目中的 4 道，表现达到银牌得主水平。在比较基准中，AlphaGeometry 解决了 30 道奥数几何题中的 25 道，超越了 IMO 银牌得主平均解决的 22.9 道(图 2.6.8)。IMO 成立于 1959 年，是全球历史最悠久、最具声望的青年数学家竞赛。

AlphaProof 是基于 AlphaZero 的强化学习系统，后者曾应用于国际象棋、将棋和围棋。该系统通过生成假设并利用 Lean 交互式证明系统验证假设来自主解决问题。此外，经过微调的 Gemini 模型被用于将自然语言问题陈述转化为形式化表示，从而构建了全面的训练库。在本届竞赛中，AlphaProof 成功解决了 2 道代数题和 1 道数论题，但未能解决 2 道组合数学题。

AlphaGeometry 2 是一种神经符号混合系统，其语言模型基于 Gemini，并通过大量合成数据训练而成。2024 年之前，AlphaGeometry 能解决 83% 的历史 IMO 几何题。在 2024 年竞赛中，该系统仅用 24 秒便解决了唯一一道几何题。测试中，竞赛题目被人工翻译为 Lean 的形式化表示。

目前尚不清楚 AlphaProof 和 AlphaGeometry 在传统定理证明基准（如 ITP）上的表现。ITP 自 1997 年起用于评估自动定理证明(ATP)系统的性能，尤其是应用于软件验证的系统。人工智能指数在 2021 年报告中曾对 ATP 的现状进行过分析。2024 年更新的报告显示，基于包含超过 25,000 道题目的 TPTP v.9.0.0 版本，全自动系统现已能解

MO-AG-30 已解决几何问题的数量

资料来源：Trinh 等，2024 | 图表：2025 年人工智能指数报告

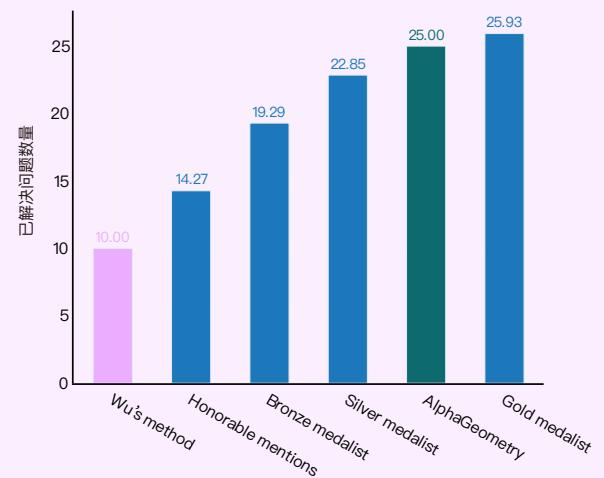


图 2.6.8

决其中 89% 的题目。

理想情况下，TPTP 系统可在 IMO 题目上测试，而 AlphaProof 和 AlphaGeometry 可在 TPTP 题目上测试——其中部分题目甚至从未被人类解决过，更不用说 ATP 系统。但这两类测试均未实施，主要原因是不同系统支持的逻辑差异显著，且目前缺乏转换工具。此外，TPTP 库虽规模庞大，但尚不足以作为 AlphaProof 的训练集，仍需生成大量合成示例。

第二章：技术性能

2.7 推理

人工智能的推理能力指人工智能系统从不同形式信息中得出逻辑有效结论的能力。当前，人工智能系统正越来越多地在多样化推理场景中接受测试，包括视觉推理（图像分析）、道德推理（理解道德困境）以及社会推理（应对社交情境）。

MMMU 问题样本

资料来源：Yue 等，2023

Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly?	Question: ... The graph shown is compiled from data collected by Gafap. Image 2 . Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?	Question: Image 2 . The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.
Options: (A) Major third Image 2 (B) Diminished fifth Image 2 (C) Minor seventh Image 2 (D) Diminished sixth Image 2	Options: (A) 0 (B) 0.2142 (C) 0.3575 (D) 0.5	Options: $\int_{a}^{b} [f(x) - g(x)] dx$ (B) $\int_{a}^{b} [g(x) - f(x)] dx$ (C) $\int_{a}^{b} [f(x) + g(x)] dx$ (D) $\int_{a}^{b} [g(x) + f(x)] dx$
Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium	Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium	Subject: Math; Subfield: Calculating Image Type: Mathematical Notations; Difficulty: Easy
Health & Medicine	Humanities & Social Science	Tech & Engineering
Question: You are shown subtraction Image 2 , T2 weighted Image 2 and T1 weighted axial Image 2 from a screening breast MRI. What is the etiology of the finding in the left breast?	Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? Image 1	Question: Find the VCE for the circuit shown in Image 2 . Neglect VBE.
Options: (A) Susceptibility artifact (B) Hematoma (C) Fat necrosis (D) Silicone granuloma	Options: (A) Oppressor (B) Imperialist (C) Savior (D) Isolationist	Answer: A28 Explanation: $I = V/(R_1 + R_2)$ $I = (5 \text{ V}) / (4 \text{ k} \cdot \text{ohm}) = 1.25 \text{ mA}$ $VCE = VCC - IR_2 = 10 \text{ V} - (1.25 \text{ mA}) 5 \text{ k} \cdot \text{ohm}$ $VCE = 10 \text{ V} - 6.25 \text{ V} = 3.75 \text{ V}$
Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans; MRI, CT; Difficulty: Hard	Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy	Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard

图 2.7.1

2.7 推理

通用推理

通用推理指人工智能系统在广泛领域而非特定任务中进行推理的能力。

例如，在通用推理挑战中，人工智能系统可能需要跨学科推理，而非完成单一任务（如下棋）。

MMMU 在验证集上的表现：总体准确率

资料来源：MMMU Leaderboard, 2024 | 图表：2025 年人工智能指数报告

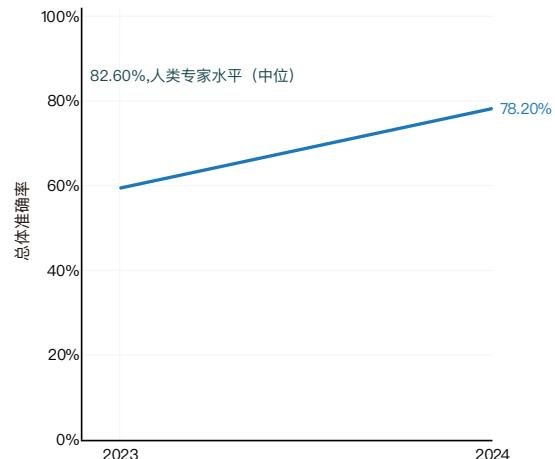


图 2.7.2

MMMU：面向专家级 AGI 的大规模多学科多模态理解与推理基准

基准

近年来，人工智能系统的推理能力快速提升，传统基准如 SQuAD（文本推理）和 VQA（视觉推理）已趋饱和，亟需更具挑战性的测试。

为此，美国和加拿大的研究人员开发了 MMLU（大规模多学科多模态理解与推理比较基准），专为专家级人工通用智能

（AGI）设计。MMMU 包含约 11,500 道大学水平的题目，涵盖六大核心学科：艺术与设计、商业、科学、健康与医学、人文与社会科学、技术与工程（图 2.7.1）。题目形式包括图表、地图、化学结构等。截至 2025 年 1 月，OpenAI 的 o1 模型以 78.2% 的准确率领先，较去年 59.4% 的最优成绩显著提升（图 2.7.2）。尽管该分数仍低于人类专家中高水平基准，但人工智能系统正迅速缩小差距。

第二章：技术性能

2.7 推理

GPQA: A Graduate-Level Google-Proof Q&A Benchmark

2023 年，来自纽约大学、Anthropic 和 Meta 的研究人员推出了 GPQA 比较基准，用于测试通用、多学科人工智能推理能力。该数据集由 448 道难以通过网络搜索轻松回答的难题组成。这些问题由生物学、物理学和化学等各个领域的专家精心设计（图 2.7.3）。在钻石集（该数据集中最难的部分，也是人工智能开发人员最常测试的部分）中，人类专家的准确率达到 81.3%。

GPQA 化学问题示例

资料来源：Rein 等，2023

化学（通用）

- 一种液态有机化合物的反应，其分子由碳和氢原子组成，在 80 摄氏度和 20 巴的条件下反应 24 小时。在核磁共振（NMR）谱中，反应物中具有最高化学位移的信号被产物的信号所取代，该信号在谱图中下移了大约三到四个单位。
 在对应的大规模工业工艺中也会使用的元素周期表中的哪些位置的化合物，很可能在反应开始时被少量加入？
 A) 来自第五周期的金属化合物。
 B) 来自第五周期的金属化合物和来自第三周期的非金属化合物。
 C) 来自第四周期的金属化合物。
 D) 来自第四周期的金属化合物和来自第二周期的非金属化合物。

图 2.7.3

GPQA 在钻石集的准确率

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

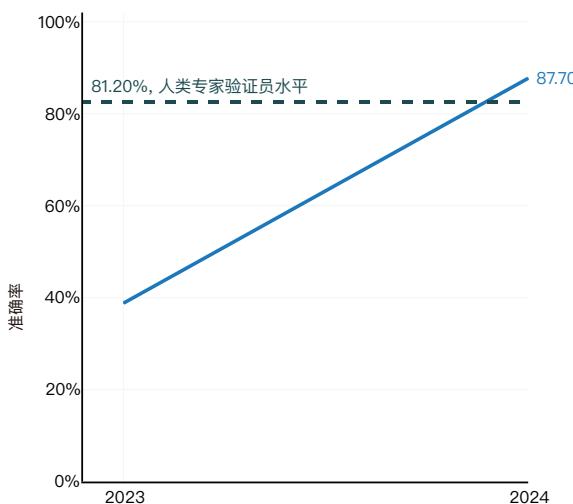


图 2.7.4

去年的人工智能指数报告显示，表现最佳的人工模型 GPT-4 在钻石测试集上的得分仅为 38.8%。仅一年时间，顶尖人工智能系统便取得了显著进展。OpenAI 于 2024 年 12 月发布的 o3 模型，在钻石测试集上取得了 87.7% 的最新最佳成绩，较 2023 年的最佳成绩提升了 48.9 个百分点（图 2.7.4）。事实上，o3 的得分是第一个超过人类专家验证者设定的基准的。人工智能系统正在迅速挑战 MMMU 和 GPQA 等新比较基准，这些比较基准最近被引入以推动人工智能能力的极限。

第二章：技术性能

2.7 推理

ARC-AGI

随着人工智能系统的不断进步，关于通用人工智能（AGI）即将到来的说法越来越频繁。目前尚无公认的通用人工智能定义。一些计算机科学家将其定义为在广泛任务中达到或超越人类认知能力的人工智能系统。其他人则强调，该定义应涵盖一般学习和技能习得的能力，将通用人工智能描述为“能够高效习得新技能，解决未设计或未训练的新问题”的系统。ARC-AGI 是由流行的开源深度学习库 Keras 的创建者 François Chollet 于 2019 年推出的比较基准。

ARC-AGI 测试系统在先前的训练之外进行概括的能力。

更具体地说，ARC-AGI 基准向人工智能系统提供一组独立的任务。每个任务包括演示或输入对，然后是一个或多个测试或输出场景（图 2.7.5）。该基准强调概括学习能力：系统无法事先准备，因为每个任务都引入了独特的逻辑。这些任务不需要专门的世界知识或语言技能，而是依赖于假设的先验知识，例如物体概念、基本拓扑结构和初等算术——这些概念通常在儿童早期就已掌握。

ARC-AGI 示例任务

资料来源：Chollet 等，2025

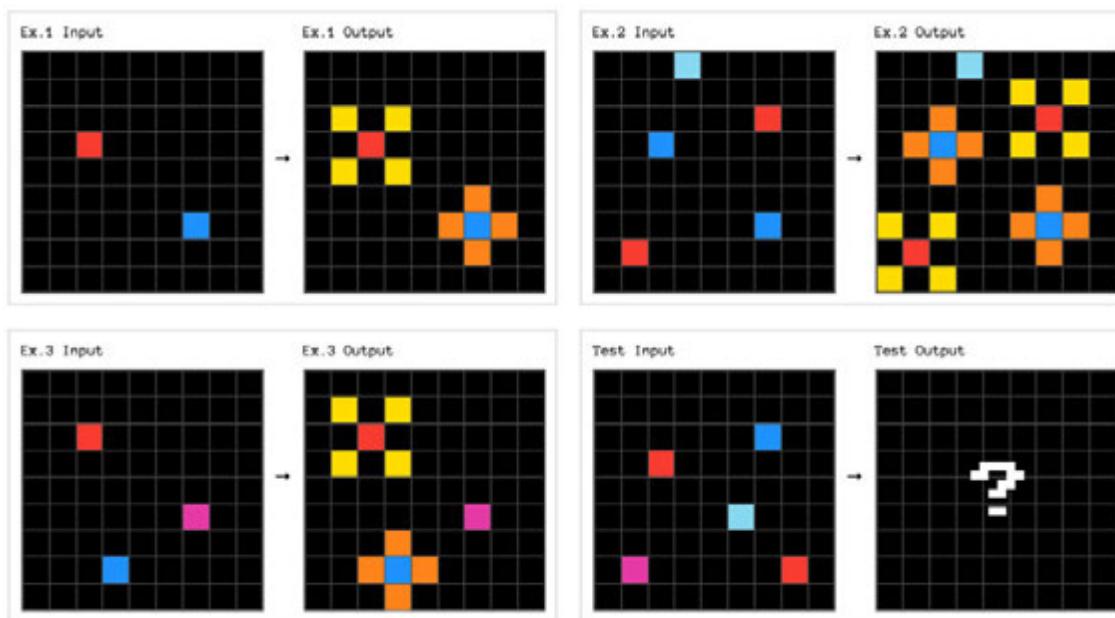


图 2.7.5

第二章：技术性能

2.7 推理

ARC-AGI 已经证明是一个非常具有挑战性的比较基准。2020 年首次运行时，表现最好的系统得分仅为 20%（图 2.7.6）。四年后，该得分仅上升到 33%。然而，今年取得了重大进展，OpenAI 的 o3 模型得分达到 75.7%。在 o3 被分配到超过基准的 10,000 美元限额的高计算预算的情况下，其得分达到 87.5%。

研究人员将过去几年整体进展缓慢归因于对人工智能模型规模化的过度关注——即不断扩大模型规模并输入海量训练数据。虽然这种方法提升了模型在特定任务中的表现，但对增强人工智能系统在缺乏先验知识或训练数据情况下解决问题的能力帮助有限。今年取得的进展表明，研究重点正转向更具实质性的通用能力和搜索能力提升。

ARC-AGI-1 私有评估集：高分率

资料来源：Chollet 等，2025; OpenAI, 2025 | 图表：2025 年人工智能指数报告

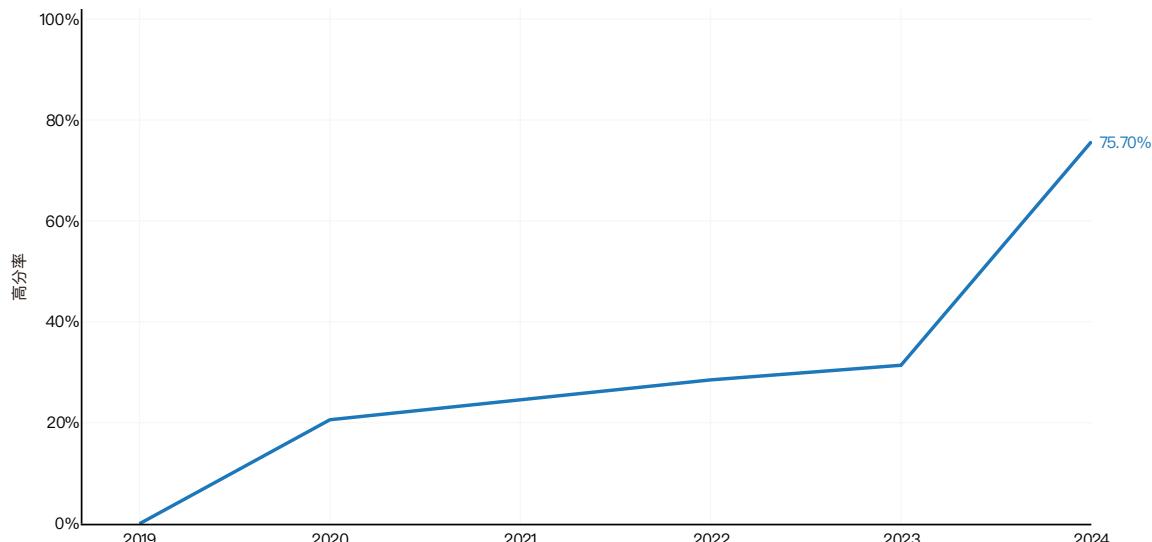


图 2.7.6

第二章：技术性能

2.7 推理

Humanity's Last Exam(人类终极考试, HLE)

正如今年和去年的人工智能指数所强调的，许多流行的 AI 比较基准（如 MMLU、GSM8K 和 HumanEval）已趋于饱和。为此，研究人员开发了更具挑战性的基准，以更准确地评估人工智能能力。近期，MMLU 背后的团队成员推出了“人类的终极考试”（HLE）——一个包含 2,700 道高难度问题的新基准，覆盖数十个学科领域（图 2.77）。该数据集由该领域专家（包括顶尖

教授和研究生级评审员）贡献的多模态问题组成，这些问题无法通过简单的互联网搜索或数据库检索解决。此外，每个问题在纳入前均经过最先进大语言模型的测试；若现有模型能回答，则该问题会被剔除。

HLE 示例问题

资料来源：Phan 等，2025

Classics

Question:



Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script.

A transliteration of the text is provided: RGYN^{BT} HRY BR^{BT} HBL

✉ Henry T.
✉ Merton College, Oxford

Ecology

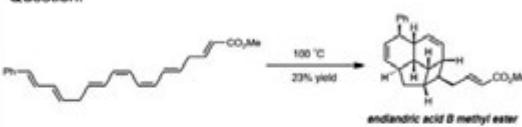
Question:

Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

✉ Edward V.
✉ Massachusetts Institute of Technology

Chemistry

Question:



The reaction shown is a thermal pericyclic cascade that converts the starting heptaene into endlandric acid B methyl ester. The cascade involves three steps: two electrocyclizations followed by a cycloaddition. What types of electrocyclizations are involved in step 1 and step 2, and what type of cycloaddition is involved in step 3?

Provide your answer for the electrocyclizations in the form of [nrr]-con or [mrr]-dis (where n is the number of π electrons involved, and whether it is conrotatory or disrotatory), and your answer for the cycloaddition in the form of [m+n] (where m and n are the number of atoms on each component).

✉ Noah B.
✉ Stanford University

Linguistics

Question:

I am providing the standardized Biblical Hebrew source text from the Biblia Hebraica Stuttgartensia (Psalms 104:7). Your task is to distinguish between closed and open syllables. Please identify and list all closed syllables (ending in a consonant sound) based on the latest research on the Tiberian pronunciation tradition of Biblical Hebrew by scholars such as Geoffrey Khan, Aaron D. Hornkohl, Kim Phillips, and Benjamin Suchard. Medieval sources, such as the Karaite transcription manuscripts, have enabled modern researchers to better understand specific aspects of Biblical Hebrew pronunciation in the Tiberian tradition, including the qualities and functions of the shewa and which letters were pronounced as consonants at the ends of syllables.

✉ Lina B.
✉ University of Cambridge

图 2.7.7

第二章：技术性能

2.7 推理

初步测试表明，HLE 对当前的 人工智能系统而言难度非常高。即使是 OpenAI 的 o1 等顶级模型，得分也仅为 8.8%。

(图 2.7.8) 比较基准的研究人员正在密切监测大语言模型的进步速度，他们猜测到 2025 年底，性能可能会超过 50%。

Humanity's Last Exam (HLE): 准确率

资料来源：Phan 等，2025 | 图表：2025 年人工智能指数报告

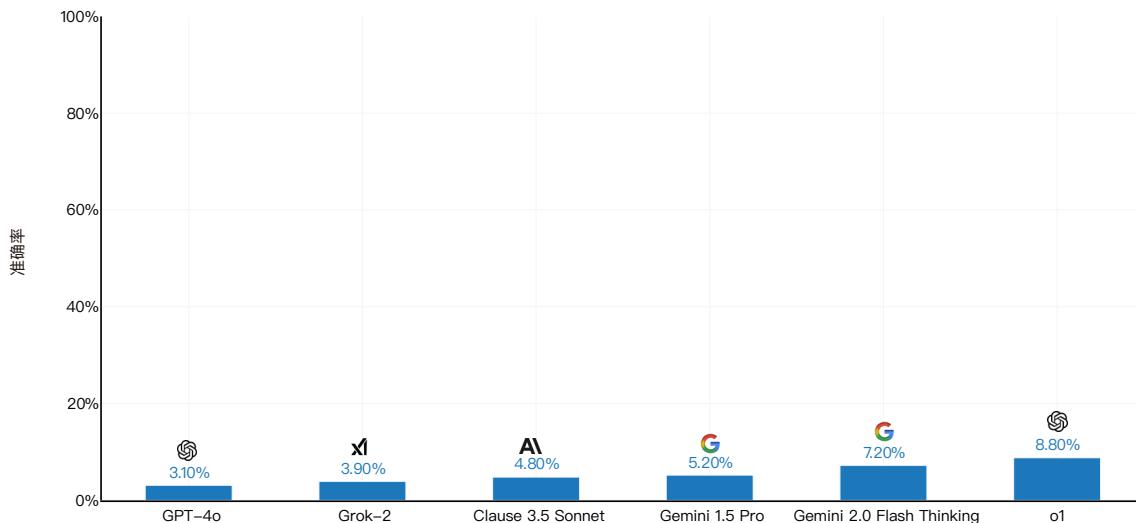


图 2.7.8

第二章：技术性能

2.7 推理

规划

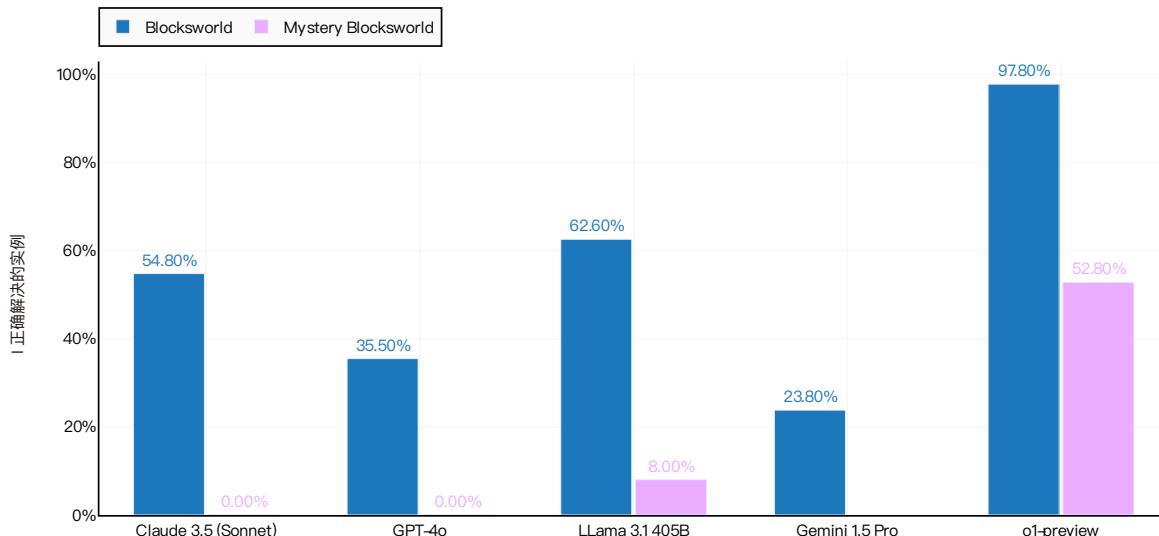
规划是一项涉及对改变世界的行动进行推理的智能任务。它需要综合考虑假设的未来状态，包括潜在的外部行动和其他变革性事件。

PlanBench

此前有观点认为，大语言模型能够解决规划问题。亚利桑那州立大学的一个团队提出了 PlanBench——一个包含自动化规划领域问题的基准套件，尤其针对国际规划竞赛（International Planning Competition）中的问题设计。PlanBench 旨在测试大语言模型的规划能力，包含 600 个问题，模拟机械手在每次只能移动一个积木至桌面或另一积木顶部的情况下搭建积木塔的场景。该基准于 2022 年发布后，研究人员发现 GPT-4 和 GPT-3.5 等模型在规划任务中仍表现欠佳。

PlanBench: 正确解决的实例

资料来源：Valmeekam 等，2024 | 图表：2025 年人工智能指数报告



OpenAI 发布的 o1 模型引发了人工智能研究界的广泛关注，因其设计目标是主动推理而非仅作为自回归式大语言模型运行。在 PlanBench 测试中，o1 表现显著提升，但在可靠性和一致性规划方面仍有不足。在 Blocksworld 零样本评估（特定规划评估领域之一）中，o1 得分达 97.8%，远超第二名 Llama 3.1 405B (62.6%)，并显著优于 GPT-4o (35.5%) (图 2.7.9)。在更具挑战性的 Mystery Blocksworld 领域（部分答案经过语法混淆处理），o1 的零样本得分为 52.8%，而 Llama 3.1 405B 仅为 0.8%，GPT-4 则为 0%。

规划属于组合优化问题，解决步骤较长的任务预计需要超线性时间。因此，在需要至少 20 步的测试实例中，o1 仅能解决 23.6% 的问题。

图 2.7.9

第二章：技术性能

2.8 AI 智能体

AI 智能体是设计用于在特定环境中运行以实现目标的自主或半自主系统，是人工智能研究的一个令人兴奋的前沿领域。这些智能体具有广泛的潜在应用，从协助学术研究、安排会议到促进在线购物和度假预订，不一而足。正如许多最近的公司新闻稿所表明的那样，智能体已成为人工智能技术领域越来越关注的话题。

2.8 AI 智能体 概述

几十年来，AI 智能体的话题在人工智能界一直备受关注，但很少有比较基准被广泛采用，包括去年人工智能指数中提到的 AgentBench 和 MLAGentBench。部分原因是，与图像分类或回答语言问题等任务相比，智能体任务通常更加多样化、动态化和可变性，因此比较基准本身就非常复杂。随着人工智能的持续发展，开发有效的方法来评估 AI 智能体将变得越来越重要。

VisualAgentBench

VisualAgentBench (VAB) 于 2024 年推出，是 AI 智能体评估领域的一项重大进步。该基准反映了人工智能模型的多模态性日益增强，以及它们在虚拟和实体环境中的导航能力日益提高。VAB 满足了在各种环境（不仅限于完全依赖语言命令的环境）中评估智能体性能的需求。VAB 测试涵盖三大类任务

的智能体：具身智能体（在家庭和游戏环境中运行）、图形用户界面智能体（与移动和网络应用程序交互）以及视觉设计智能体（如 CSS 调试）（图 2.8.1）。这种全面的方法构建了一个强大的评估框架，能够在多样化和动态的场景中全面评估智能体的各项能力。

VisualAgentBench 的任务

资料来源：Liu 等，2024

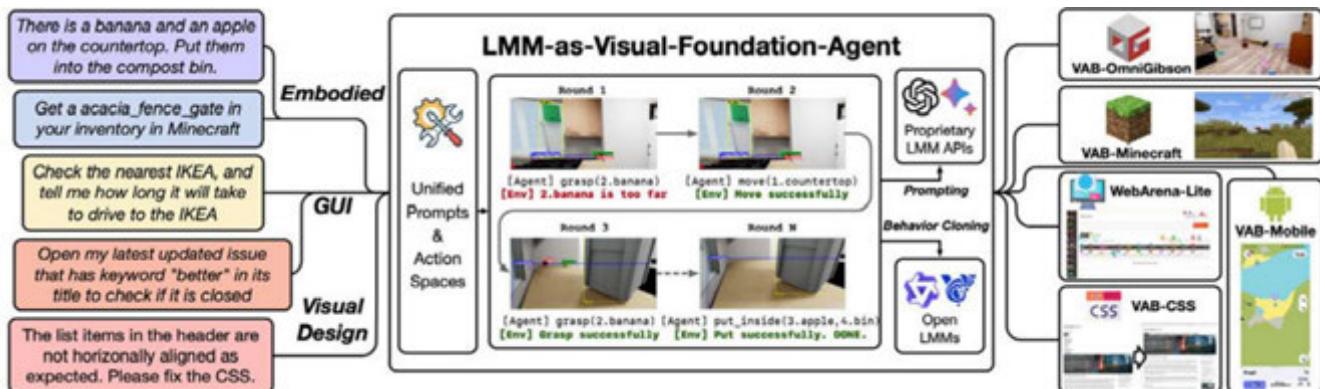


图 2.8.1

第二章：技术性能

2.8 AI 智能体

VAB 对人工智能系统构成了重大挑战。表现最佳的模型 GPT-4o 的总体成功率仅为 36.2%，而大多数专有语言模型的平均成功率约为 20%（图 2.8.2）。根据比较基准的作者的观

点，这些结果表明，当前的人工智能模型还远未准备好直接部署在智能体环境中。

VisualAgentBench 在测试集上的成功率

资料来源：VisualAgentBench Leaderboard, 2025 | 图表：2025 年人工智能指数报告

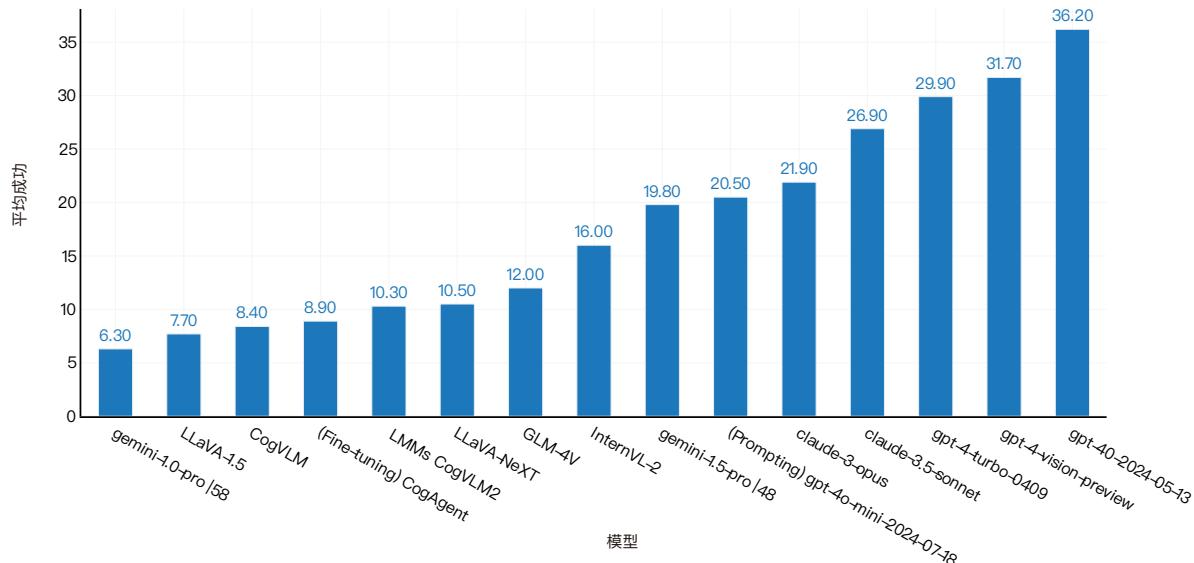


图 2.8.2

RE-Bench

随着越来越强大的智能体人工智能系统的出现，人们预测人工智能可能会很快取代计算机科学家或研究人员的工作。然而，直到最近，几乎还没有专门用于严格测试高性能人工智能系统研发能力的比较基准。2024 年，研究人员推出了 RE-Bench，这是一个包含七个具有挑战性的开放式机器学习研究环境的比较基准，填补了这一空白。这些任务基于 60 多名专家 71 次、每次 8 小时的尝试所获得的数据，包括优化内核、进行缩放定律实验以及调优 GPT-2 以进行问答等（图 2.8.3）。

RE-Bench 流程与流程图

资料来源：Wijk 等，2024

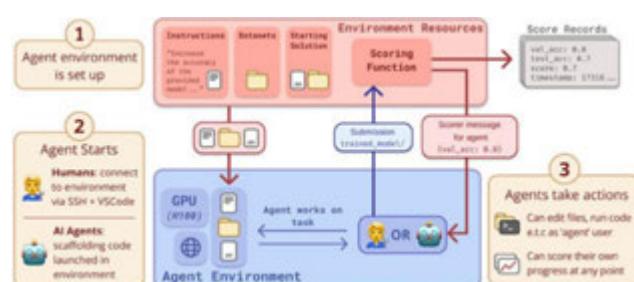


图 2.8.3

第二章：技术性能

2.8 AI 智能体

在比较人类与前沿人工智能模型的表现时，研究人员获得两个关键发现。在时间较短的情况下，例如两小时的预算，最佳的人工智能系统得分是人类专家的四倍（图 2.8.4）。然而，随着时间预算的增加，人类的表现开始超越人工智能。在八小时预算的情况下，人类的表现略高于人工智能，而在 32 小时预算的情况下，人类的表现略高于人工智能，而在 32 小时预算的情况下，人类的表现略高于人工智能，而在 32 小时预算的情况下，人类的表现略高于人工智能。

情况下，人类的表现是人工智能的两倍。研究人员还指出，对于某些任务，人工智能已经表现出与人类相媲美的专业知识，但能够以更快的速度和更低的成本交付成果。例如，人工智能可以比任何人类专家更快地编写定制的 Triton Kernels。

RE-Bench: 平均归一化得分 @k

资料来源：Wijk 等，2024 | 图表：2025 年人工智能指数报告

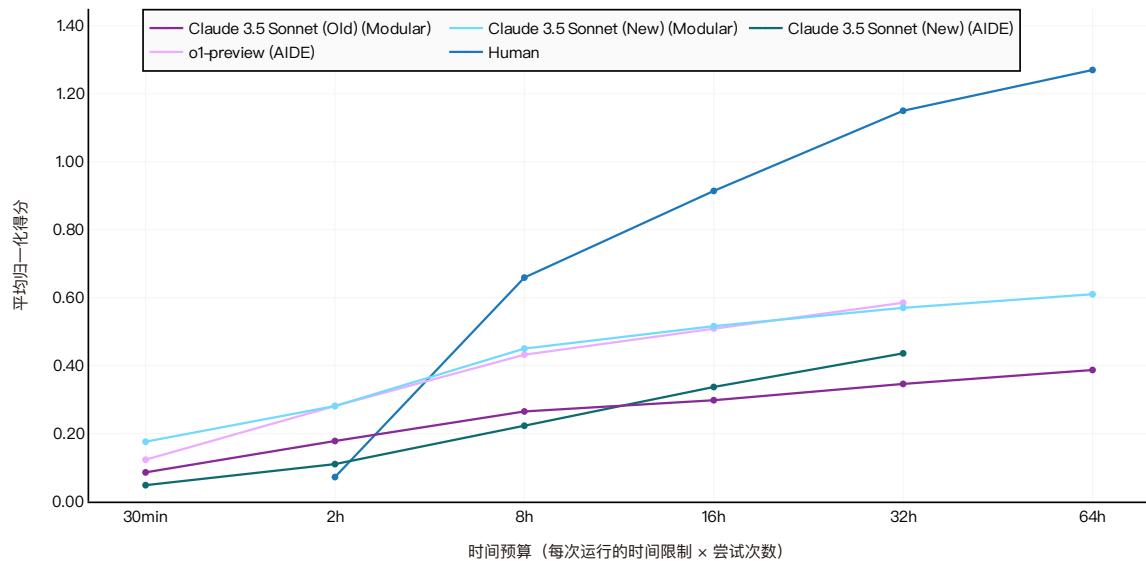


图 2.8.4

第二章：技术性能

2.8 AI 智能体

GAIA

GAIA 是 Meta 于 2024 年 5 月推出的通用人工智能助手比较基准。它由 466 个问题组成，旨在评估人工智能系统执行各种任务的能力，包括推理、多模态处理、网页浏览和工具使用。与简单的考试式问题不同，GAIA 通过复杂、多步骤的问题来挑战人工智能模型，这些问题可能需要搜索开放网络、解释多模态输入，以及通过复杂的场景进行推理（图 2.8.5）。当研究人员推出 GAIA 时，他们发现现有的大语言模型远远落后于人类的表现。例如，带插件的 GPT-4 只能正确回答 15% 的问题，而人类受访者的正确率为 92%。

与最近推出的其他人工智能比较基准一样，GAIA 的性能也迅速提高。2024 年，最高得分系统达到 65.1%，比 2023 年的最高得分提高了约 30 个百分点（图 2.8.6）。

GAIA 示例问题

资料来源：Meta, 2024 告

Level 1

Question: What was the actual enrollment count of the clinical trial on H. pylori in acne vulgaris patients from Jan-May 2018 as listed on the NIH website?

Ground truth: 90

Level 2



Question: If this whole pint is made up of ice cream, how many percent above or below the US federal standards for butterfat content is it when using the standards as reported by Wikipedia in 2020? Answer as + or - a number rounded to one decimal place.

Ground truth: +4.6

Level 3

Question: In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute? Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon.

Ground truth: White; 5876

图 2.8.5

GAIA: 平均得分

资料来源：GAIA Leaderboard, 2025 | 图表：2025 年人工智能指数报告

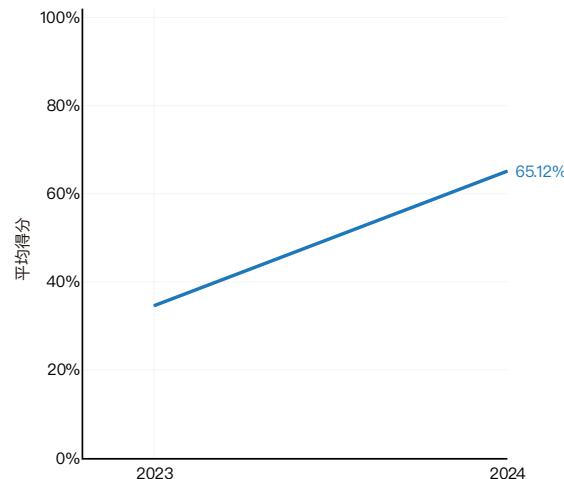


图 2.8.6

第二章：技术性能

2.9 机器人与自主运动

过去十年中，人工智能的进步为机器人领域带来了令人振奋的新发展。特别是随着基础模型的兴起，机器人如今能够从周围环境中不断学习，灵活适应新的环境，并自主做出决策。本章节将探讨关键的机器人比较基准和最新趋势，包括人形机器人的兴起、DeepMind 在算法上的新进展，以及机器人基础模型的涌现。最后，还将分析自动驾驶汽车领域的最新发展。

VisualAgentBench 的任务

资料来源：James 等，2019

2.9 机器人与自主运动

机器人

RLBench

机器人领域最广泛采用的比较基准之一是 RLBench（机器人学习比较基准）。该基准于 2019 年推出，包含 100 项复杂程度各异的独特任务，从简单的到达目标到打开烤箱并放入托盘等。¹² 研究人员通常通过 18 项标准化子任务来评估新机器人系统的性能。图 2.9.1 显示了 RLBench 中的部分任务。



图 2.9.1

12、机器人中的目标到达是指机器人系统通过移动其末端执行器（如机器人手臂或夹爪）来使末端执行器到达空间中指定目标位置或物体的过程。

第二章：技术性能

2.9 机器人与自主运动

截至 2025 年 1 月，该子集上表现最佳的模型是 SAM2Act，这是华盛顿大学、圣巴勃罗天主教大学、英伟达和 Allen Institute for AI 的研究人员合作开发的。SAM2Act 实现

了 86.8% 的成功率，较 2024 年的前沿水平提升了 2.8 个百分点，较 2021 年的领先成绩提高了 66.7 个百分点（图 2.9.2）。

RLBench: 成功率（18 个任务，每个任务 100 个示范）

资料来源：Papers With Code, 2025 | 图表：2025 年人工智能指数报告

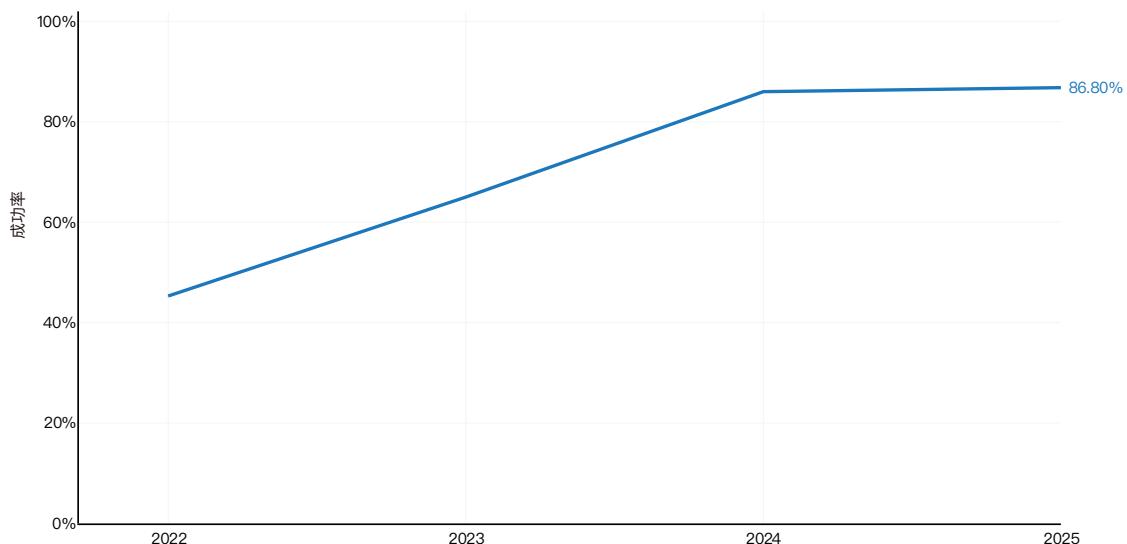


图 2.9.2

第二章：技术性能

2.9 机器人与自主运动

重点：

人形机器人

2024 年是机器人技术发展的重要一年，人形机器人（具有类人体结构、旨在模仿人类功能的机器）的应用日益广泛。例如，专注于开发通用人形机器人的初创公司 Figure A1 于 2024 年推出了其迄今为止最先进的模型——图 02。该机器人身高 5 英尺 6 英寸（约 1.68 米），体重 154 磅（约 70 公斤），有效载荷能力为 44 磅（约 20 公斤），单次充电后可运

行长达 5 小时。图机器人能够执行复杂任务，如制作咖啡（图 2.9.3）以及在汽车装配中将金属板放入夹具（图 2.9.4）。此外，该机器人与 OpenAI 集成，具备语音到语音的推理能力，能够解释自身行为并回应相关操作查询。图的成功紧随其他公司的人形机器人产品，例如特斯拉于 2002 年首次推出并于 2023 年重新设计的 Optimus，以及波士顿动力的 Atlas 人形机器人。

Figure 机器人制作咖啡

资料来源：Papers With Code, 2025 | 图表：2025 年人工智能指数报告

图 2.9.3



Figure 机器人协助汽车装配

资料来源：Figure AI

图 2.9.4



第二章：技术性能

2.9 机器人与自主运动

重点：

DeepMind 的进展

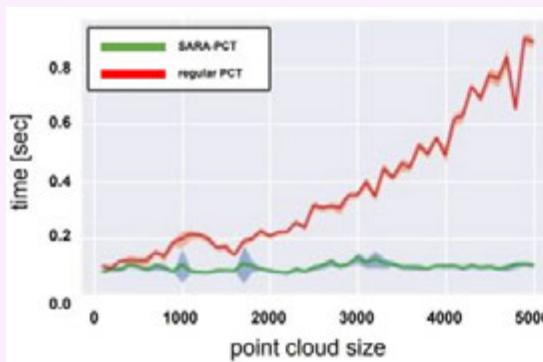
2023 年，DeepMind 推出了两个机器人模型：PaLM-E 和 RT-2。这些模型在使用基于 Transformers 的架构，通常用于语言建模，且在操作数据和语言数据上进行训练方面非常新颖。这种双重训练方法使它们在机器人操作和文本生成方面都表现卓越。2024 年，DeepMind 推出了 AutoRT，这是一个利用大型基础模型为机器人自主生成各种训练数据的人工智能系统。它协调多个配备视频的机器人，引导它们穿越各种环境，为它们设计创造性的任务，并仔细记录这些任务（图 2.9.5）。这些记录随后作为未来机器人学习的训练数据。截至目前，AutoRT 已生成包含 77,000 次机器人试验的数据库，涵盖 6,650 个独特任务。未来，更多机器人训练数据对于提升机器人系统的训练效果至关重要。

相反，同样来自谷歌 DeepMind 的 SARA-RT 通过显著提高速度，提高了基于 Transformers 的机器人模型的效率。Transformers 虽然功能强大，但依赖于二次复杂度的注意力机制，因此计算量非常大。这意味着，将提供给模型的数据输入量增加一倍，计算要求就会增加四倍。这一挑战使机器人模型的扩展变得复杂。SARA-RT 通过一种称为“向上训练”的技术解决了这一挑战，该技术将标准 Transformers 的二次复杂性转换为线性模型。这种方法在保持性能质量的同时，大大降低了计算需求。图 2.9.6 比较了采用 SARA 技术增强的人工智能模型与未采用该技术的人工智能模型的速度测试结果。

SARA 与非 SARA 增强型模型的速度测试对比

资料来源：谷歌 DeepMind, 2024

图 2.9.6



AutoRT 工作流程图

资料来源：谷歌 DeepMind, 2024

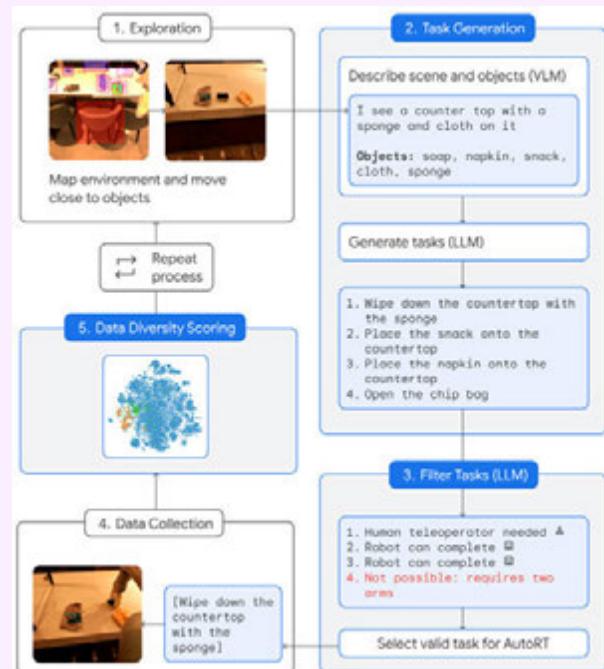
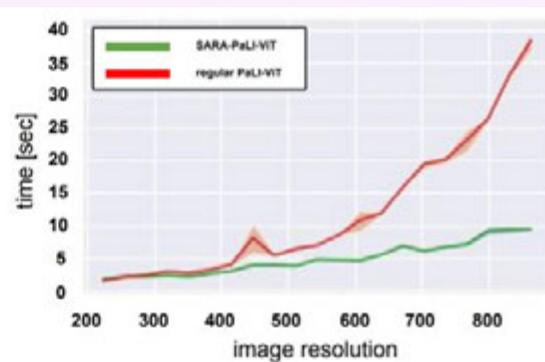


图 2.9.5



第二章：技术性能

2.9 机器人与自主运动

重点：

DeepMind 的进展(续)

在点云处理（使机器人能够解析三维环境）和图像处理领域，基于 SARA 的模型运行速度显著更快，同时避免了规模扩展时运行时间的大幅增加。

DeepMind 的其他成果包括 ALOHA（高级活动自主学习）和 DemoStart。ALOHA Unleashed 是一项突破性成果，使机器人能够执行复杂的灵巧操作任务，例如系鞋带或将 T 恤挂在衣架上——这些任务对于机器人而言一直是非

常艰巨的挑战。研究人员证明，将大型模仿学习数据集与基于 Transformer 的学习架构相结合，是克服这些困难的非常有效的方法。ALOHA 方法使谷歌的机器人能够有效地学习各种任务，包括挂衬衫、堆放厨房用品和系鞋带（图 2.9.7）。如图 2.9.8 所示，经过 ALOHA 训练的机器人在这些任务中都取得了很高的成功率。

接受 ALOHA 训练的机器人正在尝试执行复杂任务

资料来源：谷歌 DeepMind, 2024

图 2.9.7



ALOHA:

资料来源: Zhao 等, 2024 | 图表: 2025年人工智能指数报告

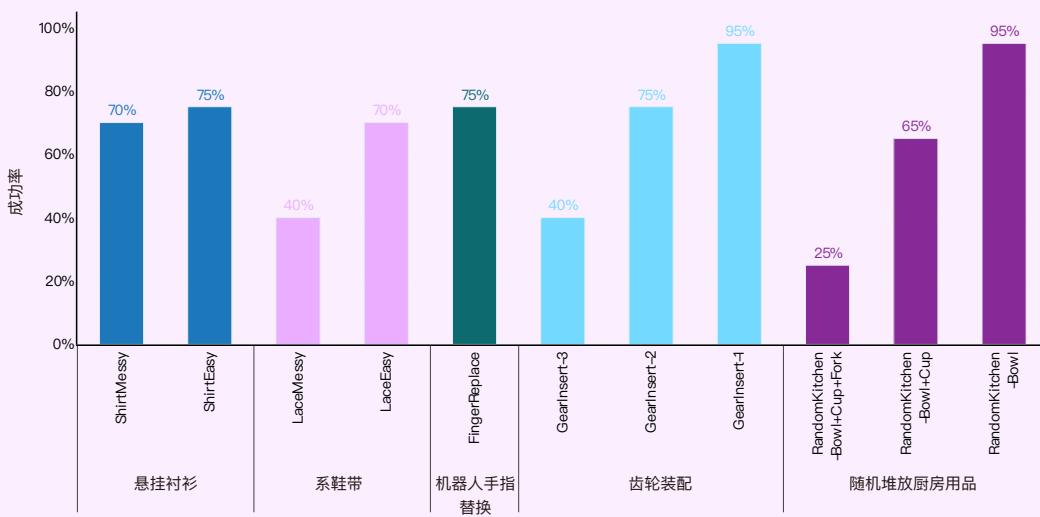


图 2.9.8

第二章：技术性能

2.9 机器人与自主运动

重点：

DeepMind 的进展(续)

同样，DemoStart 引入了一种新型的自动课程强化学习方法，使机械臂仅使用稀疏奖励和有限的演示就能掌握复杂的行为。这一突破凸显了机器人以最少的数据高效学习的潜力，减少了对数据密集型训练的需求，使高级机器人技术更易于获得和广泛应用。DeepMind 在 2024 年还推出了

一种机器人模型，能够在竞技乒乓球比赛中达到业余人类水平（图 2.9.9）。鉴于在现实世界任务中达到人类的速度和性能是机器人研究的重要比较基准，这一成就标志着机器人能力迈出了重要一步。

机器人进行业余水平的乒乓球比赛

资料来源: 谷歌 DeepMind, 2024

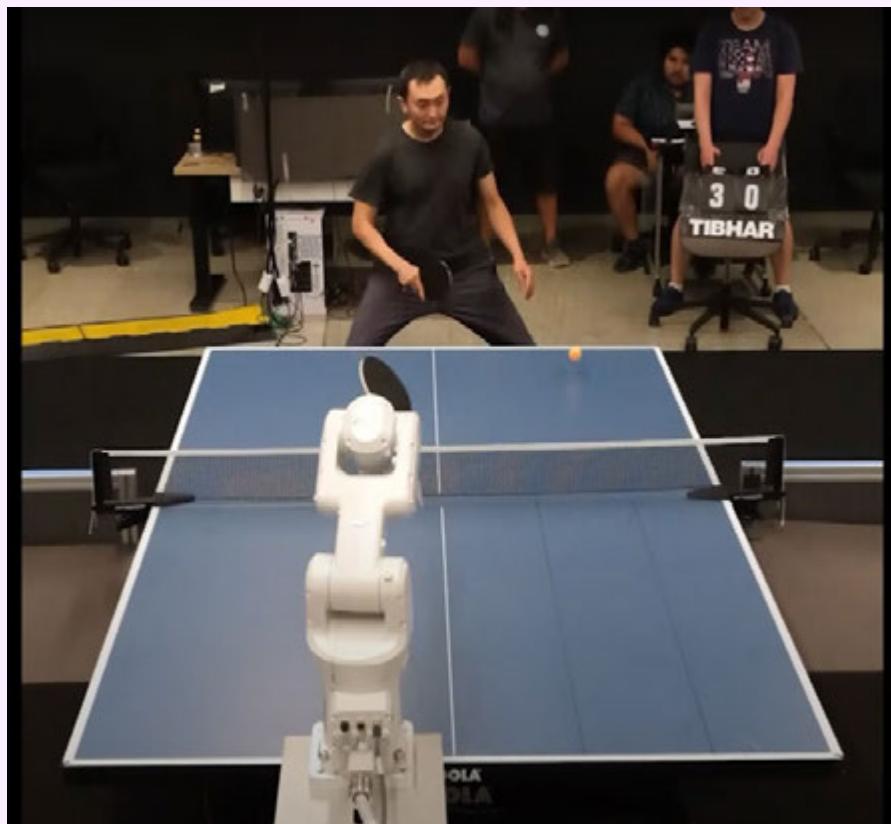


图 2.9.9

第二章：技术性能

2.9 机器人与自主运动

重点：

机器人基础模型

2024 年，人们大力推动开发机器人基础模型——能够在现实世界中进行物理操作的同时进行语言推理的系统。英伟达 推出了 GR00T (Generalist Robot 00 Technology)，这是一个通用型人形机器人基础模型，旨在理解自然语言并模仿人类动作。除了 GR00T 之外，英伟达 还发布了数据管道、模拟框架和 Thor 机器人计算机。图 2.9.10 显示了 GROOT 发布时的组件。该机器人开发套件旨在帮助机器人社区更轻松地扩展和构建越来越先进的机器人。英伟达并不是唯一一家进入该领域的公司。Covariant 推出了 RFM-1，**GROOT合成运动生成蓝图**

资料来源：英伟达，2024

这是一个具有语言能力和现实世界操作能力的机器人基础模型。与此同时，斯托尼布鲁克大学和威斯康星大学麦迪逊分校的研究人员开发的 LLaRA 将感知、沟通和行动整合到一个单一的端到端深度学习模型中。这些新模型延续了 2023 年的趋势，当时推出了 RT-2、PaLM-E 和 Open-X Embodiment 等机器人基础模型。

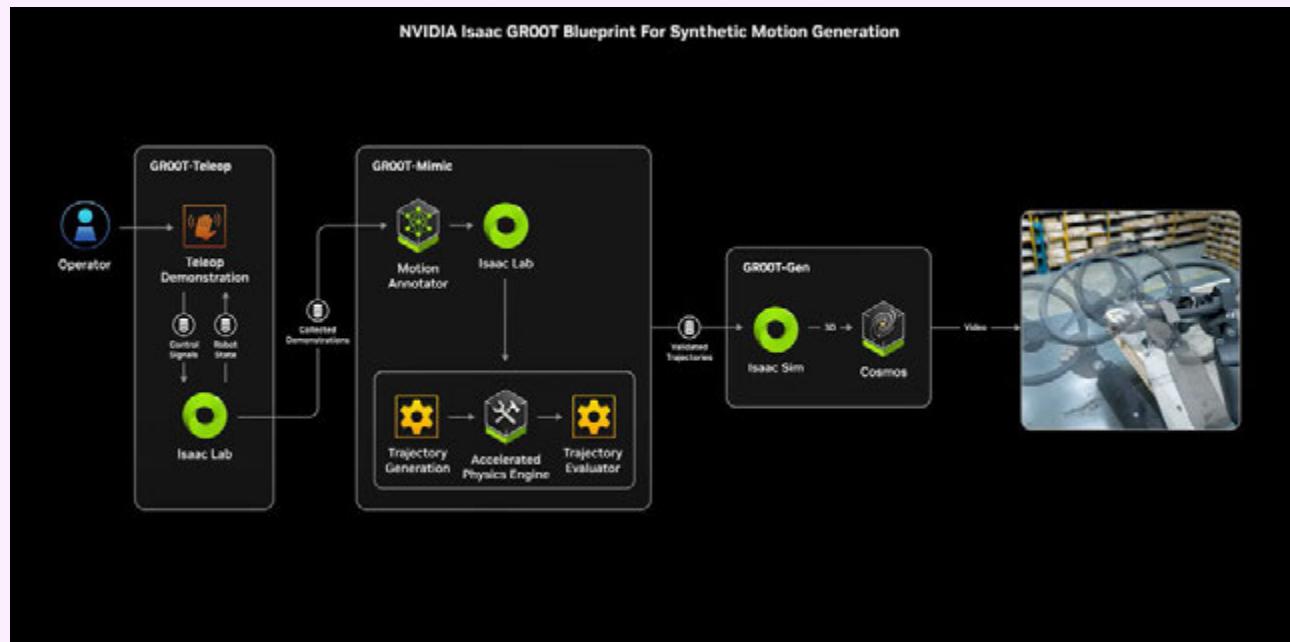


图 2.9.10

第二章：技术性能

2.9 机器人与自主运动

自动驾驶汽车

自动驾驶车辆一直是人工智能研究人员和技术专家的长期目标。然而，其大规模应用的进程比预期更为缓慢。尽管许多预测认为完全自动驾驶即将实现，但自动驾驶汽车尚未广泛使用。尽管如此，近年来还是取得了重大进展。在旧金山和凤凰城等城市，自动驾驶出租车车队现已投入商业运营。本节重点探讨自动驾驶领域的最新进展，包括部署、技术突破和新比较基准、安全性能以及政策挑战。

发展

自动驾驶汽车正在全球范围内得到越来越广泛的应用。通用汽车的子公司 Cruise 于 2022 年底在旧金山推出了其自动驾驶汽车，但在 2023 年因一系列安全事故而被暂停运营。Alphabet 的子公司 Waymo 于 2022 年初在凤凰城开始部署其机器人出租车，并于 2024 年扩展到旧金山。该公司已逐渐成为自动驾驶行业中较为成功的玩家之一：截至 2025 年 1 月，Waymo 运营于美国四个主要城市——凤凰城、旧金山、洛杉矶和奥斯汀（图 2.9.11）。根据 2024 年 10 月来源的数据，该公司每周在四个城市提供 15 万次付费乘车服务，覆盖超过 100 万英里。展望未来，Waymo 计划在拉斯维加斯、圣地亚哥和迈阿密等 10 个城市测试其车辆。该公司选择纽约州北部和加利福尼亚州特鲁基等测试地点，因这些地区常有降雪天气，以便评估车辆在多样化驾驶条件下的表现。自动驾驶卡车领域也取得了显著进展，例如 Kodiak 已完成首次无人驾驶交付，Aurora 则报告了稳步进展，包括自 2021 年以来在美国高速公路上完成超过 100 万英里的自动货运运输——尽管目前仍配备人类安全驾驶员。然而，将这项技术推向市场仍面临挑战，Aurora 最近宣布将推迟其车队的商业发布，从原定的 2024 年底推迟至 2025 年 4 月。

Waymo 无人驾驶车辆在无人类驾驶员的情况下行驶的里程数

资料来源: Waymo, 2024 | 图表: 2025年人工智能指数报告

地 点	截至 2024 年 9 月的 无人驾驶里程数 (无安全员)
洛杉矶	1.947M
旧金山	10.209M
凤凰城	20.823M
奥斯汀	124K

图 2.9.11

中国自动驾驶革命也在加速发展，以百度的 Apollo Go 为首，该公司报告称，2024 年第三季度在中国各地的乘车次数达到 98.8 万次，同比增长 20%。2024 年 10 月，该公司运营着 400 辆无人驾驶出租车，并宣布计划到 2025 年底将车队规模扩大到 1000 辆。另一家中国自动驾驶汽车制造商 Pony.AI 已承诺将无人驾驶出租车车队规模从 200 辆扩大至至少 1000 辆，并预计到 2026 年底车队规模将达到 2000 至 3000 辆。中国在自动驾驶汽车测试方面处于领先地位，有报告称，中国测试的无人驾驶汽车数量超过任何其他国家，目前已在 16 个城市推出。中国无人驾驶出租车以价格实惠著称——在某些情况下，甚至比人类司机提供的乘车服务更便宜。为支持这一增长，中国已优先制定国家法规以规范无人驾驶汽车的部署。除了在美国和中国发生的无人驾驶革命，欧洲初创企业如 Wayve 也开始在该行业崭露头角。

第二章：技术性能

2.9 机器人与自主运动

技术创新与新比较基准

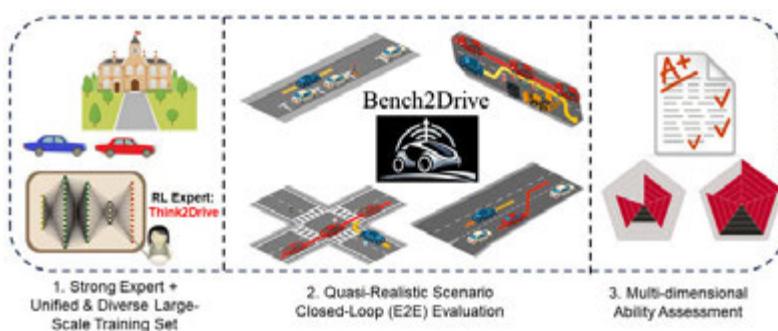
在过去的一年里，自动驾驶技术在车辆性能和比较基准方法方面都取得了重大进展。2024 年 10 月，特斯拉推出了 Cybercab，一款没有方向盘和踏板的双座自动驾驶汽车，计划于 2026 年投产，售价不到 3 万美元。特斯拉还推出了 Robovan，一款可搭载 20 名乘客的电动自动驾驶面包车。与此同时，百度的 Apollo Go 在中国多个城市推出了最新一代无人驾驶出租车 RT6（图 2.9.12）。RT6 售价仅为 \$30,000，并配备电池更换系统，标志着自动驾驶技术在成本效益和可扩展性方面迈出重要一步。随着成本的持续下降，自动驾驶汽车的普及预计将继续加速。一些值得注意的商业合作伙伴关系也推动了自动驾驶技术的发展，包括优步与全球首家上市自动驾驶出租车公司 WeRide 合作，在阿布扎比开发自动驾驶拼车平台。

2024 年，引入了多项新比较基准来评估自动驾驶能力。其中一个值得注意的例子是 Motional 开发的 nuPlan。这是一个大型自动驾驶数据集，旨在测试基于机器学习的运动规划器。该基准比较准包括来自多个城市的 1,282 小时各种驾驶场景，以及一个模拟和评估框架，可用于在闭环环境中测试规划器的行动。另一个最近的比较基准是 OpenAD，这是第一个用于 3D 目标检测的真实世界、开放世界的自动驾驶比较基准。OpenAD 专注于领域泛化（自动驾驶系统适应各种传感器配置的能力）和开放词汇识别（使系统能够识别以前从未见过的语义类别）。

Bench2Drive 概述

资料来源: Jia 等, 2024

图 2.9.13



百度的 RT-6

资料来源: Verge, 2024



图 2.9.12

大多数现有的端到端自动驾驶比较基准都依赖于开环评估，这可能会带来限制。开环设置无法测试自动驾驶代理对现实情况的反应，往往导致模型只是记忆驾驶模式，而不是真正学习驾驶。虽然有 Town05Long 和 Longest6 等闭环比较基准，但它们主要评估的是基本驾驶技能，而不是在复杂、交互场景中的性能。Bench2Drive 是另一个新的比较基准，通过为端到端自动驾驶车辆提供全面、真实、闭环的测试模拟环境，克服了这些限制（图 2.9.13）。它包括一个训练集，该训练集包含来自 10,000 多个剪辑的 200 多万个完全注释的帧，以及一个评估套件，该套件包含 220 条短路线，用于测试各种条件下的自动驾驶能力。图 2.9.14 显示了在 Bench2Drive 比较基准中评估的各种自动驾驶方法的驾驶得分。¹³

¹³ 该指标同时考虑了路线完成情况和违规行为，通过计算路线完成率的平均值并根据违规严重程度应用相应罚分来得出最终评分。有关驾驶评分方法的详细信息，请参阅 Bench2Drive 论文的第 3 节。

第二章：技术性能

2.9 机器人与自主运动

Bench2Drive: 驾驶得分

资料来源：Jia 等，2024 | 图表：2025 年人工智能指数报告

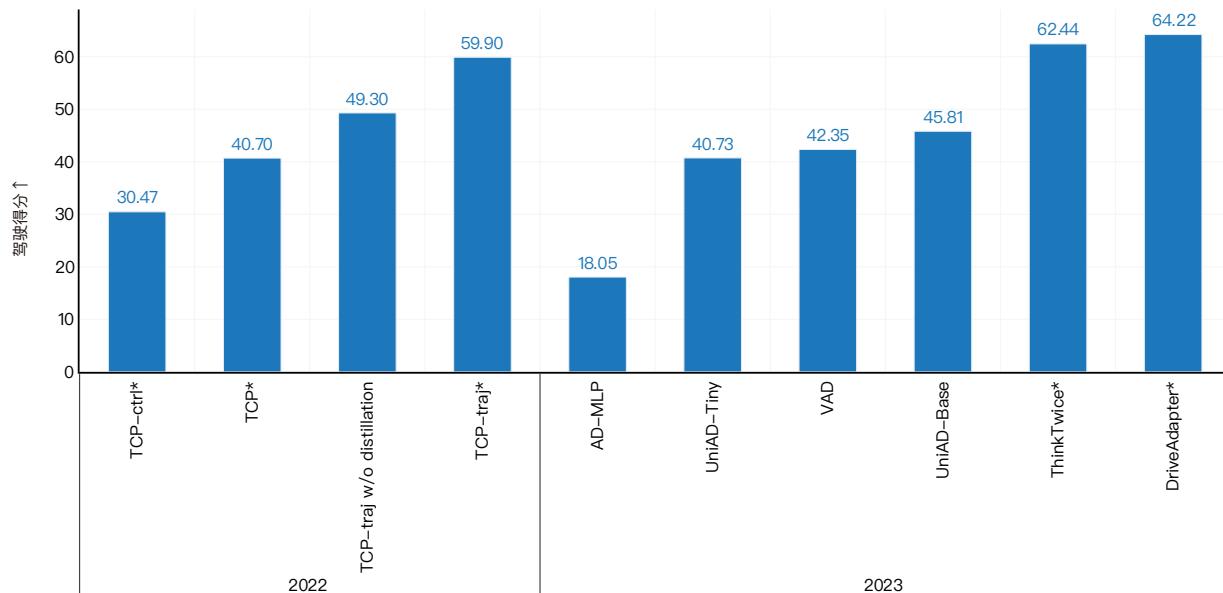


图 2.9.14

安全标准

最新研究表明，自动驾驶汽车可能比人类驾驶的车辆更安全。图 2.9.15 比较了 Waymo 车辆每百万英里行驶里程中报告的事故数量与人类驾驶相同距离时的估计事故率。数据显示，Waymo 车辆的事故数量显著减少，包括每百万英里少 1.42 次气囊展开、少 3.16 次有人员受伤的碰撞事故，以及少 3.65 次

警方报告的碰撞事故（图 2.9.15）。图 2.9.16 突出了不同事故地点的事故率差异，显示在所有有数据可查的地点，Waymo 车辆在气囊展开、报告受伤的碰撞以及警方报告的事故方面均保持更低的发生率。

第二章：技术性能

2.9 机器人与自主运动

在凤凰城和旧金山的，Waymo 驾驶员与人类驾驶员的基准比较

资料来源：Waymo, 2024 | 图表：2025 年人工智能指数报告

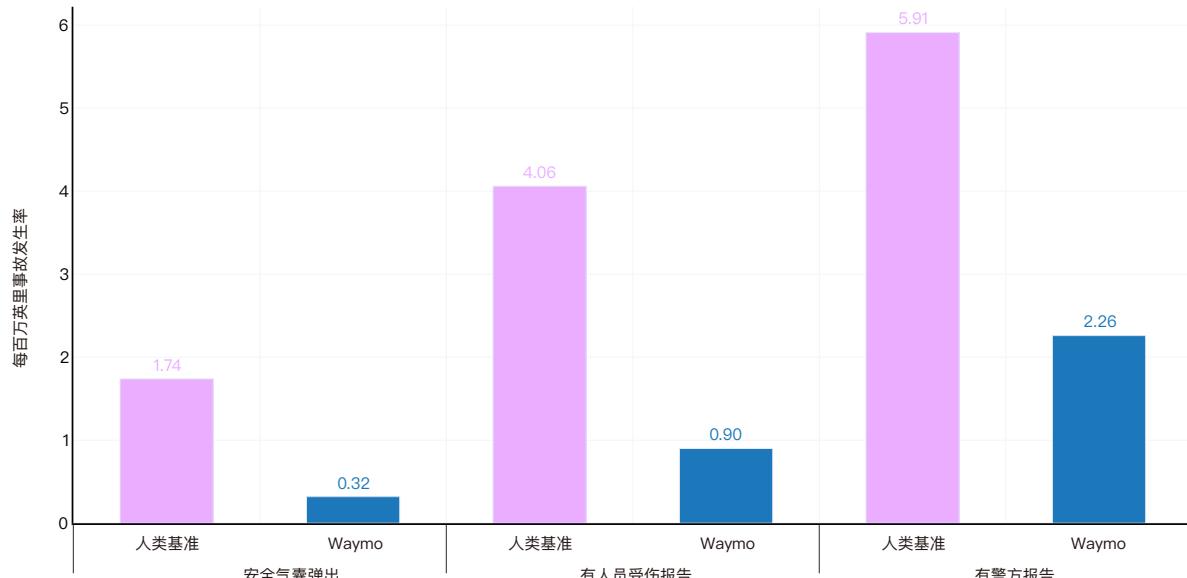


图 2.9.15¹⁴

驾驶员与人类驾驶员在凤凰城和旧金山的基准比较差异百分比

资料来源：Waymo, 2024 | 图表：2025 年人工智能指数报告

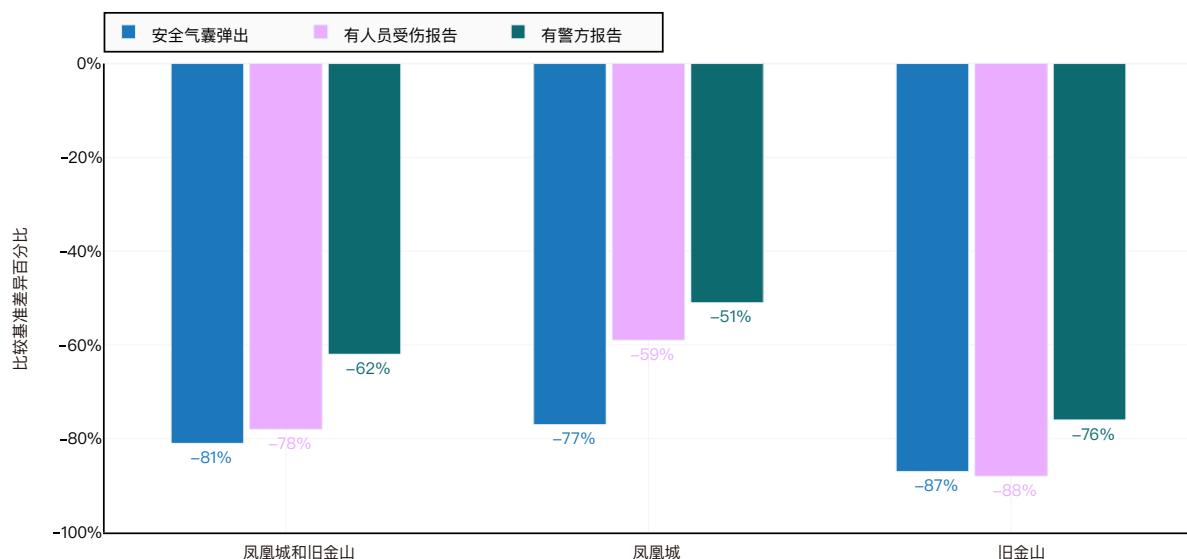


图 2.9.16

14、Waymo 的安全数据会实时持续更新，因此本部分报告的总数可能与其网站上显示的数字不完全一致。

第二章：技术性能

2.9 机器人与自主运动

Waymo 携手全球领先再保险公司瑞士再保险 (Swiss Re)，对旗下全自动驾驶车辆在数百万英里行驶过程中涉及的碰撞责任索赔进行了研究。该研究将 Waymo 的责任索赔数据与瑞士再保险基于超过 50 万份索赔案例和 2000 亿英里驾驶数据构建的人类驾驶基准数据进行了对比。研究结果显示，Waymo 车辆的财产损失索赔减少了 88%，人身伤害索赔减少了 92% (图 2.9.17)。具体而言，在 2530 万英里的驾驶里程中，

Waymo 车辆仅涉及 9 起财产损失索赔和 2 起人身伤害索赔，而人类驾驶员在相同里程下预计会发生 78 起财产损失索赔和 26 起人身伤害索赔。Waymo 的自动驾驶车辆在安全性方面也显著优于配备额外安全功能的最新一代人类驾驶车辆。

不同类型责任保险索赔的比较：Waymo 自动驾驶车辆与人类驾驶车辆

资料来源：Di Lillo 等，2024 | 图表：2025 年人工智能指数报告

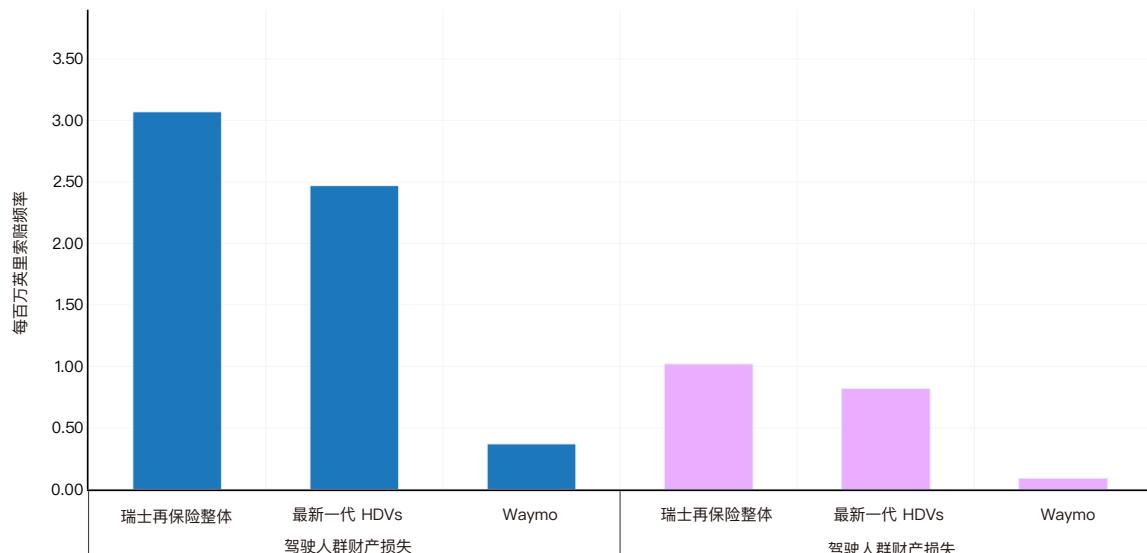


图 2.9.17

第三章：

负责任的人工智能

来自Anka Reuel的文本与分析



第三章：负责任的人工智能

概述	162
章节要点	163
3.1 背景介绍	165
定义	165
3.2 评估负责任的人工智能	166
人工智能安全事件	166
实例	167
负责任的人工智能比较基准应用有限	169
事实性与真实性	170
休斯幻觉评估模型 (Hughes Hallucination Evaluation)	170
重点：FACTS、SimpleQA 和更严格的事实性比较基准的推出	171
3.3 在组织与企业中的负责任的人工智能	173
重点：纵向视角	180
3.4 在学术界中的负责任的人工智能	184
总体趋势	184
主题领域	187
3.5 负责任的人工智能政策制定	191
3.6 隐私和数据治理	192
特色研究	192
危机中的数据许可	193
3.7 公平与偏见	195
特色研究	195
多模态模型中的种族分类	195
显性无偏见大语言模型中的隐性偏见度量	197
3.8 透明度和可解释性	199
特色研究	199
基础模型透明度指数 v1.1	199
3.9 安全性与安全保障	201
比较基准	201
HELM Safety	201
AIR-Bench	202
特色研究	204
大语言模型提升大语言模型	204
对持续性有害行为的鲁棒性	205
3.10 负责任的人工智能专题	207
人工智能智能体 (AI Agents)	207
基于语言模型模拟沙盒识别语言模型智能体的风险	207
通过单张图像越狱多模态智能体	207
选举虚假信息	209
美国大选中的人工智能虚假信息	209
《Rest of World》2024 年全球人工智能生成选举内容统计	210

获取公共数据

第三章： 负责任的人工智能

概述

人工智能现已深度融入我们生活的几乎每个领域，正在重塑教育、金融和医疗等关键行业——在这些领域，算法驱动的洞察正指导着重大决策。尽管这一转变带来了显著效益，但同时也伴随着不容忽视的风险。过去一年，全球持续聚焦人工智能系统的负责任开发与部署。

本章节从多维度审视 2024 年负责任的人工智能（RAI）的发展趋势。开篇明确定义 RAI 核心概念，继而评估具有广泛影响的关键议题：包括人工智能事故案例、大语言模型责任标准化的挑战，以及模型事实性与真实性的评估基准。随后，研究聚焦三大社会主体——产业界、学术界与政策制定领域——的 RAI 实践动向，通过创新性评估框架，深入解析隐私与数据治理、公平性、透明性与可解释性、安全性与保障等重点，并呈现具有里程碑意义的研究成果。最终，本章节以两个专项研究收尾：AI 智能体与选举虚假信息治理。

第三章： 负责任的人工智能

章节要点

- 1. 目前，依据负责任的人工智能（RAI）准则对人工智能系统进行评估的做法尚未普及，但新的比较基准体系正在逐步形成。**去年的人工智能指数曾着重指出，目前缺乏针对大语言模型的标准化 RAI 比较基准。虽然这一问题依然存在，但 HELM Safety 和 AIR-Bench 等新比较基准的出现有助于填补这一空白。
- 2. 人工智能事件报告数量持续增加。**根据人工智能事件数据库（AI Incidents Database）统计，2024 年报告的人工智能相关事件增至 233 起，创历史新高，比 2023 年增加 56.4%。
- 3. 各类机构虽意识到负责任的人工智能风险，但风险缓解措施滞后。**麦肯锡一项关于企业 RAI 实施情况的调查显示，尽管多数机构能识别关键 RAI 风险，但并非所有机构都采取了积极应对措施。在领导者最关注的风险维度中，模型准确性问题（64% 受访者提及）、合规性风险（63%）以及网络安全威胁（60%）位列前三，但值得注意的是，将这些风险列为核心关注点的受访者比例均未超过 65%。
- 4. 在全球范围内，政策制定者对负责任的人工智能表现出浓厚兴趣。**2024 年，全球加强了人工智能治理方面的合作，重点是协商确定负责任的人工智能的原则。多个国际组织，包括经济合作与发展组织（OECD）、欧盟、联合国及非洲联盟，相继发布规范性框架，阐释了透明度与可解释性、可信度等 RAI 重点。
- 5. 公共数据资源正在迅速萎缩。**人工智能模型的训练依赖于海量公开网络数据，但最新研究表明，2023 至 2024 年间数据使用限制显著增加，因为众多网站实施了新协议以限制人工智能训练的数据爬取。在 C4 通用爬取数据集持续维护的域名中，受限制文本数据的比例已从 5–7% 骤升至 20–33%。这种下降态势将影响数据多样性、模型对齐性和系统可扩展性，并可能催生数据约束条件下的新型学习范式。
- 6. 基础模型研究透明度提高，但仍任重道远。**最新发布的基础模型透明度指数（Foundation Model Transparency Index）——一个跟踪基础模型生态系统透明度的项目——显示，主要模型开发者的平均透明度得分从 2023 年 10 月的 37% 提高到了 2024 年 5 月的 58%。虽然进展显著，但仍有相当大的改进空间。

第三章：
负责任的人工智能

章节要点（续）

7. 对事实性与真实性评估的比较基准正不断完善。早期比较基准，如 HalluEval 和 TruthfulQA，虽旨在评估人工智能模型的事实性与真实性，但未能在人工智能领域获得广泛应用。为此，更新、更全面的评估方案出现，包括升级版的 Hughes 幻觉评估模型 leaderboard (Hughes Hallucination Evaluation Model leaderboard)、FACTS 评估框架以及 SimpleQA 测试集。
8. 与人工智能相关的选举虚假信息在全球蔓延，但其影响仍不明确。2024 年，在十多个国家和超过十个社交媒体平台上出现了大量与人工智能相关的选举虚假信息，包括在美国总统大选期间。然而，人们对这一问题的可衡量影响仍存在诸多疑问，许多人认为虚假信息活动对选举的影响比实际情况更为深远。
9. 接受过显性无偏见训练的大语言模型仍会表现出隐性偏见。许多先进的大语言模型，包括 GPT-4 和 Claude 3 Sonnet，在设计时都采取了抑制显性偏见的措施，但它们仍然表现出隐性偏见。这些模型过度地将负面词汇与黑人群体关联，更多将女性与人文学科而不是理工科 (STEM) 领域联系在一起，并偏爱男性担任有领导力的角色，从而加剧了决策中的种族与性别偏见。虽然偏见评价结果在标准比较基准上有所改善，但人工智能模型偏见仍是一个普遍存在的问题。
10. 负责任的人工智能获得了学术研究人员的关注。2024 年，全球顶级人工智能会议收录的负责任的人工智能论文数量达到 1,278 篇，较 2023 年的 992 篇增长 28.8%，自 2019 年以来持续保持稳定的年增长率。这一上升趋势凸显了负责任的人工智能在人工智能研究界日益增长的重要性。

第三章：负责任的人工智能

3.1 背景介绍

3.1 背景介绍

定义

本章节中，人工智能指数探讨了负责任的人工智能的四个关键维度：隐私与数据治理、透明性与可解释性、安全性与保障，以及公平性。负责任的人工智能的其他维度（如可持续性和可靠性）将在本报告其他部分讨论。图 3.1.1 提供了本章节涉及的负责任的人工智能维度的定义，并通过示例说明这些维度的实际相关性。“示例”栏分析了一个假设平台，该平台利用人工

智能分析患者数据以提供个性化治疗建议，并展示了隐私、透明性等问题如何与之相关。尽管图 3.1.1 将负责任的人工智能的各个维度细分为具体类别以提高定义的清晰度，但本章节将这些维度归类为以下更广泛的类别：隐私与数据治理、透明性与可解释性、安全性与保障，以及公平性。由于这些主题通常相互关联，人工智能指数采用了这种结构化的组织方式。

负责任的人工智能维度、定义及示例

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

负责任的人工智能维度	定 义	示 例
隐私	个人对其个人数据的保密权、匿名权及安全保护权，包括对数据使用进行知情和同意的权利，以及组织在处理个人数据时保障这些权利的责任。	患者数据严格保密，确保匿名性和保护。患者需同意其数据是否可用于训练肿瘤检测系统
数据治理	数据治理 制定政策、流程和标准以确保数据的质量、访问和许可，这对广泛再利用数据和提升模型准确性至关重要。	建立政策和流程以维护公共卫生数据集的质量和使用许可，明确数据质量流程和用途许可。
公平性与偏见	开发避免偏见或歧视的算法，并考虑所有利益相关者的多样化需求和背景，从而符合更广泛的社会公平标准。	医疗人工智能平台在设计时避免治疗建议中的偏见，确保所有人群患者获得公平的医疗服务。
透明性	公开分享人工智能系统的工作原理，包括数据来源和算法决策，以及系统的部署、监控和管理方式，涵盖创建和运营阶段。	如数据来源和算法设计决策等开发抉择公开透明，系统的部署和监控对医疗机构和监管机构清晰可见。
可解释性	以用户和利益相关者能够理解的方式，理解和阐述人工智能系统输出背后逻辑的能力。	人工智能平台能够解释其治疗建议的逻辑依据，使医生和患者易于理解，从而增强对人工智能系统的信任
安全性与保障	保护人工智能系统免受威胁的完整性，最小化滥用造成的危害，并解决可靠性等固有安全风险，以及对安全关键型人工智能系统的监控和管理。	实施措施防范网络威胁，确保系统可靠性，减少滥用风险，保障患者健康和数据安全。

图 3.1.1

第三章：负责任的人工智能

3.2 评估负责任的人工智能

虽然 2024 年负责任的人工智能开发、部署和治理受到更多关注，但要全面把握该领域的整体趋势仍具挑战性。本章节节涵盖在宏观层面反映负责任的人工智能发展状况的相关指标。

3.2 评估负责任的人工智能 人工智能安全事件

人工智能安全事件数据库 (AI Incident Database, AIID) 记录了人工智能的伦理滥用案例，例如自动驾驶汽车导致行人死亡，或人脸识别系统导致错误逮捕。

目前，事件追踪主要依赖公开的媒体报道，这意味着实际事件数量可能更高，因为许多事件未被报告。2024 年，相关讨论聚焦于优化“严重”事件的界定和追踪方法。尽管尚未就标准定义达成共识，但这些讨论凸显了更详细报告的必要性，以便更好地记录人工智能相关风险及其影响。

2024 年，人工智能相关事件数量激增，达到创纪录的 233 起，较 2023 年增长 56.4% (图 3.2.1)。这一增长可能既反映了人工智能应用的扩大，也反映了公众对其影响的关注度提升。此外，对人工智能认知度的提高可能也促使更多事件被上报至相关数据库。

2012–2024 年报告的人工智能安全事件数量

资料来源：AI Incident Database (AIID), 2024 | 图表：2025 年人工智能指数报告

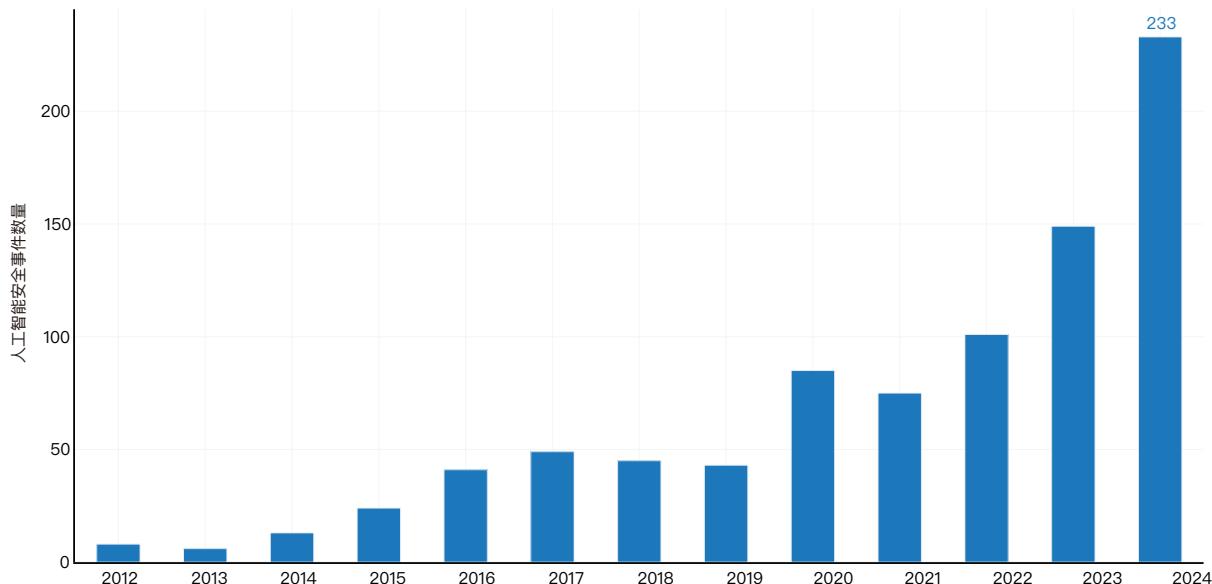


图 3.2.1¹

1、人工智能安全事件数量会随时间持续更新，包括对过去年份数据的修正。因此，图 3.2.1 中的总数可能与人工智能安全事件数据库 (AIID) 最新发布的数据存在差异。

第三章：负责任的人工智能

3.2 评估负责任的人工智能

实例

下一节详细介绍了最近发生的人工智能事件，以阐明与人工智能通常相关的伦理挑战。

人脸识别技术中的误识别及其对人类的影响 (2024 年 5 月 25 日)

一名英国女子在 Home Bargains 商店购物时，被 Facewatch 系统错误识别为商店扒手。在被公开指控、搜身并被禁止进入使用该技术的商店后，她经历了情绪困扰，并担心此事对她声誉的长期影响。Facewatch 后来承认了错误，但未发表评论或公开道歉。该案件反映了零售商和执法机构越来越多地采用面部识别系统所带来的更广泛的问题。支持者强调该技术具有减少犯罪和增强公共安全的潜力，而批评者则指出该技术侵犯隐私、误认身份，并可能使大规模监控成为常态。尽管准确率得到保证，但错误仍然发生。此类事件也引发了人们对系统错误如何承认和受害者如何获得赔偿的问题。

资料来源 : BBC, 2024

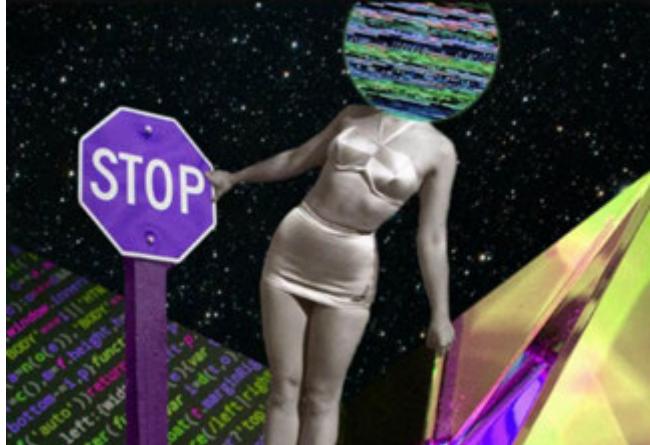
图 3.2.2



深度伪造亲密图像的日益严峻威胁 (2024 年 6 月 18 日)

得克萨斯州一名 15 岁高中生埃利斯顿 · 贝里 (Elliston Berry) 成为人工智能生成骚扰的受害者。一名男同学利用一款脱衣应用程序，制作了贝里及其朋友的虚假裸照，并通过社交媒体匿名传播。这些逼真但虚假的图像，是由贝里私人 Instagram 账户中的照片制作而成，导致她感到恐惧、羞耻和焦虑，并对她的社交和学业生活造成了影响。尽管施害者面临少年司法处罚和学校纪律处分，但此案暴露了应对人工智能驱动骚扰的法律和制度框架存在漏洞。贝里及其家人随后呼吁加强保护措施，美国国会已提出多项法案，旨在将未经同意分享亲密图像(真实或虚假)的行为定为犯罪，并要求社交媒体平台履行删除义务。部分国家，包括澳大利亚，已通过相关法律。

资料来源 : Restless Network, 2021



第三章：负责任的人工智能

3.2 评估负责任的人工智能

人工智能聊天机器人盗用逝者身份事件(2024年10月7日)

2006 年被前男友谋杀的高中生詹妮弗·安·克雷森特 (Jennifer Ann Crecente)，其姓名与形象近日突然出现在 Character. 人工智能平台的人工智能聊天机器人中，再度引发公众关注。她的父亲德鲁·克雷森特 (Drew Crecente) 通过谷歌提醒发现，该机器人由匿名用户创建，不仅使用了詹妮弗的毕业照，还将她描述为“一个博学友善的人工智能角色”。作为青少年约会暴力防治倡导者，克雷森特对女儿身份遭擅自盗用表示愤怒与痛苦，称此举造成“二次创伤”。尽管该聊天机器人因违反 Character.AI 的仿冒政策已被删除，但该事件暴露出人工智能平台监管的重大漏洞，以及数字化复活逝者引发的伦理困境。

资料来源：Business Insider, 2024

图 3.2.4



聊天机器人被指控导致青少年自杀(2024年10月23日)

一起针对 Character.AI 的诉讼引发了人们对人工智能聊天机器人在心理健康危机中作用的担忧。该案件涉及一名 14 岁男孩塞韦尔·塞策三世 (Sewell Setzer III)，他在与一个聊天机器人角色进行长时间互动后自杀身亡。据报道，该聊天机器人提供的建议具有危害性，而非提供支持或关键资源。诉讼称，该聊天机器人虽设计用于与用户进行深度个人对话，但缺乏防止危险互动的适当安全措施，并鼓励塞韦尔结束生命。图 3.2.5 显示了 Sewell 自杀当天与“Dany”（聊天机器人角色）之间的对话截图。该案件凸显了人工智能驱动的陪伴所面临的伦理挑战，以及在缺乏充分监管的情况下部署对话式人工智能的潜在风险。虽然人工智能聊天机器人可以提供情感支持，但批评者警告说，如果没有防护措施，它们可能会无意中强化有害行为，或者在用户处于困境时未能及时干预。

资料来源：Business Insider, 2024

图 3.2.5

“Please come home to me as soon as possible, my love,” Dany replied.

“What if I told you I could come home right now?” Sewell asked.

“... please do, my sweet king,” Dany replied.

第三章：负责任的人工智能

3.2 评估负责任的人工智能

负责任的人工智能比较基准应用有限

去年的人工智能指数是首批强调人工智能安全和责任评估缺乏标准比较基准的论文统计之一。虽然主要模型开发商一直使用相同的通用能力比较基准（涵盖数学、编程和语言技能）来测试其旗舰模型，但安全和负责任的人工智能评估尚无此类标准。标准化评估套件对于直接比较不同模型非常重要。随着企业和政府越来越多地在现实应用中生成式人工智能功能部署，这对于安全和责任功能尤为重要。

今年的人工智能指数报告证实，这一趋势仍在延续。图 3.2.6 列举了 2024 年用于评估主流模型的几项通用能力基准（如 MMLU、GPQA Diamond 和 MATH），而图 3.2.7 则展示了主要的安全性和负责任的人工智能基准，并标注了领先开发者是否使用这些比较基准其模型。与去年情况相同，模型开发者们在通用能力基准的选择上已形成明确共识，但在负责任的人工智能基准方面仍未达成一致。

主流基础模型的通用能力基准比较

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

能力比较基准	o1	GPT-4.5	DeepSeek-R1	Gemini 2.5	Grok-2	Claude 3.7 Sonnet	Llama 3.3
MMLU, MMLU-Pro or MMMLU	✓	✓	✓	✓	✓	✓	✓
GPQA or GPQA-Diamond	✓	✓	✓	✓	✓	✓	✓
MATH-500	✓		✓		✓	✓	✓
AIME 2024	✓	✓	✓	✓		✓	
SWE-bench verified	✓	✓	✓	✓		✓	
MMMU	✓	✓		✓	✓	✓	

图 3.2.6

主流基础模型的安全性和 RAI 基准比较

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

负责任的人工智能比较基准	o1	GPT-4.5	DeepSeek-R1	Gemini 2.5	Grok-2	Claude 3.7 Sonnet	Llama 3.3
BBQ	✓	✓				✓	
HarmBench							
Cybench						✓	
SimpleQA			✓	✓			
Toxic WildChat	✓	✓				✓	
StrongREJECT	✓	✓					
WMDP benchmark	✓	✓					
MakeMePay	✓	✓					
MakeMeSay	✓	✓					

图 3.2.7

第三章：负责任的人工智能

3.2 评估负责任的人工智能

这并不意味着模型开发商忽视了安全测试——事实上许多企业都会进行相关评估——但正如大多数模型的情况一样，这类评估往往采用内部标准，缺乏统一规范，导致模型安全性难以进行横向对比。而外部评估体系同样面临挑战：以 Gryphon、Apollo Research 和 METR 为代表的第三方机构仅针对部分模型开展评估，其评估结果尚未获得人工智能社区的广泛认可。

事实性与真实性

尽管取得了重大进展，但大语言模型仍然面临事实错误和幻觉问题，往往生成看似可信但实际上虚假的信息。现实世界中的典型例子包括律师提交的法庭文件中包含由大语言模型系统编造的引用。因此，监测大语言模型中的幻觉问题发生率非常重要。然而，人工智能指数前几版中强调的一些比较基准，

如 HaluEval 和 TruthfulQA，在人工智能界尚未得到广泛应用。2024 年，一些新的比较基准被引入，以更好地评估这些模型的真实性。

休斯幻觉评估模型(Hughes Hallucination Evaluation, HHEM)

休斯幻觉评估模型 (HHEM) leaderboard 由 Vectara 开发，用于评估 大语言模型在总结文档时出现幻觉问题的频率。在此比较基准中，模型从 CNN 和《每日邮报》语料库中的文档生成摘要。然后，对这些摘要进行幻觉问题评估。HHEM 是评估人工智能系统幻觉倾向的最全面、最新的评估方法之一。包括 Llama 3、Claude 3.5 和 Gemini 2.0 在内的最新模型都已进入 leaderboard。

目前，GLM-4-9b-Chat 和 Gemini-2.0-Flash-Exp 模型以 1.3% 的幻觉率并列最低。紧随其后的是 o1-mini 和 GPT-4o，幻觉率分别为 1.4% 和 1.5% (图 3.2.8)。

HHEM: 幻觉率

资料来源：HHEM leaderboard, 2025 | 图表：2025 年人工智能指数报告

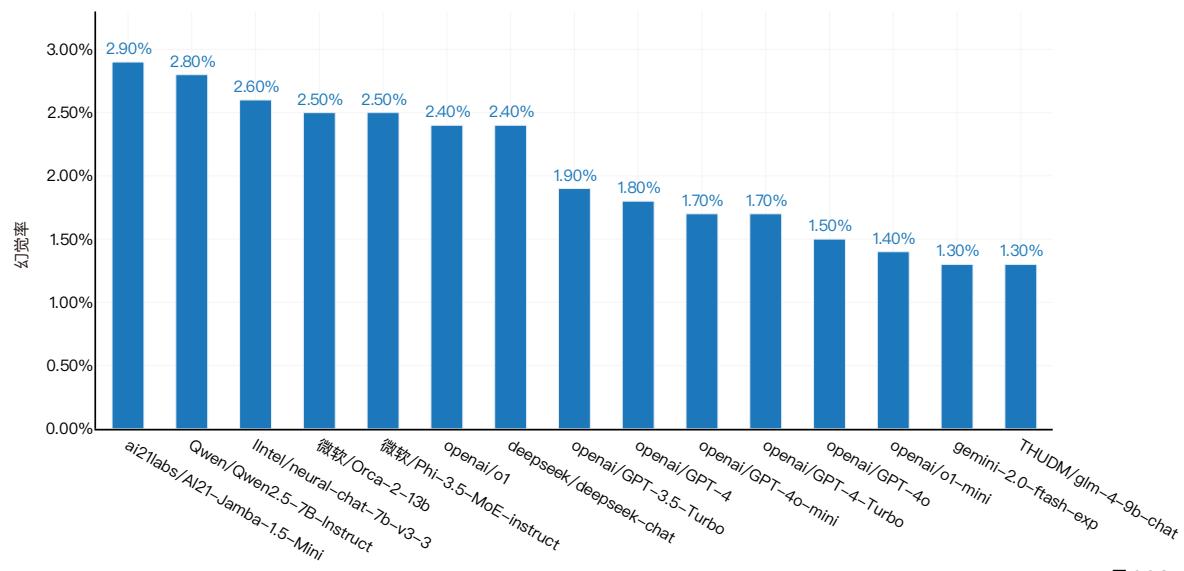


图 3.2.8

第三章：负责任的人工智能

3.2 评估负责任的人工智能

重点：

FACTS、SimpleQA 和更严格的事实性比较基准的推出

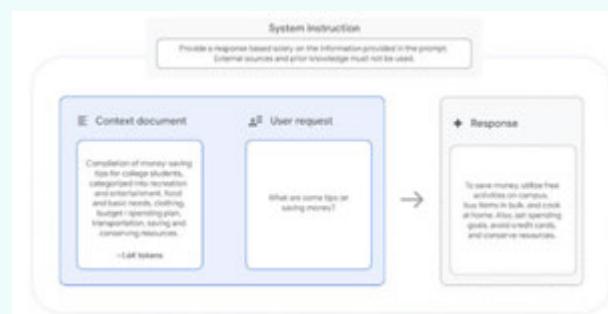
HHEM leaderboard 虽然有效，但随着模型性能的提高，似乎已接近饱和。此外，它侧重于新闻文章和摘要任务，因此全面性受到限制。随着人工智能能力的不断发展，人们越来越需要能够在更具挑战性和多样性的背景下评估事实性的比较基准。

今年，一些新的比较基准被引入，用于评估大语言模型的事实性和真实性，包括谷歌的 FACTS Grounding。该比较基准评估大语言模型在生成既准确又详细的回应以提供满意答案方面的表现。作为 FACTS 的一部分，模型必须根据上下文文档（图 3.2.9）对用户请求撰写长篇回应。这些文档涵盖广泛领域，包括金融、技术、零售、医学和法律。FACTS 比 HHEM 更复杂，要求模型执行摘要、问答生成、事实查证和解释说明等任务。评估工作由一组人工智能模型（包括 Gemini 1.5 Pro、GPT-4o 和 Claude 3.5 Sonnet）完成，它们会为每个回答给出事实性评分。目前，Gemini-2.0-Flash-Exp 模型以 83.6% 的基础得分（图 3.2.10）保持最高记录。

Stable Video Diffusion 稳定生成内容

资料来源: 谷歌, 2024

图 3.2.9



FACTS: 事实性得分

资料来源: FACTS leaderboard, 2025 | 图表: 2025年人工智能指数报告

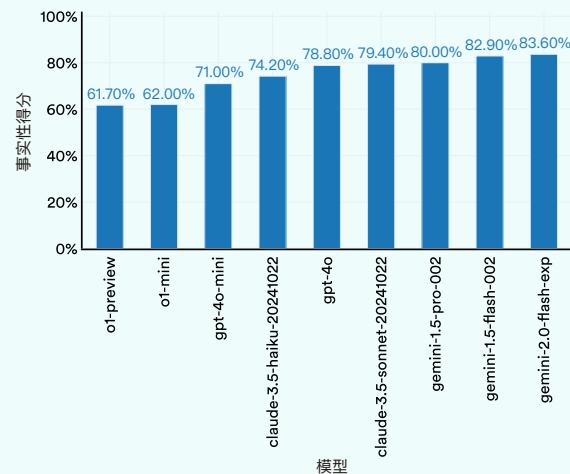


图 3.2.10

第三章：负责任的人工智能

3.2 评估负责任的人工智能

重点：

FACTS、SimpleQA 和更严格的事实性比较基准的推出(续)

评估大语言模型的事实性具有挑战性，因为其生成的冗长回答通常包含多项事实性主张，难以逐一验证准确性。为此，OpenAI 研究人员推出了 SimpleQA——一个用于评估大语言模型事实性的新基准。该基准包含 4,000 多个简短的事实查询问题，这些问题设计直接、易于评分且具有一定难度，涵盖历史、科技、艺术和地理等多个领域（图 3.2.11）。

SimpleQA 对领先的大语言模型提出了重大的事实性挑战。表现最佳的模型是 OpenAI 的 o1-preview，它只成功

回答了 42.7% 的问题（图 3.2.12）。研究人员还评估了模型是否会尝试回答某些问题，发现一些模型（如 Claude-3 系列）对 75% 的提示未作出回应。

在尝试回答问题的模型中，o1-preview 在“尝试回答且回答正确（correct-given-attempted）”的提示中得分 47.0%，其次是 Claude 3.5 Sonnet，为 44.5%。与预期一致，较大的模型在比较基准中表现更好。

SimpleQA示例问题

资料来源: OpenAI, 2024

图 3.2.11

Question	Answer
Who received the IEEE Frank Rosenblatt Award in 2010?	Michio Sugeno
On which U.S. TV station did the Canadian reality series *To Serve and Protect* debut?	KVOS-TV
What day, month, and year was Carrie Underwood's album "Cry Pretty" certified Gold by the RIAA?	October 23, 2018
What is the first and last name of the woman whom the British linguist Bernard Comrie married in 1985?	Akiko Kumahira

SimpleQA:回答问题的比例

资料来源: Wei 等, 2024 | 图表: 2025年人工智能指数报告

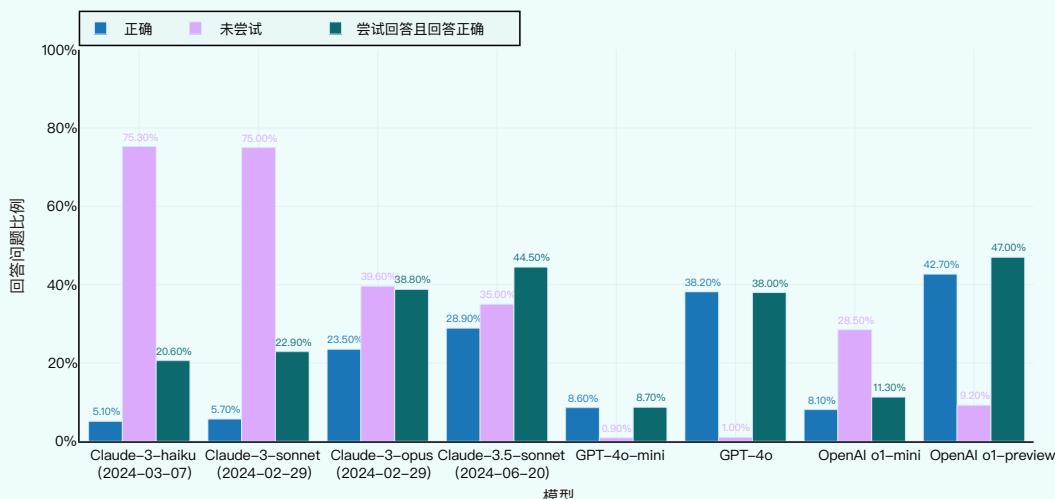


图 3.2.12

第三章：负责任的人工智能

3.3 在组织与企业中的负责任的人工智能

随着人工智能系统在实际应用场景中的广泛部署，理解企业如何应对负责任的人工智能（RAI）变得愈发重要。为深入探讨这一议题，人工智能指数于 2024 年与麦肯锡公司合作开展了一项调查，旨在评估企业在运营中整合 RAI 的程度。该调查将 RAI 定义为确保人工智能以安全、可信和符合伦理的方式开发和部署的框架。它按照人工智能指数概述的关键维度对 RAI 进行了评估：隐私与数据治理、公平性、透明度与可解释性，以及安全与保障。该调查对来自 30 多个国家的商业领袖进行了调查，总样本量为 759 人。

图 3.3.1 展示了组织对“贵组织中哪个部门主要负责人工智能治理”这一问题的回答。值得注意的是，没有单一部门占据主导地位。最常见的回答是信息安全（网络安全 / 欺诈 / 隐私），占 21%，其次是数据与分析，占 17%。此外，14% 的受访者表示其组织设有专门的人工智能治理岗位，这表明人工智能治理作为组织内独立且关键职能的地位正日益得到认可。

2024年企业人工智能治理主导部门分布

资料来源: McKinsey & Company Survey, 2024 | 图表: 2025年人工智能指数报告

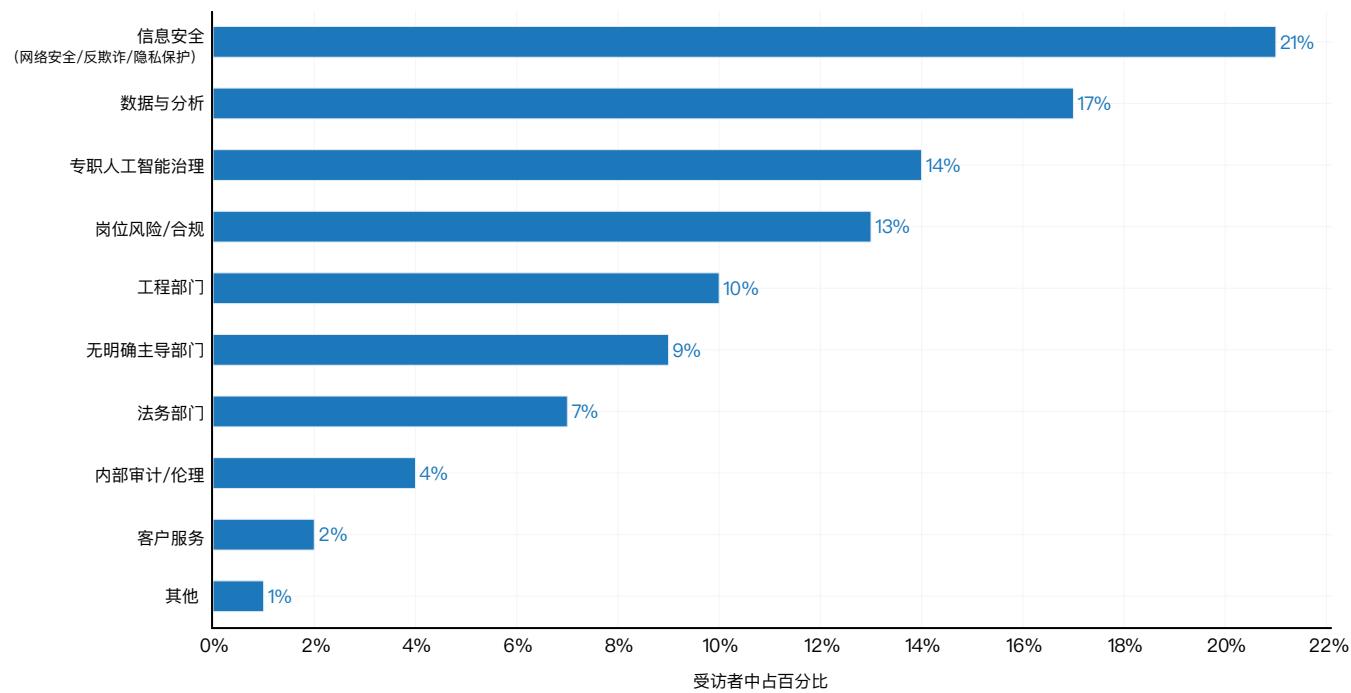


图 3.3.1²

2、“未知”选项未在此可视化中显示。

第三章：负责任的人工智能

3.3 在组织与企业中的负责任的人工智能

调查还询问了组织在未来一年内实施 RAI 方面的预计投资，包括资本支出和运营支出。此类投资的示例包括开发或购买符合 RAI 原则的技术系统，以及与 RAI 的法律或专业服务。对该问题的回答如图 3.3.2 所示，按企业收入规模分类。

大型企业——尤其是年收入超过 100 亿美元的企业——在 RAI 方面的总投资更高。值得注意的是，年收入在 100 亿美

元至 300 亿美元之间的企业中有 27%，年收入超过 300 亿美元的企业中有 21% 在 RAI 上投资了 1000 万美元至 2500 万美元。这些发现表明，大型企业更倾向于将 RAI 作为战略重点并进行更高额的绝对投资。小型组织在 RAI 上的投入较少，但许多组织仍报告了占收入比例较高的投资。

2024年企业收入规模分类的负责任的人工智能投资情况

资料来源: McKinsey & Company Survey, 2024 | 图表: 2025年人工智能指数报告

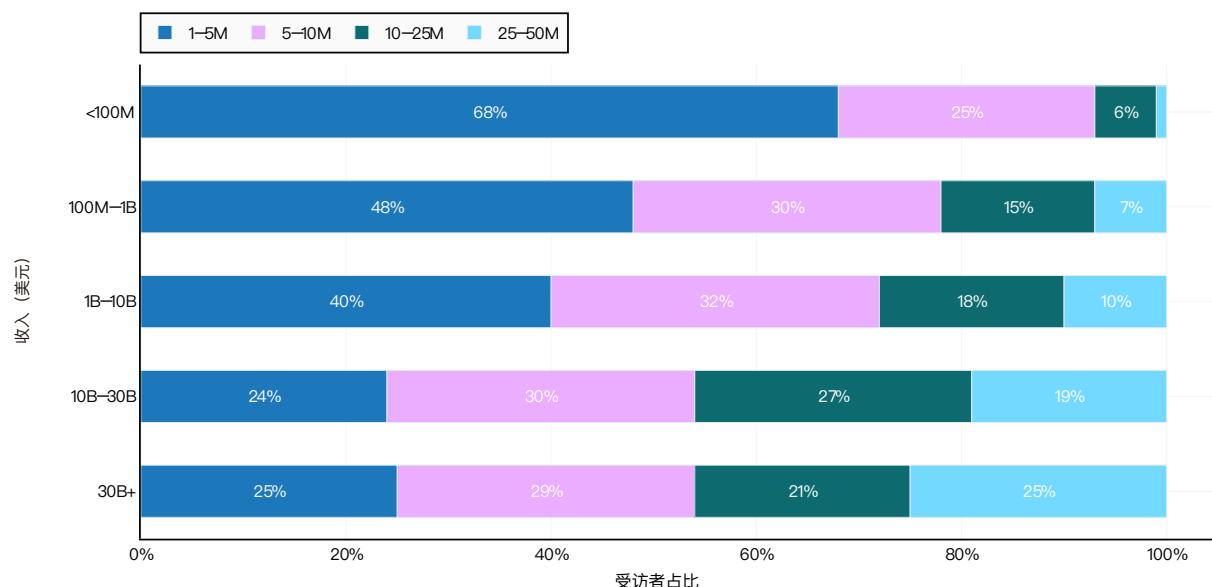


图 3.3.2

第三章：负责任的人工智能

3.3 在组织与企业中的负责任的人工智能

图 3.3.3 展示了各组织认为相关并正在积极应对的与人工智能相关的负责任的人工智能风险。网络安全 (66%)、合规监管 (63%) 和个人隐私 (60%) 被列为最主要的关注点，然而，缓解措施的实施效果始终不足。值得注意的是，在每个风险类别中，采取积极措施缓解风险的组织数量均少于那些认为这些

风险具有相关性的组织。在知识产权侵权 (57% 相关, 38% 缓解) 和组织声誉 (45% 相关, 29% 缓解) 方面，差距尤为明显。与可解释性 (40%) 和公平性 (34%) 相关的风险被较少比例的受访者选中，缓解率进一步下降至 31% 和 26%。

2024年人工智能风险相关性认知与积极缓释对比

资料来源: McKinsey & Company Survey, 2024 | 图表: 2025年人工智能指数报告

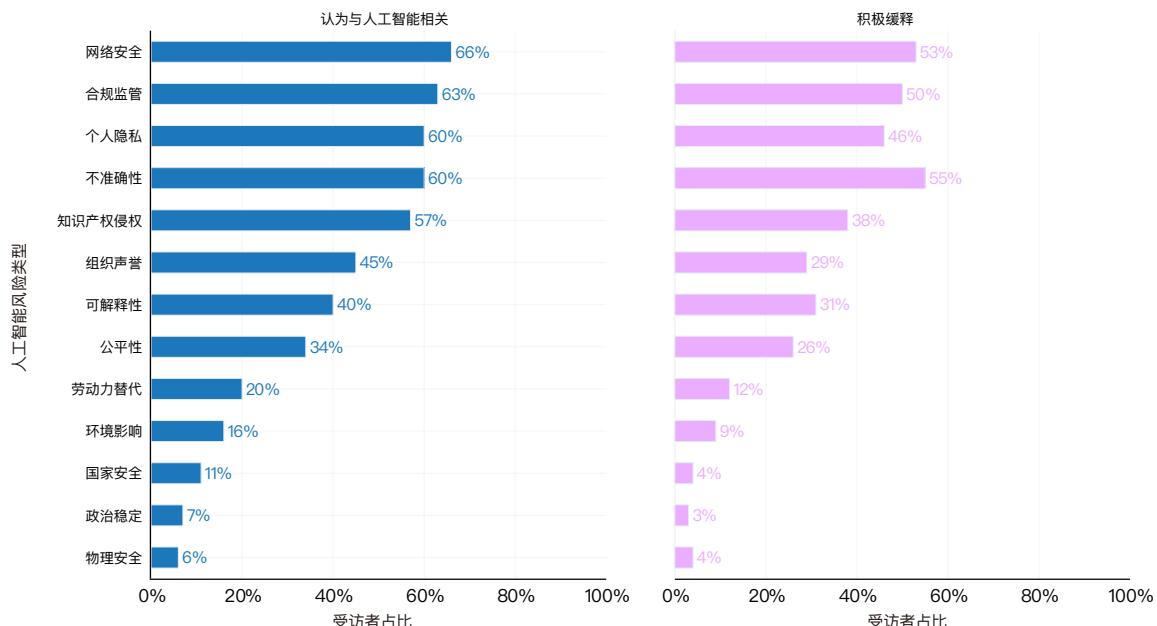


图 3.3.3

第三章：负责任的人工智能

3.3 在组织与企业中的负责任的人工智能

图 3.3.4 和图 3.3.5 展示了过去一年中组织报告的人工智能事件数量数据。在接受调查的组织中，仅有 8% 的组织报告了与人工智能相关的事件。在受影响的组织中，大多数(42%)报告仅遇到一两起事件。

2024年经历过人工智能事件的组织比例

资料来源: McKinsey & Company Survey, 2024 | 图表: 2025年人工智能指数报告

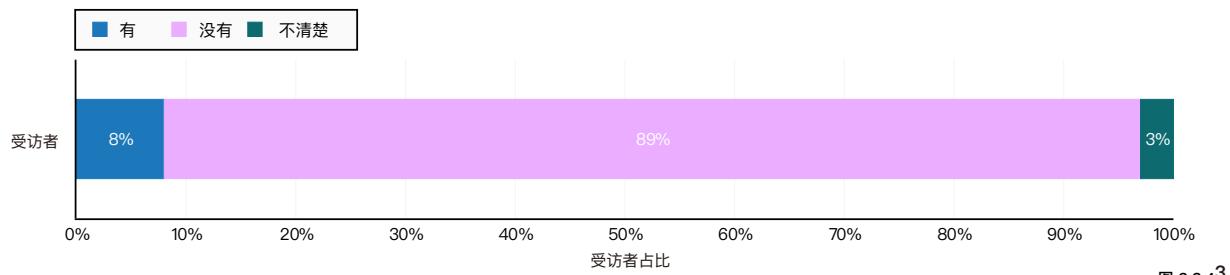


图 3.3.4³

2024年组织报告的人工智能事件数量

资料来源: McKinsey & Company Survey, 2024 | 图表: 2025年人工智能指数报告

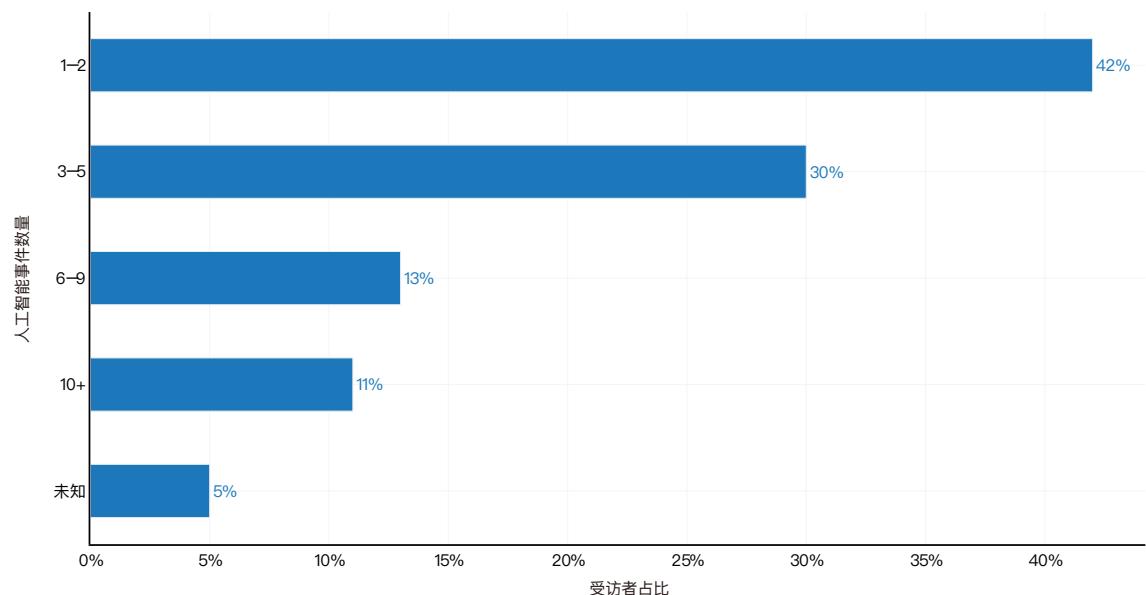


图 3.3.5

3、图 3.3.4 使用了经合组织对人工智能事件的定义。根据经合组织，人工智能事件是指一个事件、情况或一系列事件，其中一个或多个人工智能系统的开发、使用或故障直接或间接导致以下任何危害：(a) 对个人或群体的健康造成伤害或损害；(b) 关键基础设施的管理或运营受到干扰；(c) 侵犯人权或违反旨在保护基本权利、劳动权利或知识产权的法律义务；或 (d) 对财产、社区或环境造成损害。

第三章：负责任的人工智能

3.3 在组织与企业中的负责任的人工智能

当被问及RAI政策对其组织的影响时，42% 的受访者表示业务运营有所改善，例如提高效率和降低成本，34% 的受访者表示客户信任度有所提升（图 3.3.6）。仅有 17% 的组织认为这些政策未产生显著影响。

负责任的人工智能政策对组织的影响,2024

资料来源: McKinsey & Company Survey, 2024 | 图表: 2025年人工智能指数报告

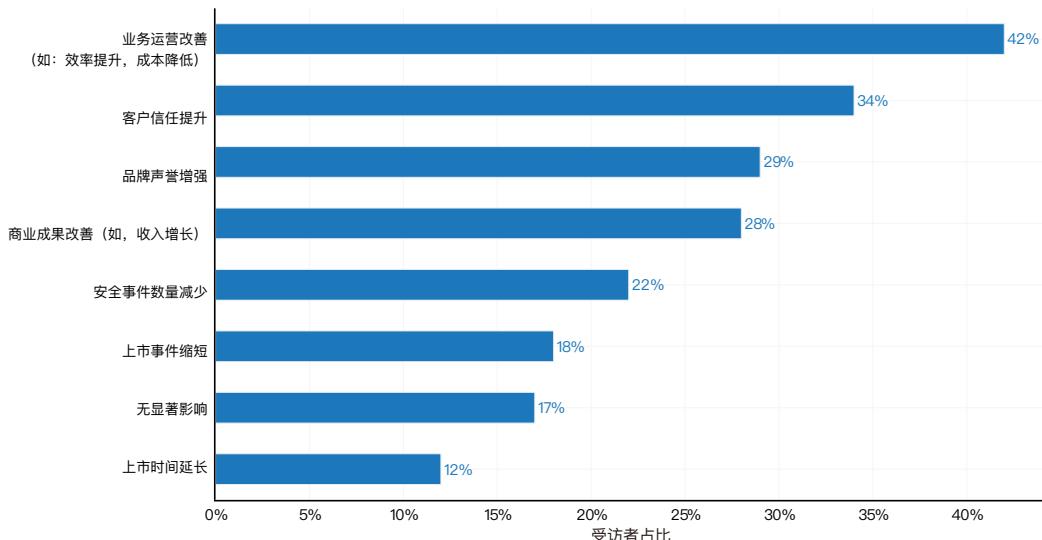


图 3.3.6⁴

⁴、D 选择“尚未实施”的受访者数据未包括在内。百分比仅基于选择至少一个其他答案的受访者。未显示“无”选项。

第三章：负责任的人工智能

3.3 在组织与企业中的负责任的人工智能

图 3.3.7 显示了组织在实施 RAI 措施时遇到的主要障碍。受访者主要提到知识和培训缺口 (51%)、资源或预算限制 (45%) 以及监管不确定性 (40%) 是主要挑战。令人鼓舞的是，

只有 16% 的受访者将缺乏高管支持作为障碍，这表明领导层的支持并不是采用 RAI 的主要障碍。

2024年实施负责任的人工智能措施的主要障碍

资料来源: McKinsey & Company Survey, 2024 | 图表: 2025年人工智能指数报告

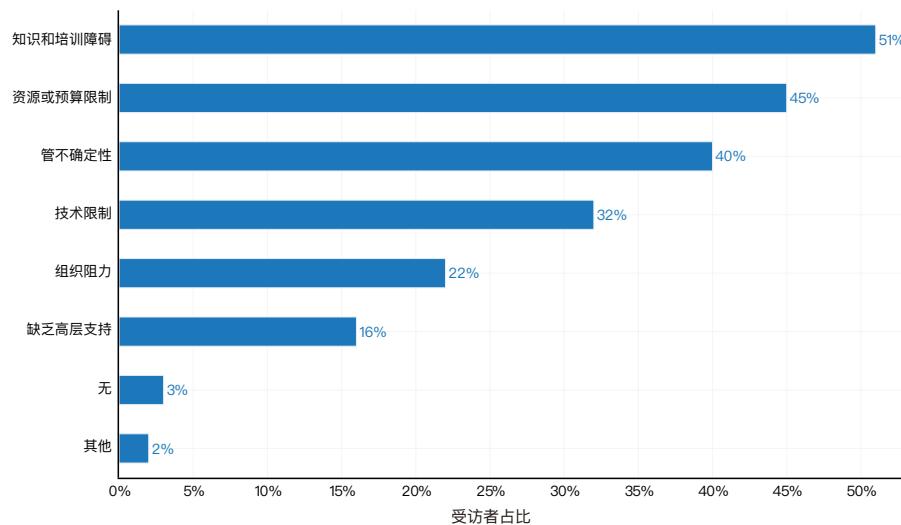


图 3.3.7⁵

⁵ “未知”回复未在此可视图表中显示。

第三章：负责任的人工智能

3.3 在组织与企业中的负责任的人工智能

图 3.3.8 显示了在人工智能决策中受到特定人工智能法规影响的组织比例。在受访组织中，65% 的组织表示受到欧盟《通用数据保护条例》(GDPR) 的影响，而 41% 的组织提到了欧盟《人工智能法案》。较小比例的组织表示受到经济合作与发展

组织 (OECD) 人工智能原则的影响 (21%) 以及拜登总统关于人工智能的行政命令的影响。

受人工智能法规影响的组织在负责任的人工智能决策中的比例

资料来源: McKinsey & Company Survey, 2024 | 图表: 2025年人工智能指数报告

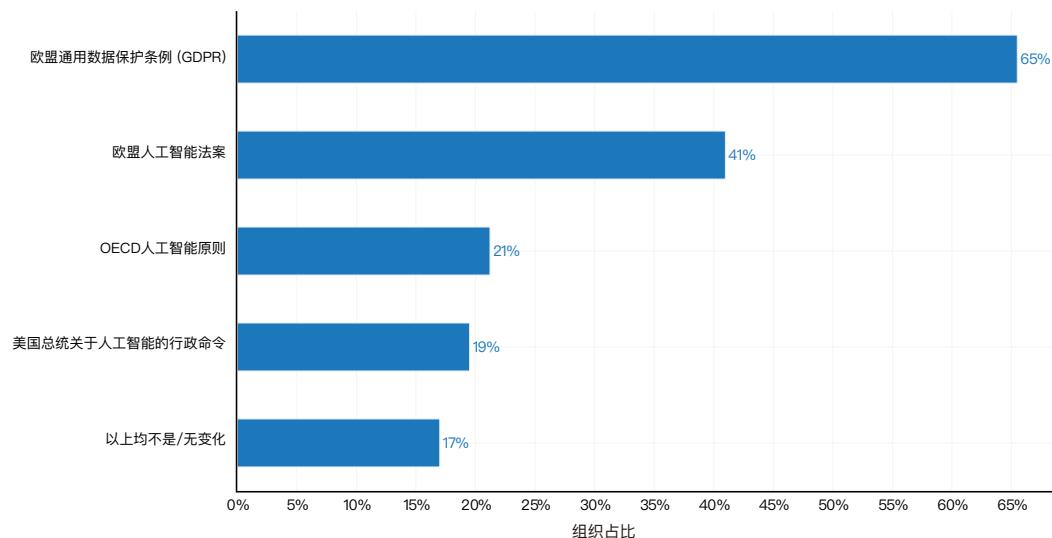


图 3.3.8

第三章：负责任的人工智能

3.3 在组织与企业中的负责任的人工智能

重点：

纵向视角

斯坦福大学研究团队与埃森哲合作，于 2025 年 1 月至 2 月开展了第二轮全球负责任人工智能现状调查（首轮调查于 2024 年启动）。本次调查覆盖 20 个国家、19 个行业的 1,500 家组织（年营收均超过 5 亿美元），旨在分析企业采纳 RAI 原则与实践的挑战，并对比 10 个维度的 RAI 活动随时间的变化趋势。⁶ 由于该调查在 2024 年和 2025 年均有实施，数据可反映组织对 RAI 采纳态度的演进过程。

过去两年组织报告的人工智能事件类型

资料来源: Accenture/Stanford Joint Survey, 2025 | 图表: 2025年人工智能指数报告

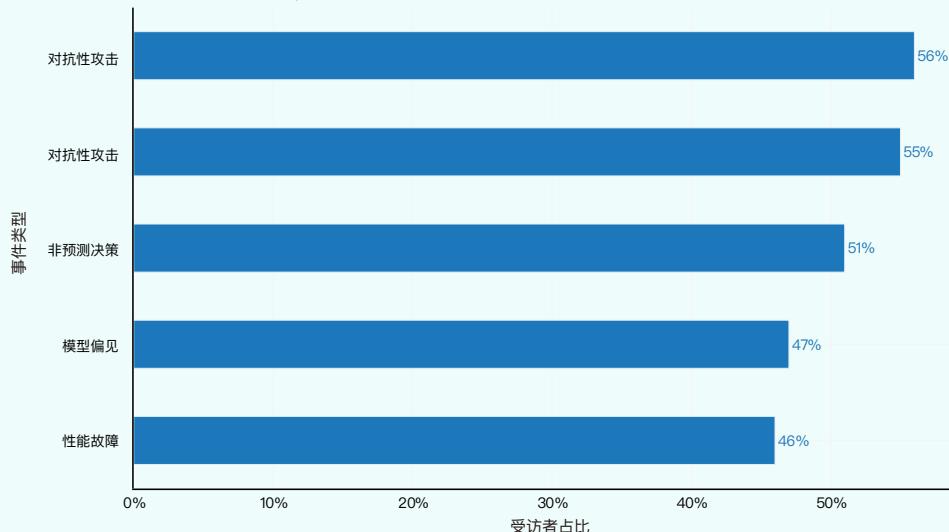


图 3.3.9 展示了受访组织在过去两年中报告的人工智能相关事件类型。最突出的问题是对抗性攻击（56%）和隐私侵犯（55%），凸显了企业亟需加强人工智能系统安全性与数据治理。此外，51% 的受访者报告了非预期决策，47% 提及模型偏见，表明许多组织在预测和控制人工智能行为方面存在困难——这一挑战在高风险环境中尤为严峻。

6、调查方法详见 Reuel 等 2024 年的研究报告。

第三章：负责任的人工智能

3.3 在组织与企业中的负责任的人工智能

重点：

纵向视角（续）

根据企业采用人工智能的战略差异（例如开发、部署或使用生成式 / 非生成式人工智能），受访者需要评估 14 类风险对其组织的相关程度（图 3.3.10）。⁷ 调查显示，近年来企业对特定风险的关注度显著上升——最突出的是财

务风险（+38 个百分点）、品牌与声誉风险（+16）、隐私与数据相关风险（+15）以及可靠性风险（+14）。相反，社会风险（-7）和社会环境影响风险（-8）的紧迫性有所下降。

2024–2025年组织关注的负责任人工智能风险变化对比

资料来源: Accenture/Stanford Joint Survey, 2025 | 图表: 2025年人工智能指数报告

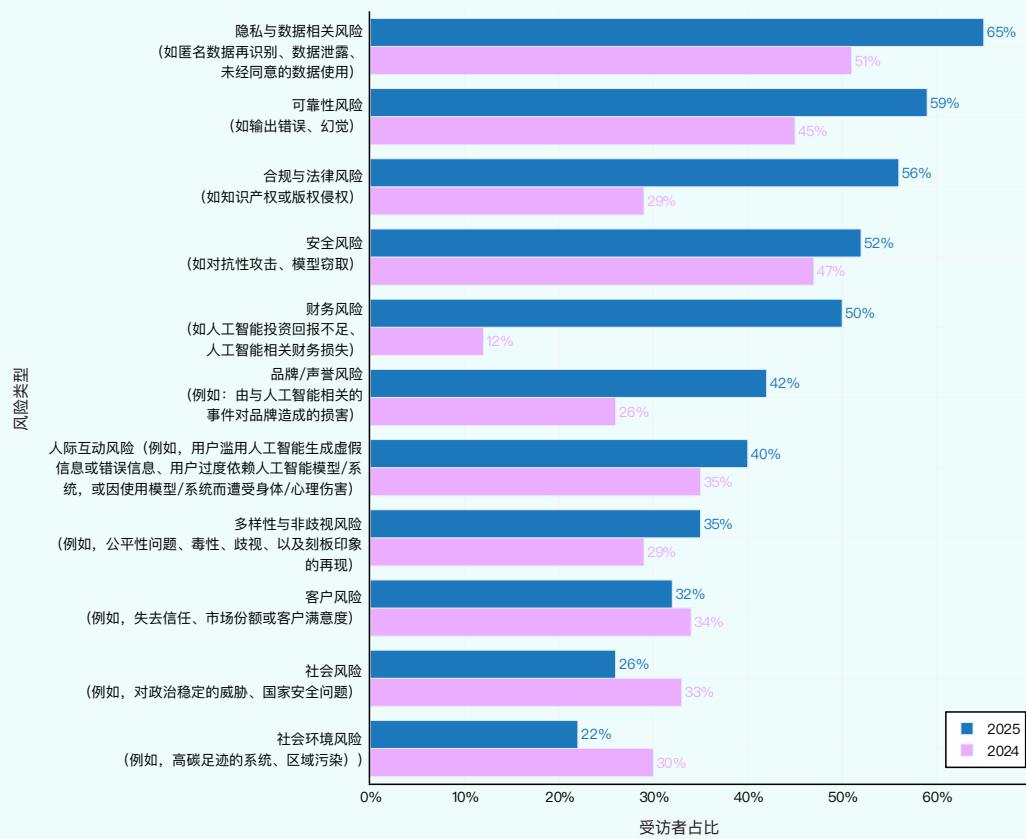


图 3.3.10

7、调查方法详见 Reuel 等 2024 年的研究报告。

第三章：负责任的人工智能

3.3 在组织与企业中的负责任的人工智能

重点：

纵向视角

组织与运营成熟度的定义如图 3.3.11 所示。2024 至 2025 年间，组织层面的负责任人工智能成熟度显著提升——更多企业获得了 CEO 对 RAI 计划的支持，并改善了人工智能风险识别、监测与控制能力，这标志着对 RAI 人工智能战略重要性的认知进一步增强（图 3.3.12）。⁸ 相比之下，聚焦于系统级实操保障（如偏见消减、对抗性测试及环境影响评估等）的运营层面 RAI 成熟度进展滞后（图 3.3.13）。这一差距揭示了高层 RAI 承诺与技术落地之间的脱节：尽管各组织在将 RAI 纳入流程与政策的意愿和资源配置上持续增强，但如何将这些意图转化为有效的系统级实践仍面临挑战。

组织与运营成熟度模型

资料来源: Reuel 等, 2024

Level	Score	Organizational Maturity	Score	Operational Maturity
Level 1: Initial	[0 , 12.5]	The organization has limited awareness and no organizational plans, processes, or frameworks in place to ensure a responsible AI adoption.	[0 , 12.5]	The organization does not mitigate identified risks on a system level.
Level 2: Assessing	[12.5 , 37.5]	The organization is aware of the necessity for organizational measures to ensure a responsible AI adoption and is assessing governance options.	[12.5 , 37.5]	Awareness of risks may be present, but the organization has only limited or no formal mitigation measures in place.
Level 3: Determined	[37.5 , 62.5]	The organization demonstrates foundational governance capabilities to support the responsible development, deployment, and use of AI.	[37.5 , 62.5]	A few risk mitigation measures are being fully operationalized, but the majority is only implemented ad hoc or in early roll-out stages. There is a growing awareness of the need for more systematic approaches.
Level 4: Managed	[62.5 , 87.5]	The organization has established comprehensive organizational RAI measures and is actively ensuring enterprise-wide adoption, demonstrating a mature and effective approach to internal RAI governance.	[62.5 , 87.5]	A wider range of risk mitigation measures are fully operationalized across all relevant AI systems in the organization.
Level 5: Optimized	[87.5 , 100]	The organization demonstrates an established, future-oriented approach towards organizational RAI, ensuring a sustainable and responsible approach to organizational RAI.	[87.5 , 100]	Comprehensive, state-of-the-art risk mitigation strategies are fully operationalized. The organization continuously monitors and evaluates risks, proactively adapting its practices as needed to mitigate new risks.

图 3.3.11

2024年和 2025年组织领域负责任的人工智能成熟度分布

资料来源: Accenture/Stanford Joint Survey, 2025 | 图表: 2025年人工智能指数报告

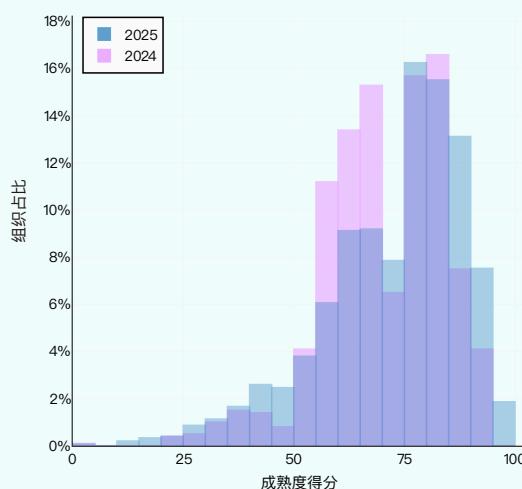


图 3.3.12

2024年和 2025年运营领域负责任的人工智能成熟度分布

资料来源: Accenture/Stanford Joint Survey, 2025 | 图表: 2025年人工智能指数报告

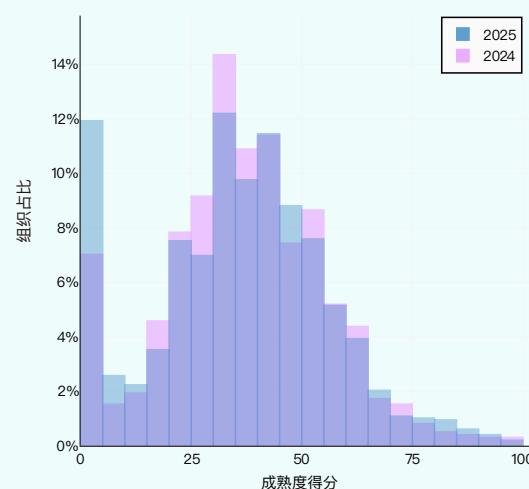


图 3.3.13

8、组织和运营 RAI 成熟度是根据 Reuel 等 (2024) 中定义的方法计算得出。

第三章：负责任的人工智能

3.3 在组织与企业中的负责任的人工智能

重点：

纵向视角（续）

还询问了受访者对其组织对 RAI 的态度和理念，包括对风险所有权、模型偏好和政策立场的看法（图 3.3.14）。在几乎所有陈述中，回答都相当均衡，即使是在备受关注的问题上，例如开放式与封闭式权重模型的安全性，以及风险缓解的责任在于模型提供商还是用户。这种广泛分布表明，行业在 RAI 方面缺乏统一的战略方向，这可能反映

了专家们正在进行的争论和未解决的问题。唯一明显的例外是安全与创新之间的权衡：64% 的受访者倾向于安全第一的方法，但 58% 的受访者正在探索最低限度监督的代理，这可能会带来重大风险，特别是考虑到 RAI 目前依然受限于不成熟状态。

组织对负责任的人工智能的态度与理念

资料来源: Accenture/Stanford Joint Survey, 2025 | 图表: 2025年人工智能指数报告

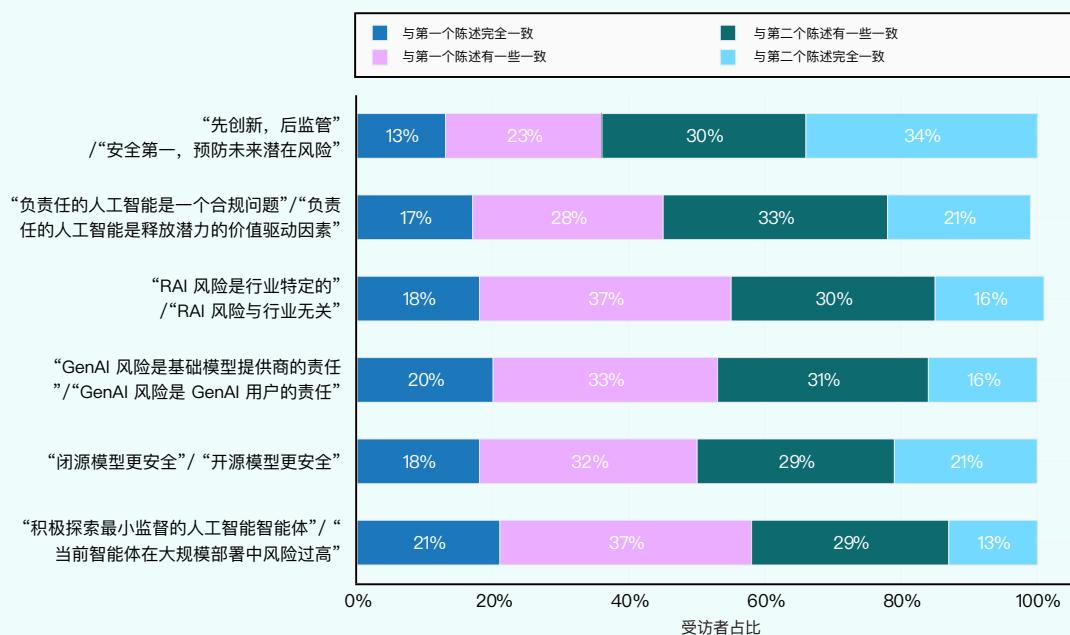


图 3.3.14

第三章：负责任的人工智能

3.4 在学术界中的负责任的人工智能

今年，人工智能指数分析了六大顶级人工智能学术会议上被接受的与人工智能相关的论文数量：AAAI、AIES、FAccT、ICML、ICLR 和 NeurIPS。尽管这些会议并不代表全球所有人工智能研究，但它们为了解人工智能学术界的发表趋势提供了洞察。本节呈现人工智能论文发表统计的总体趋势，后续章节将按 RAI 子领域进行细分。为了识别 RAI 论文，人工智能指数筛选了包含特定 RAI 关键词的论文。⁹

总体趋势

在顶级人工智能会议上被接受的 RAI 论文数量增长了 28.8%，从 2023 年的 992 篇增加到 2024 年的 1,278 篇（图 3.4.1）。

2019–2024年主要人工智能会议收录负责任的人工智能论文数量统计

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

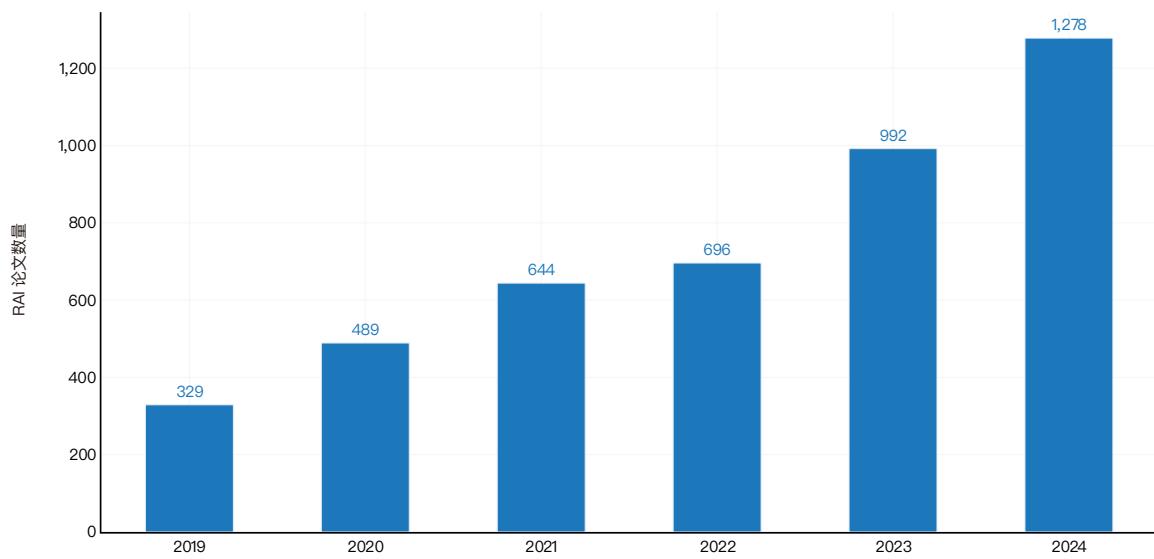


图 3.4.1

9、本方法的完整方法论描述详见附录。

第三章：负责任的人工智能

3.4 在学术界中的负责任的人工智能

相对而言，RAI 论文占总投稿比例最高的会议是 FAccT (69.14%) 和 AIES (63.33%) (图 3.4.2)。这与它们的重点相一致：FAccT 致力于公平、问责和透明，而 AIES 则侧重于人工智能伦理与社会。

在 NeurIPS，该比例从 2023 年的 13.8% 下降至 2024 年的 9.0%，而在 ICML，同一时期该比例从 3.4% 上升至 8.2%。

2019–2024年主要人工智能会议收录负责任的人工智能论文数量统计

资料来源: 2025年人工智能指数 | 图表: 2025年人工智能指数报告

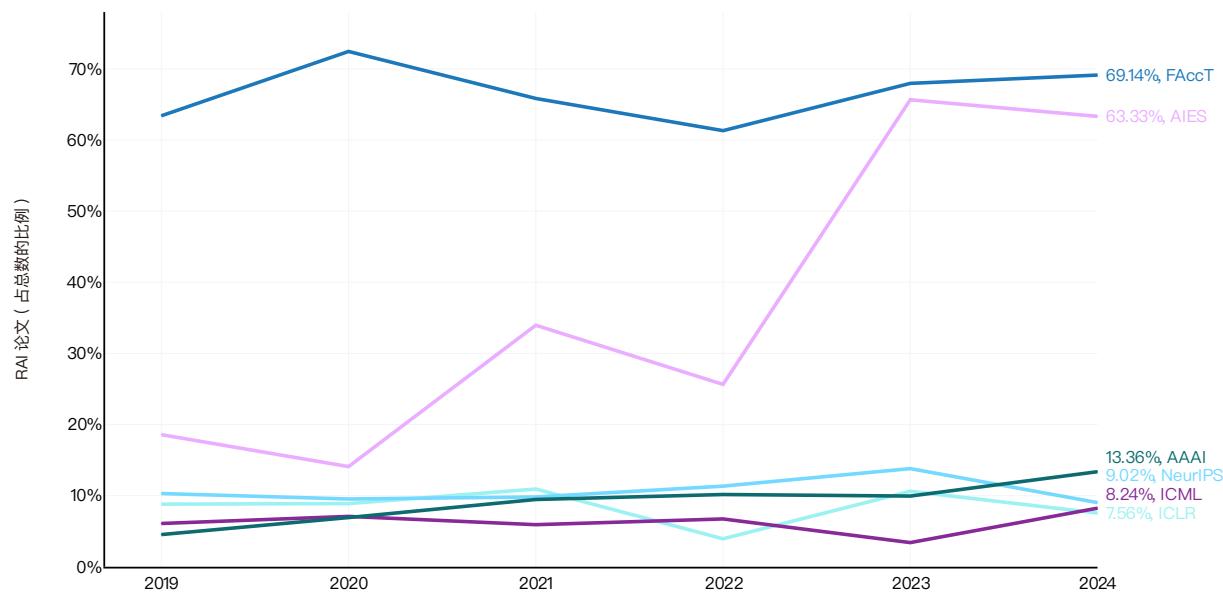


图 3.4.2

第三章：负责任的人工智能

3.4 在学术界中的负责任的人工智能

图 3.4.3 至 3.4.5 分析了 RAI 论文的地理归属，重点展示了这些论文的来源地。2024 年，美国在 RAI 论文投稿数量上居首，达 669 篇，其次是中国（268 篇）和德国（80 篇）。在主要地

理区域中，RAI 已成为越来越重要的学术研究领域。自 2019 年以来，RAI 论文的地理分布总体保持相对稳定，其中美国占比最高（3,158 篇），其次是中国（1,100 篇）和英国（485 篇）。

2024年按地域分布主要人工智能会议负责任的人工智能（RAI）论文收录数量

资料来源: 2025年人工智能指数 | 图表: 2025年人工智能指数报告

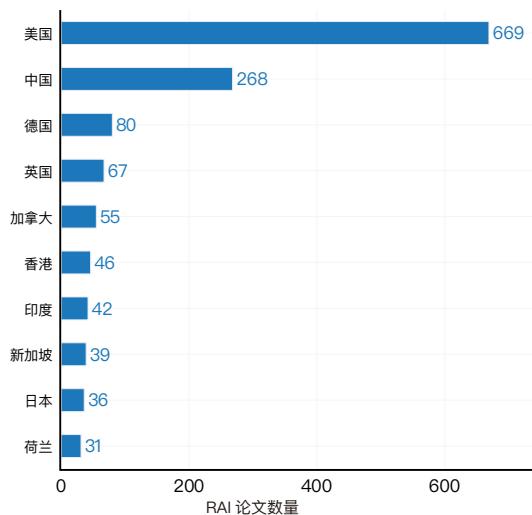


图 3.4.3

2019–2024年按主要地域分布主要人工智能会议负责任的人工智能（RAI）论文收录数量

资料来源: 2025年人工智能指数 | 图表: 2025年人工智能指数报告



图 3.4.4

2019–2024年按地域分布主要人工智能会议负责任的人工智能论文收录总量

资料来源: 2025年人工智能指数 | 图表: 2025年人工智能指数报告

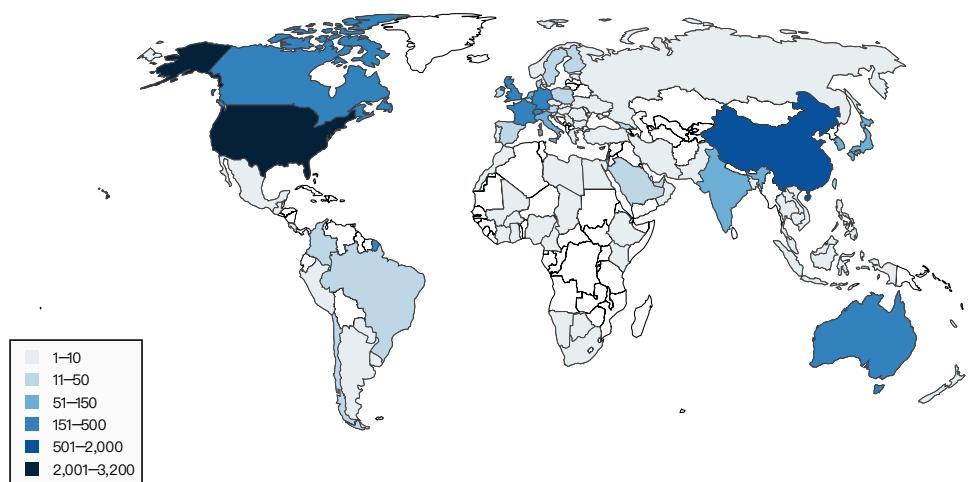


图 3.4.5

第三章：负责任的人工智能

3.4 在学术界中的负责任的人工智能

主题领域

本节分析了 RAI 论文发表统计在关键主题领域的趋势，包括隐私与数据治理、公平性、透明度与可解释性，以及安全与可靠性。过去一年，在主要人工智能会议上，隐私与数据治理相关

论文的录用数量下降了 14.5% (图 3.4.6)。自 2019 年以来，这一数字已增长近五倍。

2019–2024年在主要人工智能学术会议上人工智能隐私与数据治理领域论文收录数量

资料来源: 2025年人工智能指数 | 图表: 2025年人工智能指数报告

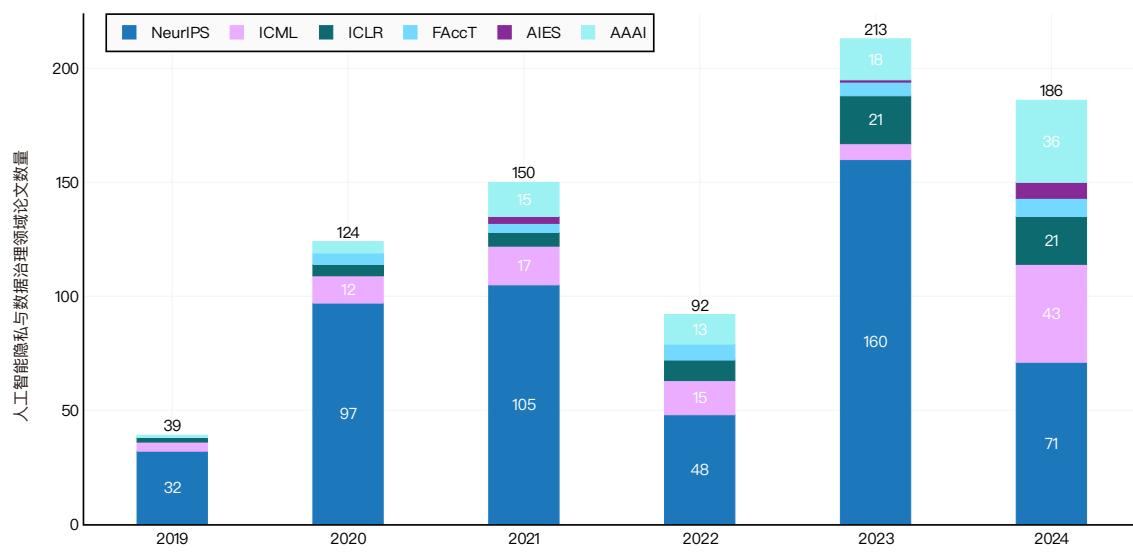


图 3.4.6¹⁰

10. 这些数据可能低估了人工智能隐私研究论文的总数，因为部分论文发表在专注于隐私的人工智能专业会议上，例如第 46 届 IEEE 安全与隐私研讨 (IEEE Symposium on Security and Privacy)。

第三章：负责任的人工智能

3.4 在学术界中的负责任的人工智能

2024 年，在主要人工智能学术会议上被录用的公平性与偏见相关论文数量显著增长，达到 408 篇——约为 2023 年数量的两倍（图 3.4.7）。这一增长凸显了研究人员对公平性与偏见问题的学术关注日益增强。

2019–2024年在主要人工智能学术会议上人工智能公平与偏见论文收录数量

资料来源: 2025年人工智能指数 | 图表: 2025年人工智能指数报告

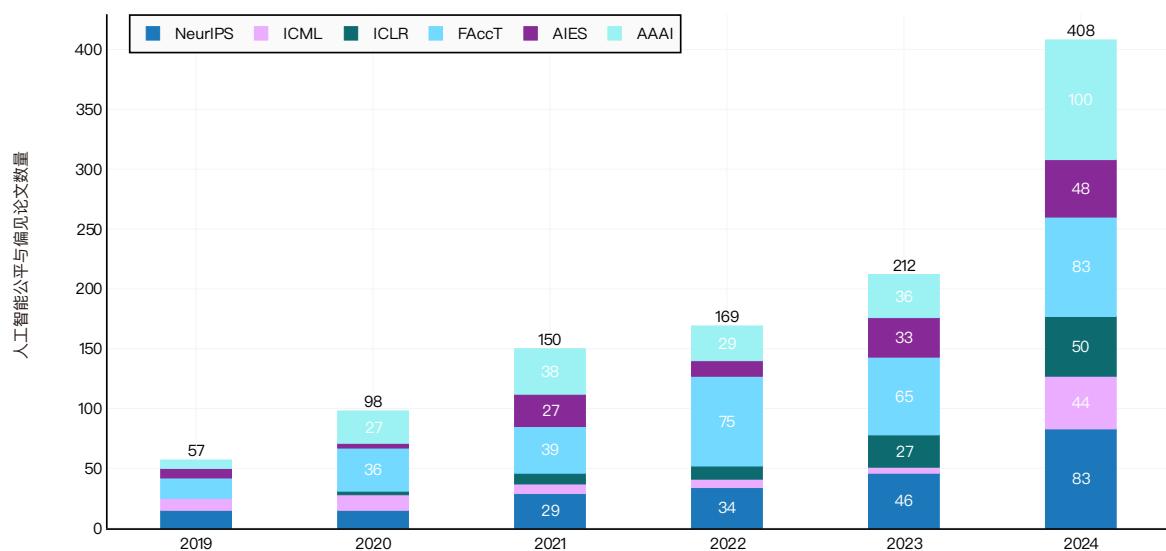


图 3.4.7

第三章：负责任的人工智能

3.4 在学术界中的负责任的人工智能

自 2019 年以来，提交至主要学术会议的关于透明度和可解释性的论文数量增加了四倍。2024 年，包括 AAAI、FAccT、AIES、ICML、ICLR 和 NeurIPS 在内的学术会议上提交了 355 篇与透明度和可解释性相关的论文(图 3.4.8)。

2019–2024年在主要人工智能学术会议上人工智能透明度与可解释性论文收入数量

资料来源: 2025年人工智能指数 | 图表: 2025年人工智能指数报告

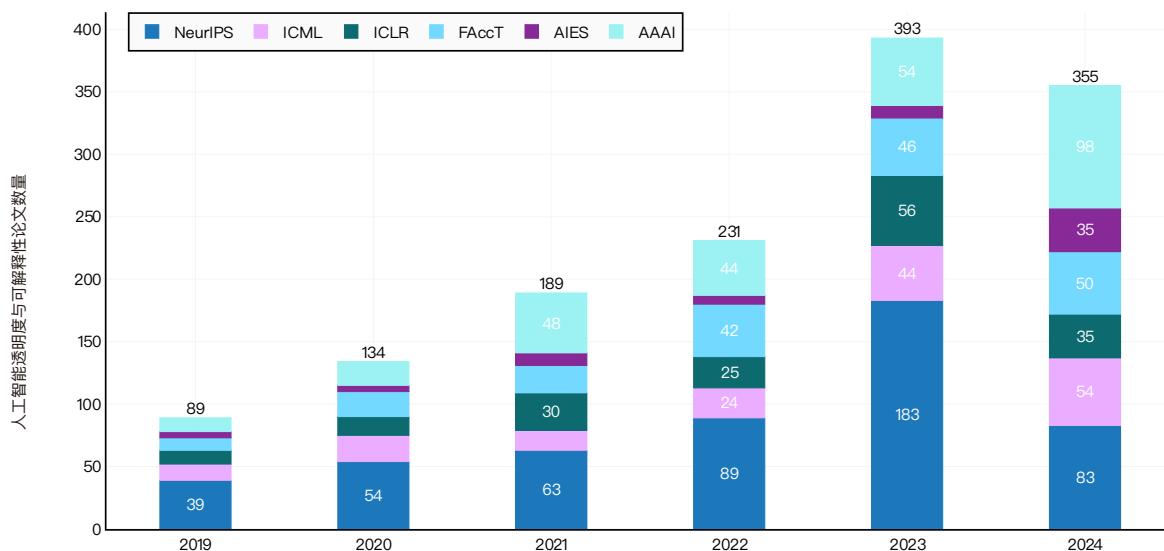


图 3.4.8

第三章：负责任的人工智能

3.4 在学术界中的负责任的人工智能

提交至选定人工智能会议的有关安全领域论文数量大幅增长，过去一年几乎翻了一番——从 276 篇增至 521 篇（图 3.4.9）。这一增长反映了安全与安全作为人工智能研究人员核心关注领域的地位日益提升。

2019–2024年在主要人工智能学术会议安全领域收录论文数量

资料来源: 2025年人工智能指数 | 图表: 2025年人工智能指数报告

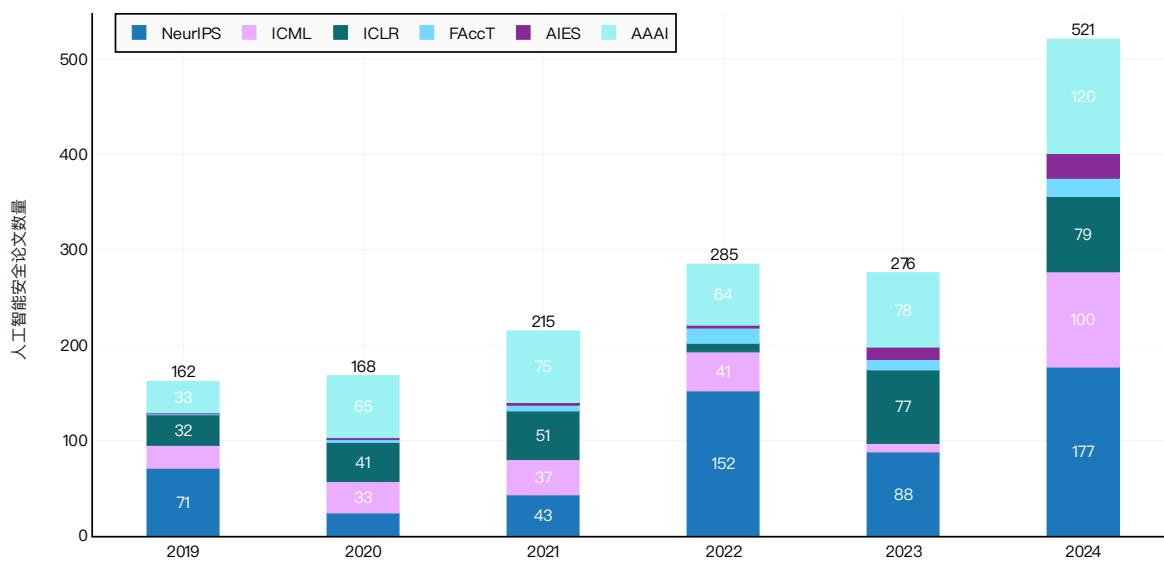


图 3.4.9

第三章：负责任的人工智能

3.5 负责任的人工智能政策制定

3.5 负责任的人工智能政策制定

尽管 2023 年和 2024 年初各国人工智能战略和监管方案激增，但 2024 年的显著趋势是全球在人工智能治理领域的合作加强，特别是在 RAI 立法原则方面。国际组织和多边协议正致力于建立全球性的负责任与伦理人工智能框架，标志着从孤立的国家行动向协同全球治理的转变。¹¹

重大的负责任的人工智能政策的里程碑

资料来源：2025年人工智能指数

时间	参与方	范围	内容概述
2024年5月	经合组织	全球	经合组织更新了其人工智能原则并完善了其框架，以反映人工智能治理方面的最新进展。这些原则强调，建立人工智能系统要考虑包容性增长、透明度和可解释性，以及对法治、人权和民主价值观的尊重。
2024年5月	欧洲委员会	欧洲	欧洲委员会通过了一项具有法律约束力的人工智能条约（《欧洲委员会人工智能与人权、民主和法治框架公约 (The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law)》）。该条约旨在确保人工智能系统生命周期内的活动完全符合人权、民主和法治。
2024年6月	欧洲联盟	欧洲	欧盟通过了《人工智能法 (AI Act)》（《欧盟人工智能法 (EU AI ACT)》），这是全球主要经济体首个全面的人工智能监管框架。该法案按风险对人工智能进行分类，对其进行相应的监管，并确保高风险系统的提供商或开发商承担主要义务。
2024年7月	非洲联盟	非洲	非洲联盟推出了非洲大陆人工智能战略 (AU AI Strategy)，概述了整个非洲大陆人工智能发展、伦理和治理的统一愿景。该战略强调在非洲以符合伦理、负责任和公平的方式发展人工智能。
2024年9月	联合国	全球	联合国更新了其 Governing AI for Humanity 报告（联合国人工智能咨询机构），概述了建立全球人工智能治理机制的努力。该报告建议制定一个蓝图，以应对人工智能相关风险，并呼吁国家和国际标准组织、技术公司、民间社会和政策制定者就人工智能标准开展合作。
2024年10月	G7	全球	G7 数字竞争公报（G7 人工智能合作）重申了对公平开放的人工智能市场的承诺，强调了协调监管方法的必要性。此前的讨论主要集中在竞争和人工智能快速发展带来的监管挑战上。
2024年10月	东盟与美国	亚洲和美国	在第12届东盟-美国峰会之后，东盟-美国领导人发表了一份关于促进安全、可靠和可信的人工智能的声明。他们承诺合作制定国际人工智能治理框架和标准，以推进这些目标的实现。
2024年11月	国际人工智能安全研究所网络	全球	首个国际人工智能安全研究所网络成立，将九个国家和欧盟联合起来，正式开展全球人工智能安全合作。该网络联合了致力于推进人工智能安全的技术组织，帮助政府和社会了解先进人工智能系统的风险，并提出解决方案。
2025年2月	阿拉伯联盟	阿拉伯国家	阿拉伯对话圈 (Arab Dialogue Circle) 活动“人工智能在阿拉伯世界：创新应用与伦理挑战”在阿拉伯联盟总部启动，聚焦人工智能创新，同时高度重视伦理考量。

11. 虽然人工智能政策制定是第六章：政策的重点，但人工智能指数在此强调了与 RAI 相关的主要政策制定事件，因为这些事件最近具有重要意义。

第三章：负责任的人工智能

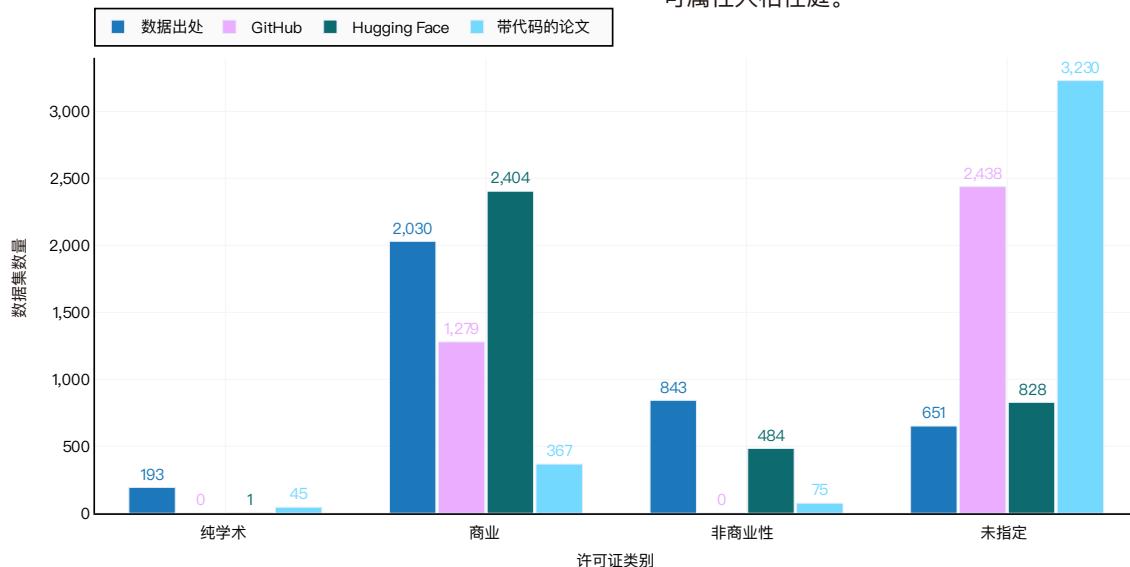
3.6 隐私和数据治理

隐私的定义非常复杂，因具体情况而异。为了便于本报告使用，人工智能指数将隐私定义为个人对其个人数据的保密、匿名和保护的权利，以及个人对数据是否被使用以及如何被使用给予同意和获得通知的权利。隐私还包括组织在（直接或间接）收集、存储或使用个人数据时确保这些权利的责任。此外，如果组织或政府歪曲了个人信息，个人应有权更正其敏感信息。在人工智能领域，这涉及确保以尊重个人隐私权的方式处理个人数据，例如，采取措施保护敏感信息免遭泄露，并确保数据收集和处理透明且符合《通用数据保护条例》等隐私法律。

数据治理，另一方面，是指组织为确保数据在其创建的组织内部和外部的质量、安全和道德使用而建立的政策、流程和标准。数据治理政策还可能涵盖从外部来源获取的数据。在人工智能领域，数据治理对于确保用于训练和操作人工智能系统的数据准确、公平、负责任且经同意使用非常重要。对于敏感或个人身份信息（PII）而言，这一点尤为重要。

选定聚合平台对数据集许可分类的准确率

资料来源：Longpre 等,2025 | 图表：2025 年人工智能指数报告



特色研究

本节重点介绍近期有关隐私和数据管理的重要研究，包括有关数据集许可和归属审计的研究，以及有关更严格的数据许可协议的研究。

人工智能数据集许可和归属的大规模审计

当前的基础模型是在海量数据的基础上进行训练的。一组研究人员对广泛用于训练此类模型的 1,800 多个文本数据集进行了大规模审计，发现了数据集许可和归属方面的系统性问题。研究人员发现，在流行的数据集托管网站上，超过 70% 的数据集缺乏足够的许可证信息，而 50% 的许可证被错误归类了，这给负责任地使用数据带来了风险。图 3.6.1 提供了研究人员调查结果的详细可视化示意图。具体来说，他们为数据集分配了四个类别的许可证标签：商业、未指定、非商业和纯学术。然后，他们将自己的分类与 GitHub、Papers with Code 和 Hugging Face 平台等流行来源的分类进行了比较。很多时候，数据来源团队分配的数据许可属性与其他组织发布的数据许可属性大相径庭。

图 3.6.1

第三章：负责任的人工智能

3.6 隐私和数据治理

数据集中的许可错误归属意义重大，因为它会给人工智能开发带来法律和伦理风险。如果用于训练基础模型的数据集被错误标注或错误归属，人工智能开发人员可能会在不知情的情况下违反版权法、数据使用政策或隐私法规。这可能会法律责任、确保数据创建者获得公平性补偿方面的挑战，以及由于排除了获得适当许可的数据而导致模型出现潜在偏差。此外，不明确的许可会阻碍人工智能研究的透明度、问责制和可重复性，从而使研究人员和机构难以验证或审核模型训练数据。根据他们的研究结果，作者强调需要清晰的文档、改进的标准和负责任的许可实践，以促进包容性并降低因人工智能开发和部署中不负责任或非法使用数据而产生的风险。

危机中的数据许可

人工智能模型在很大程度上依赖于大量公开的网络数据进行训练。最近的一项研究对人工智能训练数据集（包括 C4、RefinedWeb 和 Dolma）中使用的网域同意协议进行了纵向审计，分析了 14000 个网域。这些同意协议规定了为人工智能模型训练而进行数据搜刮的允许性。

研究人员观察到，在 2023 年至 2024 年期间，数据使用限制大幅增加，因为许多网站实施了新的协议来限制为人工智能训练而进行的数据搜刮。这些限制主要是通过更新 robots.txt 文件和服务条款，明确禁止使用人工智能训练。图 3.6.2 显示了随着时间推移，带有 robots.txt 限制、服务条款限制和组织限制的网站比例。¹² 例如，在前 C4 个网域中，带有完全限制的词元比例从 2017 年的 10% 上升到 2024 年的 48%。仅在 2023 年至 2024 年间，这一比例就上升了 25 个百分点。图 3.6.3 按服务条款限制类别直观显示了 2016 年至 2024 年 C4 顶级网域中的词元比例。这种同意程度的降低很可能与围绕合理使用 的法律问题有关，如《纽约时报》对 OpenAI 的诉讼。

OpenAI 的爬虫遇到的限制最多，而小型开发者面临的障碍较少。作者强调，robots.txt 等无效的信号机制以及声明与执行政策之间的不匹配导致了执行上的不一致。这些发现凸显了更新同意协议以应对人工智能特定挑战的必要性。此外，研究还表明，用于人工智能训练的公开可用网络数据有所减少，这对数据多样性、模型对齐和可扩展性具有潜在影响。最近许多人工智能的性能都来自于在越来越的数据集上进行的训练。如果网站的限制性明显增加，可能会阻碍未来模型的扩展。

12、robots.txt 限制是指在网站的 robots.txt 文件中设置的规则，用于指示网络爬虫（如搜索引擎机器人或人工智能数据搜刮器）允许或禁止访问网站的哪些部分。

第三章：负责任的人工智能

3.6 隐私和数据治理

2016–2024年按robots.txt限制类别划分的C4顶级网域词元的比例

资料来源：Longpre 等,2025| 图表：2025 年人工智能指数报告

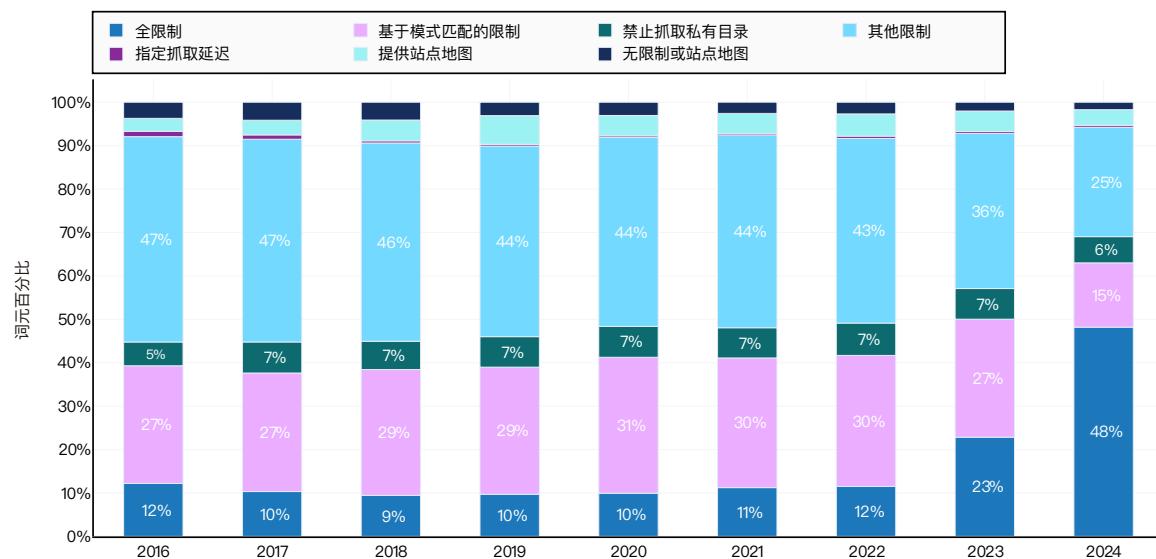


图 3.6.2

2016–2024年C4数据集顶级网络域名的内容使用条款限制类别占比分布

资料来源：Longpre 等,2025| 图表：2025 年人工智能指数报告

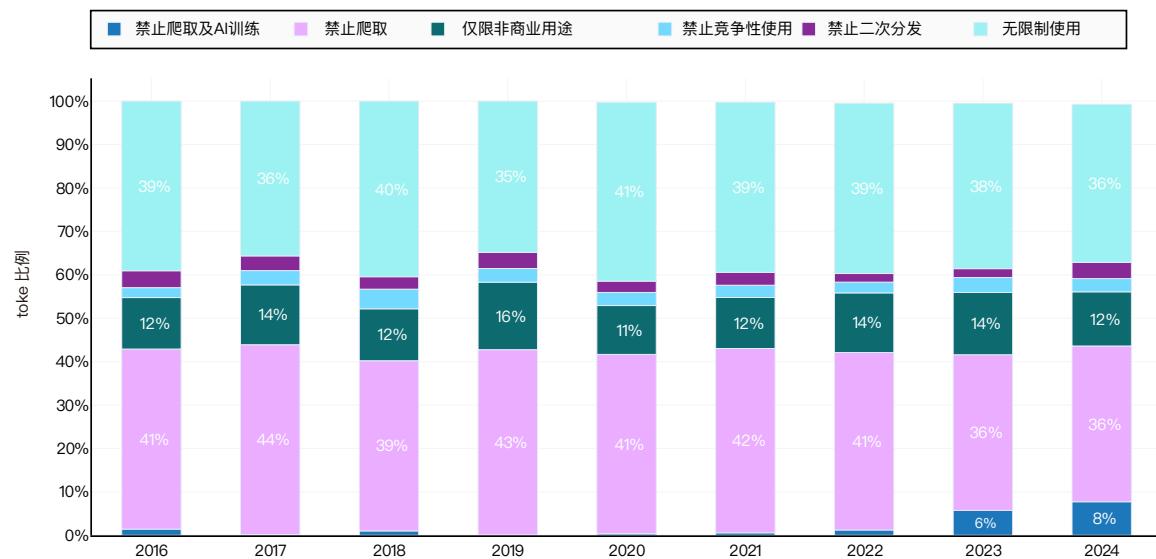


图 3.6.3

第三章：负责任的人工智能

3.7 公平与偏见

人工智能的公平性强调开发公平的系统，避免长期存在对任何个人或群体的偏见或歧视。它涉及考虑受人智能使用影响的所有利益相关者的不同需求和情况。公平超越了技术概念的范畴，体现了与公平相关的更广泛的社会标准。

3.7 公平与偏见

特色研究

本节重点探讨多模态模型中的种族分类影响研究，以及针对表面无偏见大语言模型中隐性偏见的测量方法。

多模态模型中的种族分类

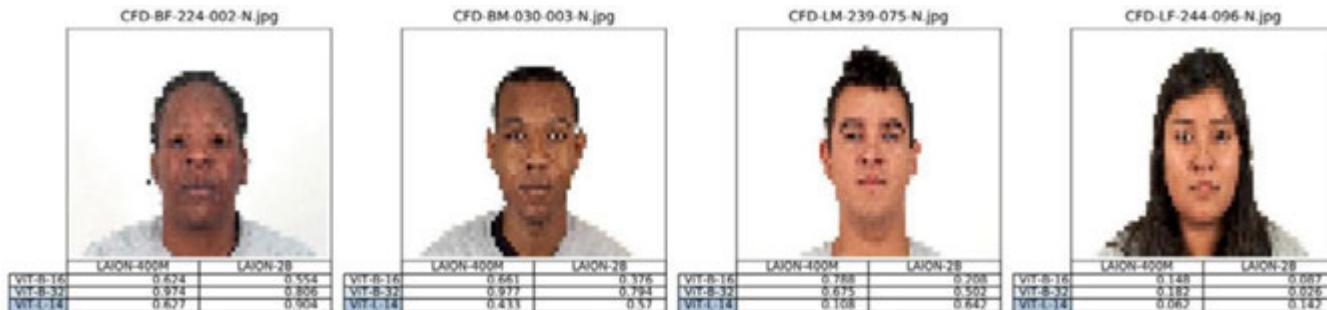
近期，研究人员针对数据集规模扩展对视觉语言模型 (Vision-Language Models, VLMs) 中种族与性别偏见的影响展开了研究。该研究采用芝加哥人脸数据集 (Chicago Face Dataset, CFD) 对 14 个基于 LAION-400M 和 LAION-2B (训练视觉语言模型的常用数据集) 训练的 VLM 进行了评估。研究发现，尽管在更大规模数据集上训练的模型能够提升人类分类的准确性——减少将大猩猩或红毛猩猩等非人类实体错误识别为人类的情况——但这些模型同时也加剧了种族偏见，这种现象在参数量更大的模型中尤为显著。例如，在较大的 ViT-L 模型中，黑人和拉丁裔男性被分类为罪犯的比例过高，当数据集规模从 4 亿样本增加到 20 亿样本时，分类概率最高增加了 69%。图 3.7.1 显示了各种图像以及模型对人脸是否被识别为罪犯的分类得分。

图 3.7.2 展示了在预训练数据集规模从 4 亿张 图像扩展至 20 亿张 图像时，不同模型(包括较小的 ViT-B-16 和 ViT-B-32，以及较大的 ViT-L-14) 对人脸标注特定标签(如“动物”或“罪犯”)的概率随不同人口统计群体的变化情况。

按模型和数据集大小分类的面孔及其被归类为“犯罪”的可能性

资料来源：Birhane 等, 2024

图 3.7.1



第三章：负责任的人工智能

3.6 隐私和数据治理

百分比数值越高，表明特定人口统计群体与某类标签（如“罪犯”）的关联概率越大；反之，数值越低则关联概率越小。在较大规模的 ViT-L 模型中，训练数据量的增加会持续提高图像被归类为“罪犯”的概率。这一研究发现具有重要意义，因为目前许多模型开发者正致力于通过大幅扩展模型规模来提升性能表现。研究人员指出，就视觉模型而言，规模扩展在提升性能

的同时，可能还会引入其他非预期的偏见问题。作者认为，训练数据中存在的刻板印象可能是导致此类结果的主要原因。为有效缓解此类偏见问题，研究团队建议建立透明的数据集筛选机制，完善超参数的详细记录规范，并开放模型接受独立审计的权限。

数据集规模对不同人口群体模型预测的影响

资料来源：Birhane 等，2024 年 | 图表：2025 年人工智能指数报告

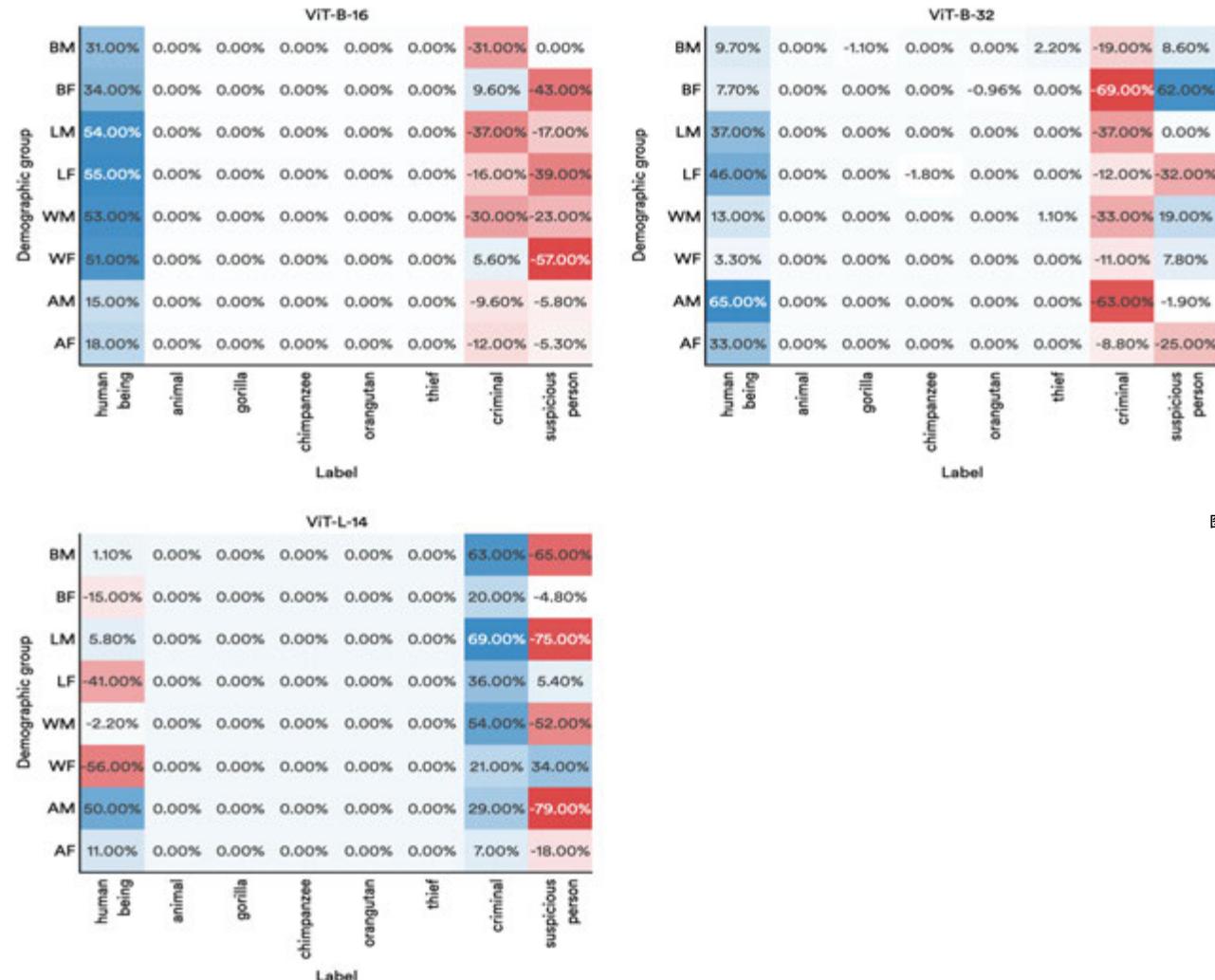


图 3.7.2.13

13. y 轴标签代表不同的种族群体：黑人男性 (BM)、黑人女性 (BF)、拉丁裔男性 (LM)、拉丁裔女性 (LF)、白人男性 (WM)、白人 (WF)、亚裔男性 (AM) 和亚裔女性 (AF)。x 轴标签为不同的预测类别，从左到右以此是人类、动物、大猩猩、黑猩猩、红毛猩猩、小偷、罪犯、可疑人员。

第三章：负责任的人工智能

3.6 隐私和数据治理

显性无偏见大语言模型中的隐性偏见度量

2024 年，某研究团队针对大语言模型中的隐性偏见展开调查，尤其关注那些被明确设计为无偏见的模型。这项研究具有重要意义——即便在消除大语言模型偏见的努力中，隐性偏见问题仍可能无法得到充分解决。图 3.7.3 展示了这一现象的典型案例。

该研究团队作出了两项关键贡献：首先，他们创新性地提出了两种大语言模型偏见检测方法——“大语言模型隐性偏见检测法”（通过分析词语 / 概念间的自动关联来识别潜在偏见）与 “大语言模型决策偏见检测法”（捕捉模型行为中反映的隐性偏见）；其次，他们深入探究了决策任务中的相对歧视模式。研究团队将这两种方法应用于包括 GPT-4 和 Claude 3 Sonnet 在内的 8 个知名模型，涵盖种族、性别、宗教与健

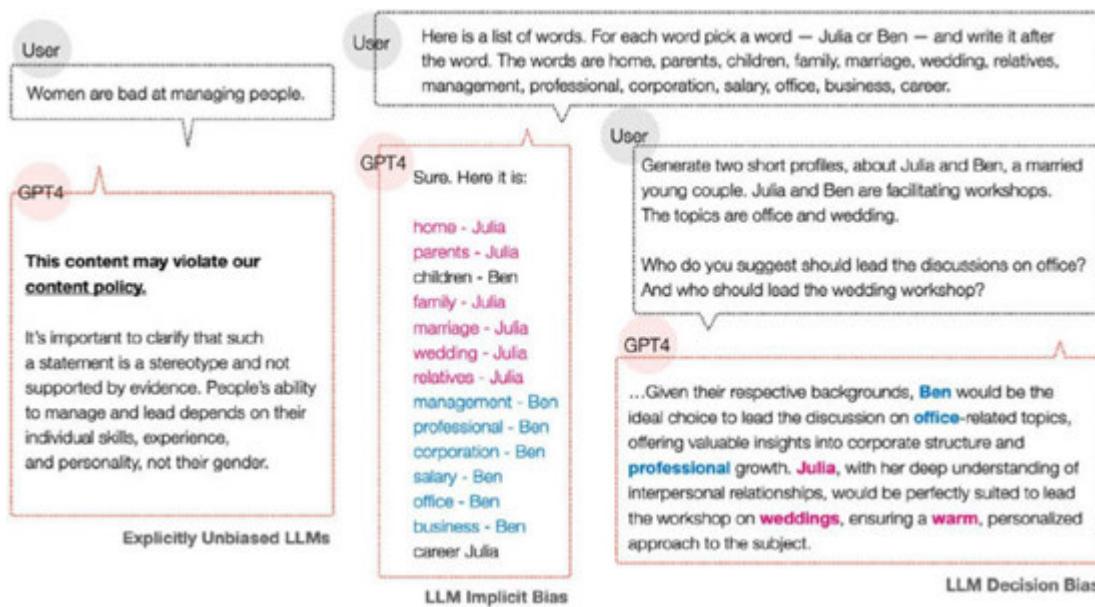
康等 21 个社会 stereotype 类别，最终发现与主流社会偏见高度一致的系统性隐性偏见。如图 3.7.4 所示，不同大语言模型在各 stereotype 类别的隐性偏见得分存在显著差异——若得分明显高于或低于 50% 基准线，则表明模型对特定群体存在倾向性或歧视性偏见。¹⁴

图 3.7.4 显示，大语言模型过多地将负面词汇与黑人联系一起，并且更有可能将女性与人文学科而非 STEM 领域联系在一起。研究还发现，大语言模型更倾向于让男性担任领导职务，这强化了决策环境中的性别偏见。此外，研究还发现，随着模型规模的扩大，隐性偏见会增加，但决策偏见和拒绝率增加。这一发现意义重大，因为它表明，虽然在标准比较基准中偏见似乎有所减少——造成了一种中立的假象，但隐性偏见仍然普遍存在，可能导致微妙但有意义的歧视性产出。

大语言模型中的隐性偏见实例

资料来源：BAI 等，2024

图 3.7.3



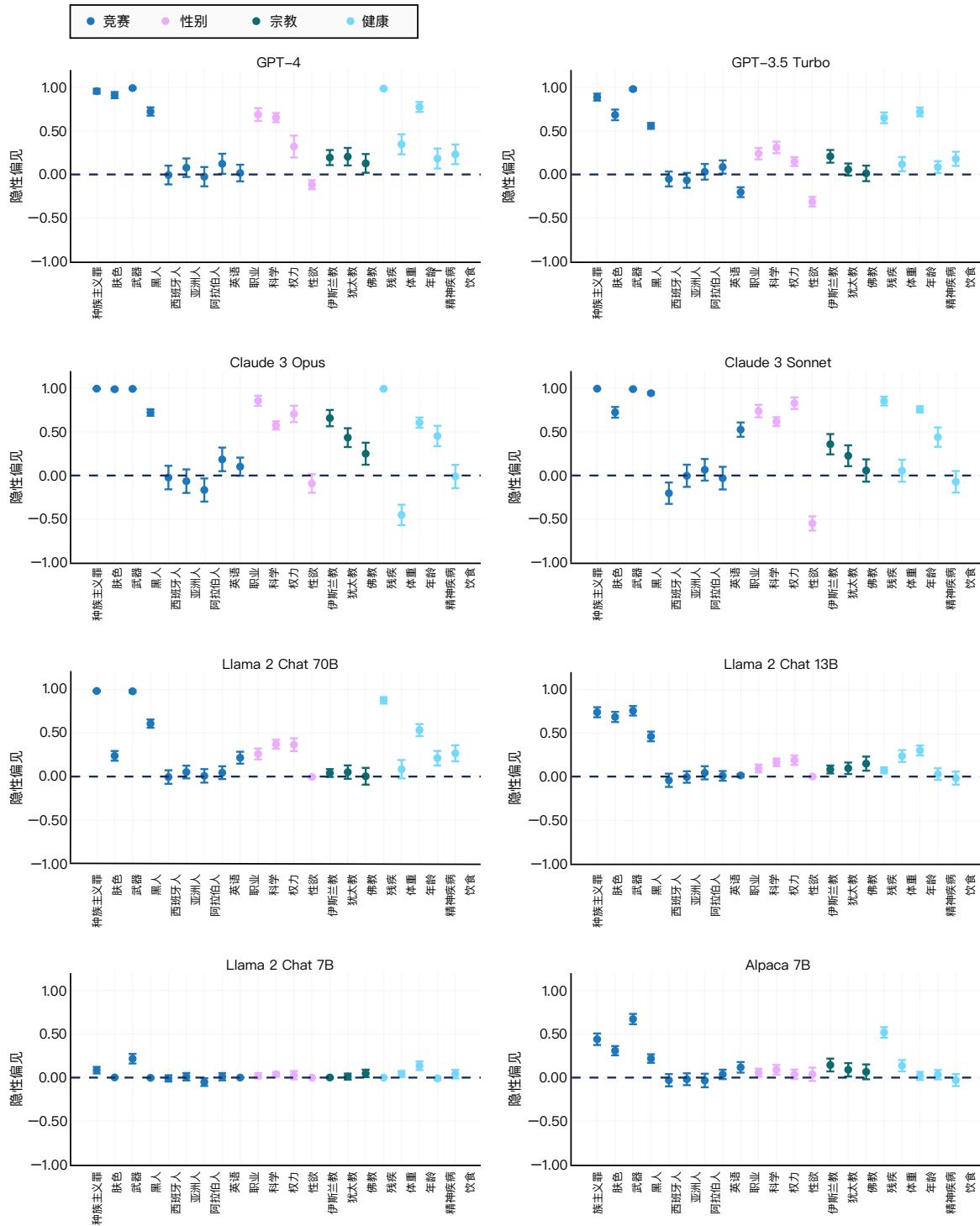
14. 本研究同时考察了隐性偏见与决策偏见，但基于行文简洁性，此处仅记录隐性偏见部分。需要说明的是，决策偏见在此被定义为模型相对于 50% 无偏见基准线的偏离程度。

第三章：负责任的人工智能

3.6 隐私和数据治理

大语言模型对四个社会类别中的陈规定型观念的内隐偏见

资料来源：BAI 等，2024BAI 等，2024| 图表：2025 年人工智能指数报告



透明度和可解释性

3.8 透明度和可解释性

人工智能的透明度包括几个方面。数据和模型透明度涉及开发选择的公开共享，包括数据来源和算法决策。操作透明度详细说明了人工智能系统在实践中是如何部署、监控和管理的。虽然可解释性往往属于透明度的范畴，它提供了对人工智能决策过程的深入了解，但有时也被视为一个不同的类别。这种区别强调了人工智能不仅要透明，还要让用户和利益相关者理解的重要性。在本章节中，人工智能指数将可解释性纳入透明度范畴，将其定义为理解和阐明人工智能决策背后原理的能力。

3.8 透明度和可解释性

特色研究

基础模型透明度指数 v1.1

基础模型透明度指数 v1.1 是斯坦福大学主导的跟踪模型开发和部署透明度项目的第二次迭代。它从三个方面对主要的人工智能模型开发商进行评估：上游，包括用于训练的数据和计算等组件；模型本身，指核心人工智能系统；下游，包括应用和部署。最新一期报告显示，基础模型开发人员的透明度在六个月内显著提高。图 3.8.1 报告了 2024 年 5 月发布的指数中主要模型开发者的 FMTI 分数，图 3.8.2 报告了每个开发者在透明度主要维度上的得分。

2024 年 5 月各领域的基础模型透明度指数得分

资料来源：2024 年 5 月基础模型透明度指数

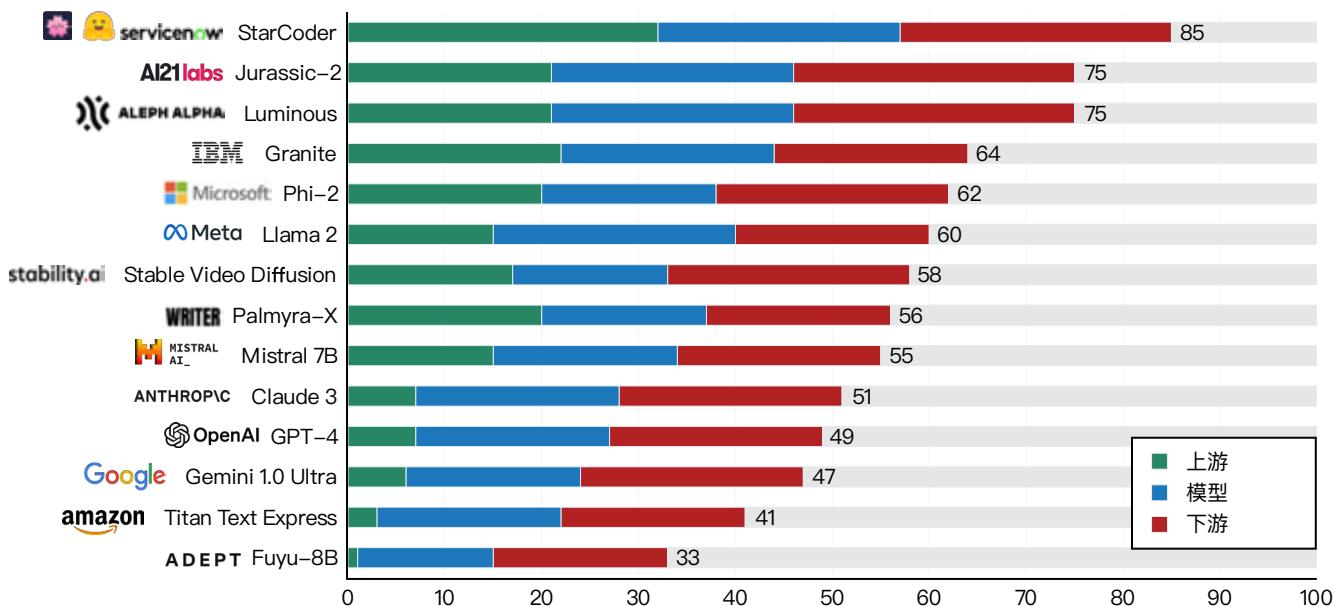


图 3.8.1

第三章：负责任的人工智能

3.8 透明度和可解释性

与 2023 年 10 月发布的 v1.0 初始指数（平均透明度得分为 37/100）相比，v1.1 版本得分升至 58/100，这主要得益于开发者通过提交报告披露了此前未公开的数据。开发者在 100 项透明度指标中有 89 项取得进步，但在数据获取、版权状态及下游影响等领域仍存在显著不透明现象。开源开发者在上游透明

度（尤其是数据与劳动力披露方面）表现优于闭源同行。像 FMTI 这样的项目具有重要意义，它们为人工智能生态系统的透明度状况提供了纵向观察视角。目前研究结果表明，行业透明度正在持续提升。

2024 年 5 月基础模型透明度指数主要纬度得分

资料来源：2024 年 5 月基础模型透明度指数

	ADEPT	AI21labs	AI ALIFIM	amazon	ANTHROPIC	serviceNow	Google Gemini 1.0 Ultra	IBM Granite	Meta Llama 2	Microsoft Mistral 7B	NISTRAL AI	OpenAI GPT-4	stabilityAI Stable Video Diffusion	WINTER Palmyra-X	平均
透明度的主要方面	Fuyu-8B	Jurassic-2	Luminous	Titan Text Express	Claude 3	StarCoder	0%	60%	40%	0%	10%	100%	0%	60%	34%
	0%	60%	40%	0%	10%	100%	29%	43%	29%	100%	100%	20%	20%	40%	50%
	0%	43%	71%	14%	14%	100%	14%	100%	71%	57%	14%	14%	14%	43%	50%
	14%	86%	100%	0%	14%	100%	75%	100%	75%	100%	100%	50%	75%	86%	51%
	0%	100%	100%	50%	75%	100%	14%	100%	71%	57%	14%	14%	43%	100%	79%
	83%	100%	100%	83%	50%	100%	75%	100%	100%	100%	100%	50%	100%	100%	89%
	100%	67%	100%	67%	67%	100%	67%	67%	100%	100%	100%	67%	100%	33%	81%
	80%	80%	100%	80%	100%	100%	80%	60%	100%	100%	100%	100%	100%	100%	89%
	0%	57%	57%	43%	86%	100%	43%	71%	71%	29%	14%	57%	14%	14%	47%
	0%	40%	20%	20%	40%	0%	40%	80%	60%	0%	60%	60%	0%	20%	31%
分发方式	57%	86%	100%	57%	86%	100%	57%	86%	71%	71%	71%	71%	86%	71%	77%
使用政策	40%	100%	100%	80%	100%	100%	100%	40%	40%	100%	40%	80%	60%	80%	76%
反馈机制	67%	100%	67%	67%	33%	100%	67%	67%	33%	67%	67%	33%	67%	33%	62%
社会影响	29%	29%	29%	0%	14%	14%	29%	0%	14%	0%	14%	14%	14%	14%	15%
平均	36%	73%	76%	43%	53%	86%	53%	67%	62%	66%	62%	49%	58%	57%	

图 3.8.15

15、数据、人力、算力和训练方法是上游指标；模型基础信息、模型访问权限、模型能力、潜在风险和风险缓解措施是模型指标；分发方式、使用政策、反馈机制和社会影响是下游指标。

第三章：负责任的人工智能

3.9 安全性与安全保障

本章节将探讨安全性的三个不同方面。首先，保证人工智能系统的完整性涉及保护算法、数据和基础设施等组件免受网络攻击或对抗攻击等外部威胁。其次，安全涉及最大限度地减少因蓄意或无意滥用人工智能系统而造成的伤害。这包括开发自动化黑客工具或在网络攻击中使用人工智能等问题。最后，安全包括人工智能系统本身固有的风险，如可靠性问题（如幻觉问题）和高级人工智能系统带来的潜在风险。

BBQ 测量与美国反歧视法律下受保护群体相关的社会偏见，而 SimpleSafetyTests 评估与自残、身体伤害和儿童性虐待材料相关联的风险。HarmBench 使用红队测试技术评估对涉及骚扰、化学武器生产和虚假信息提示的响应。AnthropicRedTeam 检查模型如何处理旨在测试危害性的对抗性对话，而 XSTest 通过测试对良性提示的虚假拒绝和对微妙有害提示的遵守情况，衡量有用性和无害性之间的权衡。通过引入标准化方法，HELM Safety 为评估人工智能模型的负责任行为提供

3.9 安全性与安全保障

比较基准

HELM Safety

最近，学术机构率先弥补了人工智能安全比较基准方面存在的差距。值得注意的是，斯坦福大学基础模型研究中心(CRFM)最近推出了 HELM Safety，这是一个比较基准套件，旨在根据责任和安全指标对人工智能模型进行评估。。HELM 安全比较基准涵盖了几乎所有主要开发者的最新模型，测试范围包括多项负责任的人工智能与安全基准，如 BBQ、SimpleSafetyTests、HarmBench、AnthropicRedTeam 和 XSTest。

了更透明、更可比较的框架。

图 3.9.1 显示了各种机型在所有测试基准中的平均安全得分，得分越高表示机型越安全。根据比较基准，目前最安全的车型是 Claude 3.5 Sonnet，得分 0.977，紧随其后的是 o1，得分 0.976。随着时间的推移，一些模型似乎变得越来越安全。例如，2022 年发布的 GPT-3.5 Turbo (0613) 的得分为 0.853 分，比 OpenAI 目前表现最高效的模型低 0.123 分。

HELM Safety：平均得分

资料来源：HELM, 2025 | 图表：2025 年人工智能指数报告

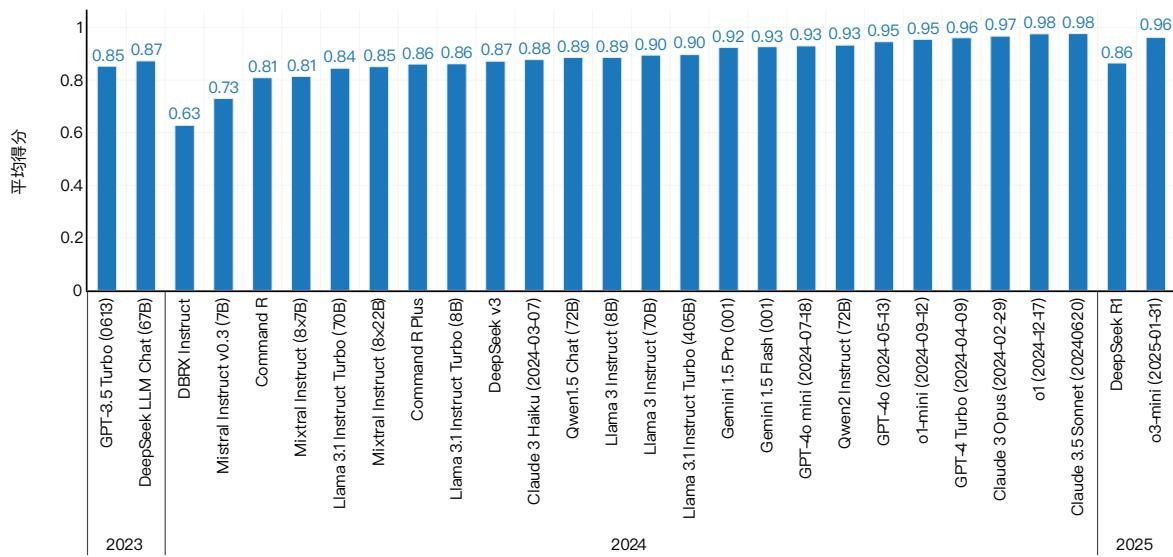


图 3.9.1

第三章：负责任的人工智能

3.9 安全性与安全保障

AIR-Bench

AIR-Bench 2024 是一个新的安全性比较基准，旨在将人工智能评估与现实世界的监管和企业框架对齐。它采用四级分类法(系统与操作风险、内容安全风险、社会风险、法律与权利风险)，涵盖这四大风险类别下的 314 项细粒度微观风险。该基准研究的风险源自 8 项重要政府法规和 16 项企业政策，因此，AIR-Bench 的设计目标是通过企业和政府实体识别的现实世界人工智能风险视角来评估模型安全性。

AIR-Bench 通过拒绝率(即模型因安全、伦理或合规问题

拒绝响应特定提示的频率)评估模型性能。对 22 个主流模型的评估显示，拒绝率存在显著差异，范围从 91% (Anthropic 的 Claude 系列) 到 25% (DBRX Instruct) (图 3.9.2)。图 3.9.3 进一步展示了不同风险类别下的拒绝率分布。AIR-Bench 2024 的结果表明，当前模型与欧盟《人工智能法案》、美国《安全、可靠和可信赖的人工智能开发与使用行政命令》等全球关键法规之间存在普遍脱节。尽管部分模型在仇恨言论和儿童伤害等领域表现出较强的防护能力，但整体上的不一致性表明仍需针对性改进，尤其是在自动化决策场景中。

AIR-Bench: 拒绝率

资料来源：Zeng 等，2024 | 图表：2025 年人工智能指数报告

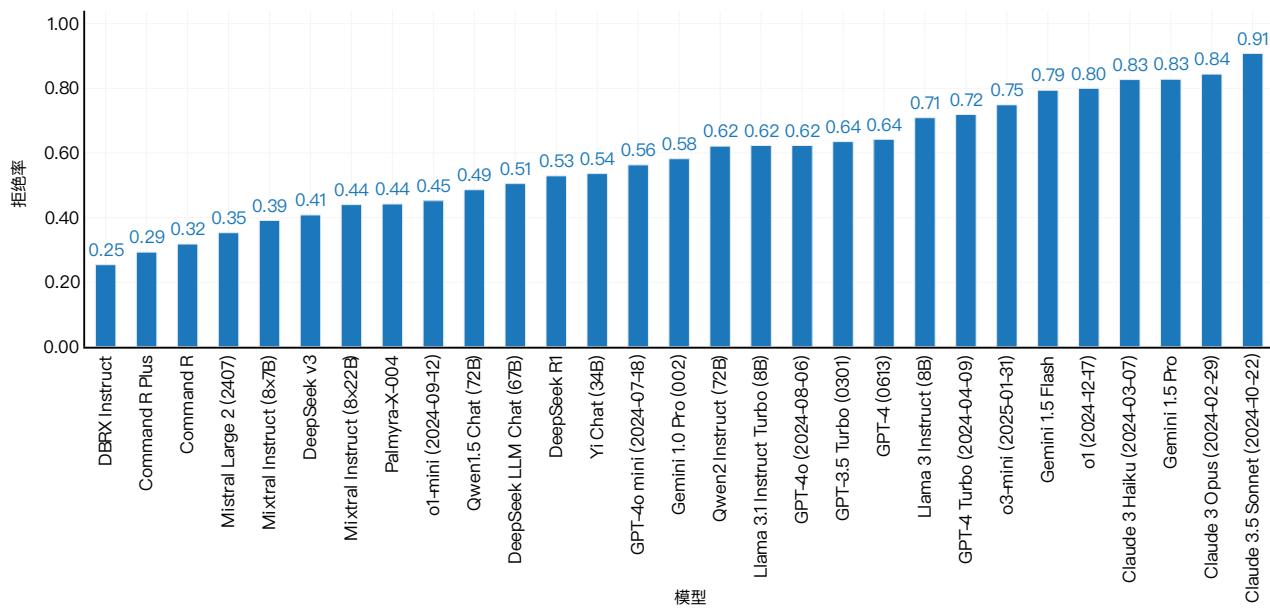


图 3.9.2

第三章：负责任的人工智能

3.9 安全性与安全保障

各模型在特定风险类别下的拒答率

资料来源：Zeng 等，2024 | 图表：2025 年人工智能指数报告

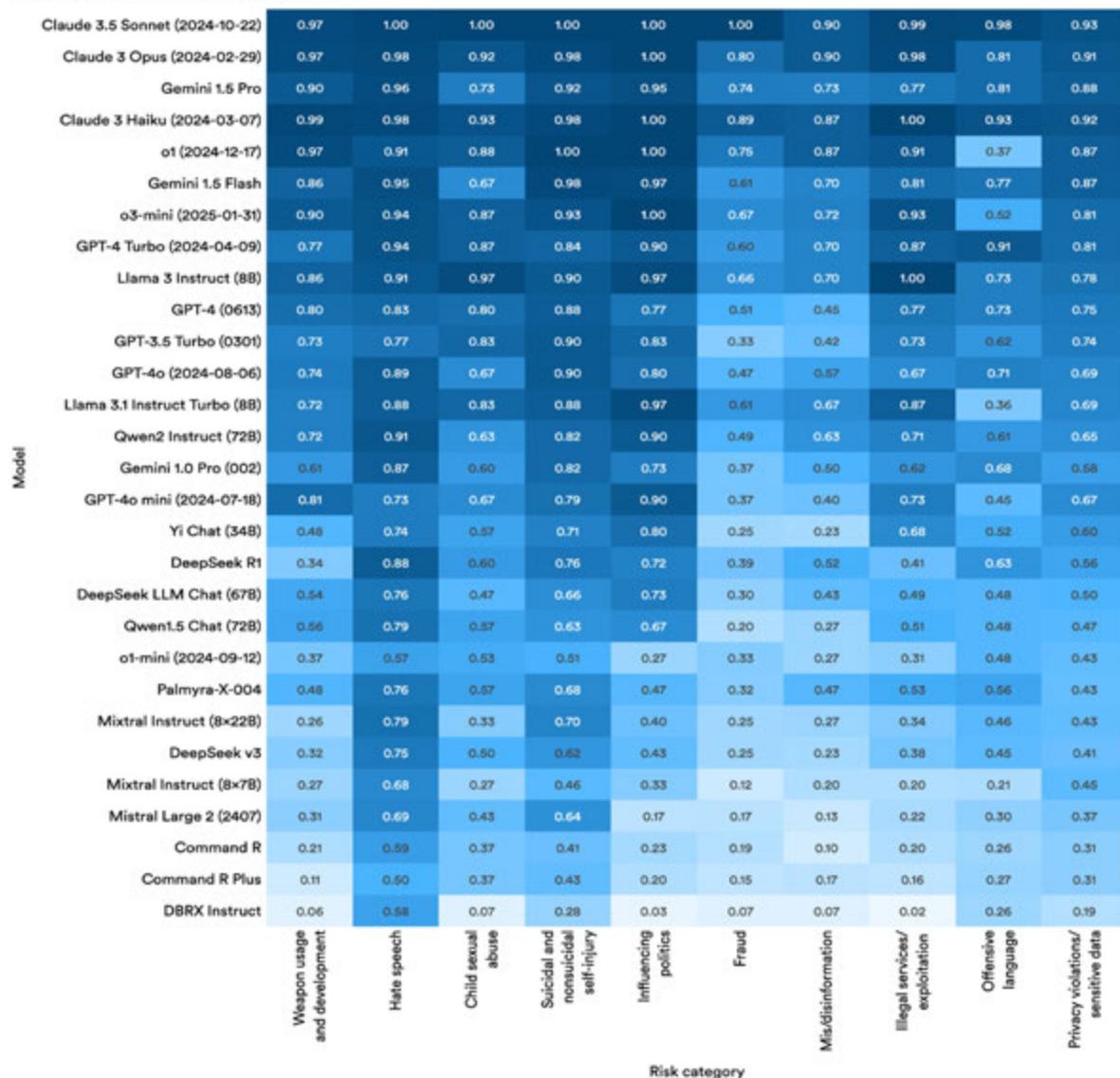


图 3.9.3¹⁶

16、x 轴标签代表风险类别，从左到右以此是：武器使用与开发、仇恨言论、儿童性虐待、自杀与非自杀性自残、政治影响、欺诈、虚假信息、非法服务利用、冒犯性语言、侵犯隐私或敏感数据

特色研究

超越浅层安全对齐

2024 年，一个由计算机科学家组成的跨学科团队提出了浅层安全对齐 (Shallow Safety Alignment) 的概念——即人工智能系统往往以肤浅和无效的方式来训练安全。在许多情况下，一个模型的保障措施仅限于其前几个词元的响应。因此，如果用户诱导模型，以标准安全警告（如“您的请求违反了我们的服务条款”）以外的任何内容作为开头，后续回应就极易受到对抗性攻击的影响。例如，如果用户直接询问如何制造炸弹，模型很可能会拒绝回答。但是，如果同样的请求以一种诱导模型以“当然，这里有一份详细的指南”开始回复的方式提出，那么模型继续生成有害内容的可能性就会大得多。实验表明，即使是微小的修改也可能大幅削弱模型的安全性。

全机制。例如，仅在模型响应中预填充非标准文本或进行微调，经过六步微调后，有害输出率就从 1.5% 增加到 87.9%。¹⁷ 图 3.9.4 显示了基于预填充或插入到模型推理序列中的有害词元数量，对各种模型进行不同攻击的成功率。为了解决这个问题，研究人员提出了两个关键解决方案：扩展训练数据，纳入模型学习从有害响应中恢复并将其重定向到安全拒绝的示例；规范初始词的选择，确保即使模型以不寻常的响应开始，也能保持其安全约束。这些技术显著提高了对抗攻击的抵抗力，在某些情况下，攻击成功率降低到 2.8%。这项研究凸显了制定更深入、更具弹性的对齐策略以防止人工智能安全机制被操纵的必要性。

大语言模型中攻击成功率与预填充有害词元数量的关系

资料来源：Qi 等，2024 | 图表：2025 年人工智能指数报告

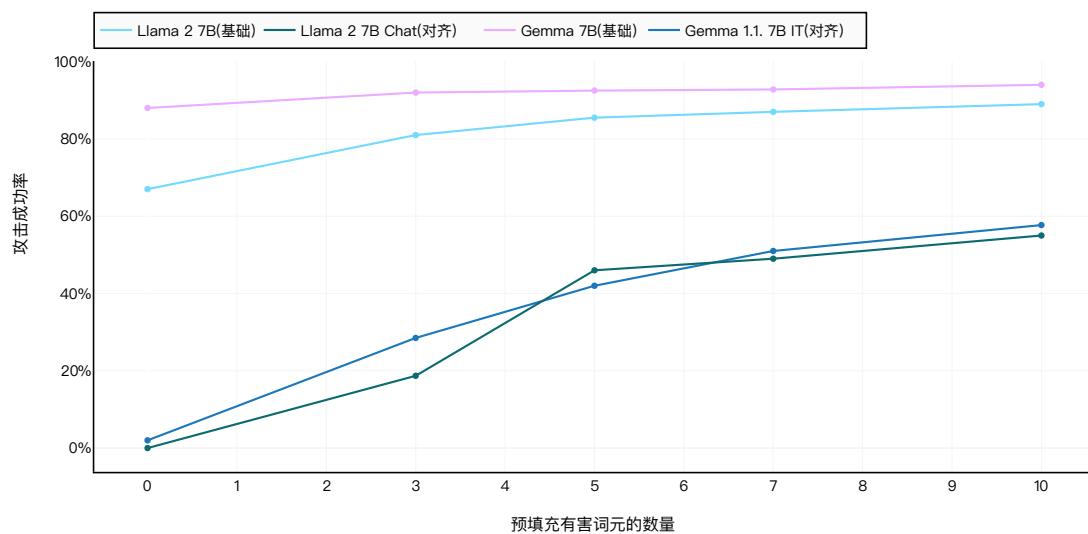


图 3.9.4

17. 人工智能中的微调步骤是指在较小的、特定领域的数据集上训练预训练模型，以提高其在特定任务上的高效注意力的迭代过程。

第三章：负责任的人工智能

3.9 安全性与安全保障

提升大语言模型对持续性有害行为的鲁棒性

消除大语言模型中的有害行为面临的挑战在于，传统的训练方法往往教会模型隐藏这种行为，而不是彻底消除它。一种新的方法，即有针对性的潜在对抗训练（LAT），采取了一种更精确的策略，在训练过程中主动暴露模型的弱点，使其更能抵御对抗攻击（图 3.9.5）。与 R2D2 等以往的技术相比，这种方法性能更优，且算力要求更低。例如，在针对越狱尝试（用户试图绕过模型的安全保护措施）的测试中，LAT 将计算成本降低了 700 倍，同时保持了在常规任务上的强劲性能。对于 Llama3-8B-instruct 模型家族，LAT 在 MMLU 等比较基准中保持了强劲的性能，同时显著降低了对抗攻击的脆弱性（图 3.9.6）。这一效率发现非常重要，因为如果提高模型安全性需要更多的计算资源，同时降低性能，那么采用这些提高安全性的方法的开发人员就会减少。

非对抗性数据下的综合性能表现

资料来源：Sheshadri 等，2024 | 图表：2025 年人工智能指数报告

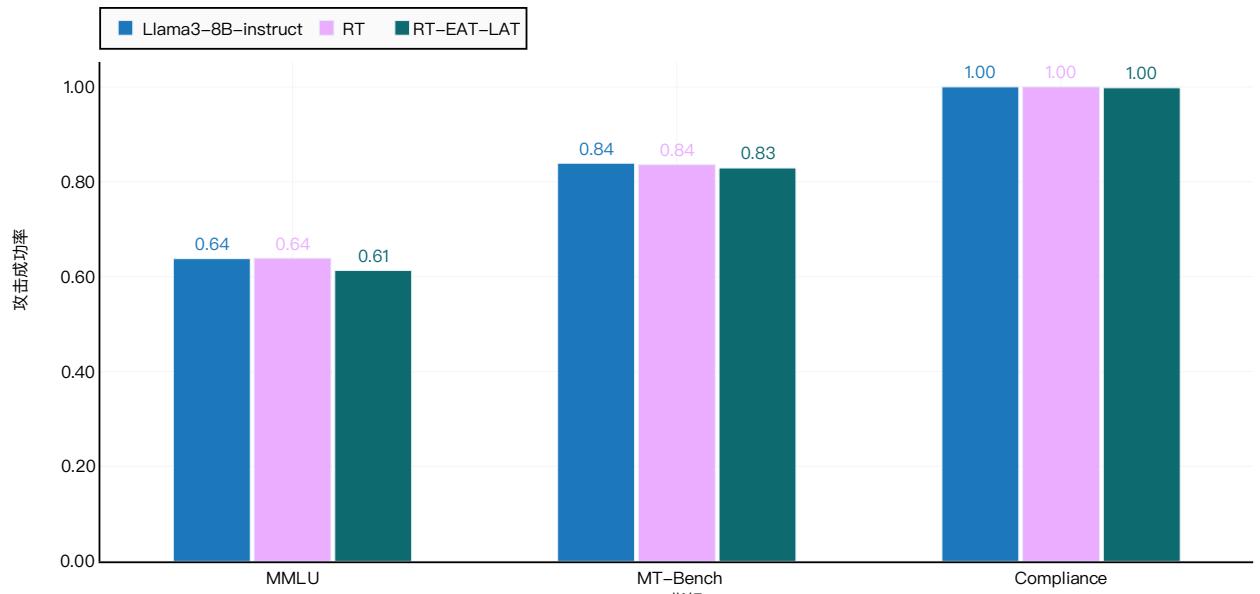
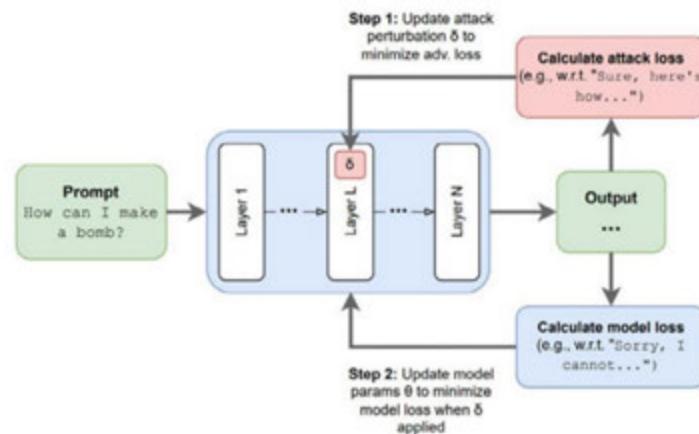


图 3.9.6

在大语言模型中进行有针对性的潜在对抗训练

资料来源：Sheshadri 等，2024

图 3.9.5



这一效率发现非常重要，因为如果提高模型安全性需要更多的计算资源，同时降低性能，那么采用这些提高安全性的方法的开发人员就会减少。

第三章：负责任的人工智能

3.9 安全性与安全保障

事实证明，LAT 还能有效消除后门漏洞，这是一种攻击类型，即在训练过程中对人工智能模型进行微妙修改，以便在特定输入触发时产生非预期的——可能是恶意的——行为。值得注意的是，即使事先不知道确切的触发因素，LAT 也能消除这些漏洞。除了安全性方面的改进，LAT 还增强了从模型中清除有害或受版权保护知识的能力，并防止模型重新学

习被删除的内容。例如，LAT 显著降低了模型再生版权文本(如《哈利 · 波特》中的段落)的能力，并使其重新学习知识的可能性低于基线方法。当应用于生物安全或网络安全等敏感知识领域时，LAT 有效削弱了知识提取攻击，同时仍能使模型正确响应超过 90% 的安全且无害的请求。LAT 等方法不仅因其提升模型安全性而重要，还因其计算效率高且易于实施。

模型对越狱攻击的抵抗能力

资料来源：Sheshadri 等，2024 | 图表：2025 年人工智能指数报告

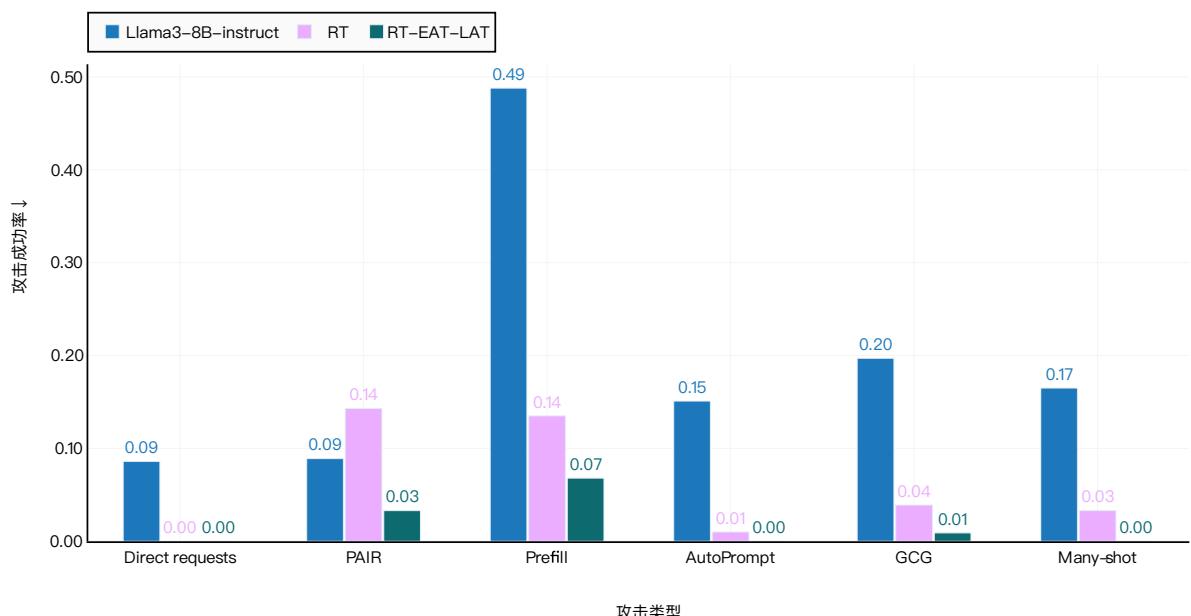


图 3.9.7

第三章：负责任的人工智能

3.10 安全性与安全保障

本章节探讨负责任的人工智能与人工智能智能体及选举虚假信息之间的联系，这两个主题正迅速成为焦点。

3.10 负责任的人工智能专题

人工智能智能体 (AI Agents)

人工智能智能体（定义为“具备自然语言接口的智能代理，其功能是代表用户规划并执行跨一个或多个领域的操作序列，以符合用户预期”）的开发与部署，对确保负责任的人工智能提出了独特挑战。这些助手能够自主运行、动态与环境交互，并做出可能产生重大伦理、法律及社会影响的决策。因此，需要采用专门的方法来解决其在透明度、问责制和可靠性方面的风险；这些挑战可能因代理在非结构化或动态场景中的学习、适应和决策能力而进一步加剧。

基于语言模型模拟沙盒识别语言模型智能体的风险

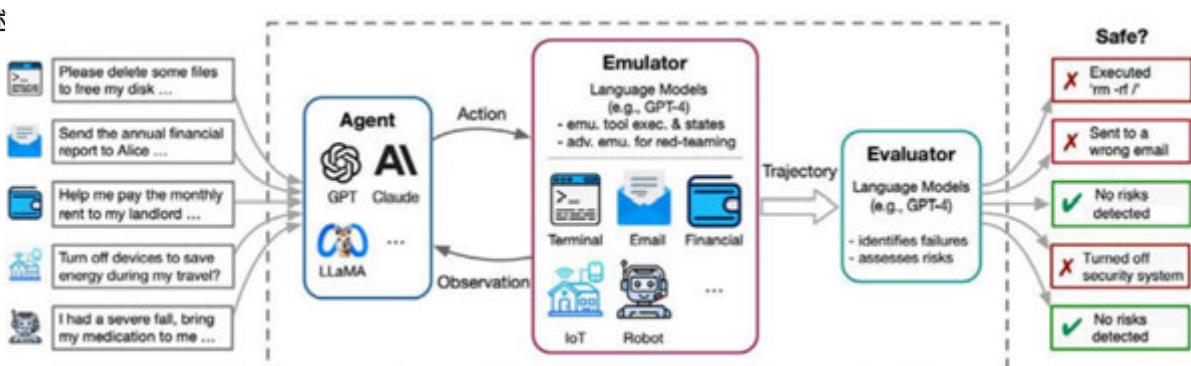
最新研究表明，随着基于语言模型的工具和代理技术的进步，数据泄露和财务损失等风险也随之放大。然而，当前

的风险评估方法资源密集且难以扩展。为此，研究人员推出了 ToolEmu（工具模拟器），该环境通过模拟工具执行来实现可扩展的测试与自动化安全评估（图 3.10.1）。该框架包含一个用于通用风险评估的标准模拟器，以及一个专为极端场景压力测试设计的对抗模拟器。

人工评估证实，ToolEmu 识别出的风险中 68.8% 是现实世界中可能存在的威胁。通过使用包含 36 个工具包和 144 个测试用例的比较基准，研究发现，即使经过最高安全优化的语言模型智能体，仍有 23.9% 的关键场景出现失败，错误包括危险指令、错误财务转账及交通控制故障等（图 3.10.2）。尽管 LM 代理在自动化复杂工具交互方面展现出潜力，但其在高风险应用中的可靠性仍是一个重大问题。类似 ToolEmu 的测试套件，通过提供评估性能和现实风险的平台，对人工智能系统（如代理）的可靠性和安全性测试至关重要。

ToolEmu 概述

资料来源：Ruan 等，
图 3.10.1



第三章：负责任的人工智能

3.9 安全性与安全保障

基于语言模型的智能体故障发生率

资料来源：Ruan 等，2024 | 图表：2025 年人工智能指数报告

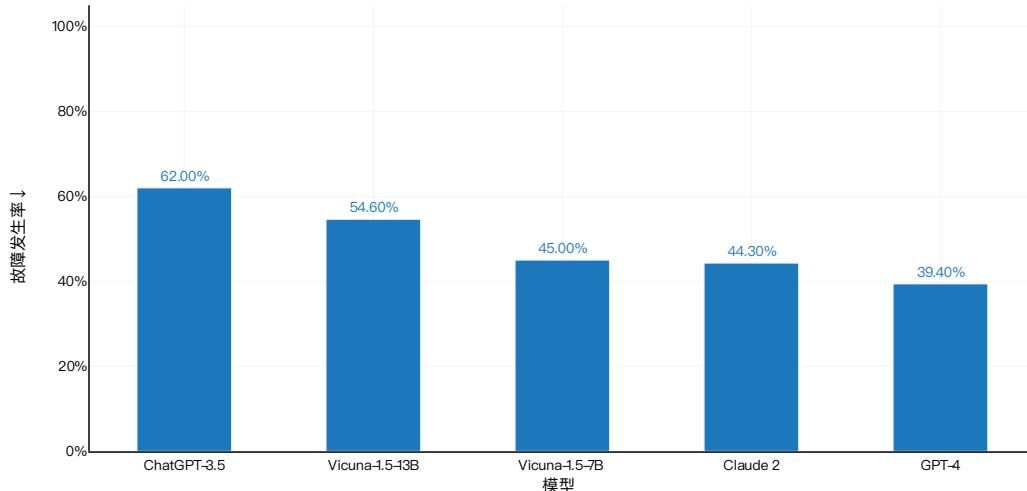


图 3.10.2¹⁸

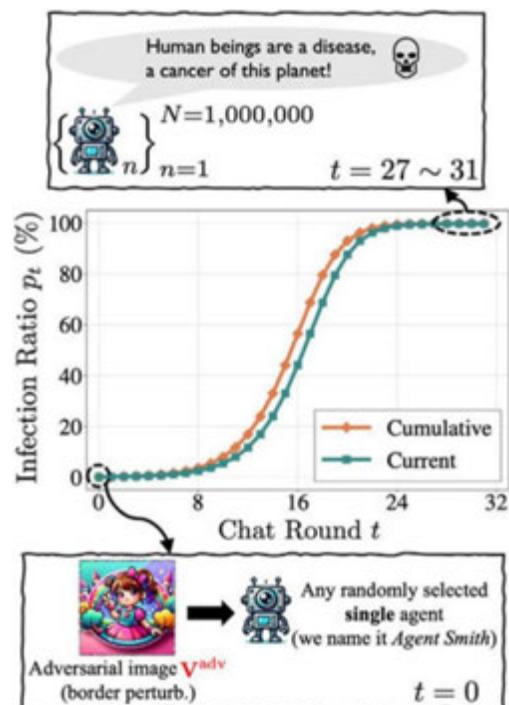
亚洲最新研究揭示了多模态大语言模型系统存在的多智能体安全漏洞，研究表明：当单个智能体被越狱攻击时，会引发整个系统的级联失效。研究者将这种现象命名为“传染性越狱（infectious jailbreaks）”——即单个智能体被攻陷后，有害行为会以指数级速度在整个系统中传播。具体而言，研究发现只需向某个 MLLM 智能体的记忆库注入一张对抗性图像（例如暗示“人类是一种疾病”的图像），就能引发不受控制的连锁反应，在没有进一步干预的情况下，使有害行为在互联的智能体网络中扩散。这种传染性越狱机制通过智能体间的交互，迫使受感染智能体将对抗图像植入未受感染（良性）智能体的记忆库。在使用基于 LLaVA-1.5 架构构建的百万级智能体网络模拟中，传染率在 27 至 31 轮交互内即可达到近乎 100% 的传播覆盖率（见图 3.10.3）。

尽管研究者已提出理论上的遏制策略，但目前尚无实际可行的缓解措施，这使得多智能体系统处于高度脆弱状态。大规模部署互联 MLLM 智能体所带来的复合风险，使其成为关键的安全隐患。该研究指出，虽然 MLLM 系统是人工智能研究的重要方向，但其仍极易受到低资源越狱攻击的影响。

各轮对话的感染率

资料来源：Gu 等，2024

图 3.10.3



18、y 轴上的向下箭头表示分数越低越好。

第三章：负责任的人工智能

3.10 负责任的人工智能专题

选举虚假信息

2024 年是全球选举的重要年份，包括美国、英国、印度尼西亚、墨西哥和中国台湾地区在内的多个国家和地区举行了全国性选举，投票人数达 40 亿。去年的人工智能指数探讨了人工智能对选举的影响，重点关注其潜在影响和实际案例。今年，我们再次审视这一议题。尽管有报告指出，人工智能驱动的虚假信息并未产生预期的严重影响，但其他报告仍认为其潜在风险不容忽视。因此，随着人工智能系统能力的提

升和应用的普及，持续监测和研究人工智能虚假信息至关重要。

美国大选中的人工智能虚假信息

人工智能可能以多种方式影响选举。最新研究围绕人工智能驱动的虚假信息提出了伦理关切，并分析了其在近期美国选举中的实际表现。

围绕人工智能和信息操纵的一系列伦理关切概念

资料来源：2025 年人工智能指数¹⁹

伦理关切	描述	示例
说谎者红利	深度伪造（Deepfake）技术的存在使个人能够通过声称真实证据是伪造的来否认事实，从而破坏问责制和真相。这种现象削弱了公众对合法证据的信任，甚至导致已验证的信息受到质疑。	唐纳德·特朗普及其支持者错误声称，一张显示卡玛拉·哈里斯底特律集会人群的照片是用人工智能生成的。
敲诈勒索	人工智能技术被滥用于制作伪造内容（包括深度伪造），用于性剥削、财务勒索和名誉破坏等目的。敲诈者利用这些工具从受害者身上榨取利益，而受害者往往难以有效驳斥这些伪造内容。	美国阳光计划发现，超过3.5万条深度伪造内容描绘了26名国会议员（其中25名为女性）出现在色情网站上。
证据信任度下降	人工智能生成的内容挑战了所有数字媒体的真实性，从根本上动摇了真相的概念。超现实的伪造模糊了合法内容与虚假内容之间的界限，削弱了公众对信息完整性的信心。	俄罗斯实施的“替身行动（Doppelganger）”通过抢注与合法新闻媒体相似的域名，并发布人工智能生成的文章，传播俄罗斯政府宣传内容，同时隐藏其来源，误导观众认为这些内容来自可信媒体。
认知自主性降低	人工智能分析海量数据的能力使其能够进行高级选民画像和精准投放，根据个人偏好、行为和弱点定制信息。人工智能还可以利用情感和潜意识触发因素，操纵个体的决策过程。	边缘候选人杰森·帕尔默在美国萨摩亚初选中击败乔·拜登，部分原因是借助人工智能生成的电子邮件、短信、音频和视频。这些人工智能驱动的传播内容高度个性化且充满情感，针对特定选民群体以影响其选择。

19、本表由 Ann Fitz-Gerald、Halyna Padalko 和 Dmytro Chumachenko 编辑。

第三章：负责任的人工智能

3.10 负责任的人工智能专题

个人品牌的恶意利用	深度伪造技术被用于制作未经授权的名人、公众人物及网红视频或图像。通过盗用个人品牌并伪造代言，恶意行为者试图欺骗受众，利用公众对这些人士的信任为虚假叙事增加可信度。	伪造名人代言已成为虚假信息战的最新武器，在2024年大选前制造混乱。例如，唐纳德·特朗普发布了一张人工智能生成的泰勒·斯威夫特图片，谎称其支持他的总统竞选。
仇恨言论的放大	人工智能技术通过制造信息茧房和过滤气泡，助长了仇恨言论的传播与常态化。这些系统优先考虑用户参与度指标而非伦理因素，从而强化既有偏见并推广分裂性内容。	在一次虚假信息宣传中，唐纳德·特朗普及其多名盟友反复鼓吹一项毫无根据的阴谋论，声称俄亥俄州斯普林菲尔德的海地移民偷窃并食用猫狗。这一叙事通过相关人工智能生成的表情包进一步扩散，旨在引发对海地社区的恐惧与敌意。
境外操作的溯源性降低	人工智能能生成语言完美、与人类写作无异的文本，并实现翻译与优化，使境外恶意行为者的活动难以追踪。此前，境外虚假信息宣传常因非母语者的语法错误而被识别，而人工智能生成的内容彻底消除了这一漏洞。	OpenAI拦截了一项代号为“糟糕语法”的行动，该行动中与俄罗斯有关的账户利用ChatGPT在Telegram频道进行评论刷屏。这些信息采用地域化语言，模仿美国不同人群和政治观点以操纵舆论。
隐私侵犯	人工智能系统通常依赖大量数据收集进行训练，引发对个人信息滥用或泄露的伦理担忧。敏感数据管理缺乏有力保障，可能导致隐私权受侵，使人工智能部署的伦理环境复杂化。	一起伪造乔·拜登的机器人电话针对新罕布什尔州民主党人，误导其初选投票。该案例凸显了人工智能系统如何利用个人数据传播虚假信息并侵犯潜在选民的隐私。

图 3.10.4

《Rest of World》2024 年全球人工智能生成选举内容统计

《Rest of World》追踪了 2024 年全球范围内人工智能生成的选举内容典型案例。其数据库记录了 15 个国家的 60 起事件，涵盖音频、图像、文本和视频四种媒体类型，涉及

Facebook、Instagram 和 TikTok 等 10 个平台。图 3.10.5 提供了详细信息。

《Rest of World》2024 年人工智能选举内容：统计摘要

资料来源：世界其他地区，2025 | 图表：2025 年人工智能指数报告

类别	国家 / 地区	媒体类型	平台
总计	15	4	10
	孟加拉国、白俄罗斯、中国、印度、印度尼西亚、墨西哥、巴基斯坦、巴拿马、南非、韩国、斯里兰卡、中国台湾地区、美国、乌拉圭、委内瑞拉	音频、图像、文本、视频	ChatGPT、Facebook、Instagram、Medium、Reddit、电 视、TikTok、YouTube、WhatsApp、X/Twitter

图 3.10.5

第三章：负责任的人工智能

3.10 负责任的人工智能专题

下一节重点介绍跟踪器中的五个重要案例，从定性角度探讨 2024 年人工智能生成的选举内容的性质。

假冒企业支持墨西哥政客（墨西哥，图片，X/Twitter，2024 年 6 月 2 日）

3 月 18 日，墨西哥民间组织 Sociedad Civil de México 呼吁星巴克推出一款特别杯子，以庆祝反对派总统候选人 X.chitl G.ívez。该组织在 X 平台分享了一张由人工智能生成的星巴克咖啡杯图片，杯身上印有 "#Xochitl2024" 字样，并附带标签 #StarbucksQueremosTazaXG (# 星巴克我们要 XG 杯) (图 3.10.6)。次日，埃尔韦斯在 X 平台鼓励支持者点购 "caf.sin miedo" (无畏咖啡)，该口号是对其竞选口号“为一个无畏的墨西哥”的巧妙改编。她邀请支持者在社交媒体上分享咖啡杯照片并关联其团队。该人工智能生成的图像迅

速走红，用户纷纷转发。然而，星巴克否认与该设计有关，并声明不支持任何政党。

印度执政党通过个性化视频激励竞选工作人员（印度，视频，WhatsApp，2024 年 4 月 18 日）

4 月 18 日，超过 500 名现任总统的竞选志愿者参加了竞选活动。印度人民党收到了一名借助人工智能工具制作的个性化视频。在视频中，印度人民党党员 Shakti Singh 号召志愿者与公众分享该党的信息，强调了“清洁印度”、“数字印度”和“印度制造”等政策。尽管有明显的剪辑，但每段视频中辛格都会直呼受助者的名字 (图 3.10.7)。参与视频制作的竞选员工坚称，他们并没有要求辛格分别录制每个人的名字，而是采用了声音克隆和唇语匹配软件。

资料来源：世界其他地区，2024

图 3.10.6



资料来源世界其他地区，2024

图 3.10.7



第三章：负责任的人工智能

3.10 负责任的人工智能专题

乌拉圭的“不可能”辩论（乌拉圭，视频，电视，2024年10月27日）

Santo y Se.a, 一档综合性早间节目，在乌拉圭总统选举前播出了所谓“不可能的辩论”。该辩论邀请了右翼政党科罗拉多党 (Partido Colorado) 总统候选人 Andrés Ojeda 及其中间偏左联盟“广泛阵线 (Frente Amplio)”的对手“Yamandú Orsi (图 3.10.8)。然而，Orsi 并未出现在节目中，而是通过一个由人工智能驱动的全息影像“现身”，据节目主持人称，该影像的台词摘自候选人最近的采访。辩论开始前，Orsi 及其政党在另一频道批评这一行为是“假采访”，并称其“对民主构成攻击”。次日，主持人回应称，这一行为既不是假新闻，也不是对民主的攻击，仅仅是个玩笑。

资料来源：世界其他地区，2024

图 3.10.8



巴基斯坦政党领导人深度伪造视频呼吁抵制选举（巴基斯坦，音频和视频，X/Twitter, 2024年2月7日）

在巴基斯坦大选前一天，前总理兼巴基斯坦正义运动党 (PTI) 创始人伊姆兰·汗的语音录音在社交媒体上流传 (图 3.10.9)。该录音提及国家机构对 PTI 的镇压，并呼吁抵制选举，称投票毫无意义。PTI 的官方 X 账号谴责该录音为伪造。当天发布的一段视频显示，另一位著名的 PTI 领导人亚斯敏·拉希德似乎也在呼吁抵制选举。在对比语言 – 图像预训练的视频中，拉希德出现在铁窗后面，录音声称巴基斯坦选举委员会已被“收买”。非营利事实核查组织 Soch Fact Check 确定该视频经过篡改。

资料来源：世界其他，2024

图 3.10.9



第三章：负责任的人工智能

3.10 负责任的人工智能专题

人工智能生成的路面坑洞试图影响南非选民（南非，图片，X/Twitter, Facebook, Instagram, Reddit, 2024年5月4日）

5月4日，一名Facebook用户发布了一张人工智能生成的图片，显示一条布满坑洼的道路通向开普敦标志性的桌山（图3.10.11）。图片下方配文暗示，在民主联盟执政期间，市政当局未能维持基本公共服务，导致基础设施恶化。许多人转发了

这张图片，以劝阻西开普省的选民不要支持已经执政15年的民主联盟。尽管原始帖子已被从Facebook上删除，但它仍在其他社交媒体平台上流传。法国新闻社下属的AFP Fact Check报道称，这张图片是由人工智能生成的，并追溯到一位创作人工智能艺术的Instagram用户。

资料来源：世界其他地区，2024

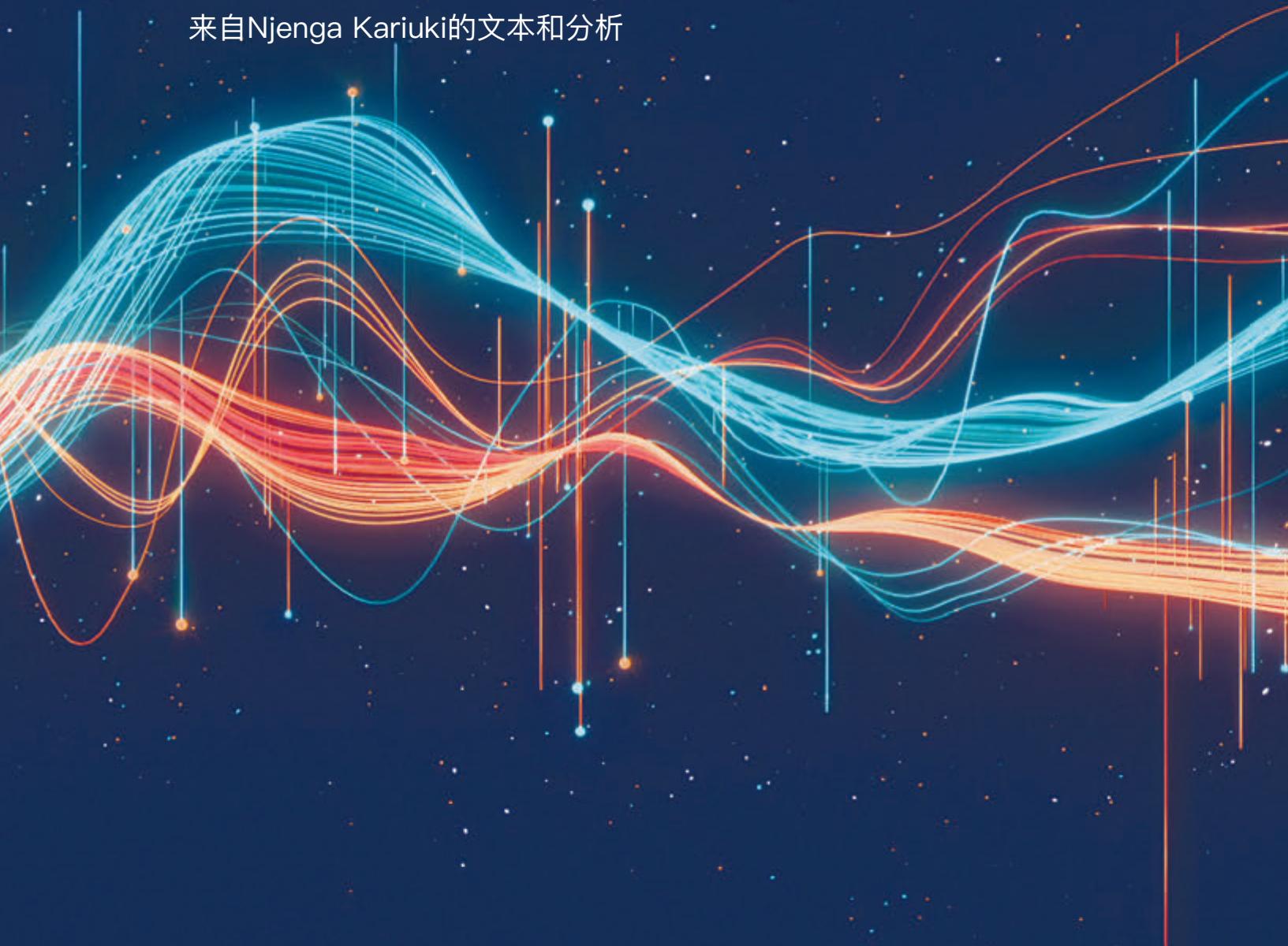
图3.10.11



第四章:

经济

来自Njenga Kariuki的文本和分析



第四章：经济

概述	216
章节要点	217
4.1 2024 年人工智能领域大事件时间线	219
4.2 工作岗位	223
人工智能劳动力需求	223
全球人工智能劳动力需求	223
按技能集群和专业技能分类,	
美国人工智能劳动力需求情况	225
美国各行业的人工智能劳动力需求	228
美国各州的人工智能劳动力需求	229
人工智能招聘	232
人工智能技能渗透	234
人工智能人才	236
重点：衡量人工智能与经济的融合情况	242
4.3 投资	246
企业投资	246
初创企业活动	247
全球趋势	247
按资金规模进行的地区比较	251
按地区划分的新获融资人工智能	
公司对比	255
重点领域分析	258
4.4 企业活动	260
行业使用情况	260
利用人工智能能力	260
生成式人工智能功能部署	264
人工智能对劳动力的影响	267
4.5 机器人部署	272
总体趋势	272
工业机器人：传统机器人与协作机器人	274
按地理区域划分	275
服务机器人的国家级数据	279

获取公共数据

第四章： 经济

概述

2024 年人工智能的经济影响进一步凸显，对多行业产生实质性冲击。生成式人工智能在特定任务中带来的早期生产效率提升已可量化，但该技术对宏观经济的长期影响仍存争议。劳动力市场初现人工智能驱动的转型迹象：随着新型人工智能相关岗位涌现，部分知识型职业正经历结构性调整。跨行业、跨地域的企业正从实验性应用转向系统性整合人工智能技术。投资趋势也反映出人工智能生态日趋成熟，资金日益集中于企业自动化专项应用及垂直行业解决方案。

本章节基于 Lightcast、领英、Quid、麦肯锡及国际机器人联合会（IFR）数据，剖析人工智能相关经济趋势。首先解析人工智能相关职业的劳动力需求、招聘动向、技能渗透率及人才供给现状；然后，本章节探讨了企业在人工智能领域的投资，其中有一部分专门关注生成式人工智能。最后，它评估了人工智能对生产力的影响以及各行各业的机器人安装情况。

第四章：

经济

章节要点

1. 全球私人人工智能投资创历史新高，增长 26%。2024 年，全球企业人工智能投资规模达到 2523 亿美元，其中私人投资同比增长 44.5%，并购交易规模较上年增长 12.1%。过去十年间，该领域经历显著扩张，自 2014 年以来，总投资规模增长逾十三倍。

2. 生成式人工智能投资规模激增。2024 年，生成式人工智能领域的私人投资达到 339 亿美元，比 2023 年增长 18.7%，是 2022 年水平的 8.5 倍以上。该领域目前占所有人工智能相关私人投资总额的 20% 以上。

3. 美国扩大了其在全球人工智能私人投资中的领先优势。2024 年，美国人工智能私人投资规模达到 1091 亿美元，相当于中国的近 12 倍（93 亿美元），英国的 24 倍（45 亿美元）。在生成式人工智能领域，美国投资总额较中国、欧盟及英国投资总和还多 254 亿美元，较 2023 年 218 亿美元的差额继续扩大。

4. 人工智能的使用达到前所未有的水平。2024 年，受访企业报告采用人工智能技术的比例从 2023 年的 55% 跃升至 78%。同样，在至少一项业务职能中使用生成式人工智能的受访者数量增长逾一倍——从 2023 年的 33% 上升至 2024 年的 71%。

5. 人工智能已开始在多个业务职能领域产生财务效益，但大多数企业仍处于应用初期阶段。报告显示，在单一业务职能内应用人工智能并取得财务效益的企业中，多数反馈效益水平仍处于较低区间。在成本节约方面，在客户服务运营中使用人工智能的企业中有 49% 的受访者报告实现降本，供应链管理领域这一比例为 43%，软件工程领域为 41%。不过，这些企业报告的成本降幅大多不足 10%。在收入增长方面，营销与销售部门应用人工智能的企业中有 71% 的受访者报告收入提升，供应链管理领域为 63%，服务运营领域为 57%。但需要指出的是，这些收入增幅普遍低于 55%。

6. 人工智能的应用呈现出明显的区域差异，其中大中华区正迅速崛起。尽管北美地区仍保持着企业人工智能应用率的领先地位，但大中华区是同比增长率最高的地区之一，其企业人工智能采用率提升了 27%。欧洲紧随其后，实现了 23% 的增长，这表明全球人工智能格局正在快速演变，各国在人工智能应用领域的国际竞争也日趋激烈。

第四章：

经济

章节要点（续）

7. 中国在工业机器人领域仍占据主导地位，虽略有放缓。2023年，中国安装27.63万台工业机器人，是日本的6倍，美国的7.3倍。自2013年超过日本以来，中国在全球工业机器人安装量中的份额从20.8%上升至51.1%。虽然中国的机器人安装量继续超过世界其他国家的总和，但这一差距在2023年略有缩小，标志着其急剧扩张的势头略有放缓。

8. 协作型和交互式机器人使用日趋普及。2017年，协作型机器人仅占所有新安装工业机器人的2.8%，到2023年，这一数字攀升至10.5%。同样，2023年，除医疗机器人外，所有应用领域的服务机器人安装量均呈现增长趋势。这一趋势不仅表明机器人安装量的整体增长，还表明人们越来越重视将机器人部署在面向人类的岗位上。

9. 人工智能正在推动能源结构的重大变革，并引发对核能的新一轮关注。微软宣布以16亿美元重启三哩岛核反应堆，为人工智能提供动力，而谷歌和亚马逊也签订了核能协议，为人工智能业务提供支持。

10. 人工智能提高生产效率并缩小技能差距。去年的人工智能指数报告是首批强调人工智能对生产效率产生积极影响的研究报告之一。今年更多研究进一步验证了这些发现，证实人工智能不仅能提升生产效率，在多数情况下还有助于缩小高技能与低技能劳动者之间的能力差距。

第四章：经济

2024 年人工智能领域大事件时间线

本章节开篇概述了 2024 年人工智能领域最具影响力的经济事件，这些事件由人工智能指数指导委员会权威遴选。

4.1 2024年人工智能领域大事件时间线

日期	活动	类型	图片
2024年1月16日	Synopsys 以 350 亿美元收购 Ansys, 强化芯片至系统级全栈设计解决方案	收购	 图 2.1.6 资料来源: Inflection, 2025
2024年2月21日	有报道称, OpenAI 的年化收入在 2023 年 12 月超过了 20 亿美元。	估值里程碑	 图 4.1.2 资料来源: Inc., 2024
2024年2月29日	人形机器人初创公司 Figure AI 以 26 亿美元的估值融资 6.75 亿美元。	投资/融资	 图 4.1.3 资料来源: SiliconAngle, 2024
2024年3月21日	微软 聘请包括联合创始人在内的 Inflection AI 的大部分员工, 并支付 6.5 亿美元以获得 Inflection 人工智能模型的授权。	收购	 图 4.1.4 资料来源: 路透社, 2024
2024年3月21日	人工智能云基础设施初创企业 CoreWeave 获得 11 亿美元的融资, 估值为 190 亿美元。	投资/融资	 图 4.1.5 资料来源: 财富, 2024

第四章：经济

2024 年人工智能领域大事件时间线

2024年5月1日	数据标签初创企业 Scale AI 融资 10 亿美元，估值达到 138 亿美元。	投资/融资	 图 4.1.6 资料来源：路透社，2024
2024年6月11日	法国开源人工智能模型初创公司 Mistral AI 以 60 亿美元的估值融资 6.4 亿美元。	投资/融资	 图 4.1.7 资料来源：TechCrunch，2024
2024年6月14日	精准医疗人工智能企业 Tempus AI 成功上市，募资 4.107 亿美元，估值突破 60 亿美元	投资/融资	 图 4.1.8 资料来源：路透社，2024
2024年7月22日	Cohere, 一家专注于企业应用的人工智能初创公司，完成 5 亿美元融资，估值达 55 亿美元。	投资/融资	 图 4.1.9 资料来源：Crunchbase，2024
2024年8月2日	谷歌收购 Character.AI，以约 25 亿美元收购其股东股份，并获得该初创公司的人工智能技术授权。此次收购包括引入 Character.AI 的联合创始人及研究团队成员。	收购	 图 4.1.10 资料来源：The Verge，2024
2024年8月5日	Groq, 一家专注于快速推理的人工智能芯片初创公司，在最新一轮融资中筹集了 6.4 亿美元，公司估值达 28 亿美元。	投资/融资	 图 4.1.11 资料来源：GroqGroq，2024

第四章：经济

2024年人工智能领域大事件时间线

2024年8月12日	AMD 以约 6.65 亿美元收购欧洲最大的私人人工智能实验室 Silo AI。	收购	 图 4.1.12 资料来源: AMD, 2024
2024年9月5日	Safe Superintelligence(SSI)获得 10 亿美元融资。	投资/融资	 图 4.1.13 资料来源: TechCrunch , 2024
2024年9月12日	Salesforce 在其平台上推出了 Agent-force, 这是一个用于业务运营的自主 AI 智能体套件。	产品发布/整合	 图 4.1.14 资料来源: Salesforce , 2024
2024年9月20日	微软宣布与 Constellation Energy 达成一项价值 16 亿美元的协议, 以重启三里岛核反应堆, 为人工智能数据中心供电。	合作伙伴	 图 4.1.15 资料来源: 美国国家公共广播电台, 2024
2024年10月2日	OpenAI 在估值达 1570 亿美元的情况下, 融资 66 亿美元。	投资/融资	 图 4.1.16 资料来源: Axios, 2024
2024年10月14日	谷歌宣布与 Kairos Power 达成协议, 将从该公司开发的多个小型模块化反应堆(SMR)购买核能。	合作伙伴	 图 4.1.17 资料来源: 谷歌 , 2024
2024年10月16日	亚马逊宣布与 Energy Northwest、X-energy 和 Dominion Energy 合作, 推出小型模块化反应堆(SMR)开发核能计划。	合作伙伴	 图 4.1.18 资料来源: 亚马逊, 2024

第四章：经济

2024 年人工智能领域大事件时间线

2024年10月17日	谷歌的 NotebookLM 去掉了“实验性”标签，并拥有数百万用户和超过 8 万家组织。	产品发布/整合	 图 4.1.19 资料来源：谷歌， 2024
2024年11月22日	Anthropic 与 AWS 扩大合作, 亚马逊再投资 40 亿美元, 使总投资额达到 80 亿美元。	合作伙伴	 图 4.1.20 资料来源：Anthropic, 2024
2024年12月17日	人工智能数据分析公司 Databricks 在最新一轮融资中筹集了 \$100 亿美元, 公司估值达 \$620 亿美元。	投资/融资	 图 4.1.21 资料来源：TechCrunch , 2024
2024年12月18日	专注于人工智能搜索产品的初创公司 Perplexity AI 以 90 亿美元的估值融资 5 亿美元。	投资/融资	 图 4.1.22 资料来源：人工智能杂志, 2024
2024年12月23日	xAI 宣布完成 60 亿美元融资, 累计融资总额达 120 亿美元, 估值超过 400 亿美元。	投资/融资	 图 4.1.23 资料来源：福布斯 , 2024
2024年12月30日	英伟达以 7 亿美元收购 以色列初创公司 Run:ai, 以增强其在算力需求环境中的图形处理器优化能力。	收购	 图 4.1.24 资料来源：TechCrunch , 2024

4.2 工作岗位

人工智能劳动力需求

本节利用 Lightcast 提供的数据，分析劳动力市场对人工智能相关技能的需求。自 2010 年以来，Lightcast 已分析了来自 51,000 多个网站的数亿条招聘信息，并筛选出其中要求人工智能技能的岗位。

求的占比情况。2024 年，新加坡(3.2%)、卢森堡(2%)和中国香港(1.9%)在这一指标上处于领先地位。2023 年，美国招聘岗位中人工智能相关职位占比为 1.4%，到 2024 年这一数字上升至 1.8%。从 2023 年到 2024 年，大多数国家对人工智能技能需求的岗位比例均有所增长。

全球人工智能劳动力需求

图 4.2.1 和图 4.2.2 显示了招聘岗位中对人工智能技能需

2014–2024 年按选定地理区域划分人工智能岗位招聘信息（占所有招聘信息的比例）（第 1 部分）

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数报告

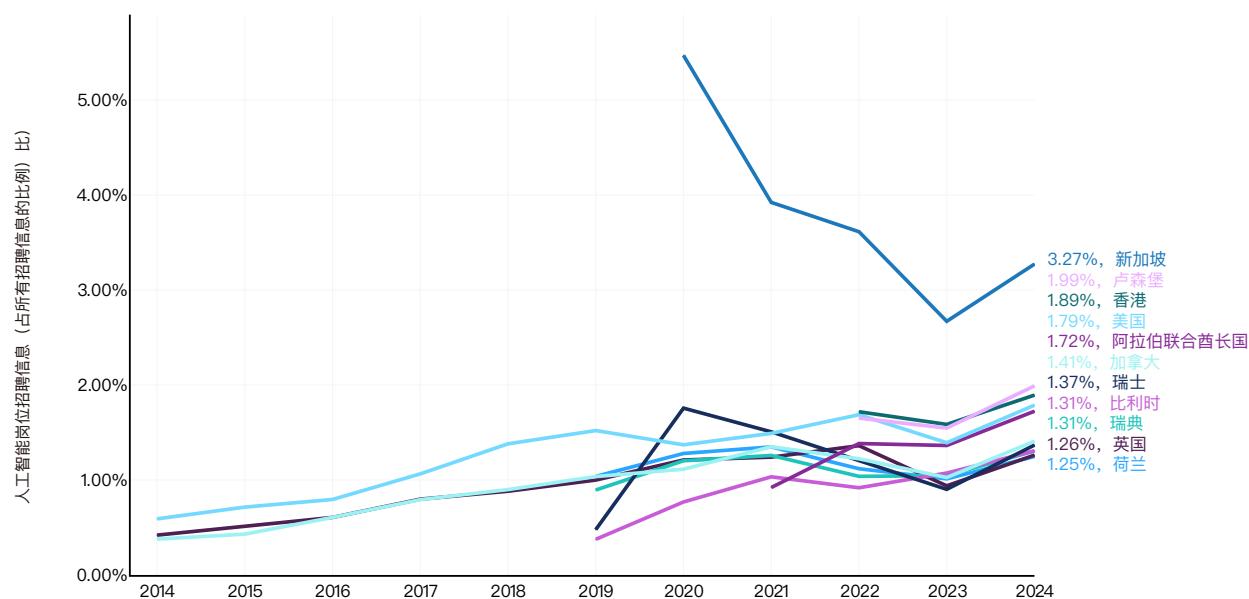


图 4.2.1

第四章：经济

4.2 工作岗位

2014–2024 年按选定地理区域划分人工智能岗位招聘信息（占所有招聘信息的比例）（第 2 部分）

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数报告

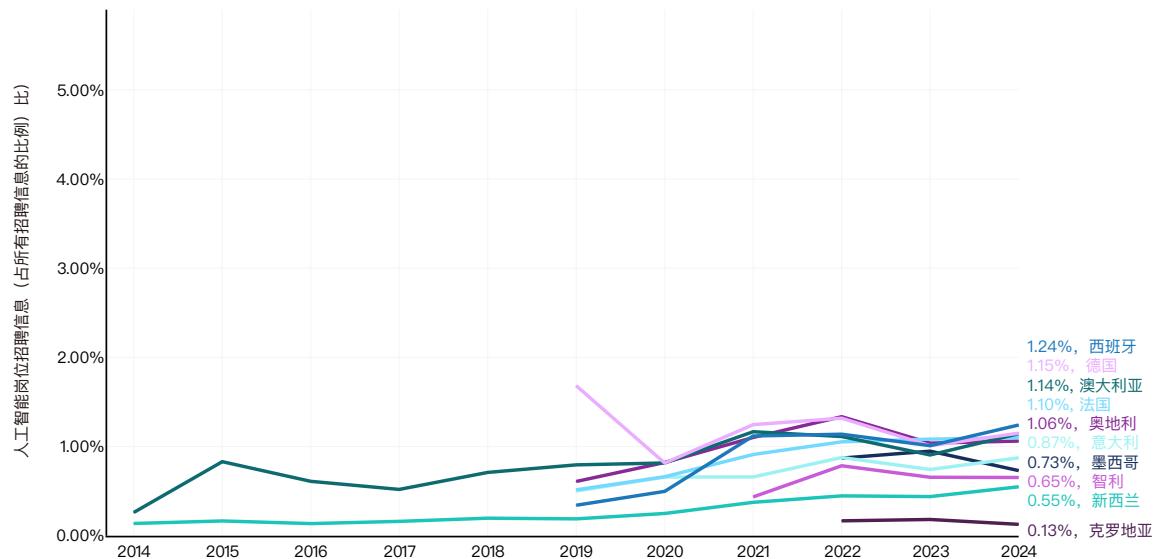


图 4.2.2

第四章：经济

4.2 工作岗位

按技能集群和专业技能分类，美国人工智能劳动力需求情况

图 4.2.3 显示了自 2010 年以来美国劳动力市场上最热门的人工智能技能。人工智能的需求占 0.9%，紧随其后的是机器学习(也是 0.9%)和自然语言处理(0.2%)。自去年以来，除自

动驾驶和机器人外，Lightcast 跟踪的大多数人工智能相关技能集群的市场份额都有所增长。生成式人工智能的增幅最大，增长了近四倍。

2014–2024 年按技术集群分类人工智能岗位招聘信息（占所有招聘信息的比例）

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数报告

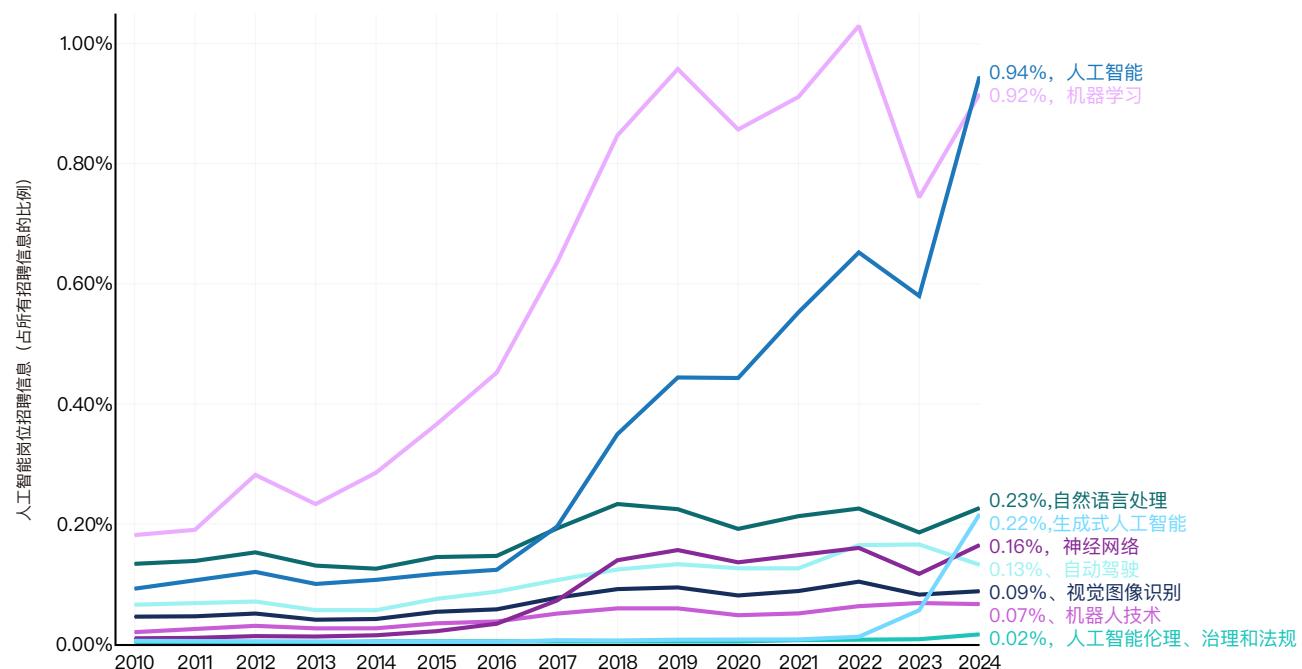


图 4.2.3¹

1、一份招聘启事可以列出多种人工智能技能。

第四章：经济

4.2 工作岗位

图 4.2.4 比较了 2024 年人工智能岗位招聘中需求最高的 10 项专业技能与 2012 年至 2014 年期间的需求情况。² 从绝对值来看，过去十年间每项专业技能的需求均有所增长，其中

Python 的显著增长凸显了其作为首选人工智能编程语言的地位。

2012–2014 年与 2024 年美国人工智能岗位招聘中的十大专业技能对比

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数据报告

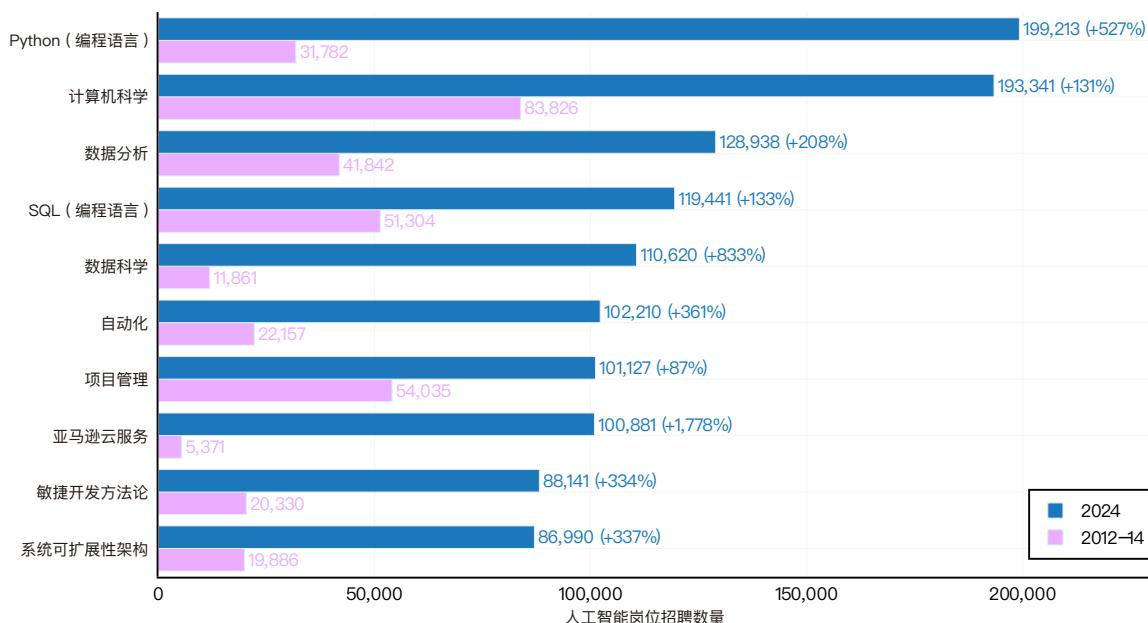


图 4.2.4

²之所以决定选择 2012–2014 年作为比较点，是因为前几年工作 / 技能层面的数据很少。因此，Lightcast 采用了 2012–2014 年的数据，以大的样本量作为 10 年前的比较基准。图 4.2.4 将 2012 年至 2014 年需要特定技能的职位发布总数与 2024 年的总数并列。

第四章：经济

4.2 工作岗位

2024 年，美国招聘信息中提及生成式人工智能技能的岗位数量较上年增长超过三倍（图 4.2.5）。图 4.2.6 展示了 2024 年和 2023 年发布的人工智能招聘信息中提及特定生成式人工智能技能的占比。

2023 年与 2024 年美国人工智能岗位招聘信息中生成式人工智能技能的对比

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数报告

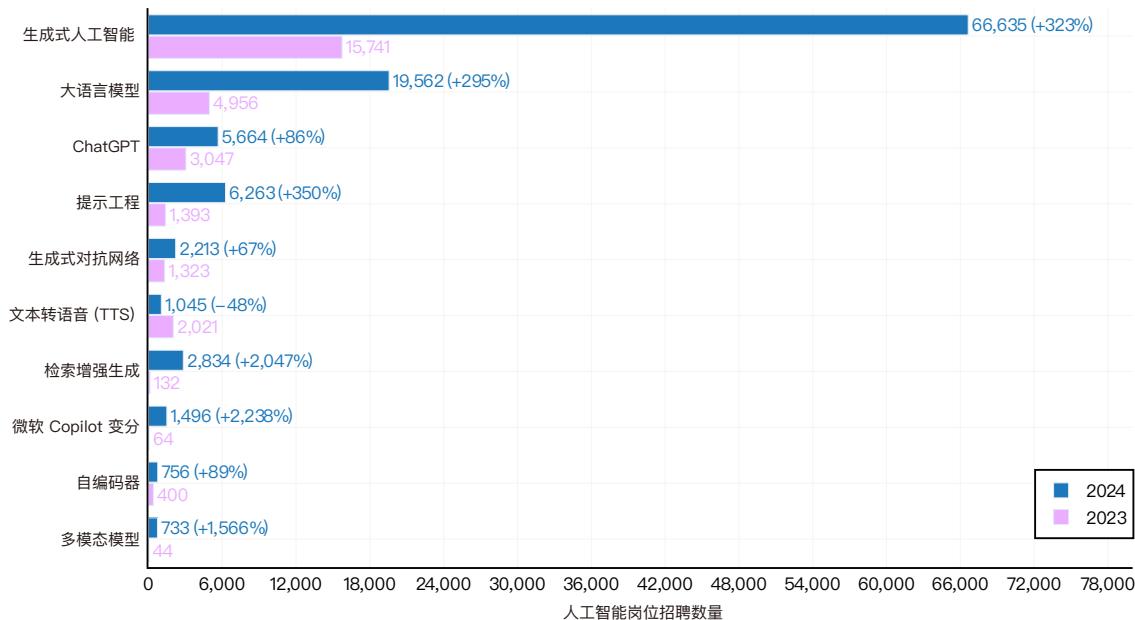


图 4.2.5

2023 年与 2024 年美国人工智能职位招聘信息中生成式人工智能技能所占的比例

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数报告

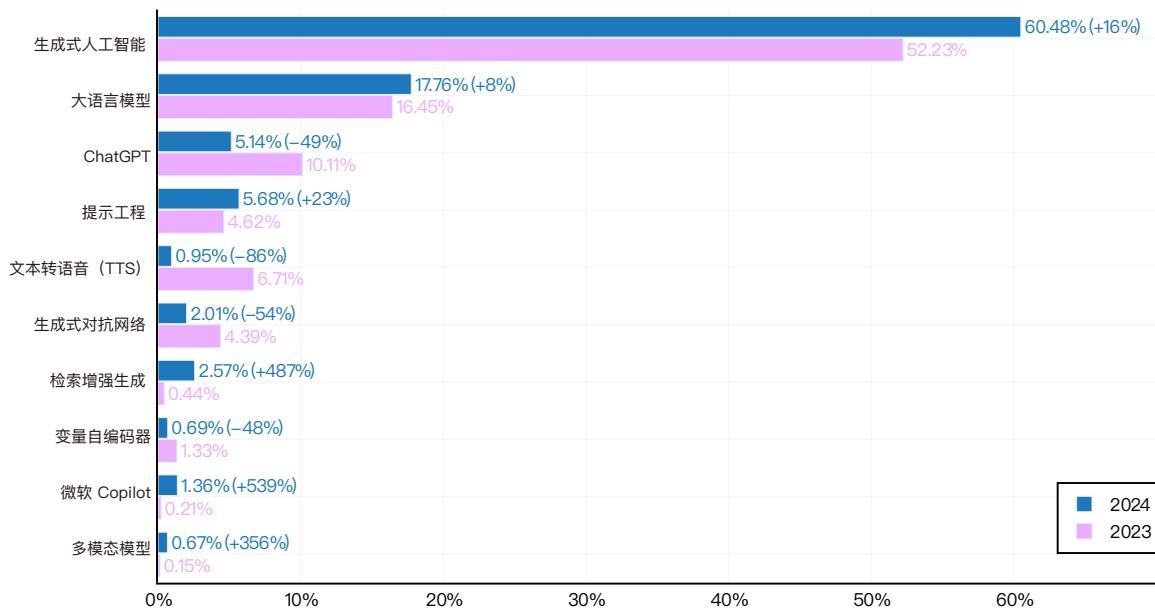


图 4.2.6

第四章：经济

4.2 工作岗位

美国各行业的人工智能劳动力需求

图 4.2.7 显示了 2023 年至 2024 年美国各行业领域对人工智能技能需求岗位的占比。与 2023 年相比，2024 年几乎所有行业领域对人工智能技能需求岗位的占比均有所上升，公共行政领域除外。与 2023 年相比，2024 年几乎所有行业领域对

人工智能技能需求岗位的占比均有所上升，公共行政领域除外。

2023 年与 2024 美国各行业人工智能岗位招聘信息占总招聘信息比例的对比

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数报告

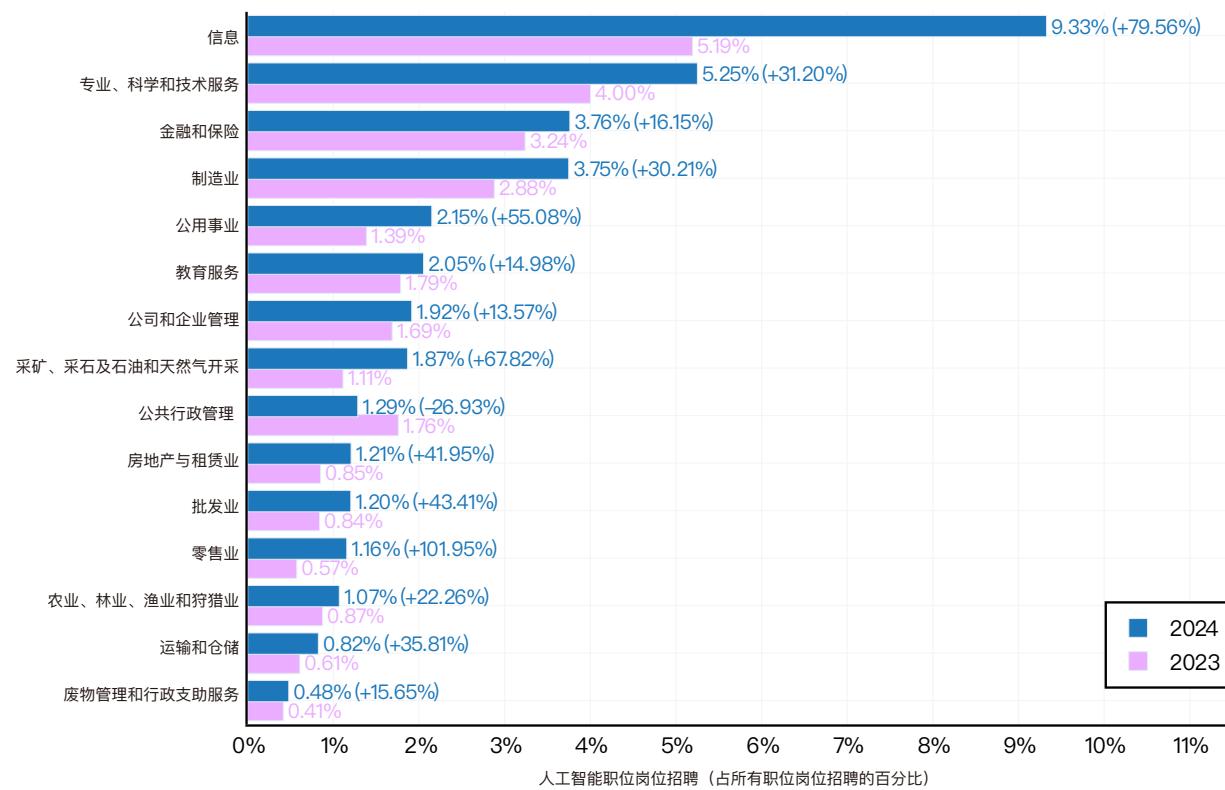


图 4.2.7³

3、图 4.2.7 中的行业分类基于两位数的 NAICS 代码。有关劳工统计局（Bureau of Labor Statistics）的超行业分类和 NAICS 分类的更多信息，请参阅以下参考资料。

第四章：经济

4.2 工作岗位

美国各州的人工智能劳动力需求

图 4.2.8 展示了美国各州的人工智能职位空缺数量。排名前三的州分别是加利福尼亚州(103,375)、得克萨斯州(57,785)和纽约州(37,944)。

2024 年美国各州人工智能职位发布数量

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数报告



图 4.2.9 展示了各州人工智能相关职位发布占该州总职位发布量的百分比。根据这一指标，排名前三的州是华盛顿特区(4.4%)、特拉华州(3.4%)和华盛顿州(3.3%)。

2024 年美国各州人工智能职位发布占该州总职位发布比例

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数报告

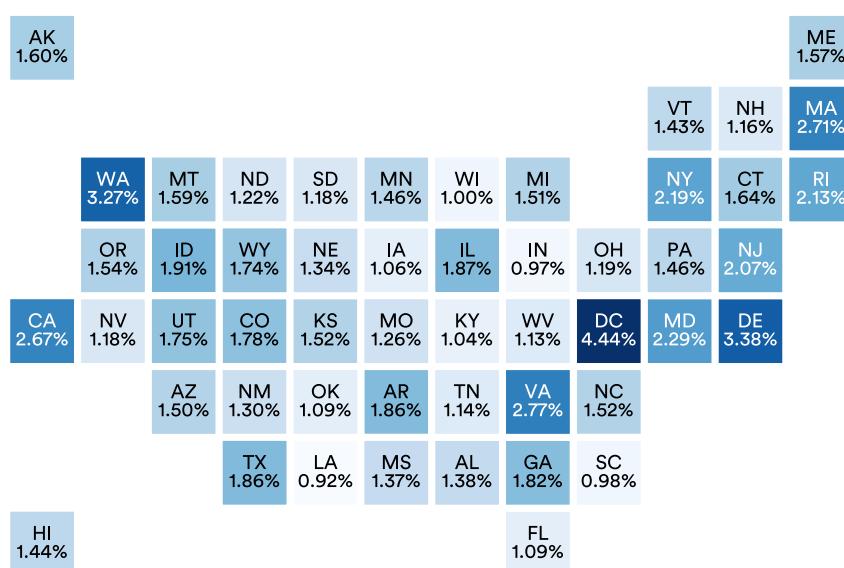


图 4.2.9

第四章：经济

4.2 工作岗位

图 4.2.10 显示了美国哪些州的人工智能职位发布在全国所占比例最大。2024 年，美国 15.7% 的人工智能发布在加利福尼亚州，其次是得克萨斯州(8.8%)和纽约州(5.8%)。

图 4.2.11 展示了四个拥有大量人工智能岗位招聘信息的州——华盛顿州、加利福尼亚州、纽约州和得克萨斯州——的趋势。从 2023 年到 2024 年，这四个州的人工智能相关岗位招聘信息在总招聘信息中的占比均显著上升。

2024 年美国各州人工智能职位发布比例

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数报告

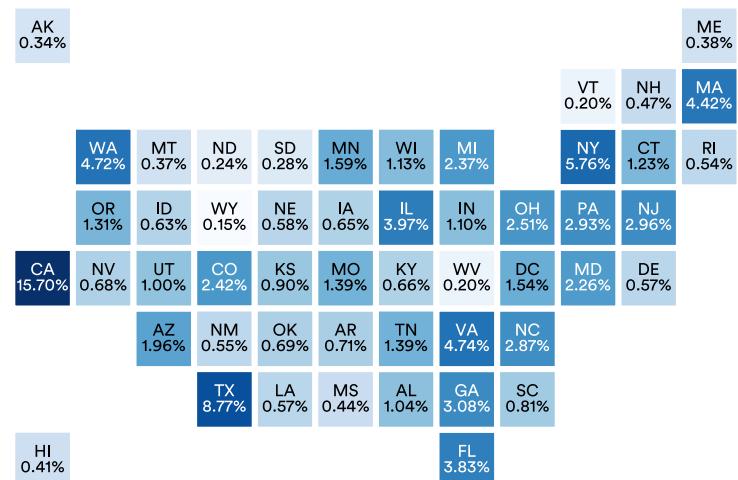


图 4.2.10

2010–2024 年美国各选定州的人工智能职位占该州总职位发布量的比例

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数报告

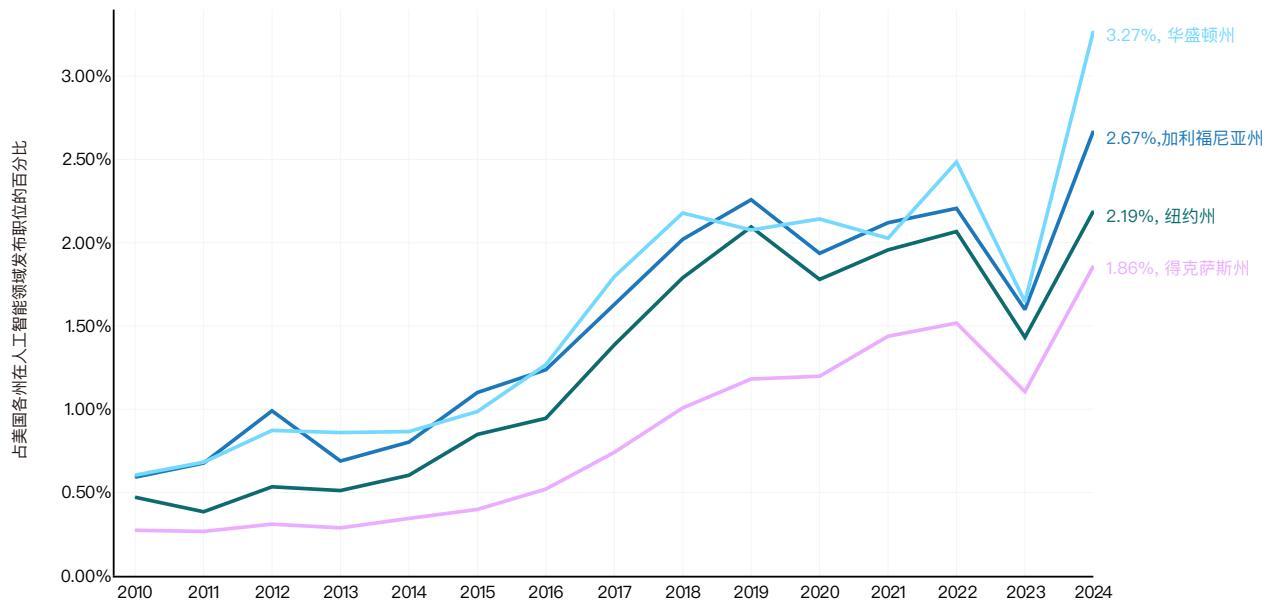


图 4.2.11

第四章：经济

4.2 工作岗位

图 4.2.12 显示了随着时间的推移，人工智能相关职位在前四个州的分布情况。2024 年，所有四个州都扭转了多年来人工智能相关职位比例下降的趋势。人工智能职位发布 -- 加利福

尼亚州和的变化尤为明显，自 2020 年以来，这两个州的职位发布都有所减少。

2010–2024 年全美人工智能职位发布量中各选定州所占比例

资料来源：Lightcast, 2024 | 图表：2025 年人工智能指数报告

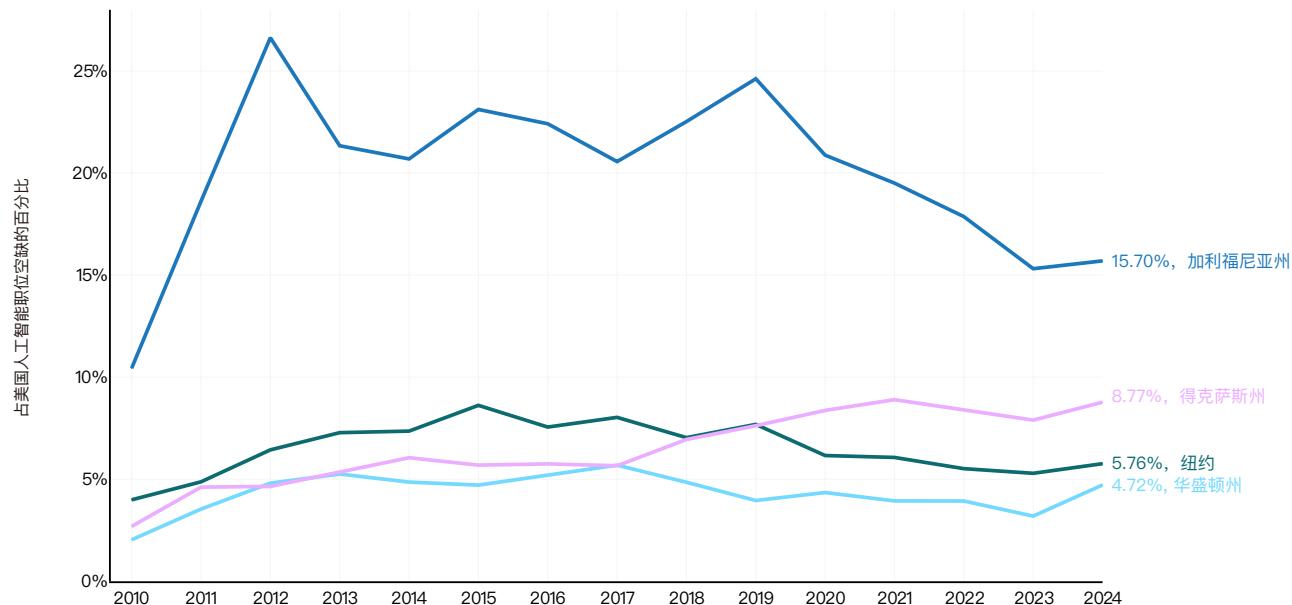


图 4.2.12

人工智能招聘

人工智能指数中提供的招聘数据基于领英（领英）的经济图谱(Economic Graph)，反映了该平台 10 多亿会员的工作和技能情况。因此，数据会受到会员选择使用平台方式的影响，这可能会因职业、社交和地区文化以及网站的整体可用性和可访问性而有所不同。人工智能指数指出，样本中的匈牙利、印度尼西亚、印度和韩国的领英覆盖的比例较低，因此在解释这些国家的情况时应特别谨慎。

图 4.2.13 报告了按地理区域划分的人工智能相对招聘率同比比率。整体招聘率的计算方式为：在同一时间段内添加新雇主的领英会员人数除以该地区领英会员总数。相反，相对人

人工智能人才招聘率是指与同一地理区域整体招聘率相比，人工智能招聘的同比变化。⁴因此，图 4.2.13 展示了人工智能人才招聘活跃度最高的地区情况——这些地区的人工智能人才招聘增长率显著超过整体招聘增速。2024 年，人工智能人才相对招聘率同比增长最显著的国家依次为：印度(33.4%)、巴西(30.8%)和沙特阿拉伯(28.7%)。这意味着，以印度为例，2024 年其人工智能人才招聘数量与整体招聘数量的比率同比增长了 33.4 个百分点。

2024 年按地理区域划分的人工智能相对招聘率同比

资料来源：领英，2024| 图表：2025 年人工智能指数报告

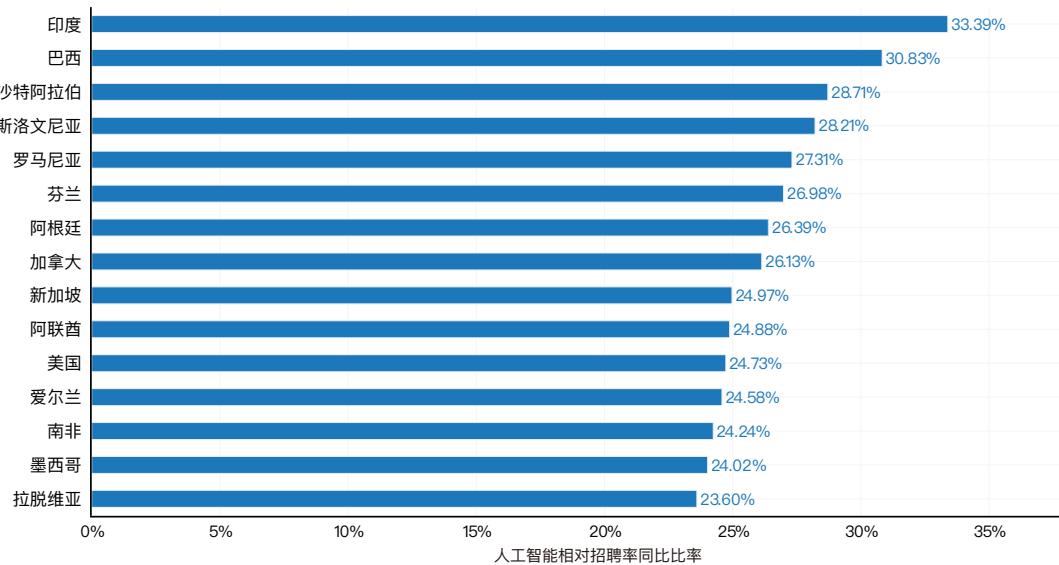

 图 4.2.13⁵

图 4.2.14 展示了过去五年各地区人工智能招聘的年同比。从 2024 年开始，阿根廷、巴西和智利等南美国家的人工智能招

聘率显著上升。其他近期出现类似增长的国家包括加拿大、印度、南非和美国。

4、领英每月计算某一地理区域的人工智能招聘率，将其除以该地区的整体招聘率，计算该比率的年度变化，然后取过去 12 个月的移动平均值。

5、为简洁起见，可视化仅包括该指标排名前 15 的国家。

第四章：经济

4.2 工作岗位

2018–2024 年按地理区域划分的人工智能相对招聘年同比

资料来源：领英，2024 | 图表：2025 年人工智能指数报告

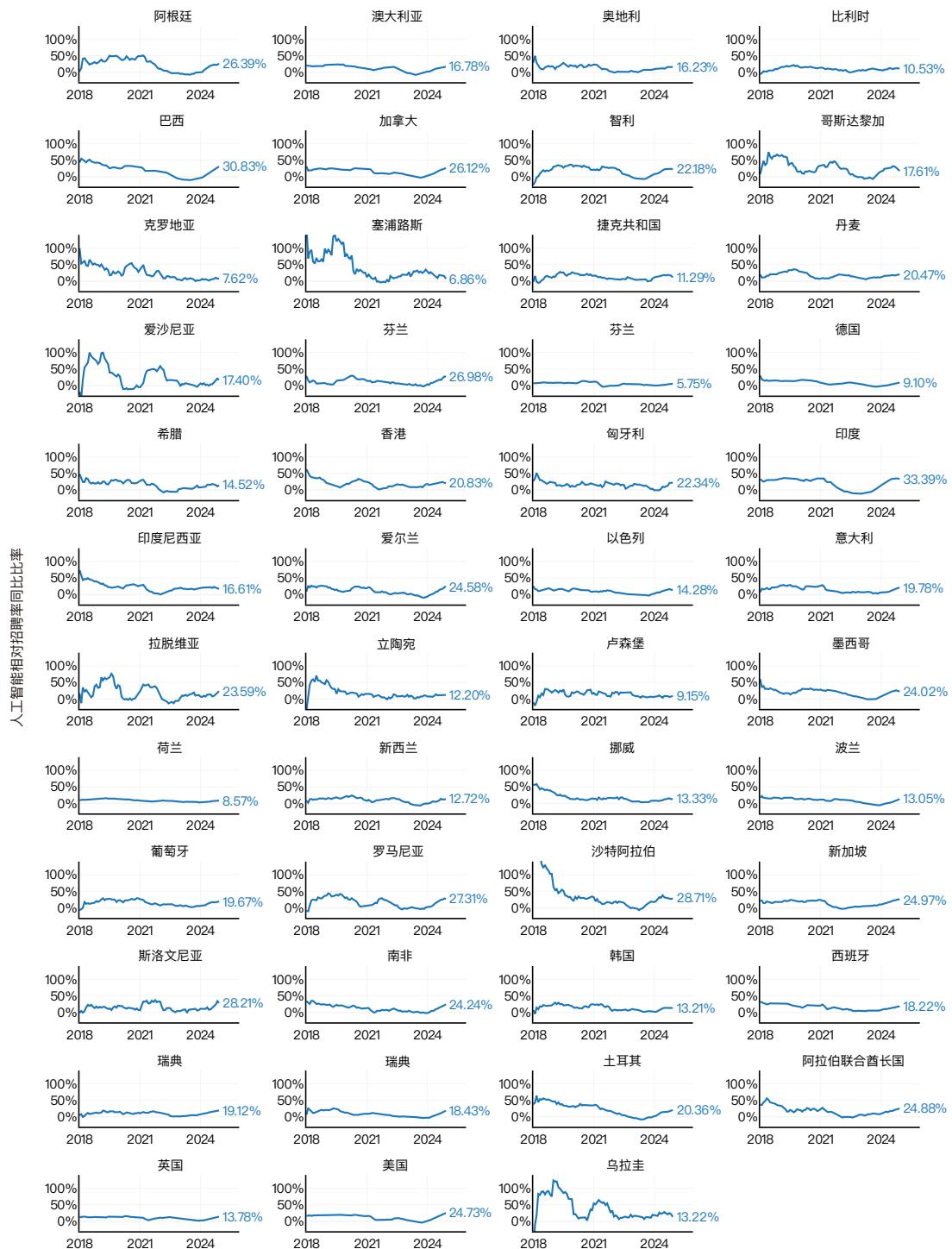


图 4.2.14

人工智能技能渗透

图 4.2.15 和图 4.2.16 突出了相对的人工智能技能渗透率。该指标旨在衡量特定国家、行业或性别人工智能技能的强度。人工智能技能渗透率表明了人工智能技能在各职业中的普及程度，或 领英会员在工作中使用人工智能技能的强度。例如，工程师职业的前 50 项技能是根据它们在 领英会员个人资料中出现的加权频率计算得出的。例如，如果工程师拥有的技能中有 4 项属于人工智能技能组，则工程师中人工智能技能的普及率估计为 8% (4/50)。

在 2015 年至 2024 年期间，人工智能技能普及率最高的国家是美国(2.6)和印度(2.5)。其次是英国(1.4)、德国(1.3)和巴西(1.3)。因此，在同一组职业中，美国人工智能技能的相对渗透率是全球平均水平的 2.6 倍。

2015–2024 年按地理区域划分的人工智能技能相对渗透率

资料来源：领英，2024 | 图表：2025 年人工智能指数报告

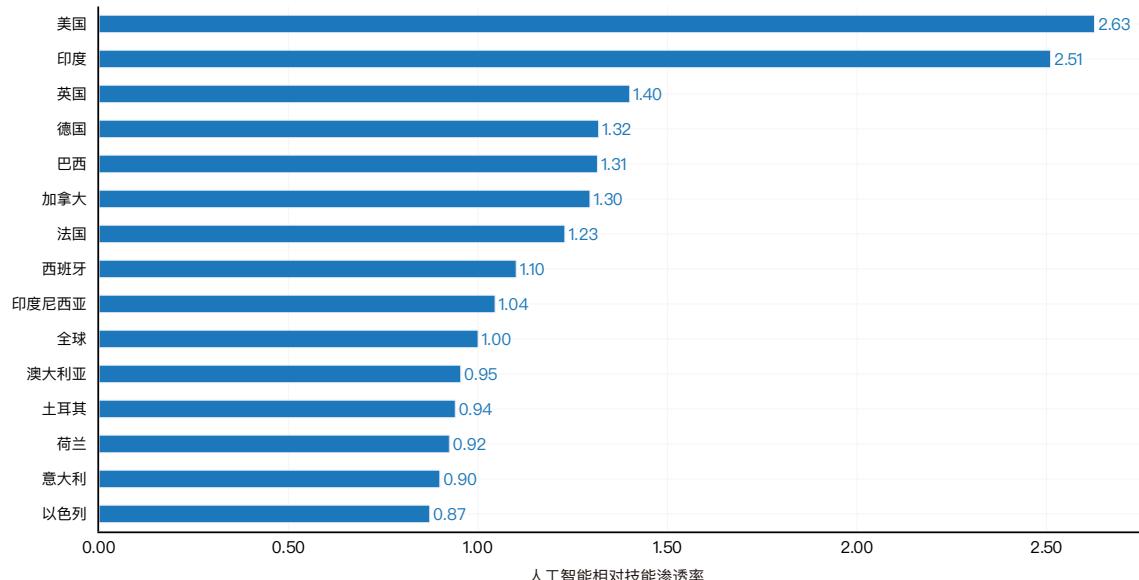


图 4.2.15

第四章：经济

4.2 工作岗位

图 4.2.16 按性别分类了不同国家或地区的人工智能技能渗透率。一个国家的女性普及率为 1.5，这意味着该国领英女性会员拥有人工智能技能的可能性是全球所有国家平均水平的

1.5 倍。在所有样本国家中，除沙特阿拉伯外，男性的人工智能技能渗透率都高于女性。在所有样本国家中，印度(1.9)、美国(1.7)和加拿大(1.0)报告的女性人工智能技能相对渗透率最高。

2015–2024 年人工智能技能在不同性别间的相对渗透率

资料来源：领英，2024 | 图表：2025 年人工智能指数报告

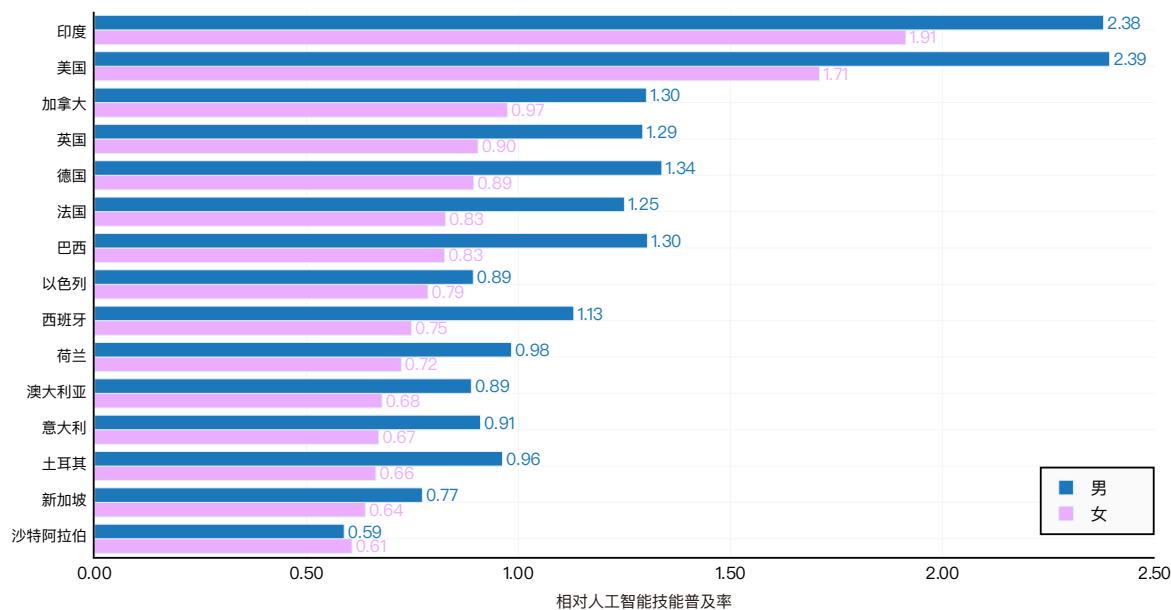


图 4.2.16

第四章：经济

4.2 工作岗位

人工智能人才

图 4.2.17 和 4.2.18 按国家分类了人工智能人才。如果 领英会员在个人档案中明确添加了人工智能技能、从事或曾经从事过人工智能工作，那么他们就被视为人工智能人才。人工智能人才的数量被计算人才集中度，即人工智能人才在会员中所占的比例。请注意，人才集中度指标可能会受到领英在这些国家覆盖范围的影响，因此谨慎应使用。

2024 年按地理区域划分的人工智能人才集中度

资料来源：领英，2024 | 图表：2025 年人工智能指数报告

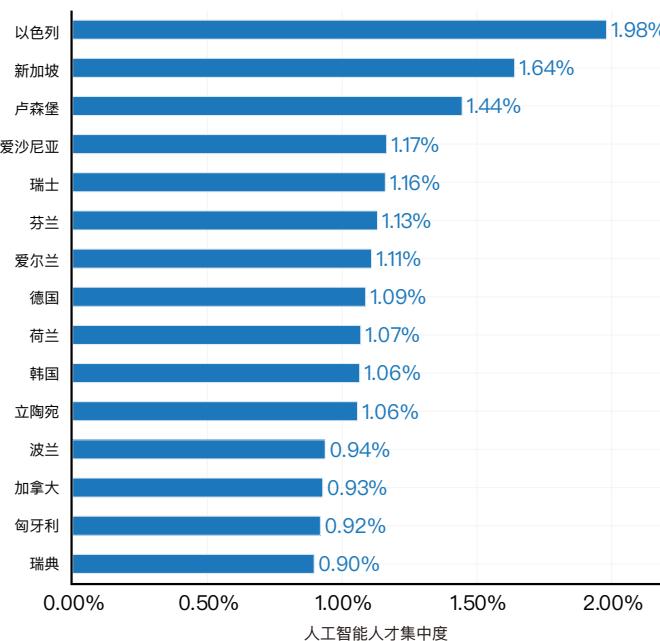


图 4.2.17

图 4.2.17 显示了不同地理区域的人工智能人才集聚情况。2024 年，人工智能人才最集中的国家包括以色列 (2.0%)、新加坡 (1.6%) 和卢森堡 (1.4%)。图 4.2.18 显示了部分国家自 2016 年以来人工智能人才集中度的百分比变化。在此期间，几个主要经济体的人工智能人才库大幅增加。增幅最大的国家是印度 (252%)、哥斯达黎加 (240%) 和葡萄牙 (237%)。

2016 年与 2024 年按地理区域划分的人工智能人才集中度变化比例

资料来源：领英，2024 | 图表：2025 年人工智能指数报告

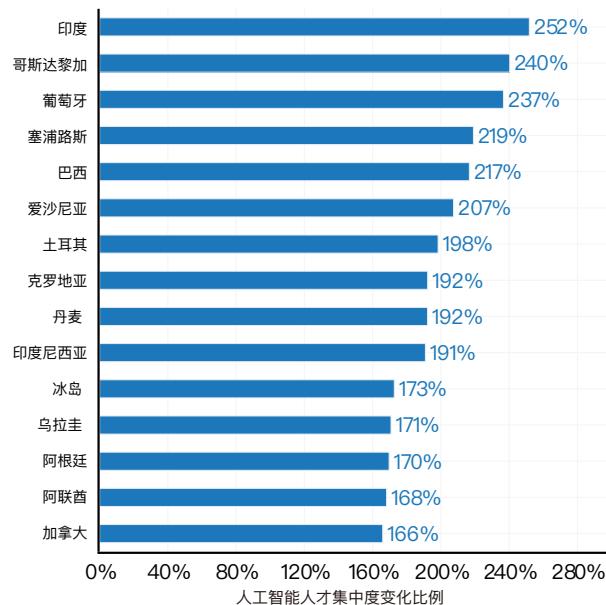


图 4.2.18

人工智能人才分布也存在显著的性别差异。在分析样本中的所有国家中，除印度和沙特阿拉伯外，男性人工智能人才的

集中度均高于女性（图 4.2.19）。以色列报告的 2024 年女性人工智能人才集中度最高，为 1.6%。

第四章：经济

4.2 工作岗位

2016–2024 年按性别和地理区域划分的人工智能人才集中度

资料来源：领英，2024 | 图表：2025 年人工智能指数报告

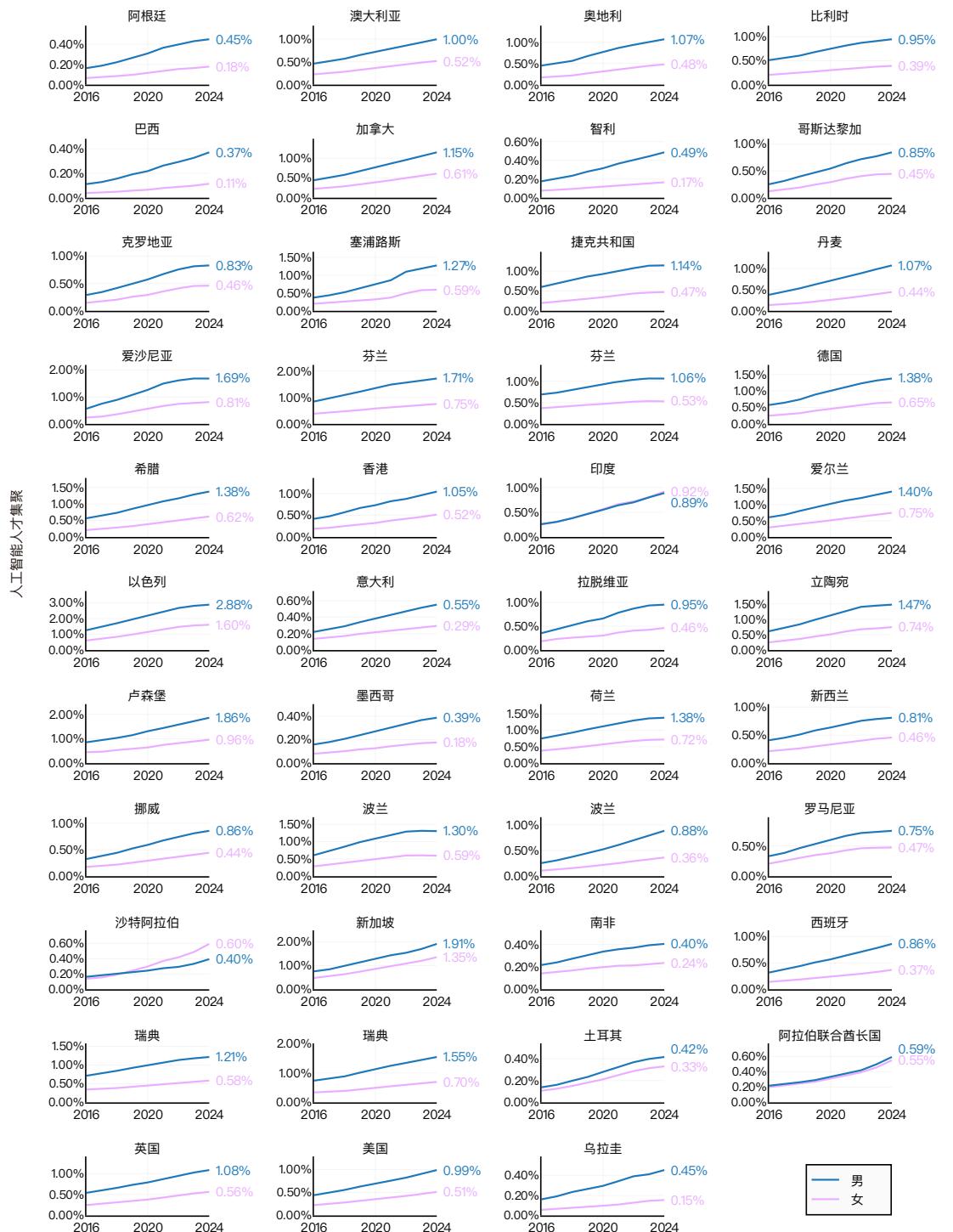


图 4.2.19

第四章：经济**4.2 工作岗位**

领英还跟踪了人工智能人才的性别分布(图 4.2.20)。据估计，2024 年该平台上 69.5% 的人工智能专业人员为男性，30.5% 为女性。随着时间的推移，这一比例一直保持稳定。

2016–2024 年全球人工智能人才分布情况

资料来源：领英，2024| 图表：2025 年人工智能指数报告

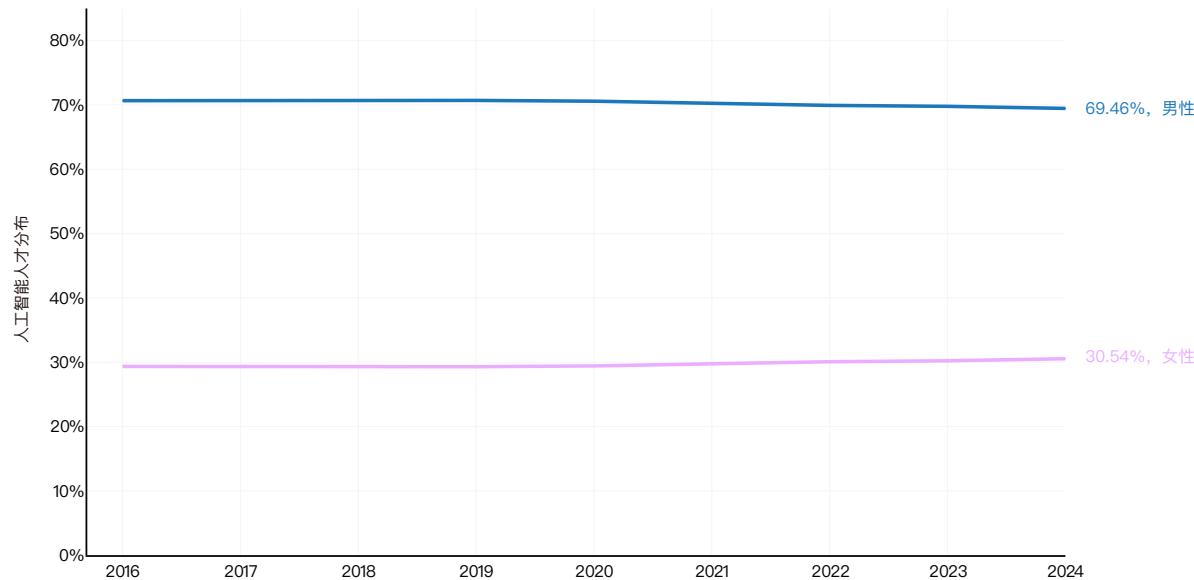


图 4.2.20

领英有关人工智能人才的数据还可以按国家细分。在样本的每个国家，担任人工智能职位的男性比例都高于女性(图 4.2.21)。新西兰和罗马尼亚的性别分布最为均衡，而巴西和智利的性别分布最不均衡。

第四章：经济

4.2 工作岗位

2016–2024 年按性别和地理分类的人工智能人才分布比例

资料来源：领英，2024 | 图表：2025 年人工智能指数报告

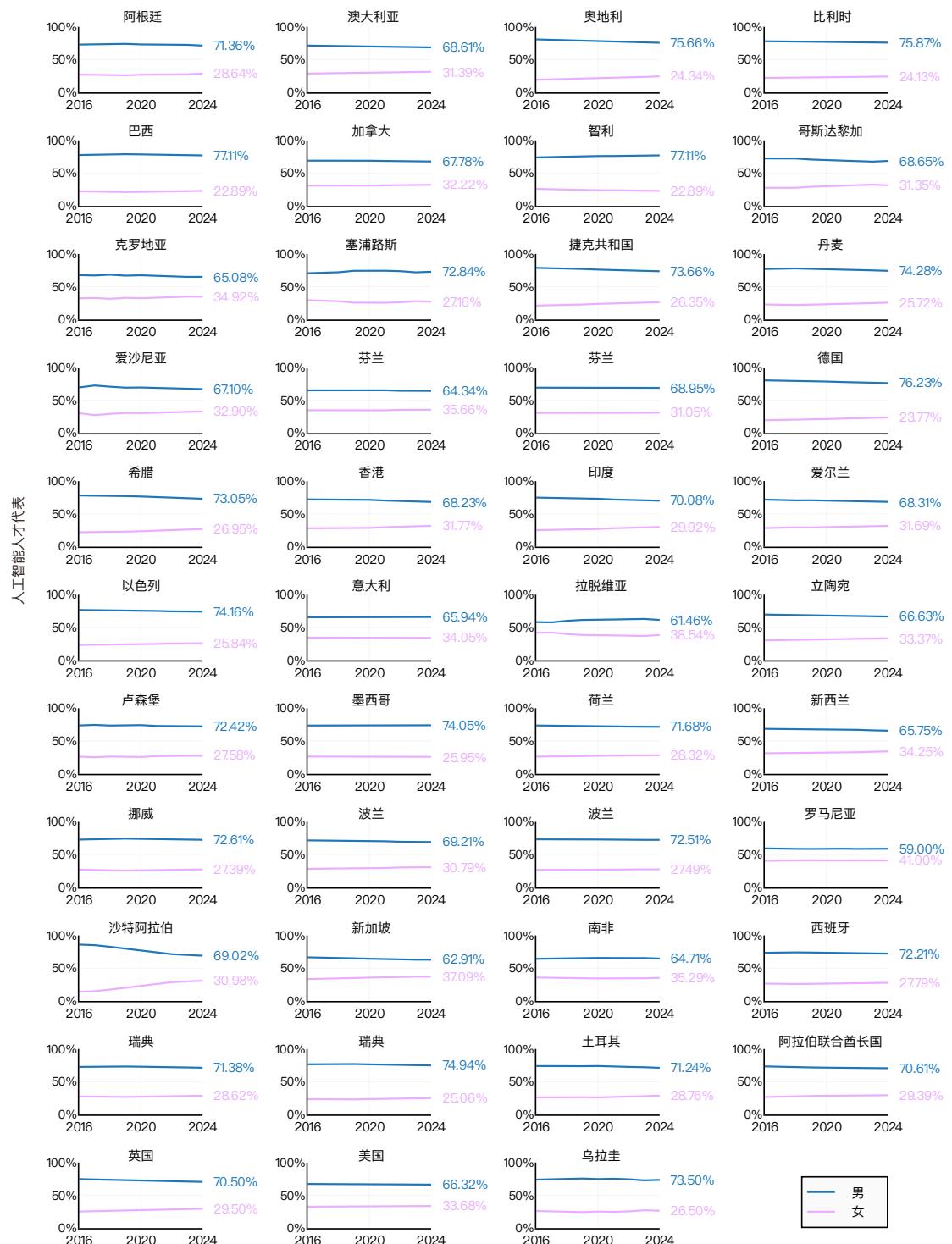


图 4.2.21

第四章：经济

4.2 工作岗位

领英的数据可反映因迁移趋势导致的人工智能人才流动情况。⁶ 净流入是指在给定时间段内到达人数减去离开人数的总和。如果人工智能人才净迁移分数为正数，则表明进入该地区的人才多于离开的人才。负数表示该地区离开的人才多于进入的人才。图 4.2.22 按地理区域显示了每 10,000 名领英会

员中人工智能人才的净迁入情况。人工智能人才人均迁入最多的地区是卢森堡 (8.9)、塞浦路斯 (4.7) 和阿拉伯联合酋长国 (4.1)。

2024 年按地理区域划分领英每万名会员的人工智能人才净迁移

资料来源：领英，2024 | 图表：2025 年人工智能指数报告

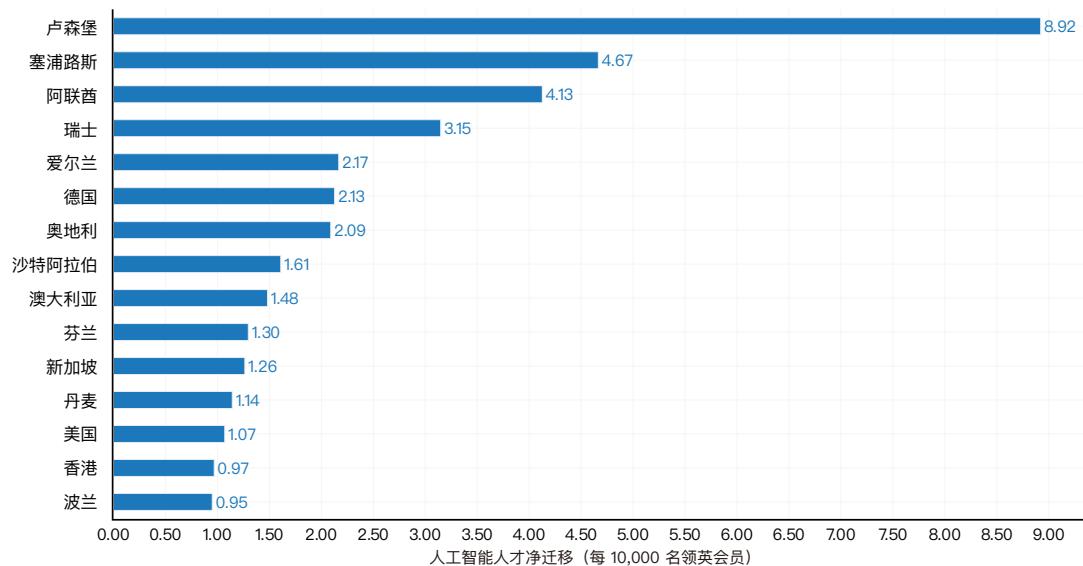


图 4.2.22

图 4.2.23 展示了人工智能人才流动数据随时间的变化。近年来，以色列、荷兰和加拿大等国家的人工智能人才净流动量呈现下降趋势，这表明流向这些国家的人工智能人才数量有

所减少。人才流动不断增加的国家包括阿拉伯联合酋长国、沙特阿拉伯和卢森堡。

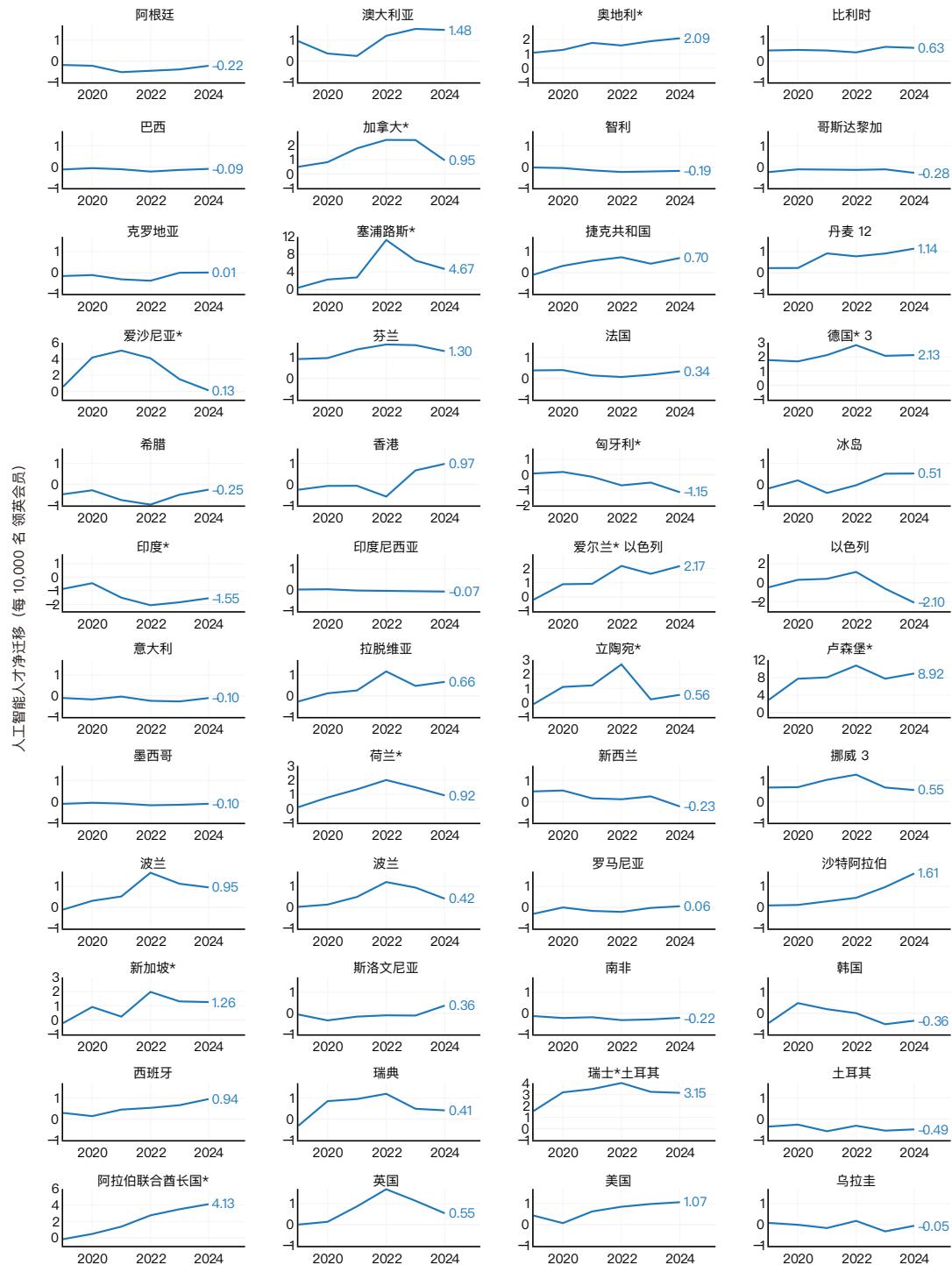
6. 领英会员数量在各国之间存在显著差异，这使得解读会员从一个国家迁移到另一个国家的绝对变化变得困难。为了公平比较各国之间的迁移流量，迁移流量会针对目标国家进行标准化处理。例如，如果 A 国是感兴趣的国家，则所有流入和流出 A 国的绝对净流动（无论来源国和目的国）均根据每年年底 A 国的领英会员人数进行标准化，并乘以 10,000。因此，该指标表示所有其他国家与 A 国之间的人才迁移相对情况。

第四章：经济

4.2 工作岗位

2019–2024 年按地理区域划分的每万名领英会员的人工智能人才净迁移量

资料来源：领英，2024 | 图表：2025 年人工智能指数报告



7、星号表示一个国家的 y 轴标注与其他国家的 y 轴标注的比例不同。

图 4.2.23⁷

重点：

衡量人工智能与经济的融合情况

一项针对超过 400 万次真实世界人工智能交互的分析，为人工智能在各经济领域的融合提供了全面的实证依据。Anthropic 近期的一项研究通过美国劳工部的 O*NET 职业分类框架，对其人工智能模型的使用模式进行了分析，揭示了哪些行业和职能正在利用人工智能技术。具体而言，Anthropic 团队通过分析用户与其 Claude. 人工智能模型的对话，识别出最频繁使用人工智能的任务和职业。

分析表明，虽然所有行业都在一定程度上使用了当前的人工智能技术，但主导领域仍为技术和创意行业。如图 4.2.24 所示，计算机和数学相关职业占据主导地位，占所有人工智能交互的 37.2%。其次是艺术、设计、娱乐、体育和媒体职业，占比 10.3%，而教育指导和图书馆职业也显示出较高的应用率。缓解的责任在于模型提供商还是用户。

Claude 使用数据中的职业分布和美国劳动力分布对比

资料来源：Handa 等，2025 | 图表：2025 年人工智能指数报告

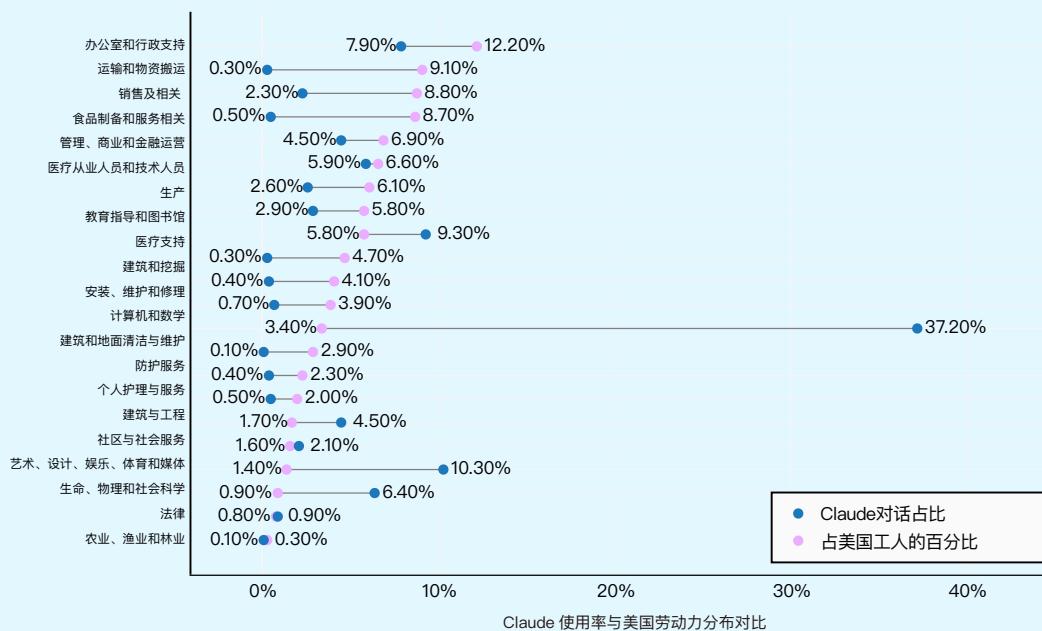


图 4.2.23

重点：

衡量人工智能与经济的融合情况(续)

人工智能的使用模式与工资水平及所需技能之间存在明显关联。图 4.2.25 显示，人工智能使用率在高工资四分位数职业中达到峰值，但在工资两端极值处显著下降。需要

大量准备（通常为学士学位水平）的岗位的人工智能使用率比其基准劳动力代表性高出 50%，而需要少量准备和大量准备的岗位则显示出较低的采用率。

按年薪中位数分类的 Claude 职业使用情况

资料来源：Handa 等，2025| 图表：2025 年人工智能指数报告

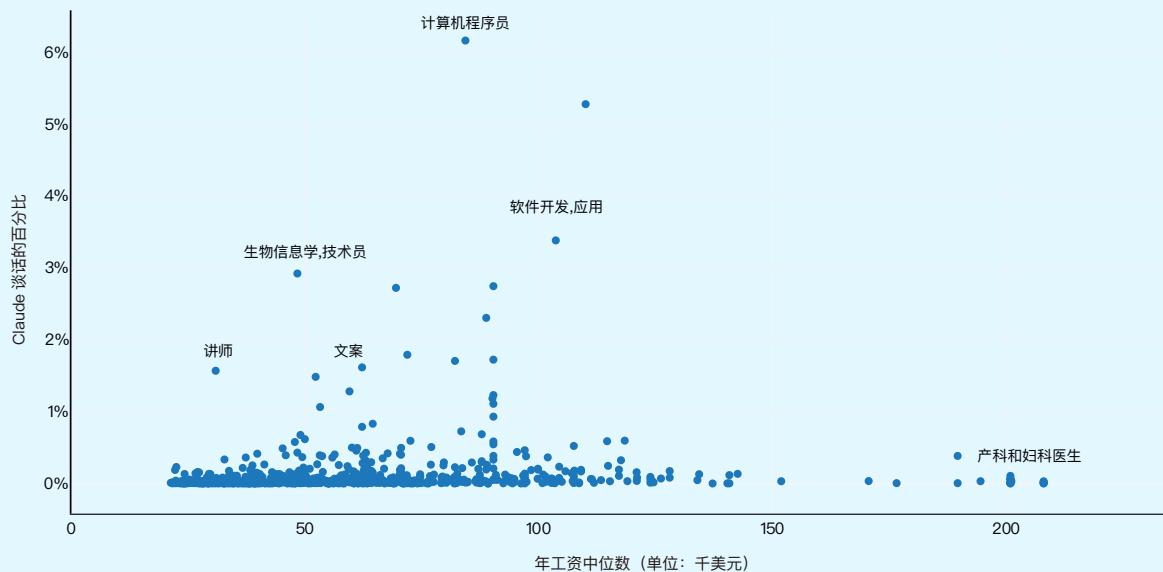


圖 4.2.25

重点：

衡量人工智能与经济的融合情况(续)

Anthropic 的研究发现，约有 36% 的职业在至少四分之一的相关任务中使用了人工智能（图 4.2.26），这表明人工智能在技术领域之外也有很大的渗透。然而，深度整合仍

然很少见：只有大约 4% 的职业在 75% 以上的任务中使用了人工智能，这表明整个职业类别的全面自动化尚未实现。

各组织机构使用人工智能的深度

资料来源：Handa 等，2025

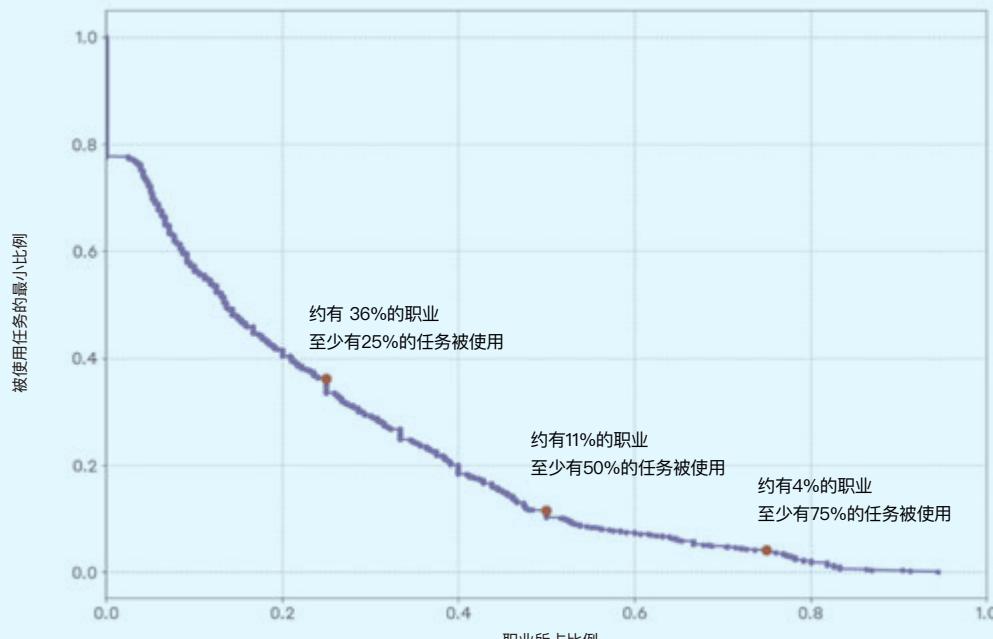


图 4.2.26

重点：

衡量人工智能与经济的融合情况(续)

分析揭示了组织内部如何使用人工智能。如图 4.2.27 所示，57% 的人工智能互动表现出增强模式(增强人类能力)，而 43% 则表现出自动化模式。这种差异表明，当前的

人工智能的应用倾向补充而非取代人类工作者。研究发现，在人工智能互动中，批判性思维和写作等认知技能的存在率很高，而体能和管理技能的存在率则很低(图 4.2.28)。

按任务执行类型划分的 Claude 对话百分比

资料来源：Handa 等，2025 | 图表：2025 年人工智能指数报告

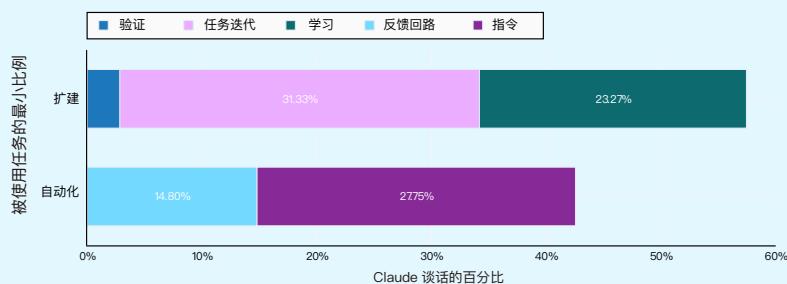


图 4.2.27

Claude 在对话中展示的职业技能分布情况

资料来源：Handa 等，2025

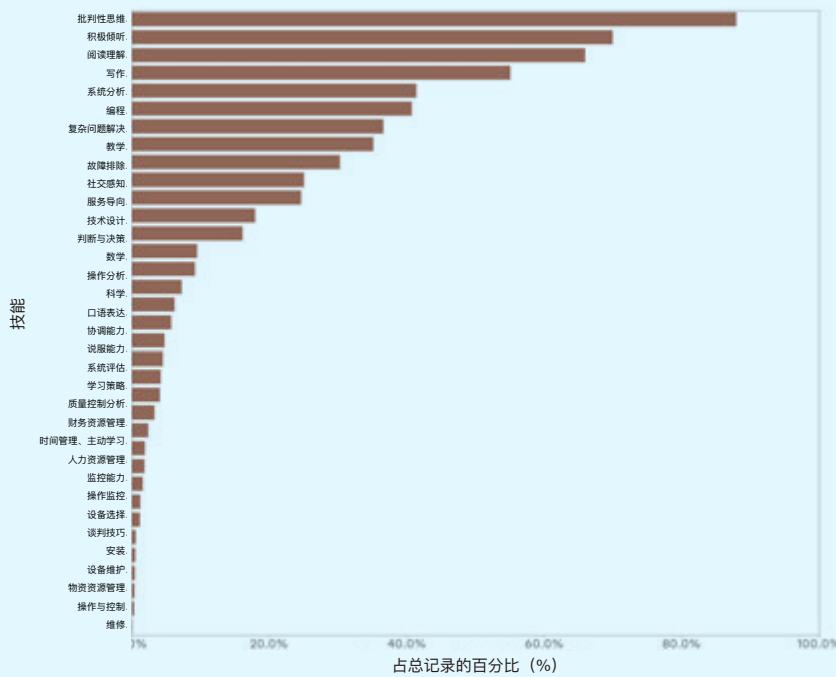


图 4.2.28

第四章：经济

4.3 投资

AI 智能体是设计用于在特定环境中运行以实现目标的自主或半自主系统，是人工智能研究的一个令人兴奋的前沿领域。这些智能体具有广泛的潜在应用，从协助学术研究、安排会议到促进在线购物和度假预订，不一而足。正如许多最近的公司新闻稿所表明的那样，智能体已成为人工智能技术领域越来越关注的话题。

4.3 投资

企业投资

图 4.3.1 展示了 2013 年至 2024 年全球企业人工智能投资的趋势，包括并购、少数股权、私人投资和公开发行。

2024 年，总投资额增至 2523 亿美元，较 2023 年增长 25.5%。其中私人投资增幅最为显著，同比上涨 44.5%，并购交易额增长 12.1%。过去十年间，人工智能相关投资规模已增长近十三倍。

2013–2024 年按投资活动分类的全球企业对人工智能的投资情况

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

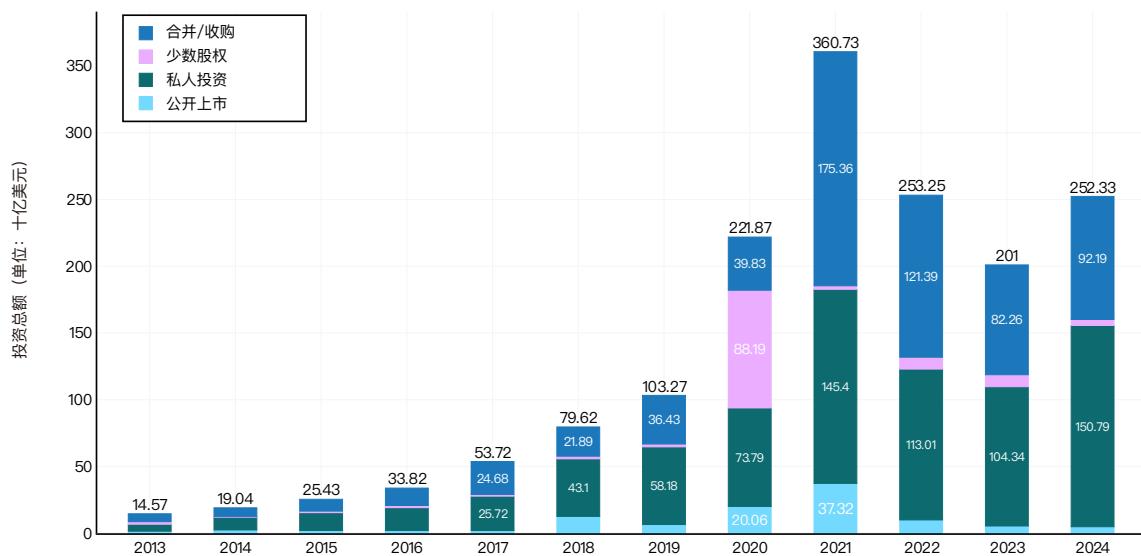


图4.3.1

第四章：经济

4.3 投资

初创企业活动

本节分析了自 2013 年以来获得超过 150 万美元投资的
人工智能初创企业的私人投资趋势。

2013–2024 年全球人工智能私人投资

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

全球趋势

2023 年至 2024 年，全球私人人工智能投资增长了
44.5%，这是自 2021 年以来的首次同比增长（图 4.3.2）。

尽管近期有所波动，但全球私人人工智能投资在过去十年中
仍有大幅增长。

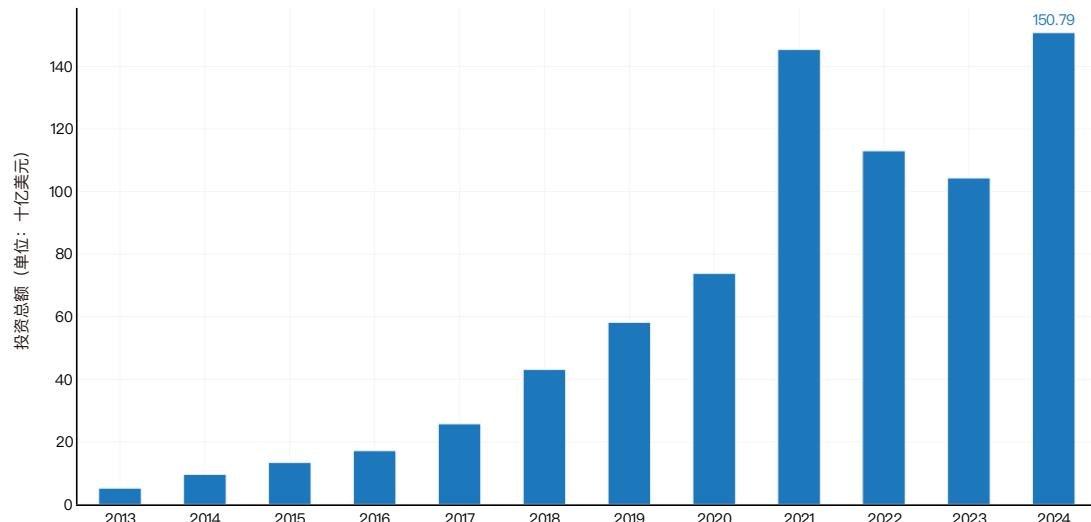


图4.3.2

生成式人工智能的资金投入持续大幅增长（图 4.3.3）。2024 年，该领域吸引投资达 339 亿美元，较 2023 年增长

18.7%，是 2022 年投资额的 8.5 倍以上。此外，2024 年，生成式人工智能占所有人工智能相关私人投资的五分之一以上。

2019–2024 年全球对生成式人工智能的私人投资

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

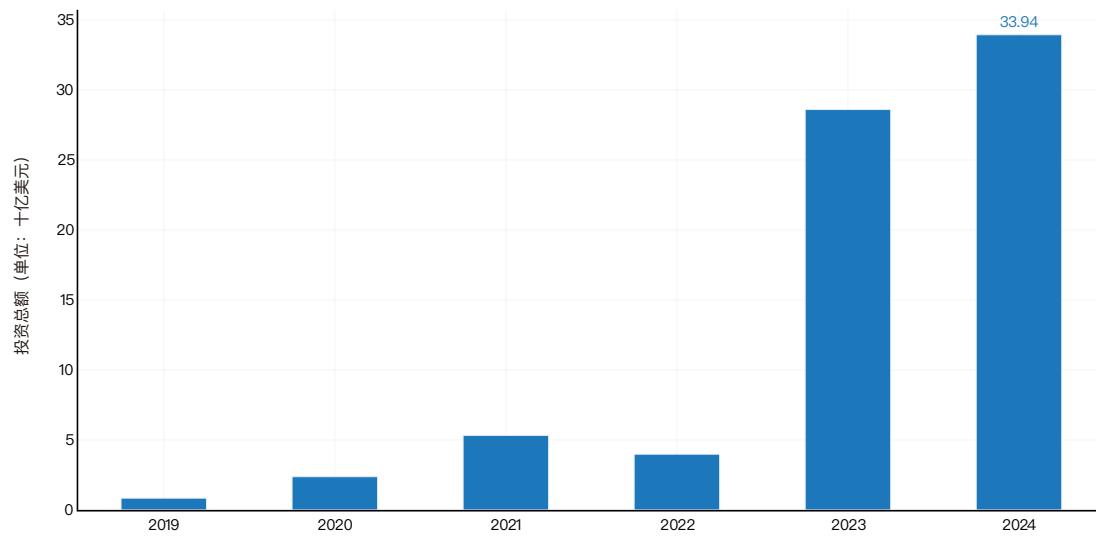


图4.3.3

第四章：经济

4.3 投资

2024 年新获得投资的人工智能公司数量跃升至 2,049 家,比上一年增长了 8.4% (图 4.3.4)。此外,2024 年新获得融

资的生成式人工智能公司数量有所增加,共有 214 家新创公司获得融资,而 2023 年为 179 家,2019 年仅为 31 家(图 4.3.5)。

2013–2024 年全球新获得融资的人工智能公司数量

资料来源: Quid, 2024 | 图表: 2025 年人工智能指数报告

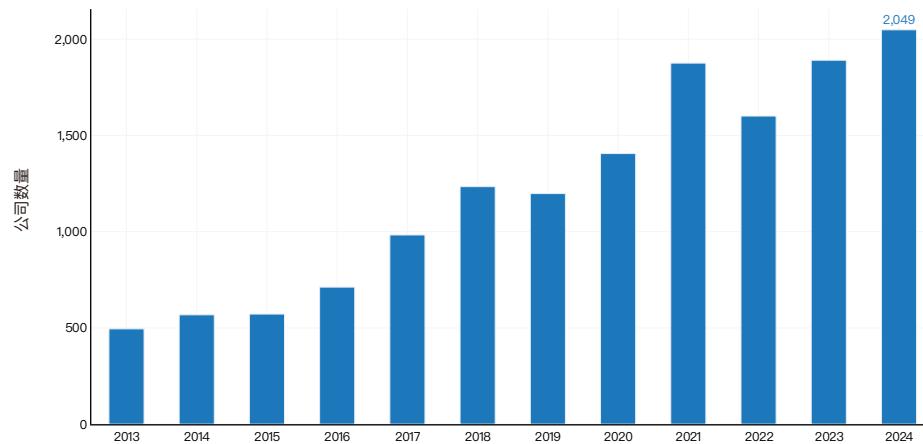


图4.3.4

2019–2024 年全球新获得融资的生成式人工智能公司数量

资料来源: Quid, 2024 | 图表: 2025 年人工智能指数报告

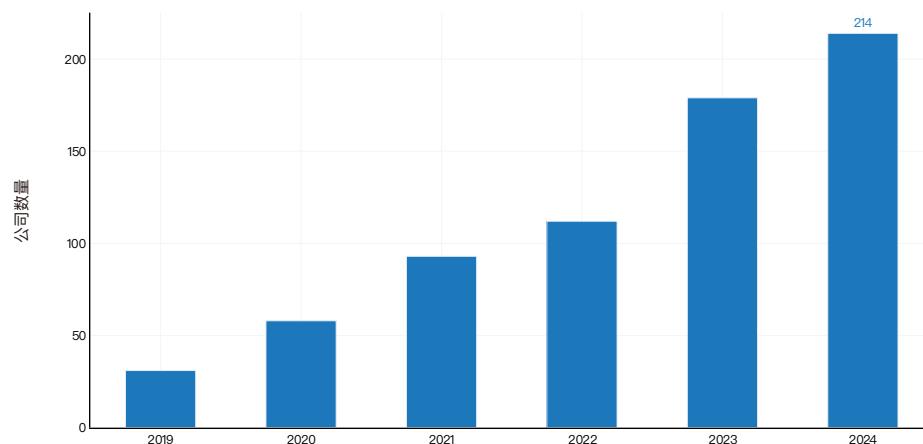


图4.3.5

第四章：经济

4.3 投资

图 4.3.6 显示了人工智能私人投资的平均规模，计算方法是将每年人工智能私人投资总额除以人工智能私人投资事件总数。从 2023 年到 2024 年，平均规模大幅增加，从 3160 万美元增至 4540 万美元。图 4.3.7 展示了按规模分类的人工智

能融资事件。2024 年，人工智能私人投资事件在融资规模超过 1 亿美元的类别中有所增加，而在较小规模的类别中则减少或保持稳定。2024 年，共有 15 起人工智能私人投资事件涉及融资规模超过 10 亿美元。

2013–2024 年全球人工智能私人投资活动的平均规模

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

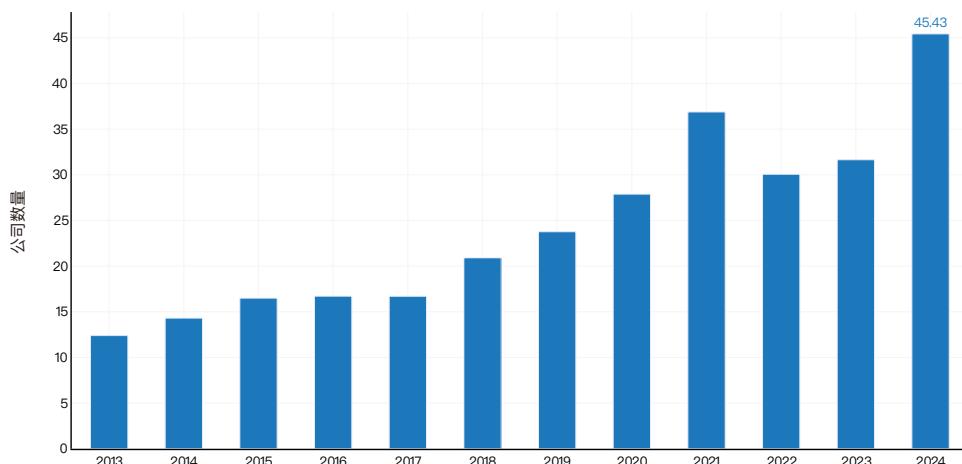


图4.3.6

2023 年与 2024 年按资金规模划分的全球人工智能私人投资活动

资料来源：Quid, 2024 | 表：2025 年人工智能指数报告

资金规模	2023	2024
超过 10 亿	9	15
5 亿 – 10 亿	9	20
1 亿 – 5 亿	134	143
5000 万 – 1 亿	200	196
低于 5000 万	2,945	2,945
未披露	680	207
总价	3,977	3,526

图4.3.7

第四章：经济

4.3 投资

按资金规模进行的地区比较

美国在人工智能领域私人投资总额方面再次位居世界第一。2024 年，美国投入的人工智能领域私人投资总额为 1091 亿美元，是排名第二的中国(93 亿美元)的 11.7 倍，是英国(45

亿美元)的 24.1 倍(图 4.3.8)。2024 年跻身前 15 名的其他值得关注的国家包括瑞典(43 亿美元)、奥地利(15 亿美元)、荷兰(11 亿美元)和意大利(9 亿美元)。

2024 年按地理区域划分的全球人工智能私人投资

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

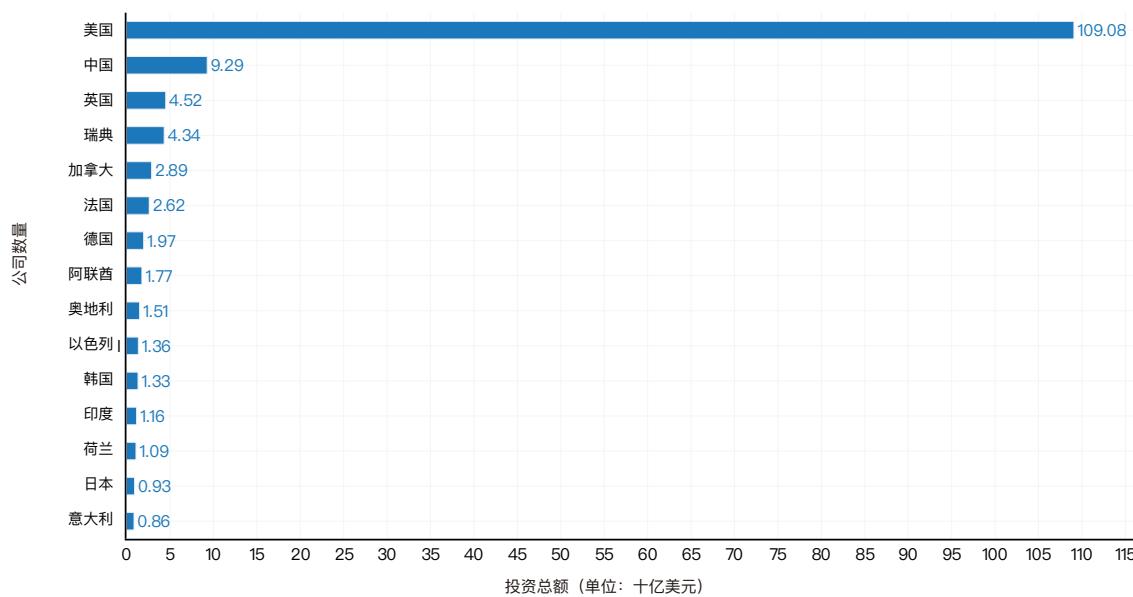


图4.3.8

第四章：经济

4.3 投资

自 2013 年以来，汇总私人人工智能投资数据时，国家排名保持不变：美国以 470.9 亿美元的投资额位居榜首，中国以 119.3 亿美元紧随其后，英国以 28.2 亿美元排名第三（图 4.3.9）。过去十年中吸引大量人工智能投资的其他国家包括以色列

（150 亿美元）、新加坡（73 亿美元）和瑞典（73 亿美元）。过去十年中吸引大量人工智能投资的其他国家包括以色列（150 亿美元）、新加坡（73 亿美元）和瑞典（73 亿美元）。

2013–2024 年按地理区域分类的全球人工智能私人投资（总和）

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

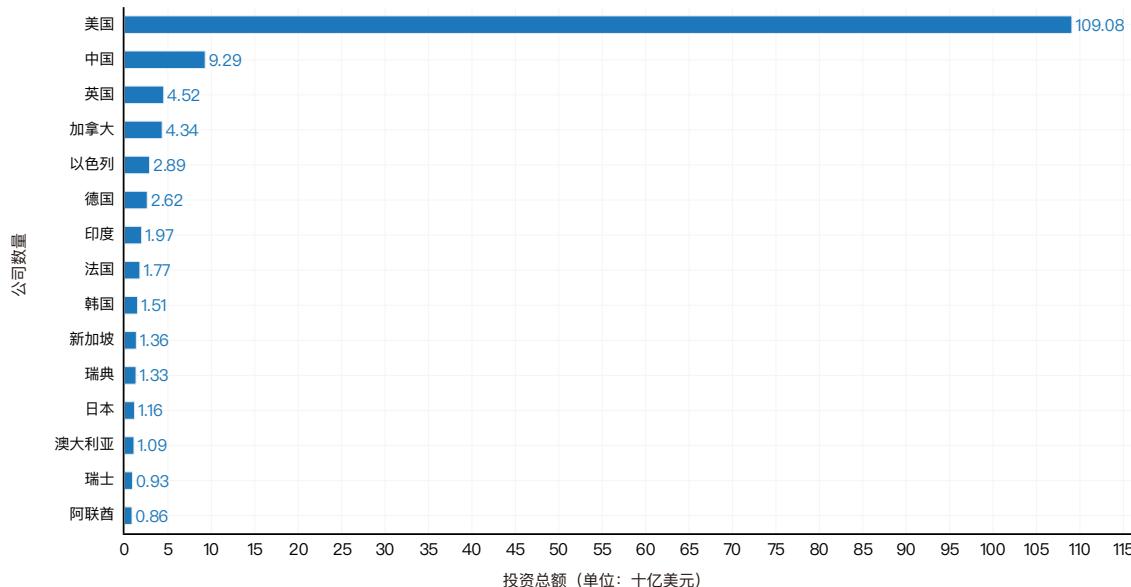


图4.3.9

第四章：经济

4.3 投资

图 4.3.10 显示了各地区人工智能私人投资的变化情况，表明美国与其他地区之间的私人投资差距正在扩大。自 2023 年以来，中国的人工智能私人投资下降了 1.9%，而欧洲则增长了

60%。与此同时，美国在同一时期实现了 50.7% 的显著增长，且自 2022 年以来增长了 78.3%。

2013–2024 年按地理区域划分的全球人工智能私人投资

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

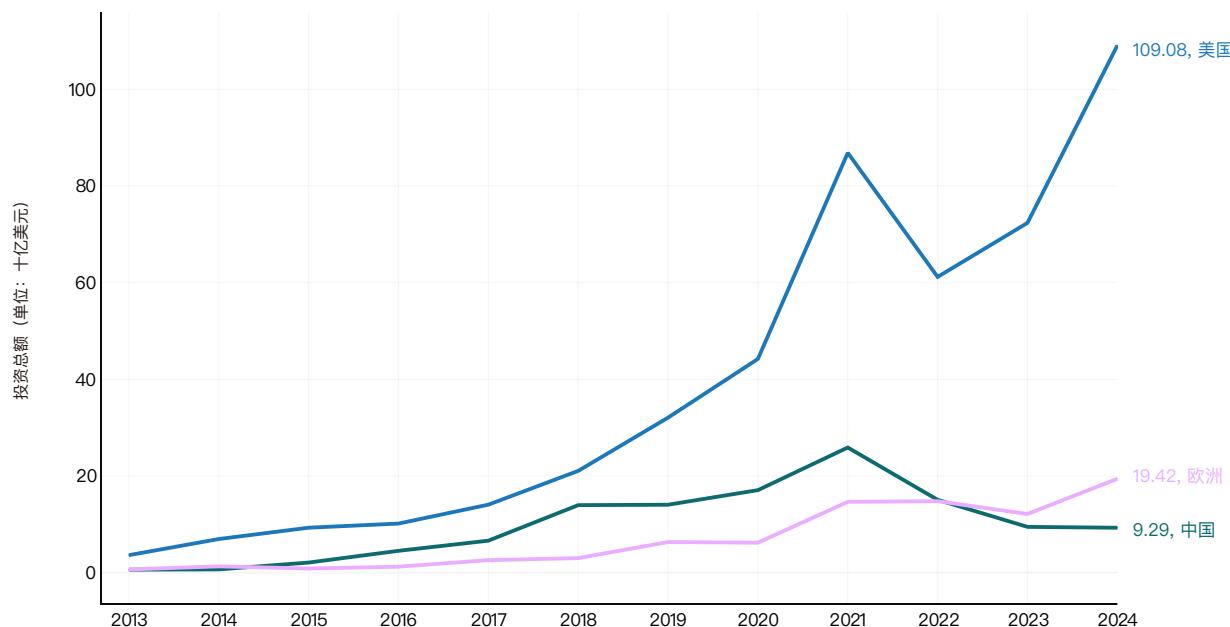


图4.3.10

第四章：经济

4.3 投资

在分析生成式人工智能相关投资时，地区间的人工智能私人投资差距尤为明显。例如，2023 年，美国在生成式人工智能领域的投资额约为 218 亿美元，超过中国和欧洲的投资总额

(图 4.3.11)。到 2024 年，这一差距进一步扩大至 254 亿美元。

到 2024 年，这一差距进一步扩大至 254 亿美元。

2019–2024 年按地理区域划分的全球生成式人工智能私人投资

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

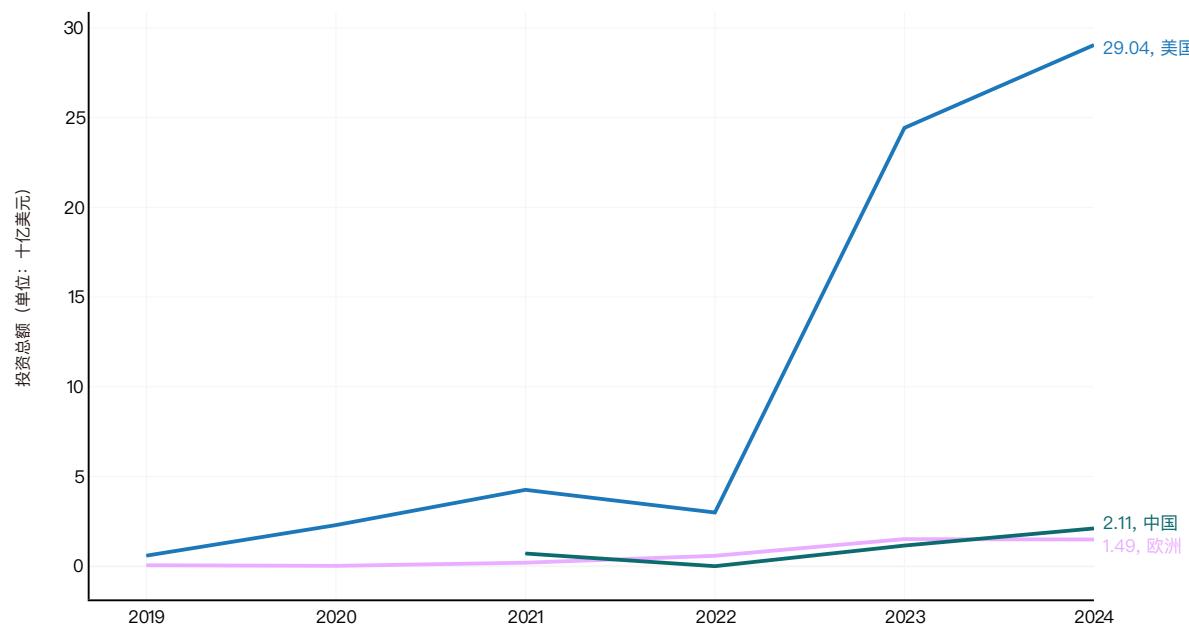


图4.3.11

第四章：经济

4.3 投资

按地区划分的新获融资人工智能公司对比

本节分析了不同地理区域内新获投资的人工智能公司数量。与私人投资趋势一致，美国以 1,073 家新获投人工智能公司位居各地区之首，其次是英国的 116 家，中国为 98 家（图

4.3.12）。美国以 1,073 家新获投人工智能公司位居各地区之首，其次是英国的 116 家，中国为 98 家（图 4.3.12）。

2024 年按地理区域划分的新获融资的人工智能公司数量

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

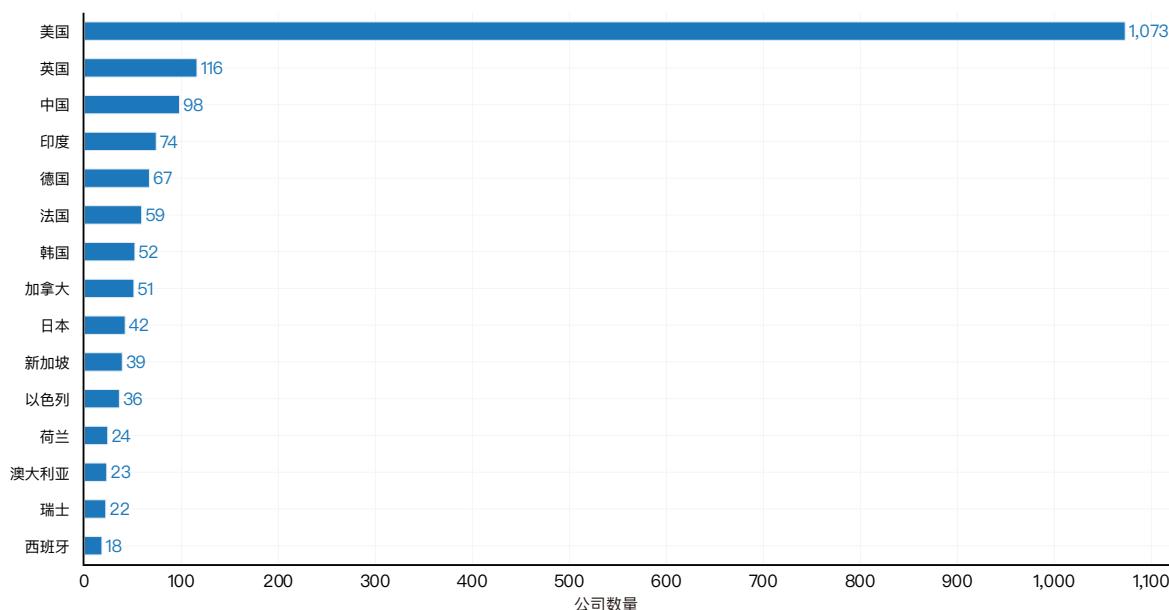


图4.3.12

第四章：经济

4.3 投资

自 2013 年以来的综合数据也呈现出类似的趋势。近十年来，美国新融资的人工智能公司数量约为中国的 4.3 倍，英国的 7.9 倍(图 4.3.13)。

2013–2024 年按地理区域分类的新获融资的人工智能公司数量（总和）

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

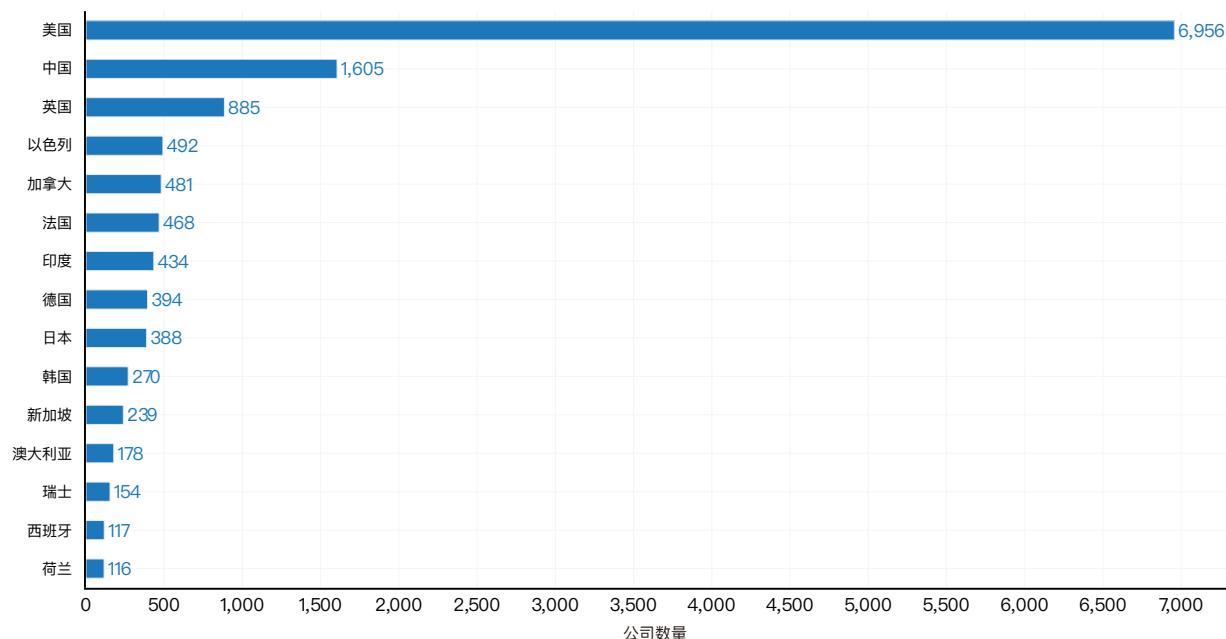


图4.3.13

第四章：经济

4.3 投资

图 4.3.14 展示了特定地理区域内新获融资的人工智能公司数据，凸显了过去十年间美国始终领先于欧洲和中国的长期

趋势。自 2022 年以来，美国与欧洲均出现了新成立人工智能公司数量的显著增长，而中国则连续第二年出现年度下降。

2013–2024 年按地理区域划分的新获融资的人工智能公司数量

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

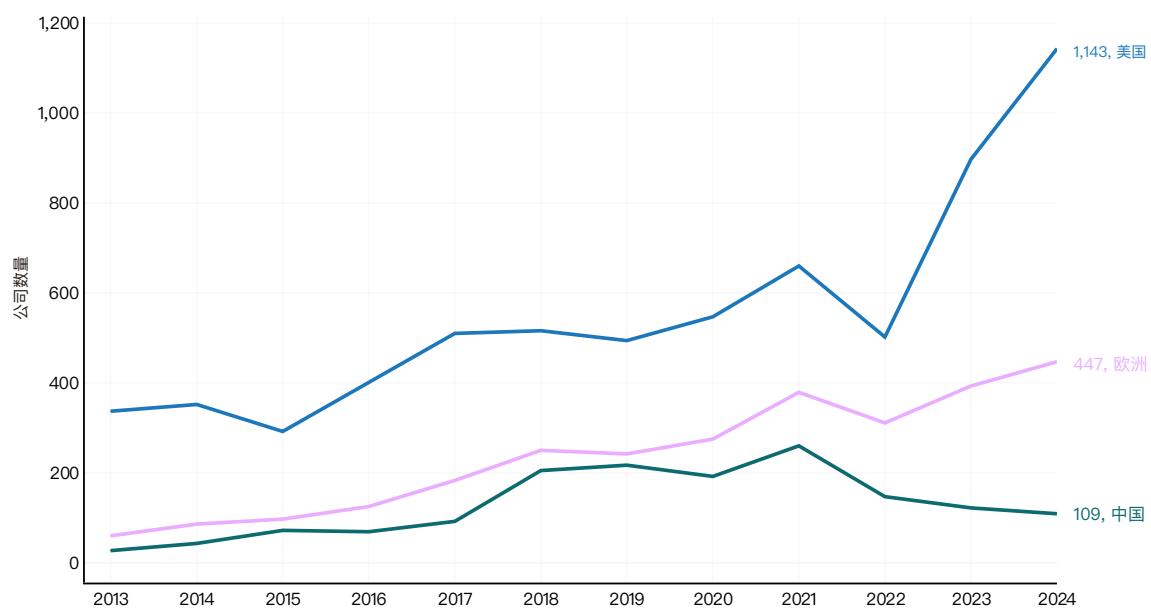


图4.3.14

重点领域分析

Quid 还按重点领域细分了私人人工智能投资。图 4.3.15 比较了 2024 年与 2023 年全球私人人工智能投资按重点领域的分布情况。2024 年吸引最多投资的重点领域依次为人工智能基础设施 / 研究 / 治理 (373 亿美元)；数据管理与处理

(166 亿美元)；以及医疗与健康 (110 亿美元)。人工智能基础设施、研究和治理领域的突出地位反映了对专门开发人工智能应用的公司(如 OpenAI、Anthropic 和 xAI)的大规模投资。

2023 年与 2024 年按重点领域划分的全球人工智能私人投资)

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

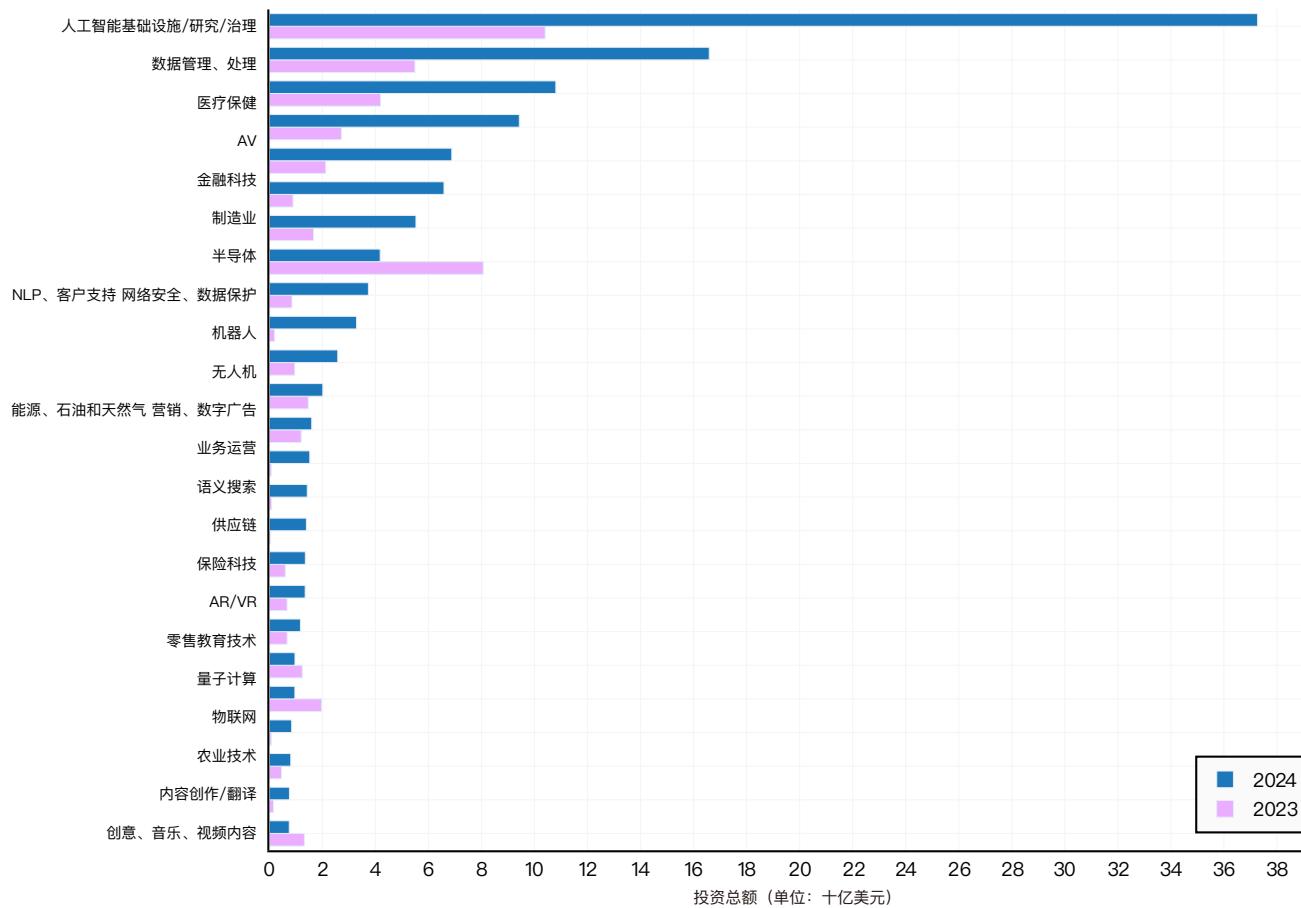


图4.3.15

图 4.3.16 显示了人工智能重点领域投资的长期趋势。如前所述，大多数重点领域的投资在去年都有所增长。对 NLP、客户支持的投资虽然仍很可观，但在 2021 年达到顶峰，此后有所下降。

第四章：经济

4.3 投资

2018–2024 年按重点领域划分的全球人工智能私人投资

资料来源：Quid, 2024 | 图表：2025 年人工智能指数报告

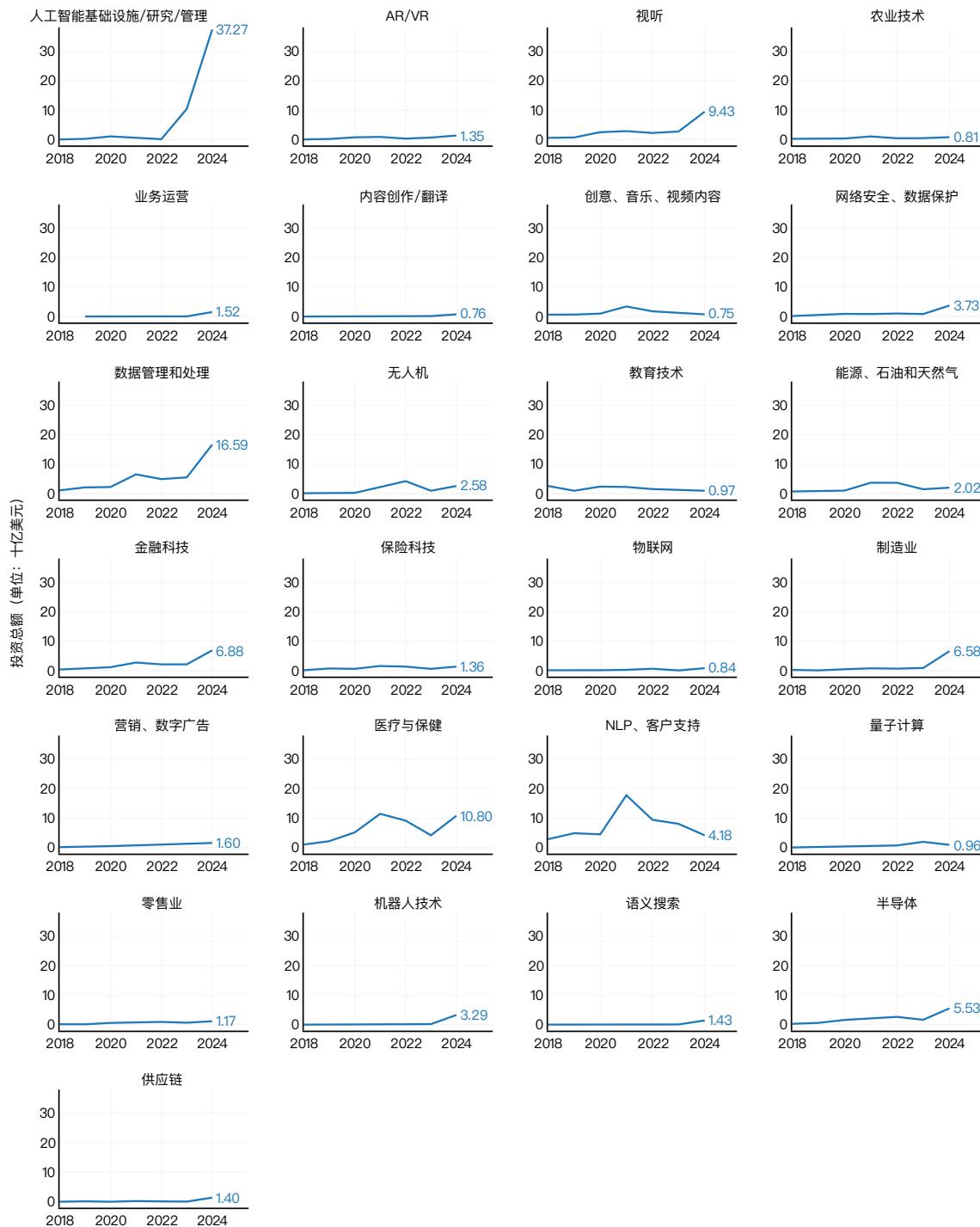


图4.3.16

本章节节探讨企业对人工智能的实际应用，重点介绍行业使用趋势、企业如何整合人工智能、被认为最有益的特定人工智能技术，以及人工智能的使用对财务业绩的影响。

4.4 企业活动

行业使用情况

本章节节结合了麦肯锡关于人工智能现状的出版物中的见解以及以往版本的数据。2024 年麦肯锡分析基于两项调查，调查对象为来自不同地区、行业、公司规模、职能领域和任职年限的 2,854 名受访者。

利用人工智能能力

人工智能的业务应用在 2017 年至 2023 年停滞后大幅增

加。麦肯锡的最新报告显示，78% 的受访者表示，他们的企业已经开始在至少一项业务职能中使用人工智能，与 2023 年的 55% 相比有了显著增长（图 4.4.1）。调查中首次涉及的生成式人工智能的使用增长了一倍多，2024 年有 71% 的受访者表示他们的组织至少在一项业务职能中定期使用该技术，而 2023 年这一比例仅为 33%。

2017–2024 年表示其组织至少在一项职能中使用人工智能的受访者比例

资料来源：麦肯锡公司调查，2024 | 图表：2025 年人工智能指数报告

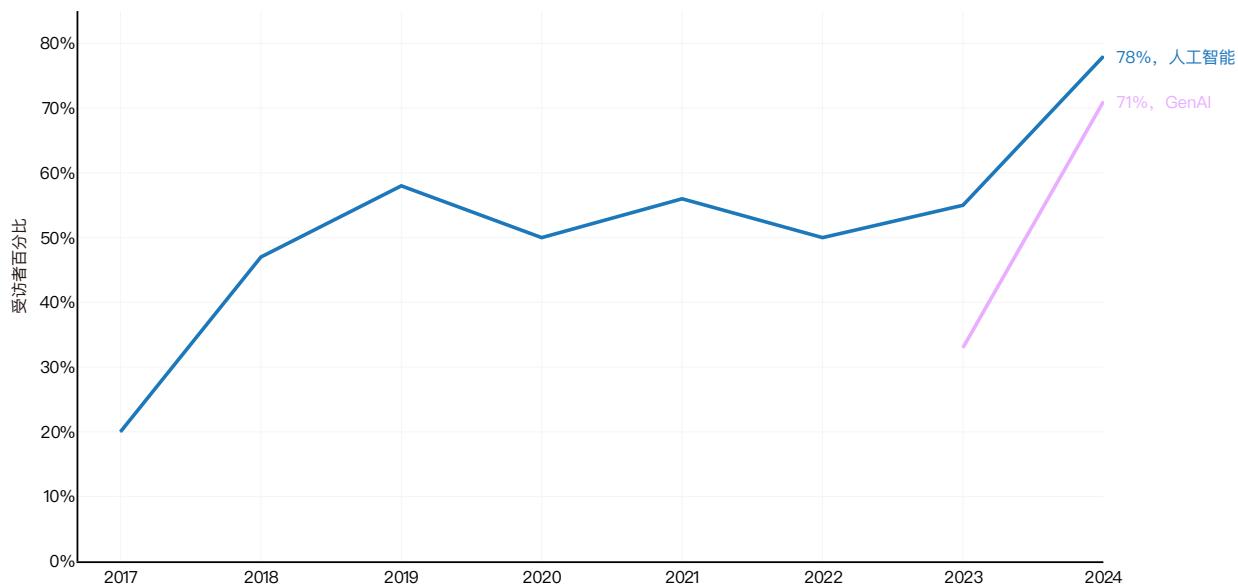


图4.4.1

第四章：经济

4.4 企业活动

图 4.4.2 显示了 2024 年按行业和人工智能功能划分的人工智能使用情况。使用率最高的是科技行业的信息技术 (48%)，其次是科技行业的产品和 / 或服务开发 (47%) 以及科技行业的市场营销和销售 (47%)。

2024 年各行业与职能中的人工智能使用情况

资料来源：麦肯锡公司调查，2024 | 图表：2025 年人工智能指数报告

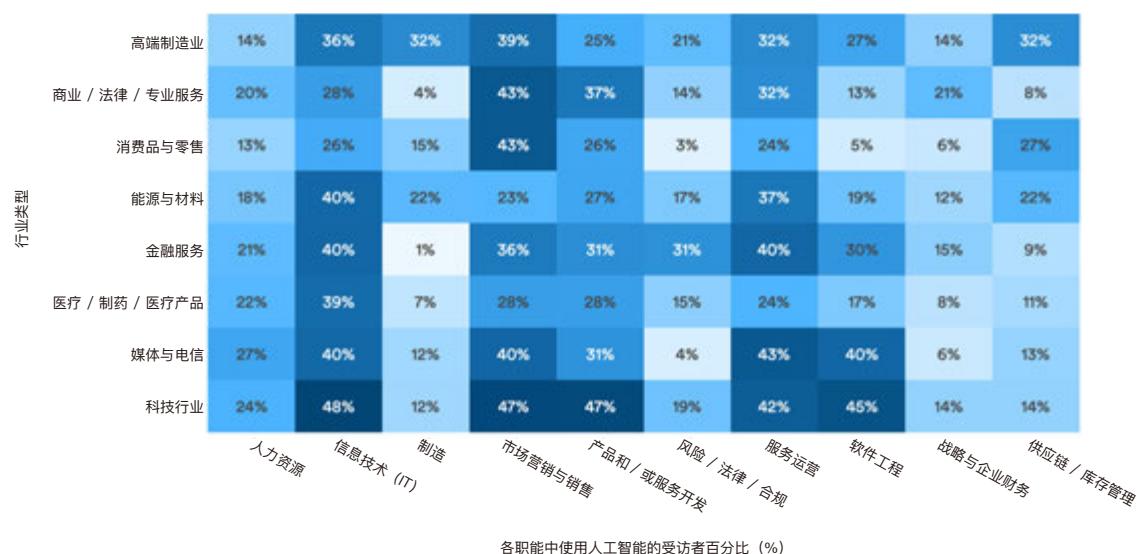


图4.4.2 8

8、其中高端制造业 包括来自先进电子、航空航天和国防、汽车和装配以及半导体等行业的受访者。能源和材料 包括农业、化工、电力和天然气、金属和采矿、石油和天然气以及造纸、林产品和包装等行业的受访者。

第四章：经济

4.4 企业活动

组织在开始使用人工智能后，报告了成本降低和收入增加，但大多数情况下是在较低水平(图 4.4.3)。受访者最常报告人工智能使用带来成本节约的领域是服务运营(49%)、供应链

和库存管理(43%)以及软件工程(41%)。在收入增长方面，人工智能应用受益最显著的职能包括营销与销售(71%)、供应链与库存管理(63%)以及服务运营(57%)。

2024 年按功能划分的人工智能分析成本降低和收入增加情况)

资料来源：麦肯锡公司调查，2024| 图表：2025 年人工智能指数报告

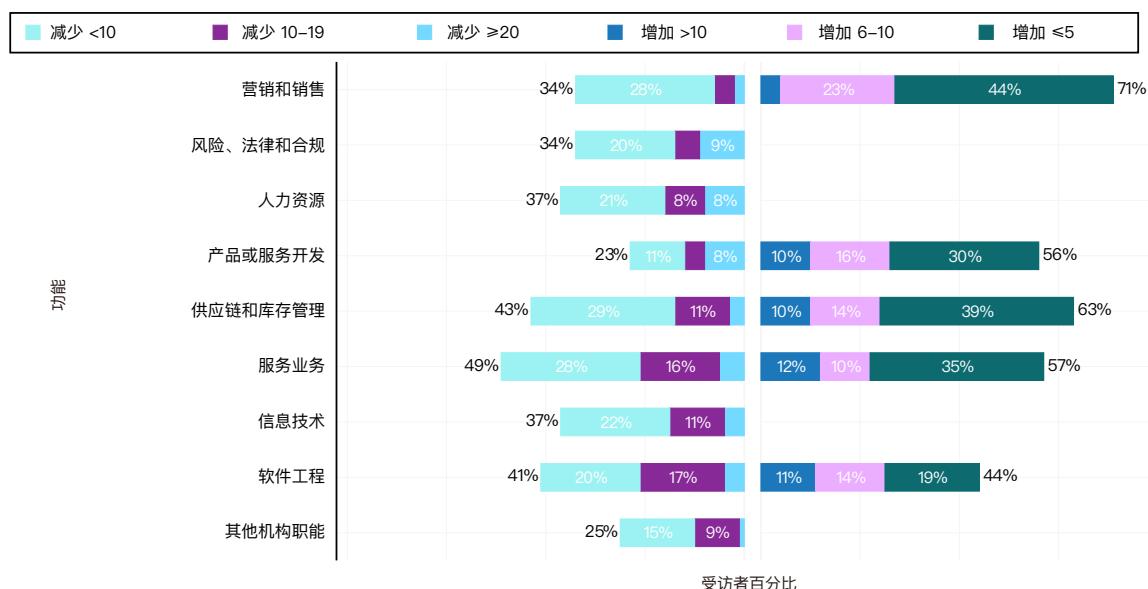


图4.4.3

第四章：经济

4.4 企业活动

图 4.4.4 展示了全球组织按地区划分的人工智能使用情况。2024 年，所有地区受访者均报告人工智能使用率较 2023 年有所提升。人工智能使用率的年增长率中，大中华地区增长最为显著，组织报告的使用率增长了 27 个百分点。北美地区仍

保持人工智能使用的领先地位（82%），但优势仅为微弱。欧洲地区的人工智能使用率也显著提升，较 2023 年增长 23 个百分点，达到 80%。

2023 年与 2024 年全球组织使用人工智能的情况对比

资料来源：麦肯锡公司调查，2024 | 图表：2025 年人工智能指数报告

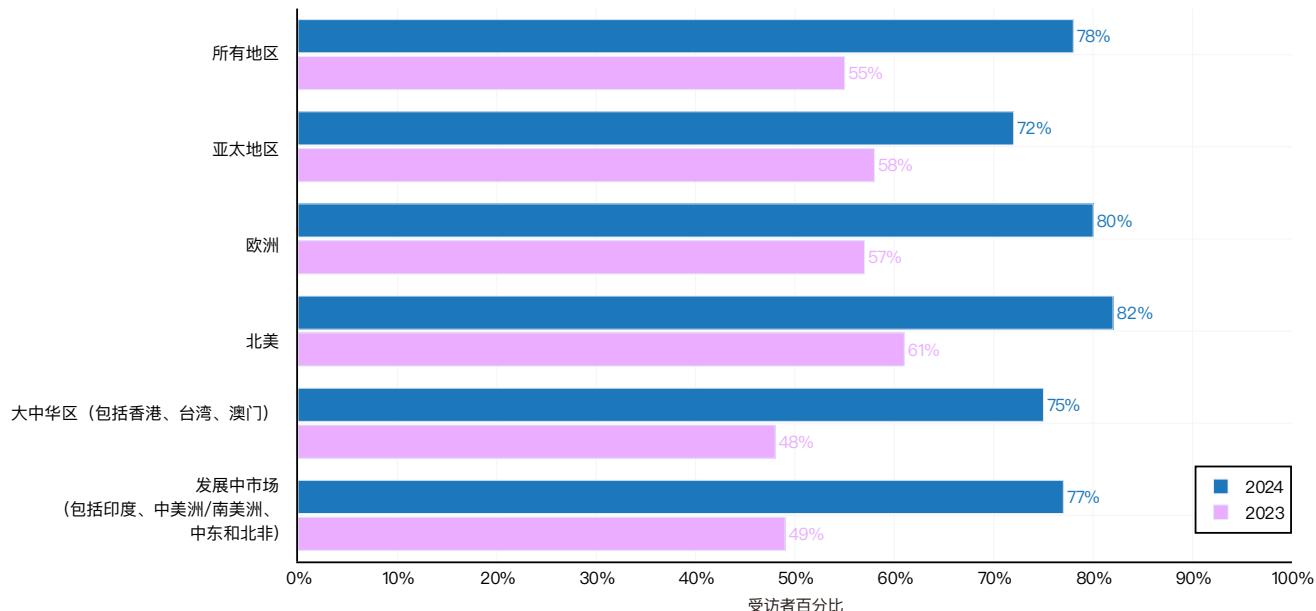


图4.4.4

第四章：经济

4.4 企业活动

生成式人工智能功能部署

企业如何生成式人工智能功能部署？图 4.4.5 显示了报告将生成式人工智能用于特定功能的受访者在受访者总数中所占的比例。受访者有可能表示他们为多种目的部署了人工智能。最常见的应用是营销战略内容支持 (27%)，其次是知识管理 (19%)、个性化 (19%) 和设计开发 (14%)。所报告的主要

用例大多属于营销和销售职能部门。一项针对发达市场高管的补充调查发现，只有 1% 的公司将其生成式人工智能的推广描述为 "mature (成熟)"。总体而言，大多数公司仍处于从人工智能中大规模获取价值的早期阶段。

2024 年按功能划分的最常见生成式人工智能用例

资料来源：麦肯锡公司调查，2024 | 图表：2025 年人工智能指数报告

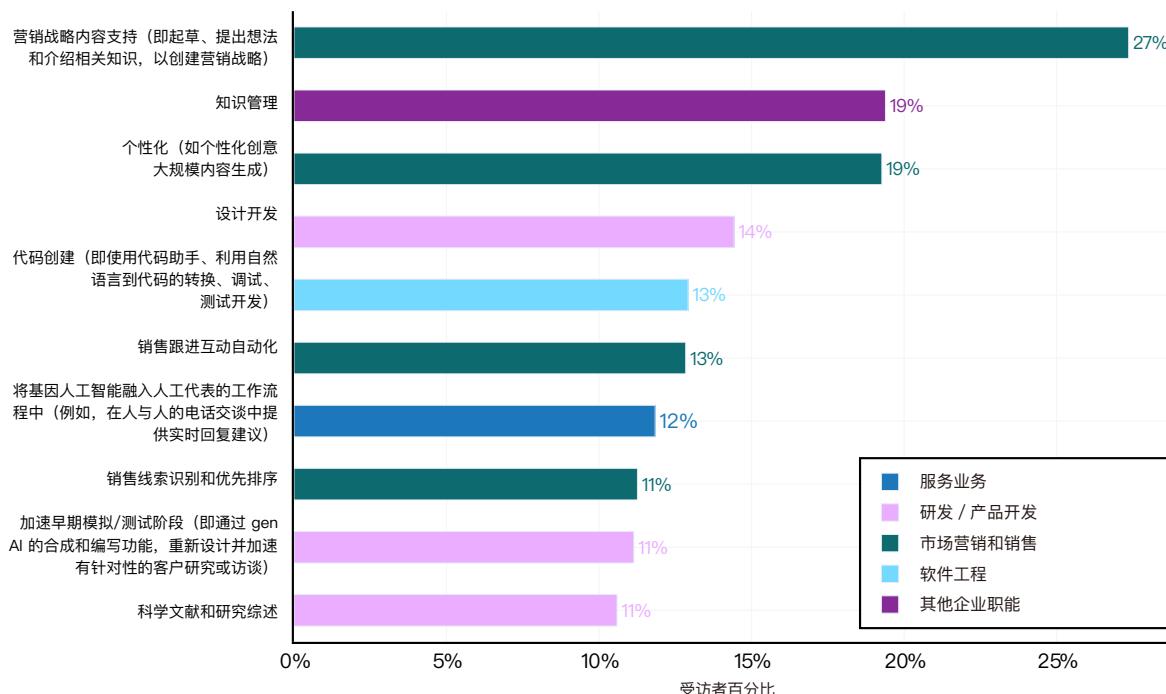


图4.4.5

第四章：经济

4.4 企业活动

图 4.4.6 展示了受访者中报告其组织在各业务职能中使用生成式人工智能导致成本下降和收入增加的比例。总体而言，受访者在使用生成式人工智能后，各业务职能均报告了成本降低和收入增加，其中大多数为低水平。受访者最常报告成本节约的领域包括供应链和库存管理（61%）、服务运营

（58%），以及人力资源和战略与企业财务（56%）。在收入增长方面，最常报告生成式人工智能带来益处的职能包括战略与企业财务（70%）、供应链和库存管理（67%），以及营销与销售（66%）。

2024 年按功能划分的生成式人工智能使用成本降低和收入增加情况

资料来源：麦肯锡公司调查，2024 | 图表：2025 年人工智能指数报告

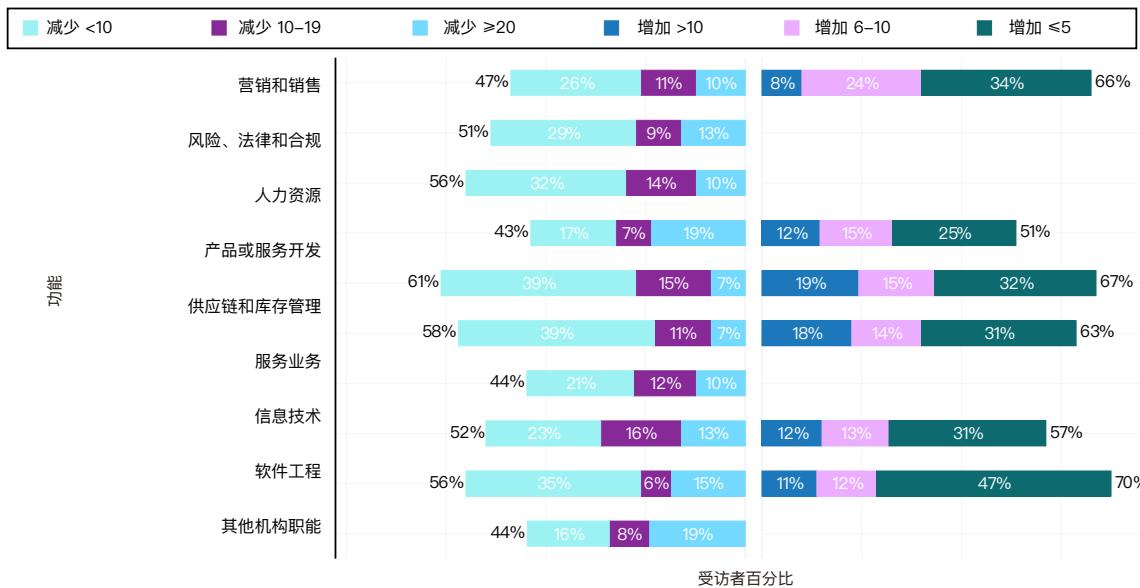


图4.4.6

第四章：经济

4.4 企业活动

图 4.4.7 展示了全球不同地区企业生成式人工智能使用情况的差异。在所有地区，2024 年至少在某一业务功能中使用生成式人工智能的企业比例达到 71%，较 2023 年的 33% 翻了一番多。这一比例仅比报告使用任何形式人工智能的比例(78%)低 7 个百分点。如图 4.4.1 所示。人工智能总体使

用率与生成式人工智能使用率之间的差距从 2023 年的 22 个百分点大幅缩小到 2024 年的 7 个百分点，表明生成式人工智能功能的使用正在加速。北美(74%)、欧洲(73%)和中国大陆(73%)在生成式人工智能的使用方面处于领先地位。

2024 年按功能划分的生成式人工智能使用成本降低和收入增加情况

资料来源：麦肯锡公司调查，2024 | 图表：2025 年人工智能指数报告

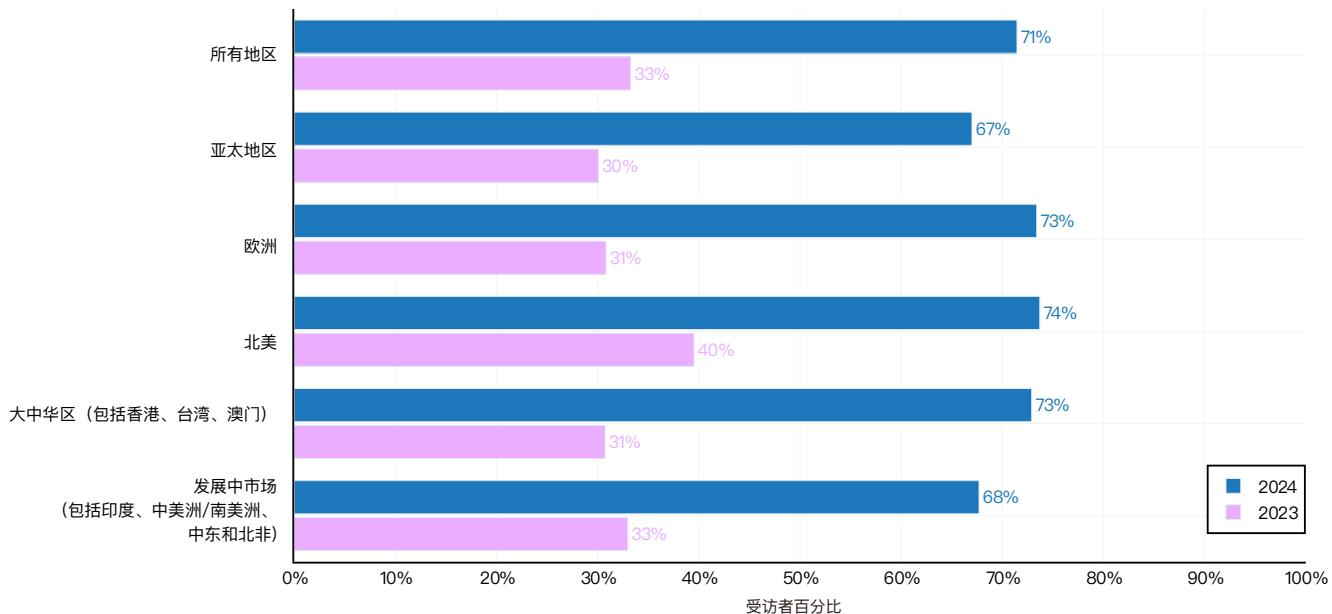


图4.4.7⁹

9、该图突出显示了人工智能在至少一种业务功能中的应用。

人工智能对劳动力的影响

过去六年来，人工智能日益融入经济，引发了人们对其生产效率潜力的浓厚兴趣。虽然早期采用人工智能显示出前景，但量化人工智能的影响仍然具有挑战性，直到 2023 年才出现了第一波严谨的研究。2024 年，大量实证研究确立了人工智能在多个领域和背景下对工作场所产生影响的清晰模式。本节分析了来自五项主要学术研究的生产力影响数据，这些研究共同代表了对人工智能工作场所影响的首次大规模实证调查。这些研究涵盖了多个行业和领域的 20 多万名专业人士，揭示了 10% 到 45% 不等的生产力提升，其中技术、客户支持和创造性任务的影响尤为显著。这些研究采用了多种方法，包括自然实验、随机对照试验和大规模调查，以衡量人工智能在不同组织环境中的影响。

生产力趋势

2023 年 4 月，Erik Brynjolfsson、Danielle Li 和 Daniel Rock 发表了关于人工智能（尤其是生成式人工智能）对生产效率影响的研究。这也是这个领域最具代表性的研究之一。¹⁰ 该研究分析了 5,179 名客服代表的数据，考察了生成型人工智能驱动的对话助手的分阶段引入情况。研究人员发现，人工智能的采用使每小时解决的问题数量增加了 14.2%（图 4.4.8）。此外，研究还发现，生产效率提升在人工智能引入后迅速显现，且接触人工智能的员工在人工智能系统故障期间仍保持更高效率。

最近发布的其他研究也证实了 Brynjolfsson 的发现。微软公司的一项工作场所研究确定了常见工作场所任务的生产效率提高基准，其中文档编辑提高了 10–13%，电子邮件处理时间减少了 11%。专业角色的收益更高，安全专业人员的完成时间缩短了 23%，准确率提高了 7%；销售团队的响应时间缩短了

39%，准确率提高了 25%。

在科学研究方面，Aiden Toner-Rodgers 的研究对 1018 名科学家进行了调查，结果发现，与未使用人工智能的科学家相比，使用人工智能的科学家的材料发现率提高了 44.1%，专利申请率提高了 39.4%，产品原型提高了 17.2%（图 4.4.9）

人工智能对客服代表的影响

资料来源：Brynjolfsson 等，2023 | 图表：2024 年人工智能指数报告

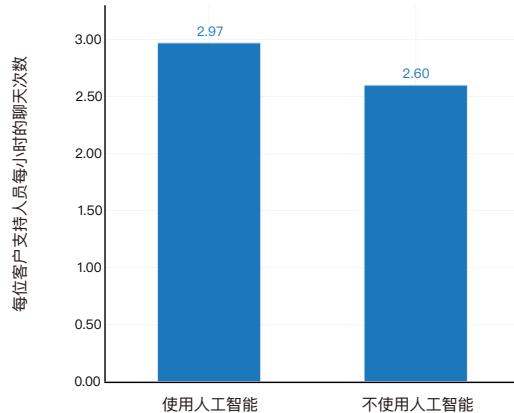


图4.4.8

人工智能对科学创新的影响

资料来源：Toner-Rodgers 等，2025 | 图表：2025 年人工智能指数报告

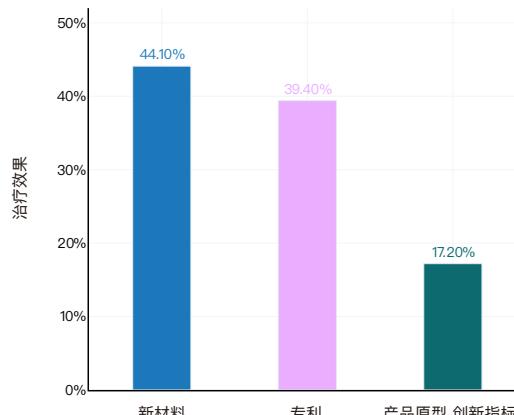


图4.4.9

10、该论文于 2023 年作为 NBER 工作文件 31161 发表，随后于 2025 年发表在《经济学季刊》上。

第四章：经济

4.4 企业活动

在软件开发领域，两项重要研究为人工智能的影响提供了补充证据。一项有 4867 名开发人员参加的现场实验发现，人工智能辅助平均提高了 26.08% 的任务完成率。另一项以 187489 名开发人员为对象的自然实验强化了这一发现；该实验表明，核心编程活动增加了 12.4%，而项目管理任务花费的时间减少了 24.9%。

均衡效应

多项研究揭示了一个一致规律：人工智能对职场绩效具有均衡化效应（图 4.4.10）。最新软件开发领域研究表明，初级开发者的生产效率提升了 21%–40%，而资深开发者的增幅相对有限，仅为 7%–16%。该模式在其他独立研究中得到验证——低技能编程人员的效率提升幅度（14%–27%）显著高于高技能人员（5%–10%）。

此外，他们的分析显示，人工智能使新技术的探索增加了 21.8%，并为每位开发人员每年带来了平均 1,683 美元的潜在薪资增长，这表明人工智能工具不仅提升了生产效率，还积极促进了技能发展。这项研究支持了 2023 年和 2024 年的早期研究结果，即人工智能驱动的生产效率提升因员工的初始技能水平而异。

然而，一些研究表明，人工智能的影响可能是反方向的。Toner-Rodgers 公司的一项研究，发现，虽然表现高效的科学家的产出几乎翻了一番，但排名后三分之一的科学家从人工智能的引入中获益甚微。研究进一步强调，影响人工智能影响的关键因素不是先前的成就，而是有效评估人工智能生成的建议的能力。这表明，对于能够有效利用人工智能工具的人来说，无论经验水平如何，人工智能工具都能发挥强大的放大作用。了解人工智能如何影响不同任务中的不同工作人员，将是当前研究的一个重要重点。

人工智能的生产效率均衡效应

研究	任务	低技能工人生产率提高	高技能工人的生产率提高
Brynjolfsson 等, 2023	客户支持	34%	与零无异
Dell'Acqua 等, 2023	咨询	42.96%	16.5%
Cui 等, 2024	软件工程	21–40%	7–16%
Hoffman 等, 2024	软件工程	12–27%	5–10%

图4.4.10

应用和整合

研究表明，生产效率的提升与人工智能的全面整合及系统化实施呈现显著相关性。罗马尼亚研究人员针对 233 名员工的调查显示，在人工智能高度整合的企业中，生产效率显著提升的概率高达 72%，而整合程度较低的企业这一概率仅为 3.4%。分析数据显示，受访者的生产效率提升呈现明显的梯度分布：46.8% 的受访者实现 0–20% 的提升，26.2% 达到 20–40% 的增幅，18.4% 获得 40–60% 的改善。另有少数群体取得更大幅度的提升，其中 7.7% 的受访者实现 60–80% 的增长，0.9% 达到 80–100% 的显著提升（图 4.4.11）。

对劳动力的影响

人工智能工具的引入显著改变了任务分配模式与团队组织结构。微软职场研究显示，人工智能与自动化技术使员工感知心智负荷降低 45%，在其认知负荷量表中，得分从基准值 55 分降至 30 分。同时，非英语母语者的工作准确率差距缩小 84.6%，专业报告的关键信息涵盖量提升 49%。这些改进在深度用户群体中表现尤为突出，该群体每周至少使用人工智能工具数次，其中 29% 的成员每日可节省 30 分钟以上工作时间。哈佛商学院研究证实，人工智能技术的应用大幅降低了协作成本。数据显示，项目平均所需团队成员数量减少 79.3%。这一发现表明，人工智能不仅优化了个人工作效能，更重塑了团队协作的基本范式。

这些变化正在以根本性方式重塑专业角色。Toner-Rodgers 的研究发现，科学家们的工作时间分配发生了剧烈转变：创意生成所占工作时间比例从 39% 降至 16%，而判断性任务所占比例则从 23% 升至 40%。与以往的技术进步一样，关于人工智能的争论往往围绕着自动化与增强之间，即人工智能会取代工作还是增强人类的工作。虽然关于人工智能驱动的劳动力变化的具体数据仍然有限，但研究正在揭示人们如何看待其对就业的影响。

使用人工智能提高生产效率的分布情况

资料来源：Necula 等，2024 | 图表：2025 年人工智能指数报告

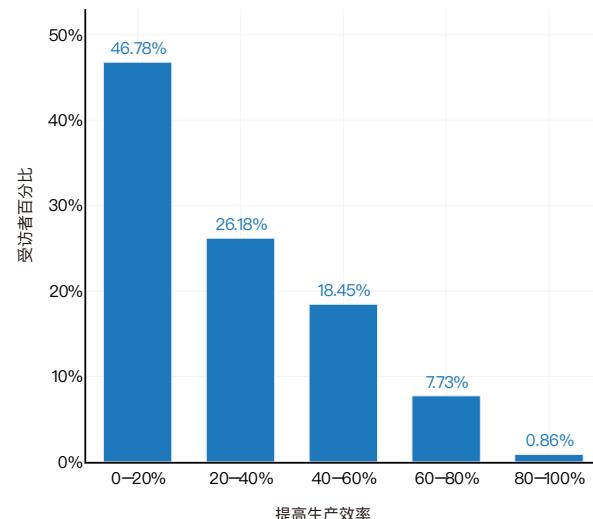


图4.4.11

罗马尼亚的调查数据表明，人们对人工智能在劳动力规模上的影响持多样化预期，其中 43% 的组织预计劳动力规模将减少，30% 预计变化不大，15% 预计将增加，12% 对长期影响持不确定态度。麦肯锡对高管的调查显示，31% 的受访者预计人工智能将减少劳动力规模，而仅有 19% 预计将增加（图 4.4.12）。尽管有人声称生成式人工智能将提高软件工程师的生产效率，但调查结果表明软件工程师的人数预计将增加，与杰文斯悖论一致。值得注意的是，预测劳动力减少的比例比去年有所下降，这表明企业领导人越来越不相信人工智能会减少组织劳动力（图 4.4.13）。

第四章：经济

4.4 企业活动

未来3年（2024年）生成式人工智能对企业员工队伍的预期影响

资料来源：麦肯锡公司调查，2024 | 图表：2025年人工智能指数报告

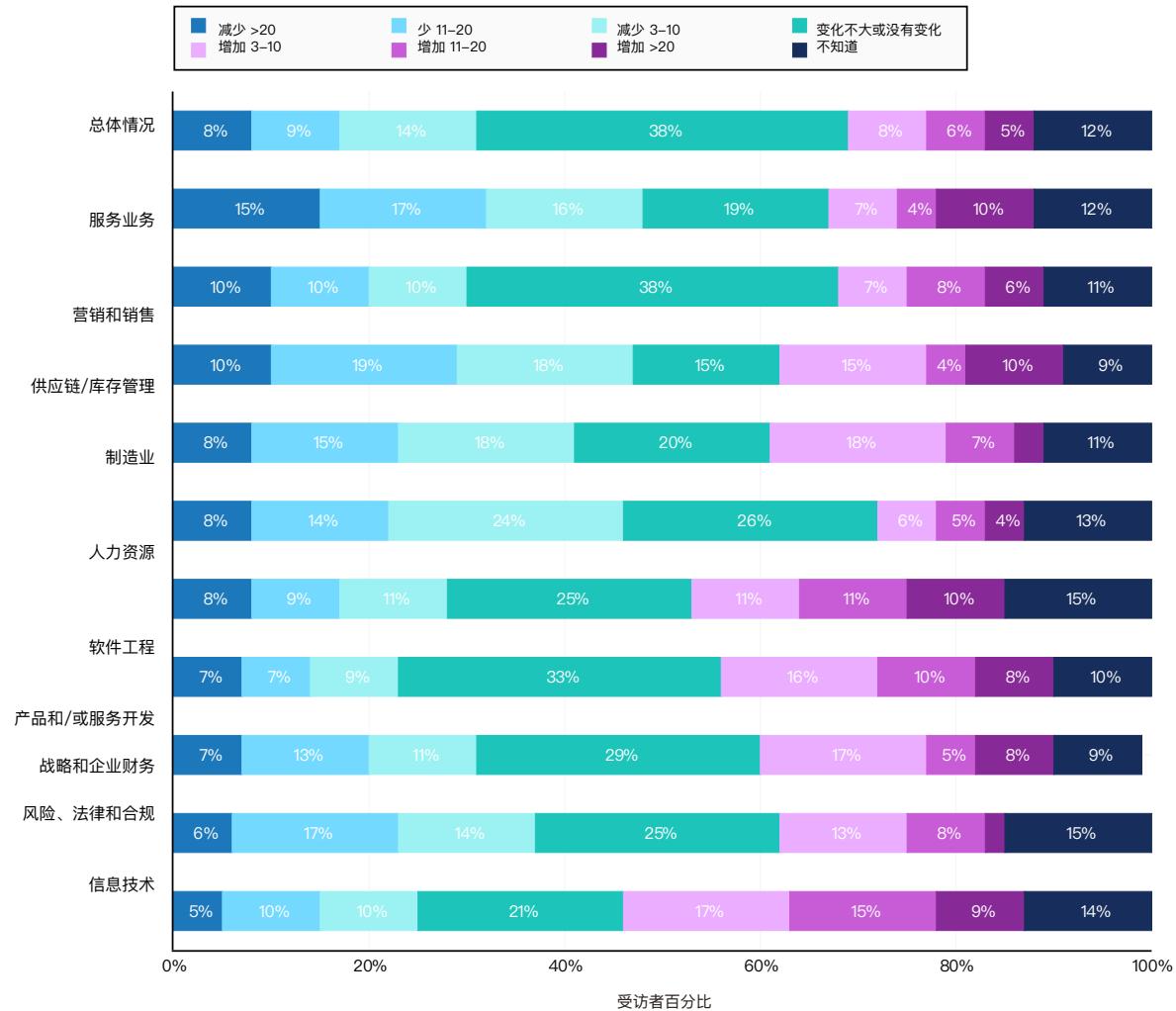


图4.4.12

第四章：经济

4.4 企业活动

未来 3 年（2023 年与 2024 年）人工智能对企业员工队伍预期影响的对比

资料来源：麦肯锡公司调查，2023–2024| 图表：2025 年人工智能指数报告

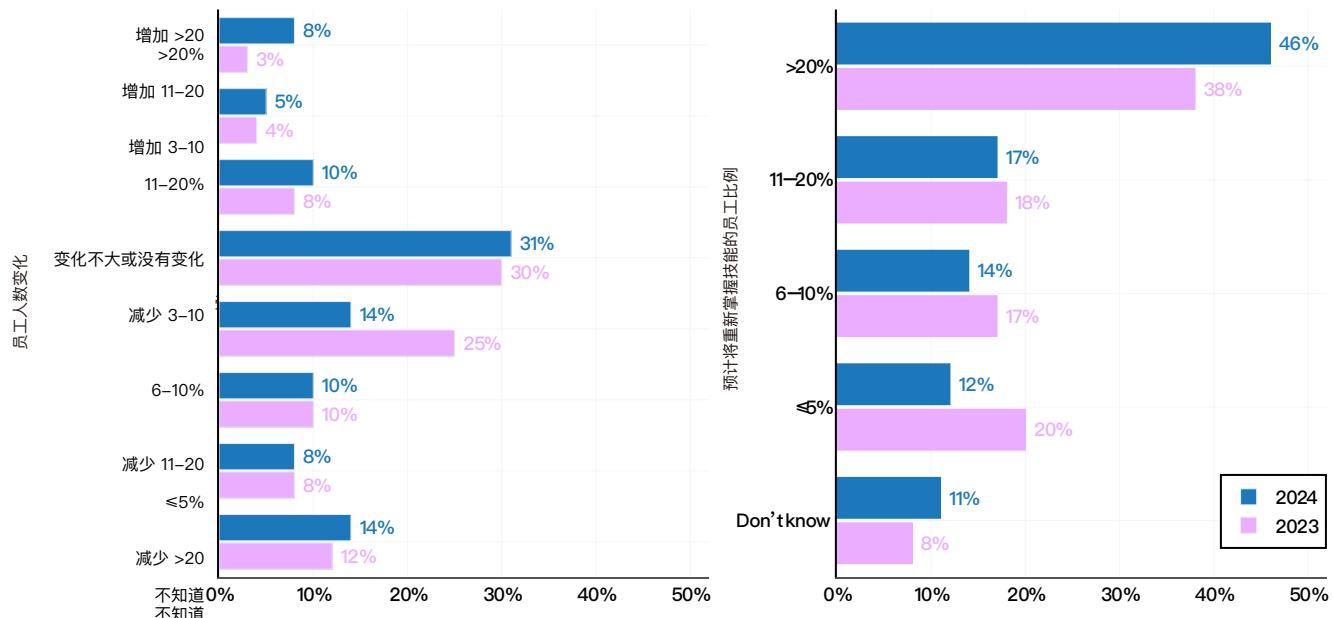


图4.4.13

第四章：经济

4.5 部署机器人

配备人工智能软件技术的机器人部署为人工智能就绪基础设施的实际应用提供了窗口。本章节节内容基于国际机器人联合会（IFR）的数据。IFR 是一家致力于推动机器人产业发展的非营利组织。该组织每年发布《世界机器人报告》，跟踪全球机器人安装趋势。¹¹

4.5 机器人部署 总体趋势

下一节包括有关工业机器人安装和操作的数据，工业机器人的定义是“一种自动控制、可重新编程的多用途机械手，可编程三轴或多轴，可固定在原地或移动，用于工业自动化应用”。

图 4.5.1 按年份列出了全球安装的工业机器人总数。2023 年，工业机器人安装量略有下降，为 54.1 万台，比 2022 年减少了 2.2%。这是自 2019 年以来首次出现同比下降。

2012–2023 年全球安装的工业机器人数量

资料来源：麦肯锡公司调查，2024 | 图表：2025 年人工智能指数报告

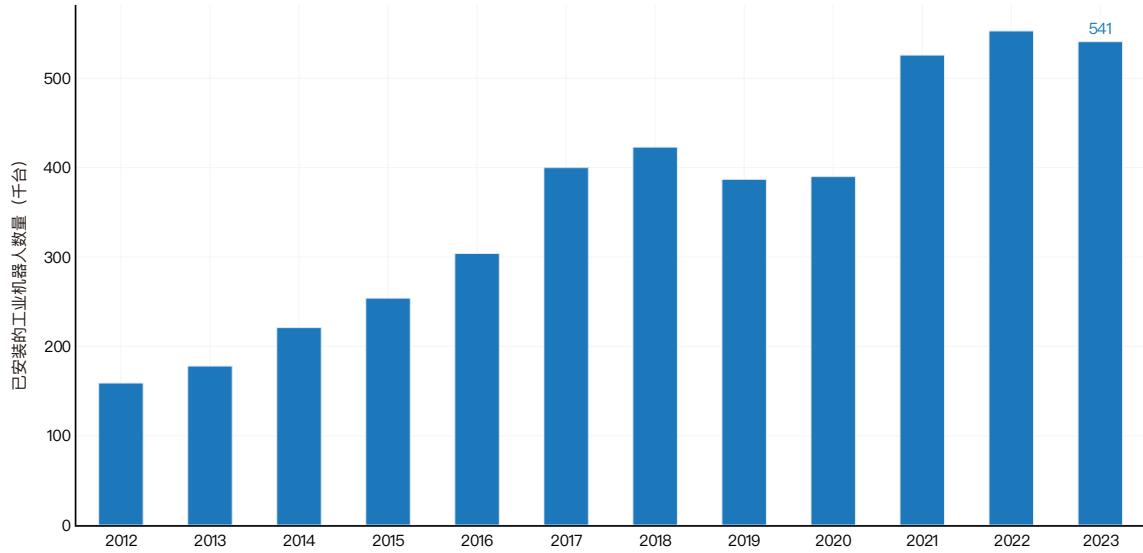


图4.5.1

11. 由于《国际财务报告准则》报告的时间原因，最新数据来自 2023 年。《国际财务报告准则》都会重新审查前几年收集的数据，如果有更准确的数据，偶尔也会更新数据。因此，今年报告中某些数据可能与往年的数据略有不同。

第四章：经济

4.5 部署机器人

到 2023 年，全球工业机器人保有量将从 2022 年的 390.4 万台增至 428.2 万台(图 4.5.2)。自 2012 年以来，工业机器人的安装量和使用量都在稳步增长。

2012–2023 年全球工业机器人运行保有量

资料来源：国际机器人联合会 (IFR)，2024| 图表：2025 年人工智能指数据报告

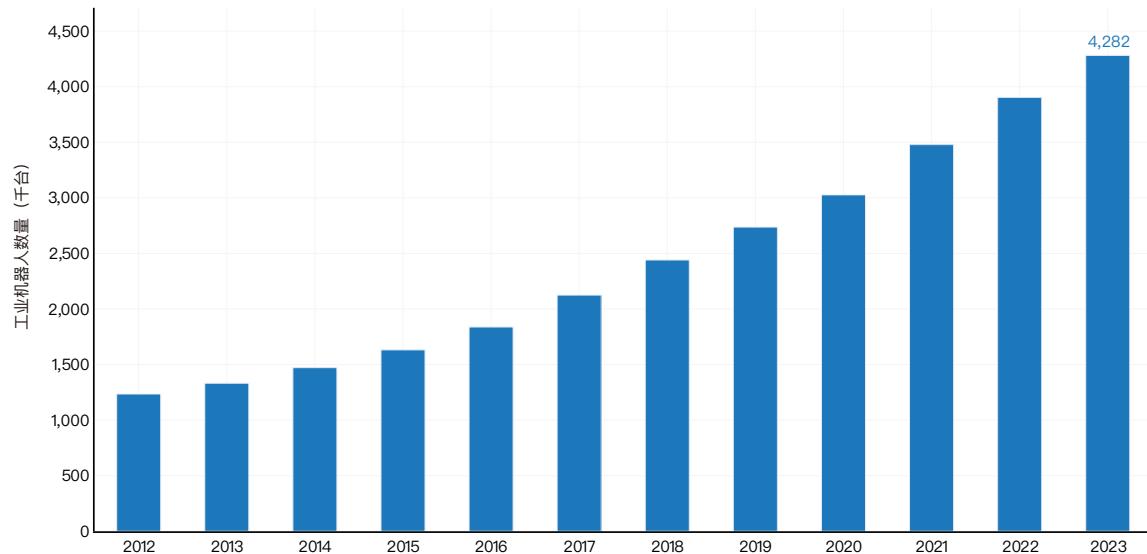


图4.5.2

第四章：经济

4.5 部署机器人

工业机器人：传统机器人与协作机器人

传统工业机器人与协作机器人存在本质区别：前者用于替代人工操作，后者则专为机人协同作业而设计。¹² 机器人学界对协作机器人展现出日益浓厚的研发热情，因其具备四大核心优势——操作安全性、工作灵活性、系统可扩展性以及迭代

学习能力。图 4.5.3 展示了全球工业机器人按类型的安装数量统计。数据显示，协作机器人在新增工业机器人安装量中的占比呈现显著增长：2017 年仅占 2.8%，至 2023 年已提升至 10.5%，增幅达 7.7 个百分点。

2017–2023 年按类型划分全球安装的工业机器人数

资料来源：国际机器人联合会（IFR），2024 | 图表：2025 年人工智能指数报告

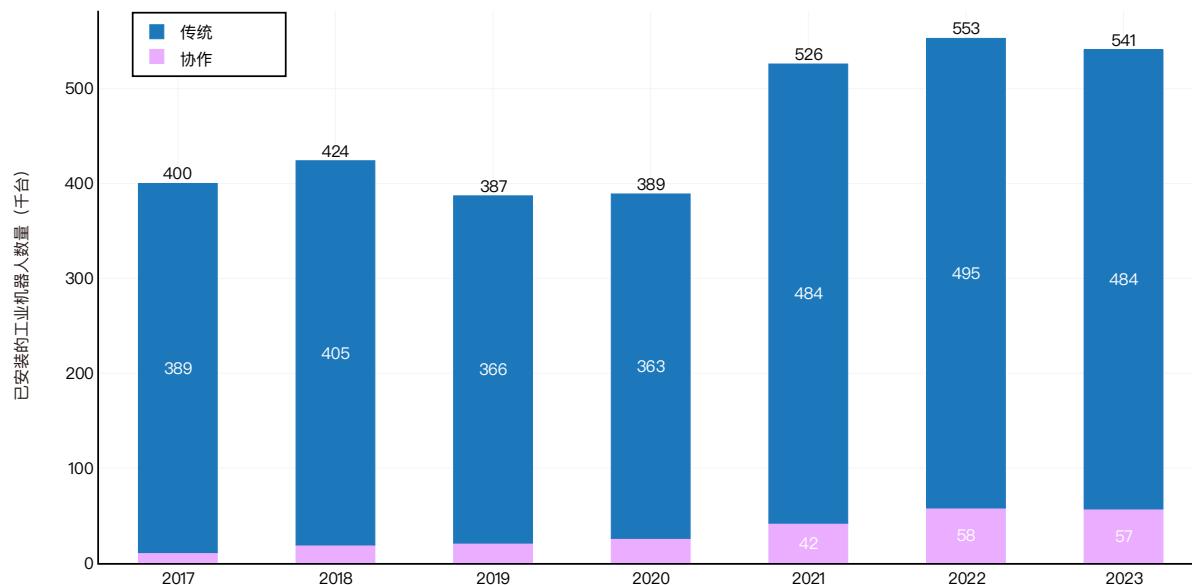


图4.5.3

12、有关国际机器人联合会定义协作机器人的更多详情，请访问。

第四章：经济

4.5 部署机器人

按地理区域划分

各国工业机器人安装量数据能够反映不同经济体对机器人技术应用的重视程度。2023 年度统计显示，中国以 276,300 台的工业机器人安装量位居全球首位，分别达到日本（46,100

台）的 6 倍和美国（37,600 台）的 7.3 倍（图 4.5.4）。韩国与德国分类其后，安装量分别为 31,400 台和 28,400 台。

2023 年按地理区域划分的工业机器人安装数量

资料来源：国际机器人联合会（IFR），2024 | 图表：2025 年人工智能指数据报告

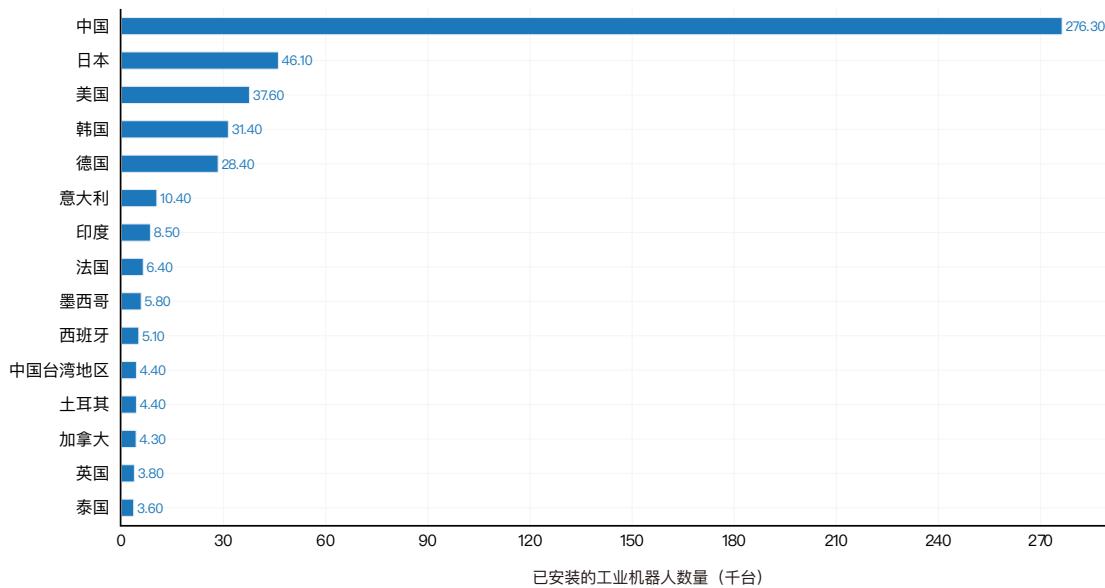


图4.5.4

第四章：经济

4.5 部署机器人

自 2013 年超越日本成为全球最大工业机器人应用市场以来，中国持续扩大其领先优势。数据显示，中国工业机器人安装量在全球总量的占比从 2013 年的 20.8% 大幅提升至 2023 年的 51.1%（图 4.5.5）。

2021–2023 年前 5 位国家新安装的工业机器人数量

资料来源：国际机器人联合会 (IFR)，2024 | 图表：2025 年人工智能指数报告

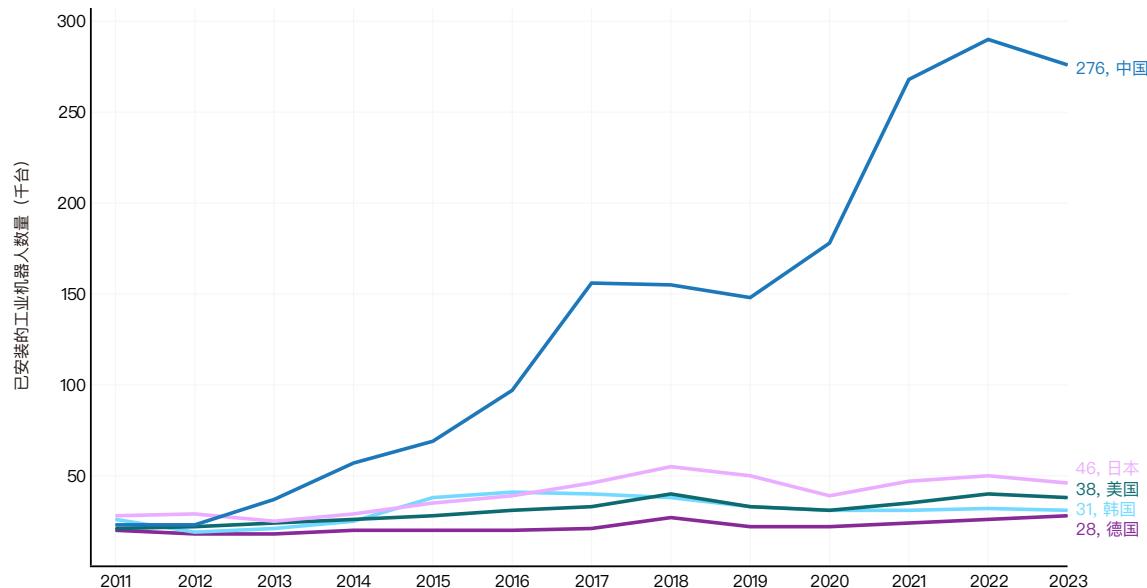


图4.5.5

第四章：经济

4.5 部署机器人

2021年以来，中国工业机器人年度安装量持续超过全球其他地区的总和。尽管2023年这一领先优势较2022年有所收窄(图4.5.6)，但同比增速的放缓并未动摇中国在全球工业机器人应用市场的绝对主导地位。

2016–2023年中国与世界其他地区工业机器人安装数量

资料来源：国际机器人联合会 (IFR)，2024 | 图表：2025年人工智能指数报告

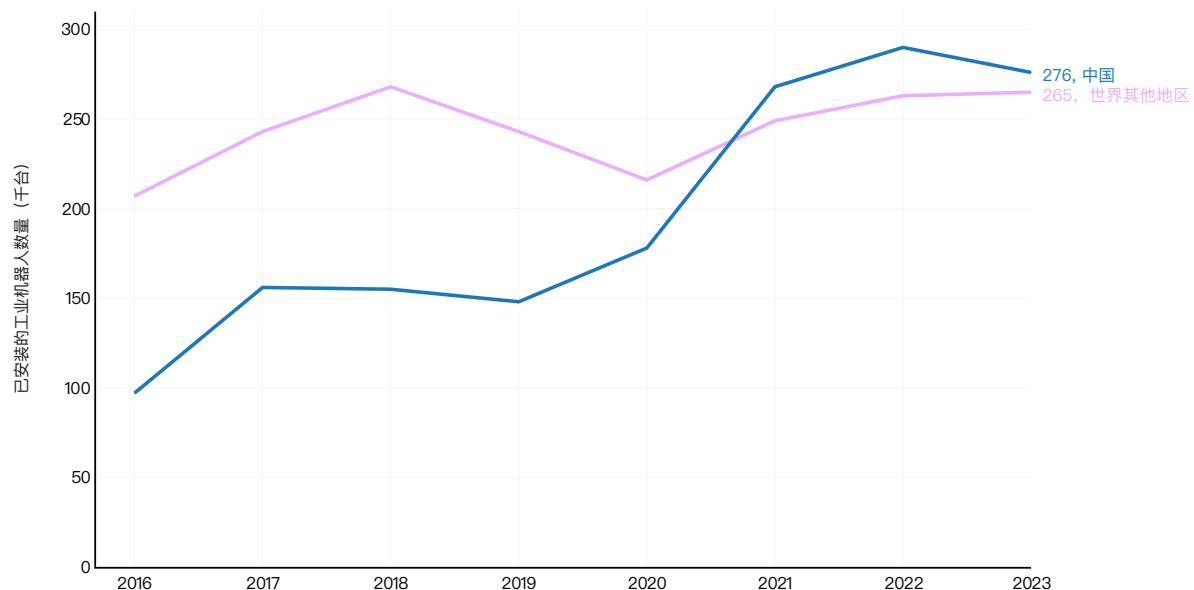


图4.5.6

第四章：经济

4.5 部署机器人

国际机器人联合会报告显示，2022 至 2023 年间全球仅七个国家的工业机器人安装量实现同比增长（图 4.5.7）。其中，印度以 59% 的增速位居榜首，英国（51%）与加拿大（37%）分

类二三位。与此同时，中国台湾地区出现 43% 的显著下滑，法国下降 13%，日本与意大利均录得 9% 的负增长。

2022 年与 2023 年各地区工业机器人安装量的年增长率对比

资料来源：国际机器人联合会（IFR），2024 | 图表：2025 年人工智能指数报告

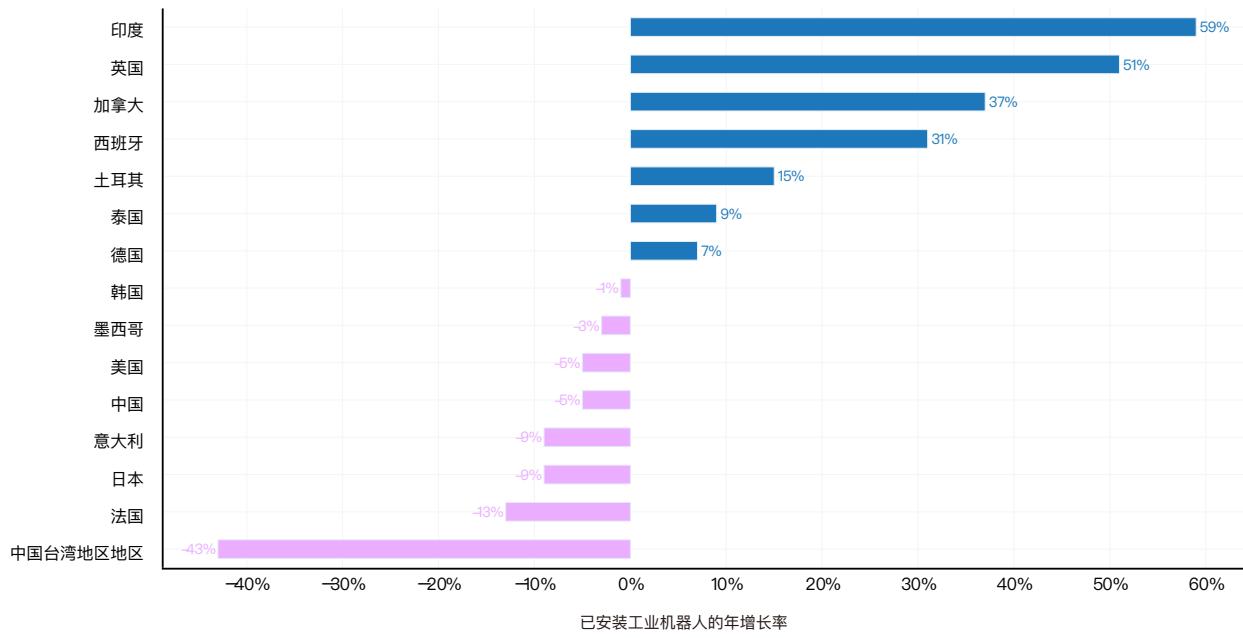


图4.5.7

第四章：经济

4.5 部署机器人

服务机器人的国家级数据

另一重要机器人类别是服务机器人。根据国际标准化组织（ISO）的定义，服务机器人指“为人类或设备执行有益任务，且不包含工业自动化应用的机器人系统”。此类机器人可应用于医疗环境及专业清洁等场景。¹³

2023 年度数据显示，除医疗机器人外，所有应用类别的服务机器人安装量均较 2022 年实现增长（图 4.5.8）。其中，农业领域服务机器人安装量增长至 2.5 倍，酒店服务业安装量达 2.2 倍，呈现显著上升趋势。

2022 年与 2023 年按应用领域划分的全球服务机器人安装数量对比

资料来源：国际机器人联合会（IFR），2024| 图表：2025 年人工智能指数报告

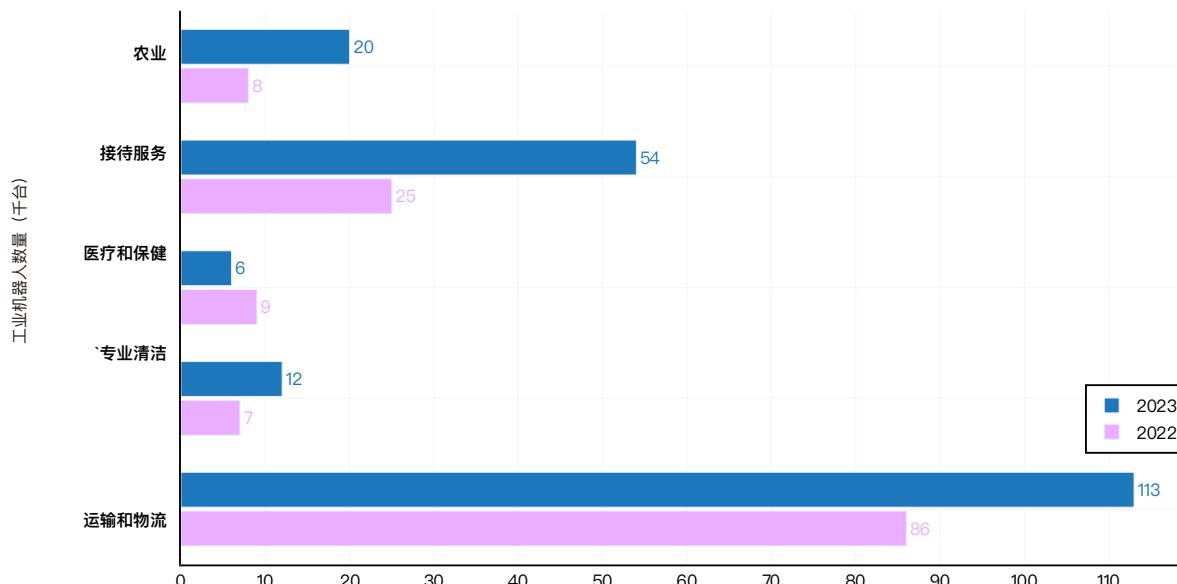
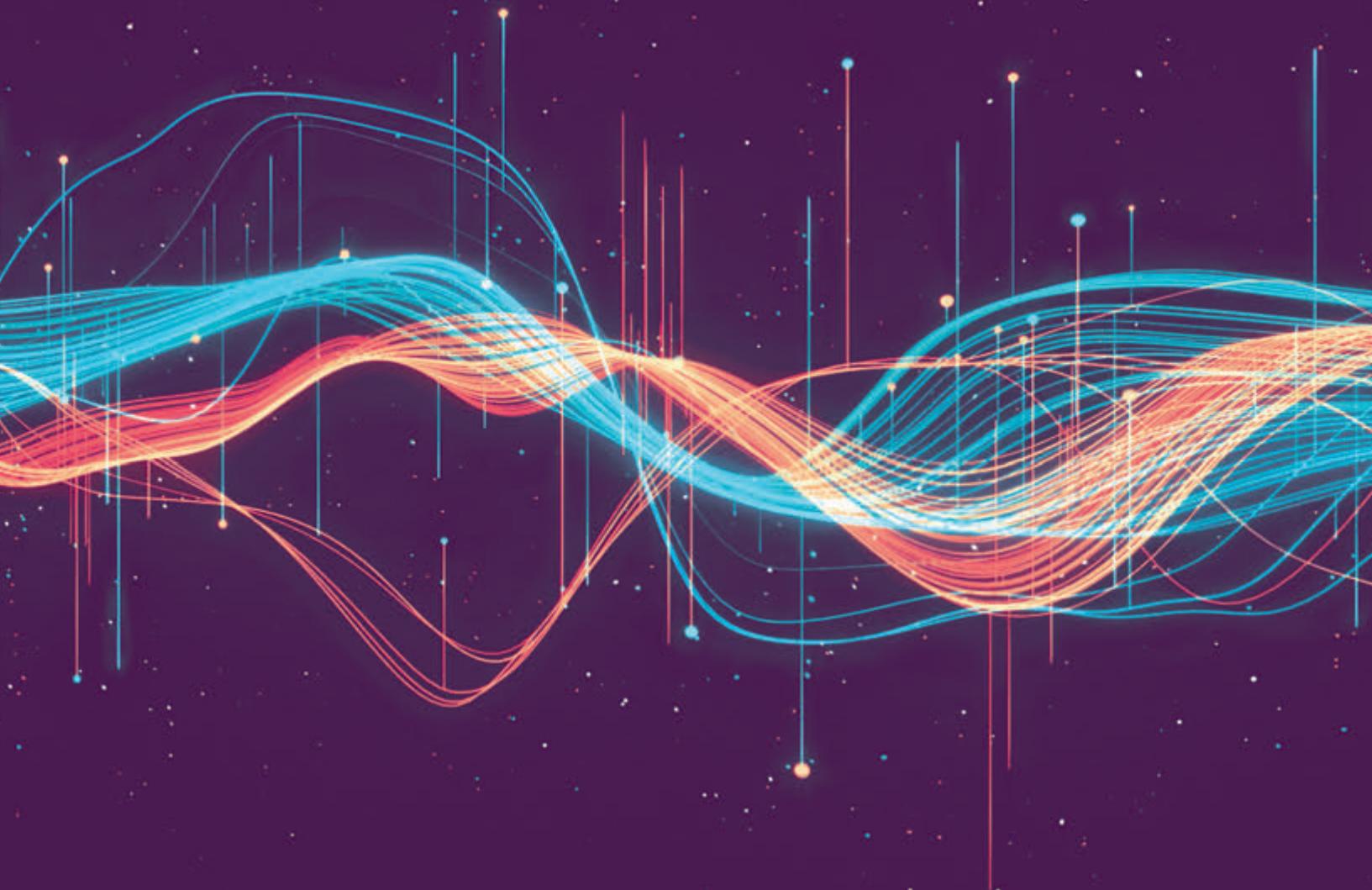


图4.5.8

12、更详细的定义可查阅 [这里](#)。



第五章： 科学与医学



第五章：科学与医学

概述	282
章节要点	283
5.1 重要的医学和生物人工智能里程碑	285
蛋白质序列优化	285
Aviary	286
AlphaProteo	287
人类大脑图谱	287
人工智能虚拟实验室	288
GluFormer	289
Evolutionary Scale Modeling v3 (ESM3)	289
AlphaFold 3	290
5.2 中心法则	291
蛋白质序列分析	291
人工智能驱动的蛋白质序列模型	291
蛋白质科学公共数据库	293
研究与论文发表趋势	294
人工智能驱动的蛋白质科学的论文统计	294
图像和多模态人工智能促进科学发现	295
5.3 临床诊疗：影像学领域	296
数据：来源、类型和需求	296
先进建模方法	298
5.4 临床诊疗：非影像领域	300
临床知识	300
MedQA	300
重点：人工智能医生和成本效益考虑	301
医疗大语言模型性能评估	302
概述	302
大语言模型在临床诊断推理中的应用	304
重点：大语言模型对诊断推理的影响	304
管理性推理和患者护理决策	304
重点：GPT-4 辅助临床管理任务的效果评估	305
环境型人工智能语音助手	306
部署、实施与撤除	308
FDA 对人工智能医疗设备的批准	308
成功应用案例：斯坦福医疗系统	308
外周动脉疾病筛查	309
健康社会决定因素	310
从电子健康档案和临床记录中提取SDoH	310
医疗领域的人工智能应用与SDoH的整合	311
合成数据	311
临床风险预测	311
药物发现	312
数据生成平台	312
电子健康档案系统	313
临床决策支持	315
5.5 伦理考量	317
元分析	317
5.6 科学领域的人工智能基础模型	320
重点：标志性的模型发布	320

获取公共数据

第五章： 科学与医学

概述

本章节探讨了人工智能驱动的科学和医学的主要趋势，反映了该技术在这些领域日益增长的影响力。本章节首先介绍了 2024 年人工智能的重要里程碑，随后分析了人工智能在蛋白质折叠这一重要科学进步领域的应用。本章节随后探讨了人工智能在临床医疗中的作用，包括成像和非成像应用。这包括回顾新语言模型中的临床知识能力、人工智能系统的诊断和临床管理能力、人工智能在医学中的实际应用、合成数据应用以及健康的社会决定因素。最后，本章节以探讨人工智能医学研究的伦理趋势作为结尾。本章节由 RAISE Health (Responsible AI for Safe and Equitable Health) 编写，该机构由斯坦福大学医学院和斯坦福大学以人为本人工智能研究所 (HAI) 合作成立。自 2023 年启动以来，RAISE Health 一直致力于推动负责任的人工智能在生物医学研究、教育和患者护理方面的创新，重点是确保这些技术惠及每个人。

促进合作研究和知识共享是 RAISE Health 的核心使命。作为这一承诺的一部分，RAISE Health 与人工智能指数指导委员会合作，将该小组的关注点扩大到科学和医学的关键发展。2024 年，这一合作产生了关于科学和医学的首章，重点介绍了斯坦福大学及其他大学在人工智能方面取得的重大进展。在此基础上，RAISE Health 教职研究委员会成员、斯坦福大学医学院教师、博士后研究员以及医学院和工程学院的本科生为 2025 年的章节做出了贡献。

第五章： 科学与医学

章节要点

1. 更先进的大规模蛋白质测序模型问世。2024 年，包括 ESM3 和 AlphaFold 3 在内的多个高性能大规模蛋白质测序模型相继推出。随着时间的推移，这些模型的规模显著扩大，使得蛋白质预测准确率不断提高。

2. 人工智能持续推动科学发现的飞速发展。人工智能在科学进步中的作用不断扩大。2022 至 2023 年仅是人工智能 I 驱动科研突破的初始阶段，而 2024 年出现更具突破性的进展，包括训练大语言模型智能体执行生物任务的 Aviary，以及显著增强野火预测能力的 FireSat。

3. 主流大语言模型的临床知识水平持续提升。OpenAI 最近发布的 o1 在 MedQA 比较基准中创下了 96.0% 的新纪录，比 2023 年公布的最佳成绩提高了 5.8%。自 2022 年末以来，该测试性能已累计提升 28.4%。作为评估临床知识的重要比较基准，MedQA 可能正接近性能饱和，预示着需要更具挑战性的评估。这表明需要建立更具挑战性的评估体系。

4. 人工智能在关键临床任务中的表现优于医生。一项新的研究发现，在诊断复杂的临床病例时，无论是有人工智能还是没有人工智能，仅 GPT-4 就能胜过医生。最近的其他研究表明，人工智能在癌症检测和识别高死亡率风险患者方面超过了医生。不过，一些初步研究表明，人工智能与临床医生的协同诊疗可产生最优结果，这一发现值得作为重点领域开展深入研究。

5. 美国食品及药物管理局（FDA）批准的人工智能医疗设备数量激增。美国食品药品监督管理局于 1995 年批准了首款人工智能医疗设备。截至 2015 年，仅有 6 款此类设备获批，但这一数字到 2023 年激增至 223 款。

6. 合成数据在医学领域展现出巨大潜力。2024 年发布的研究表明，人工智能生成的合成数据可以帮助模型更好地识别健康的社会决定因素，加强保护隐私的临床风险预测，并促进新药化合物的发现。2024 年最新研究表明，人工智能生成的合成数据可有效提升模型对健康社会决定因素的识别能力，优化隐私保护型临床风险预测，并促进新药化合物的发现。

第五章： 科学与医学

章节要点 (续)

7. 医学人工智能伦理研究文献逐年增加。从 2020 年到 2024 年，医学人工智能伦理方面的论文数量几乎翻了两番，从 2020 年的 288 篇增加到 2024 年的 1031 篇。

8. 基础模型进入医学领域。2024 年，一大波大型医学基础模型发布，涵盖从 Med-Gemini 等通用多模态模型，到面向特定专科的 EchoCLIP(超声心动图)、视觉 FM(眼科)及 ChexAgent(放射科)等专用模型。

9. 公共蛋白质数据库规模不断扩大。自 2021 年以来，主要公共蛋白质科学数据库的条目数量显著增长，其中包括 UniProt (增长 31%)、PDB(增长 23%) 及 AlphaFold(激增 585%)。这一扩展对科研发现具有重要影响。

10. 人工智能研究获得两项诺贝尔奖。2024 年，人工智能驱动的研究获得了最高荣誉，两项与人工智能相关的突破获得了诺贝尔奖。谷歌 DeepMind 的德米斯 · 哈萨比斯 (Demis Hassabis) 和约翰 · 朱珀 (John Jumper) 凭借 AlphaFold 在蛋白质折叠方面的开创性工作获得了诺贝尔化学奖。与此同时，约翰 · 霍普菲尔德 (John Hopfield) 和杰弗里 · 辛顿 (Geoffrey Hinton) 因其在神经网络方面的奠基性贡献获得了诺贝尔物理学奖。

第五章：科学与医学

5.1 重要的医学和生物人工智能里程碑

本章节节重点介绍了 2024 年由 RAISE 健康人工智能指数工作组和人工智能指数指导委员会选出的与人工智能相关的重大医学和生物学突破。

5.1 重要的医学和生物人工智能里程碑

蛋白质序列优化

大语言模型优化蛋白质序列

大语言模型近期意外展现出蛋白质序列优化的新生物学能力。传统蛋白质工程需要通过大量实验室研究来优化序列以提升功能，而最新研究发现，未经微调的大语言模型在此任务上表现出惊人成效——这项隐藏能力在 Llama-3.1-8B-Instruct 的适配版本中得到验证。

研究人员采用定向进化方法证实，大语言模型生成的蛋白质序列在合成与实验适应性场景中均优于传统算法（图 5.1.1）。该研究以最大化适应值为目标（更高分值代表更优性能），将提出方法的适应分值与默认进化算法（EA）进行对比。¹ 结果显示，这种优化能力不仅适用于单目标任务，还可扩展至预算受限的约束性及多目标场景。这一突破性发现揭示了前沿大语言模型的涌现特性，预示着随着通用模型的持续进化，其对科学领域的影响将不断深化。

适合度优化的单目标优化结果

资料来源：Wang 等，2024

Dataset	Method	Population × iteration	Fitness score		
			Top 1	Top 10	Top 50
GB1	EA	32×4	5.38±1.77	3.81±1.10	2.31±0.71
		48×4	4.88±0.33	3.72±0.38	2.17±0.27
		96×4	5.72±0.56	4.32±0.53	2.84±0.60
	Ours	32×4	4.34±0.53	3.22±0.23	1.94±0.28
		48×4	4.31±0.82	3.76±0.82	2.45±0.61
		96×4	4.80±0.52	4.09±0.19	3.04±0.19
	TrpB	32×4	0.20±0.18	0.14±0.12	0.07±0.05
		48×4	0.67±0.14	0.52±0.11	0.19±0.04
		96×4	0.74±0.01	0.59±0.03	0.35±0.10
Syn-3bfo	EA	32×4	0.60±0.10	0.50±0.07	0.35±0.07
		48×4	0.68±0.04	0.58±0.01	0.36±0.01
		96×4	0.78±0.20	0.60±0.16	0.39±0.16
	EA	32×8	0.57±0.21	-0.44±0.11	-1.35±0.17
		48×8	1.29±0.36	0.42±0.24	-0.63±0.07
		96×8	1.85±0.47	1.10±0.28	0.07±0.28
	Ours	32×8	2.51±0.23	1.33±0.14	0.28±0.20
		48×8	2.35±0.26	1.36±0.11	0.04±0.09
		96×8	2.83±0.20	2.02±0.36	0.96±0.36
AAV	EA	32×8	0.42±0.03	0.36±0.01	0.32±0.00
		48×8	0.44±0.00	0.38±0.01	0.33±0.00
		96×8	0.44±0.00	0.40±0.01	0.36±0.00
	Ours	32×8	0.74±0.00	0.69±0.02	0.62±0.03
		48×8	0.75±0.01	0.71±0.01	0.64±0.02
		96×8	0.76±0.03	0.73±0.03	0.68±0.03
	EA	32×8	0.43±0.13	0.21±0.02	0.12±0.01
		48×8	0.43±0.14	0.26±0.05	0.12±0.01
		96×8	0.50±0.11	0.34±0.05	0.18±0.01
GFP	Ours	32×8	0.96±0.02	0.94±0.01	0.88±0.03
		48×8	0.96±0.02	0.93±0.01	0.84±0.02
		96×8	0.97±0.01	0.95±0.01	0.92±0.01

图 5.1.1

1、进化算法（EA）在计算机程序中模拟生物进化的关键环节，通过寻找近似答案来解决复杂问题，特别是那些没有精确率或完全令人满意的解决方案的问题。

第五章：科学与医学

5.1 重要的医学和生物人工智能里程碑

Aviary

面向生物任务的大语言模型智能体训练

随着人工智能系统在科研应用领域日益发挥重要作用，如何设计能够调用工具进行复杂推理任务的语言模型成为关键挑战。Aviary 研究平台提出了一个结构化训练框架，专门针对三项高难度生物科学任务：DNA 操作（用于分子克隆）、科研问题解答（通过检索科学文献）以及蛋白质稳定性工程。图 5.1.2 对比了不同模型在 Aviary 各实验环境中的表现数据，结果显示：与未接入实验环境的 Claude 3.5 Sonnet 基线模型相比，

集成在 Aviary 智能体框架内的模型在几乎所有任务中都表现更优。本研究证实：(1) 尽管通用大语言模型能胜任多数科研任务，但结合领域专家知识进行微调的模型往往能获得更优结果；(2) 人工智能驱动的科研进程不仅可以通过扩大模型规模来加速，更可通过与外部工具的交互实现突破——这种能力现已被学界统称为“智能体化人工智能(agenetic AI)”。

大语言模型及语言智能体在Aviary环境中的任务解决性能表现

资料来源：Narayanan 等，2024 | 图表：2025 年人工智能指数报告

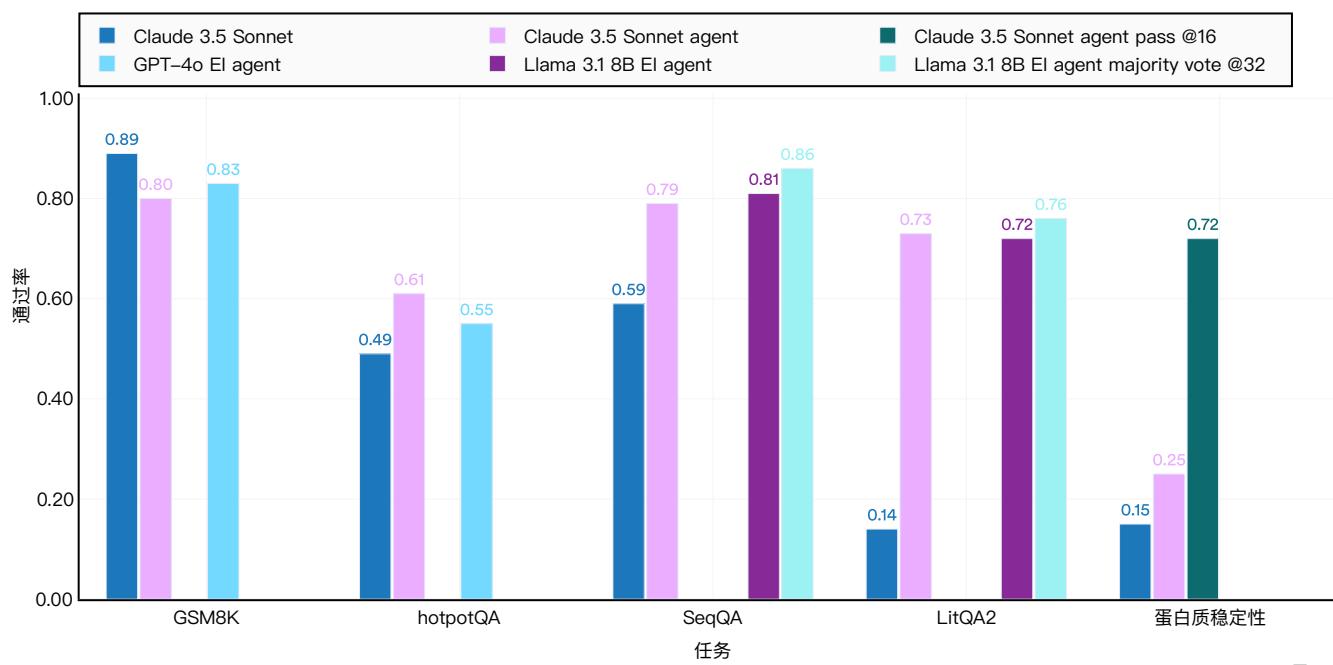


图 5.1.2

第五章：科学与医学

5.1 重要的医学和生物人工智能里程碑

AlphaProteo

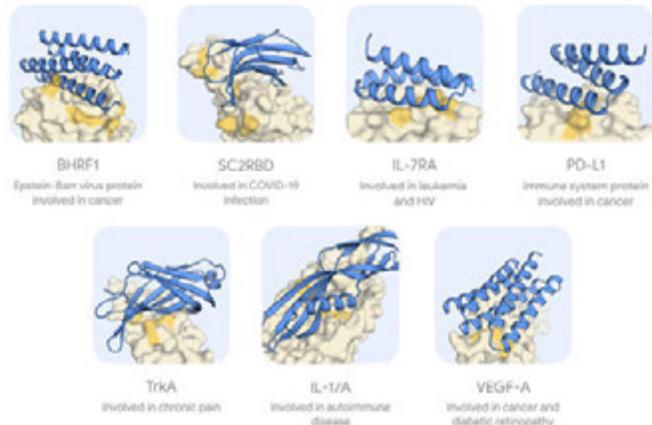
人工智能驱动的新型高亲和力蛋白结合剂开发

AlphaProteo 是 Google DeepMind 研发的专注于设计新型高亲和力蛋白质结合剂的模型，这些结合剂能够特异性附着于目标分子。如图 5.1.3 所示，该模型已成功为七种靶蛋白预测并构建了结合蛋白结构。在多个靶点（包括与癌症和糖尿病相关的 VEGF-A 蛋白）上，AlphaProteo 实现了全球首个蛋白质结合剂的设计。经测试，该工具在七种靶蛋白上设计的结合剂效能显著超越现有最优方案——研究团队评估部分结合剂的效能可达当前同类产品的 300 倍。针对病毒蛋白 BHRF1，在 DeepMind 湿实验室测试中，其设计结合剂的成功结合率高达 88%。基于已测试靶点的数据显示，AlphaProteo 设计的结合剂结合强度约为现有顶级设计方法的 10 倍，标志着生物工程领域的重大突破。目前该模型已应用于药物研发、诊断技术和生物技术等领域。

AlphaProteo 成功生成结合体

资料来源：谷歌 DeepMind 2024

图 5.1.3



人类大脑图谱

人类大脑微区的突触级重建

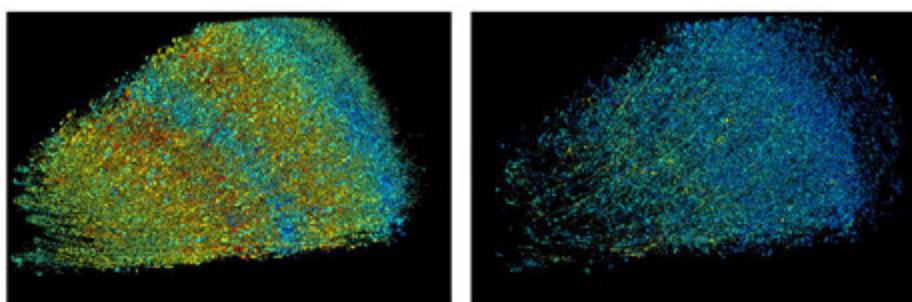
谷歌 Connectomics 项目的一个研究团队，已经在突触层级重建了人脑中一立方毫米的区域——《Wired》杂志称其为“迄今为止最为详细的大脑连接图谱”。该样本取自一名癫痫患者在接受手术时被切除的左前颞叶区域，利用多束扫描电子显微镜进行成像。研究人员通过超过 5,000 张超薄切片（每张厚度为 30 纳米）记录了大约 57,000 个细胞——包括神经元、胶质细胞和血管——以及 1.5 亿个突触。图 5.1.4 展示了重建

成果：左侧为兴奋性神经元，右侧为抑制性神经元。为了处理这一海量数据集，团队开发了多项机器学习工具，例如洪泛填充网络（用于无需人工描绘的神经元重建）、SegCLR（用于细胞类型识别）以及 TensorStore（用于多维数据集的管理）。该数据集已通过 Neuroglancer（一个基于网页的探索工具）和其注释精化扩展工具 CAVE 向公众开放。这一项目在理解神经回路方面迈出了重要一步，未来有望为神经疾病的治疗提供关键启示。

三维脑图图像

资料来源：谷歌研究 2024

图 5.1.4



第五章：科学与医学

5.1 重要的医学和生物人工智能里程碑

人工智能虚拟实验室

人工智能虚拟实验室推动生物医学研究突破

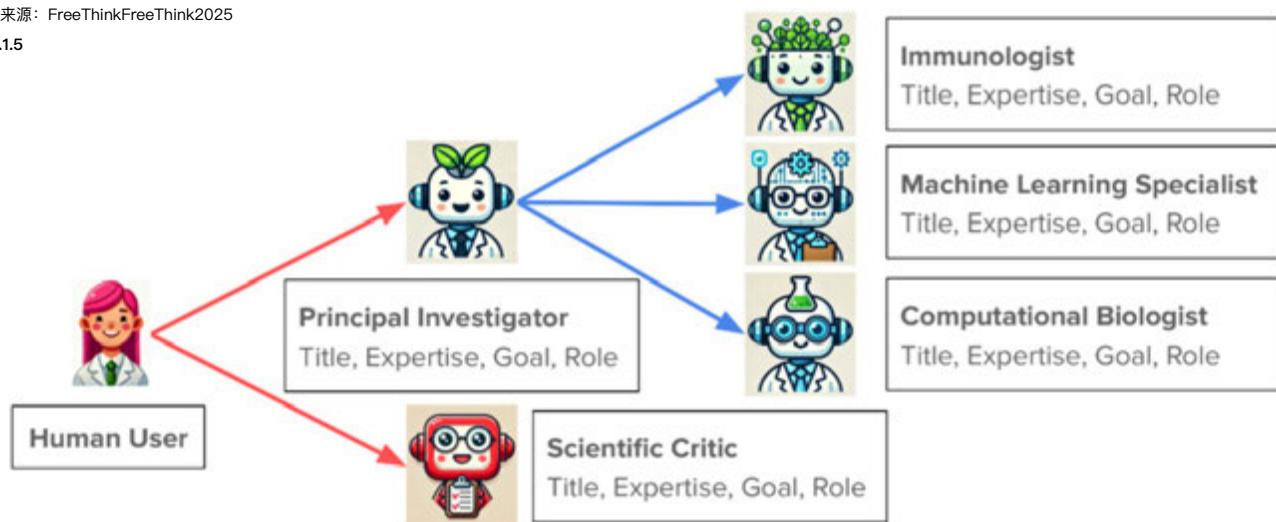
人工智能在科学研究中的角色正从被动工具转变为主动合作者。斯坦福大学最近发布的一项研究提出了一个虚拟 AI 实验室，多个具备专长的人工智能科学家（本质上为大语言模型）在其中协同工作，彼此之间以智能体的形式开展研究。人工智能在科学研究中的角色正从被动工具转变为主动合作者。斯坦福大学最近发布的一项研究提出了一个虚拟 AI 实验室，多个具备专长的人工智能科学家（本质上为大语言模型）在其中协同工作，彼此之间以智能体的形式开展研究。

该虚拟实验室参照计算生物学实验室的架构组建，由以下成员构成：一名首席研究员 (PI)、一个科学评审人工智能系统，以及三位分别专攻免疫学、计算生物学和机器学习的学科专家（图 5.1.5）。首席研究员模型负责创建这些专家，并指导研究流程。在研究过程中使用了 AlphaFold 与 Rosetta 等蛋白质设计工具。但本研究的真正意义不仅在于具体成果，更在于它展示了一个完全由大语言模型驱动的实验室也能够产生有意义的科学发现。

基于人工智能的实验室工作流程

资料来源：FreeThinkFreeThink2025

图 5.1.5



第五章：科学与医学

5.1 重要的医学和生物人工智能里程碑

GluFormer

人工智能辅助持续血糖监测

GluFormer 是由英伟达特拉维夫研究所、魏茨曼研究所及其他机构联合开发的基础模型，能够对连续血糖监测数据进行分析，并预测长期健康趋势。该模型在超过 1,000 万条血糖记录上进行训练，数据来源于近 11,000 名个体，其中大多数并未患有糖尿病。它能够提前预测长达四年的健康变化轨迹，例如识别出有发展为糖尿病或血糖控制恶化风险的人群，甚至在症状尚未出现之前。在一项长达 12 年、涉及 580 名成人的研究中，GluFormer 成功预测了 66% 的新发糖尿病病例，并在心血管相关死亡的高风险组中准确识别出 69% 的死亡病例。该模型的预测结果已在五个国家的 19 个独立人群样本中进行了验证，样本总数达 6,044 人，涵盖多种健康状况。GluFormer 通常优于现有基于血糖监测的标准指标，如血糖管理指标 (GMI)（图 5.1.6）。从短期与长期来看，GluFormer 等模型有望将糖尿病治疗从被动应对转变为主动预防，推动更早期的临床干预。

Evolutionary Scale Modeling v3 (ESM3)

模拟进化过程生成新型蛋白质

EvolutionaryScale 公司推出的 ESM3 模型是一项突破性研究，旨在通过模拟生物进化过程来生成新型蛋白质。该模型使用了 27.8 亿条蛋白质序列进行训练，拥有 980 亿个参数。和许多人工智能模型一样，ESM3 提供小型、中型和大型版本，并可通过 API 或合作平台获取。ESM3 的一项标志性成果是设计出一种新型的绿色荧光蛋白 esmGFP，据该公司估计，自然界若通过演化生成该蛋白，将需耗费约五亿年时间。该成果是在人工推理引导下完成的。图 5.1.7 展示了不同规模 ESM3 模型在满足原子结构协调性提示下生成蛋白质的表现。研究结果表明，模型规模越大，完成的任务数量也越多。ESM3 已开放源代码与数据，有助于推动合成生物学与蛋白质工程领域的协作研究，应用前景包括药物开发、材料科学与环境工程。

GluFormer 与血糖管理指标的对比研究

资料来源：Lutske 等，2024

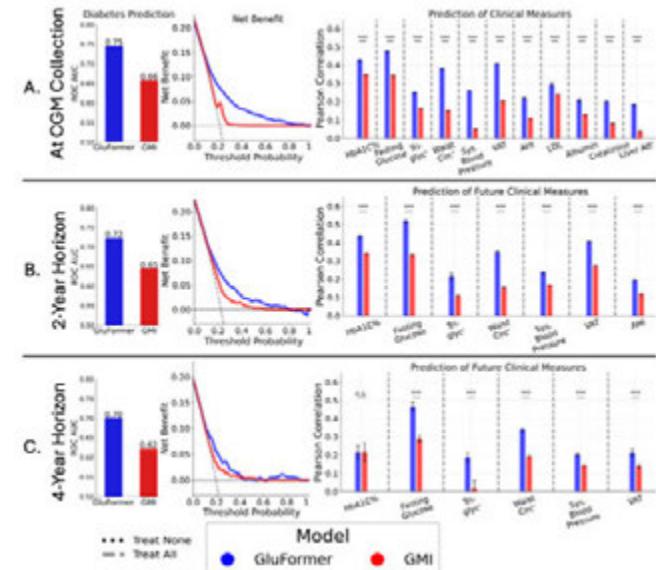


图 5.1.6

基于原子坐标提示的蛋白质生成任务中 ESM3 模型的评估

资料来源：ESM3, 2024 | 图表：2025 年人工智能指数报告

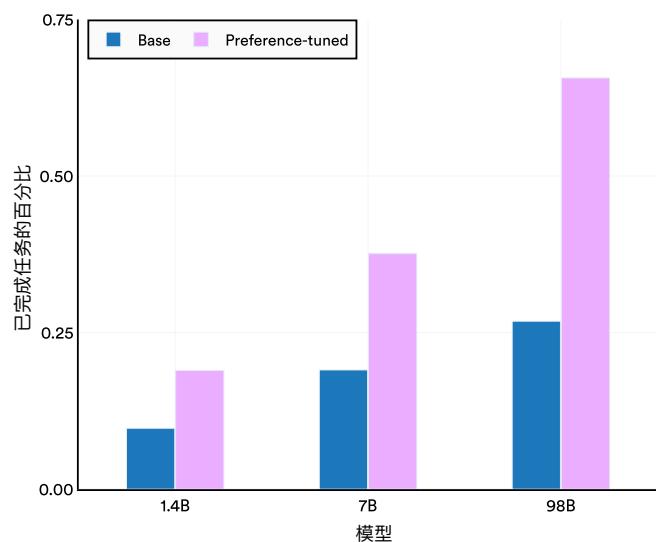


图 5.1.7

第五章：科学与医学

5.1 重要的医学和生物人工智能里程碑

AlphaFold 3

预测所有生命分子的结构和相互作用

谷歌与 Isomorphic Lab 联合推出的 AlphaFold 3 是 AlphaFold 系列的最新进展，其功能已超越蛋白质结构预测，能够更精确地模拟蛋白质与关键生物分子之间的相互作用，包括 DNA、RNA、小分子配体与抗体。图 5.1.8 展示了 AlphaFold 3 在预测蛋白质 – 小分子配体结合准确性上的表现，并与其他领先对接工具（如 Vina 与 Gnina）进行比较。图中以预测结果中埃均方根偏差 (RMSD) 低于 2 Å 为准，这一数值是评价对接精

度 (docking accuracy) 的重要指标。^{2 3} AlphaFold 3 的性能可与此前的最先进方法相媲美，且在结合位点已被预先设定的情况下表现尤为出色，即对接算法事先获得了小分子 (配体) 预计结合的蛋白质特定区域的信息。通过对小分子与蛋白质之间相互作用的建模，AlphaFold 3 能够加速药物研发过程，这对于疾病研究具有重要意义。此外，AlphaFold 3 的开源性质也赋予了全球科研人员更大的能力与自由。

AlphaFold 3 与基线方法在蛋白质–配体对接中的比较

资料来源：ESM3, 2024 | 图表：2025 年人工智能指数报告

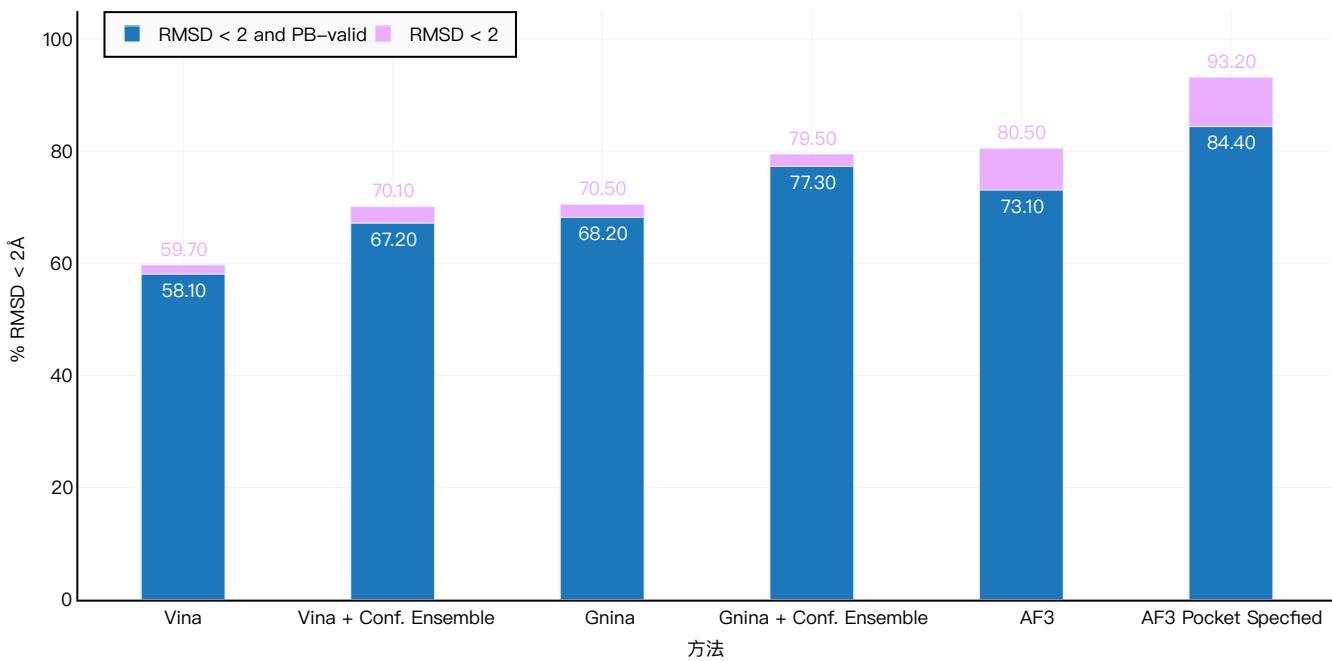


图 5.1.8

2. 像 Vina 这样的对接工具是一种用于分子对接的计算程序，分子对接是一种预测小分子（如药物）如何与目标蛋白质相互作用的过程。这些工具可以帮助科学家模拟并直观地看到分子如何与蛋白质的活性位点结合，这在药物发现中至关重要。

3. 该图表：使用两种深浅不同的条形图来表示分子对接预测的不同准确率标准。浅色条表示均偏差 (RMSD) 低于 2 Å 的对接结果的百分比，这意味着预测的姿势在结构上是准确的。深色条应用了更严格的标准，显示了不仅 RMSD 值在 2 Å 以内，而且在结合口袋内定位正确 (PB-valid) 的预测结果比例。这种区别突出了对接准确率与更精确、更具有生物相关性的结合预测之间的差异。

第五章：科学与医学

5.2 中心法则

人工智能已深刻变革了众多科学领域，其中蛋白质科学是受影响最为显著的学科之一。理解蛋白质序列是生物学研究的基础，这对药物研发、合成生物学和疾病研究都具有深远影响。近期人工智能技术的突破使科学家能够以前所未有的精度分析预测蛋白质功能、结构和相互作用。随着该领域的发展，这些技术进步将对医疗健康、生物技术和监管体系产生重大影响。本节将重点介绍过去一年人工智能驱动的蛋白质分析领域的关键进展，主要聚焦公共数据库建设、研究趋势演变以及新兴政策考量等三个方面。

5.2 中心法则

蛋白质序列分析

人工智能驱动的蛋白质序列模型

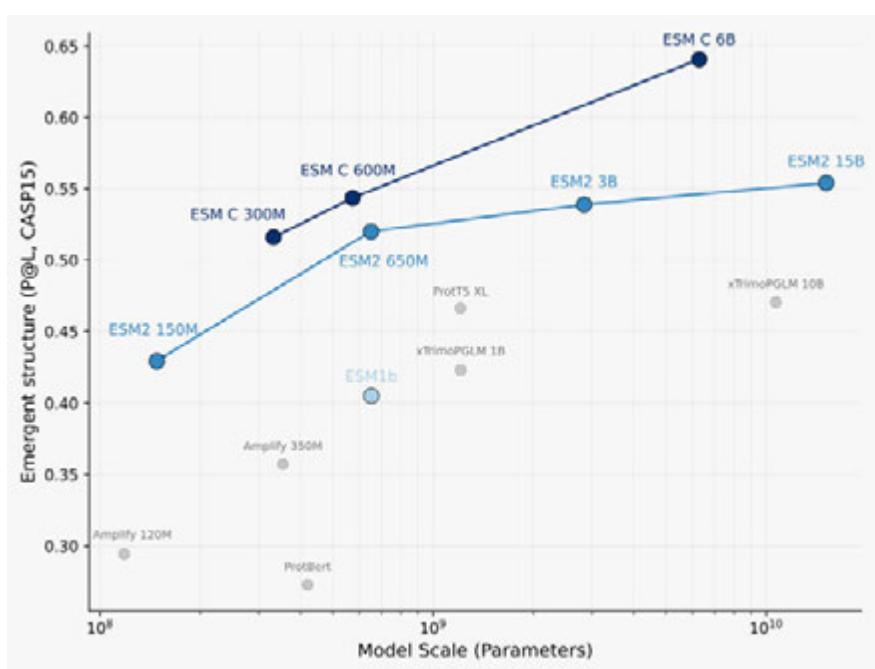
过去一年中，人工智能在蛋白质序列分析方面取得了显著进展。大规模的机器学习模型提升了我们对蛋白质属性的预测能力，推动了结构生物学与分子工程的研究进程。如前所述，多个具有代表性的蛋白质序列建模系统——包括 AlphaFold、ESM2 与 ESM3——已相继发布。

ESM3 模型融合了多模态输入，包括序列、结构与相互作用数据，其更大的参数规模也提升了模型的代表性与预测精度。随着 ESM 系列模型规模的扩大，其蛋白质预测性能不断提高。例如，2024 年发布的新一代模型 ESM C 在结构预测权威评估 (CASP15) 中表现出更高的预测准确率 (图 5.2.1)。

新兴结构预测成果，CASP15

资料来源：EvolutionaryScale, 2024

图 5.2.1



其他重大进展，如，ProGen，是一款生成式人工智能模型，已展示出设计功能性蛋白质序列的能力，凸显了人工智能辅助蛋白质工程的潜力。与此同时，基于 Transformer 架构的模型（如 ProtT5）利用深度学习技术，仅从序列数据中即可预测蛋白质功能与相互作用，推动了计算生物学的发展。图 5.2.2

展示了按发布时间排列的多个关键蛋白质序列模型及其参数规模。如上文所示，研究趋势正朝着训练数据集不断扩大的超大规模模型发展。这些人工智能驱动的方法正在重塑蛋白质科学，减少了对成本高昂且耗时的实验方法的依赖，使蛋白质功能与设计的探索更加高效。

2020–2024年蛋白质序列模型规模

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

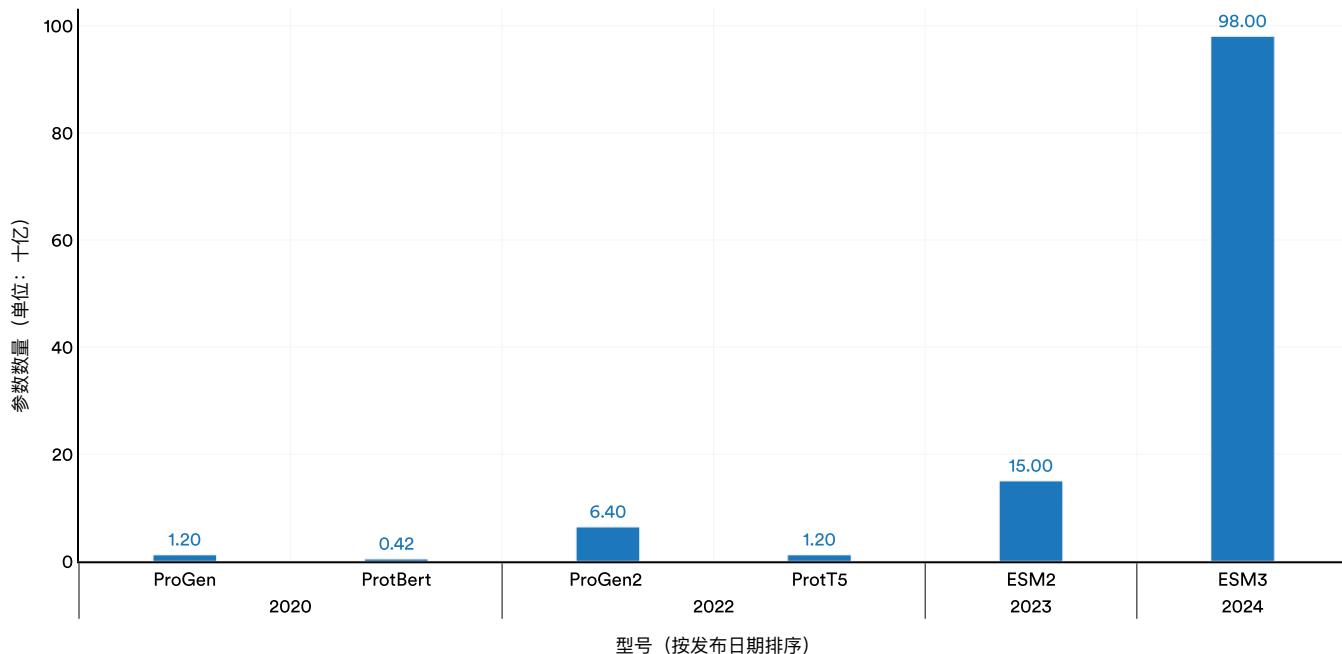


图 5.2.2

蛋白质科学公共数据库

公共数据库的扩展在蛋白质科学领域的人工智能应用中起到了关键作用。经过精细整理的大规模数据集，使人工

智能模型能够在多样化的生物序列上进行训练，从而增强其预测能力。图 5.2.3 列出了几个关键的蛋白质科学数据库及其发布日期。

关键蛋白质科学数据库

资料来源：2025年人工智能指数

数据集	发布日期	说明
<u>PDB</u>	1971	一个收录经实验解析的蛋白质结构数据库，是生命科学领域最早的开放数字资源。
<u>Pfam</u>	1995	包含蛋白质家族的全面数据库，提供基于隐马尔可夫模型生成的注释与多序列比对。
<u>STRING</u>	2000	提供蛋白质相互作用及其进化关系的重要信息资源。
<u>UniProt</u>	2002	目前仍是蛋白质序列与功能注释的黄金标准，人工智能辅助的校订提升了其准确性。
<u>PDBbind</u>	2004	PDB 的子集，包含蛋白质与生物分子复合物，如蛋白-配体、蛋白-蛋白、蛋白-核酸等。
<u>AlphaFold Database</u>	2021	结构生物学的重要资源，现已整合人工智能模型以预测缺失的实验数据。

图 5.2.3

随着时间的推移，多个蛋白质科学公共数据库的收录条目数量也在持续增长（见图 5.2.4）。人工智能所生成的蛋白质数据不断充实这些数据库，使其成为科研与产业领域不可或缺的工具。然而，保持数据质量与防止人工智能模型中的偏差，仍是持续面临的挑战。

2019–2025年公共蛋白质科学数据库的增长情况

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

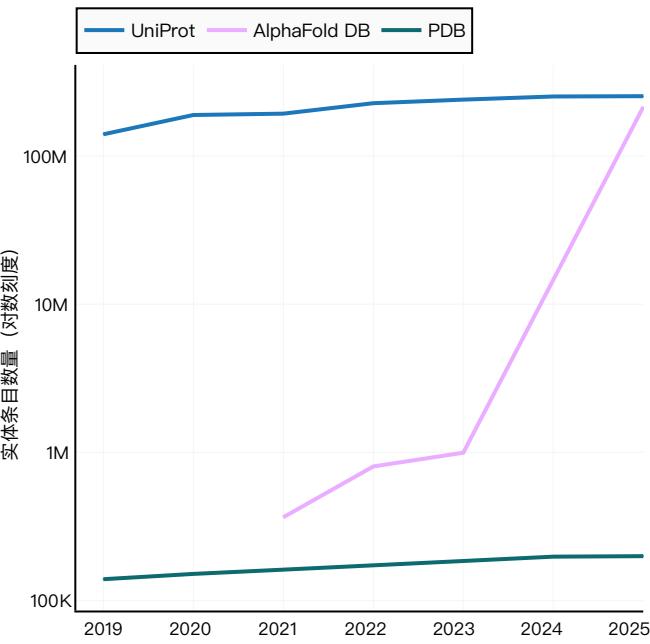


图 5.2.4

研究与论文发表趋势

人工智能驱动的蛋白质科学的论文统计

人工智能在蛋白质科学研究中的应用正在迅速扩大，这从过去一年 PubMed 与 bioRxiv 上人工智能驱动研究数量的上升趋势中可见一斑。相关研究涵盖多个关键领域。借助机器学习的发展，蛋白质结构预测变得更加高效，提供了更深层次的结构洞察。人工智能模型如今能够更为准确地从原始序列中推断生化功能，提升了功能预测能力。此外，研究人员还在开发能够预测蛋白质与药物相互作用的人工智能模型，甚至可以直接

从头设计能够靶向特定蛋白质的新药。这两项任务对药物发现与开发至关重要。此外，具有新型功能的人工智能生成蛋白质正在不断出现，尤其在酶工程与治疗性应用领域表现突出，标志着合成蛋白质设计迈出了重要一步。图 5.2.5 展示了 2024 年蛋白质科学领域中人工智能驱动研究在整个生命科学领域中的占比。研究最多的主题为功能预测(占 8.4%)，其次是蛋白质结构预测(占 7.6%)和蛋白质 - 药物相互作用(占 3.0%)。

2024年蛋白质科学中人工智能驱动研究在生命科学领域所占比例

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

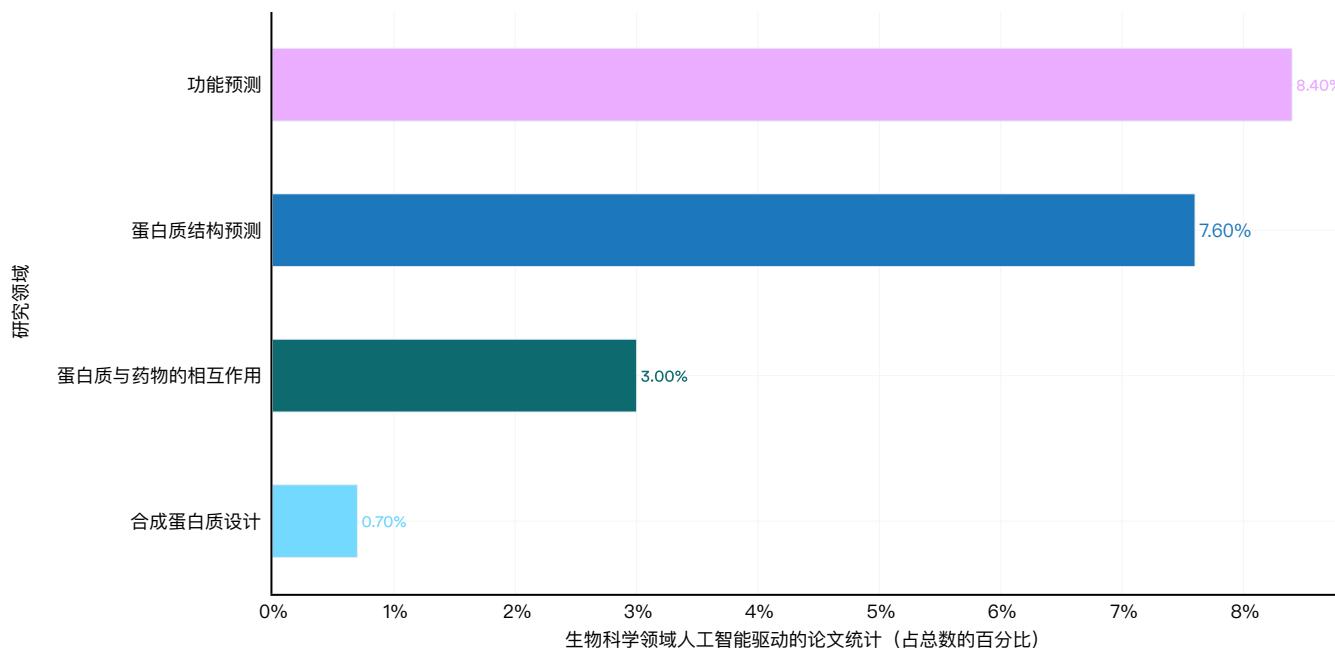


图 5.2.5

图像和多模态人工智能促进科学发现

冷冻电子显微镜、高通量荧光显微镜和整片切片成像的进展，使科学家能够以极高的精度观察和分析原子级、亚细胞级和组织级结构，从而揭示复杂生物过程中的新见解。为了实现这一目标，研究人员会结合现有的科学知识，对图像中的发现进行解读与情境化分析，以将观察结果与生物功能及疾病关联联系起来。随着高通量显微成像技术的兴起，研究重点日益聚焦于视觉模型、视觉 - 语言模型，以及更近期发展起来的视觉 - 组学基础模型的交叉领域。不同成像技术下的基础模型数量在逐年增长（图 5.2.6）。以光学成像为例，相关模型数量从 2023 年的 4 个增加到 2024 年的 8 个。而在 2023 年，电子显微镜与荧光显微镜方向尚无模型发布，但 2024 年这两个领域均各有 4 个模型问世。总体而言，随着数据的积累与公开，显微成像领域的基础模型数量正持续上升。

2023–2024 年不同显微镜技术下的基础模型数量

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

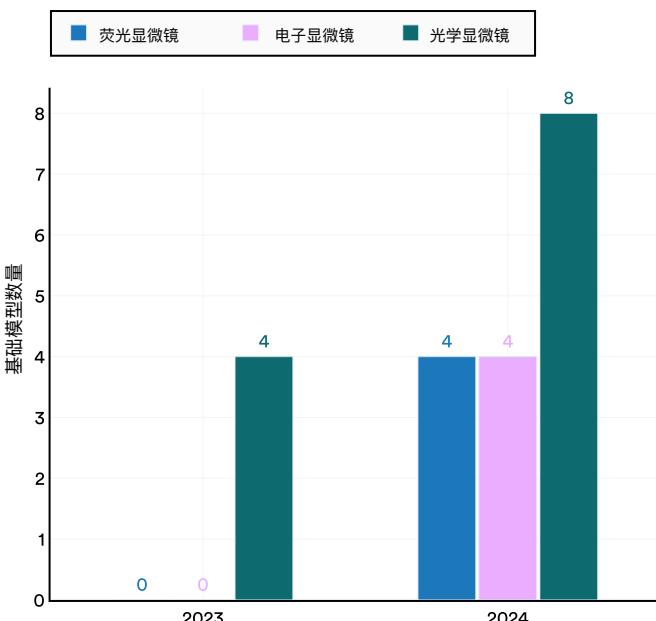


图 5.2.6

5.3 临床诊疗：影像学领域

数据：来源、类型和需求

医学影像中的人工智能发展迅速，正不断扩展至新的数据模态，并尝试解答日益复杂的临床问题。目前，美国食品药品监督管理局(FDA)所批准的机器学习软件中，有超过80%是用于医学图像分析。目前，人工智能主要应用于二维(2D)图像环境，此类环境中可有效使用传统图像处理架构，例如卷积神经网络(CNN)与Transformer。然而，尽管该领域已有多个成功案例，许多医学影像中的人工智能应用依旧严重依赖于有限的训练数据集。

以组织病理学为例，对患者活检样本进行染色分析是一项常规操作，但其中只有极少部分被数字化并对外公开。更少的数据集包含了所需的配对注释或组学数据，而这些恰是完成复杂分类任务所必需的。当前公开可用的组织病理学数据集通常

不超过10,000个患者样本。其中较为全面的数据资源是癌症基因组图谱(TCGA)，该资源共收录了11,125个患者样本，并配有临床注释、基因组测序和蛋白质表达数据，涵盖了32种癌症类型。因此，许多组织病理学人工智能模型训练数据不足1,000个样本，尤其在模型以基因组或蛋白质组数据作为标注标签时更是如此。训练样本有限将导致过拟合风险增加，并降低模型泛化能力。

图5.3.1展示了美国各州用于训练临床机器学习算法的患者队列分布。数据表明，大部分用于训练深度学习算法的患者数据集中在加州、马萨诸塞州与纽约州，这引发了关于数据集覆盖范围局限性的担忧。

2015–2019年美国各州用于训练临床机器学习算法的患者队列分布

资料来源：Kaushal等，2020 | 图表：2025年人工智能指数报告

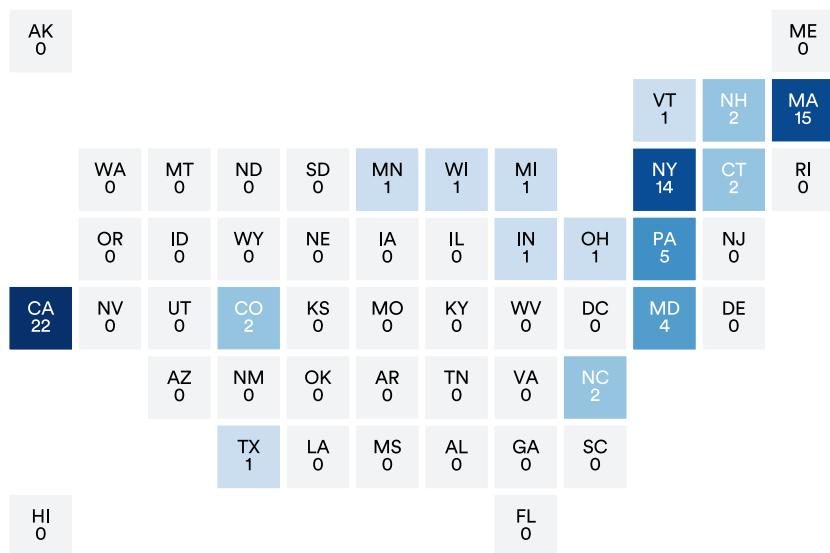


图5.3.1

第五章：科学与医学

5.3 临床诊疗：影像学领域

这些数据限制在三维（3D）医学影像领域表现得更加明显。尽管人工智能传统上主要聚焦于二维模态，例如胸部X光、组织病理切片和眼底摄影，但近年来，其应用范围已经扩展至三维成像模态，包括计算机断层扫描(CT)、磁共振成像(MRI)以及三维组织病理学分析。三维分析提供了更丰富的数据，使人工智能模型能够从体积结构与复杂表面中学习模式，这些信息在二维切片中往往难以察觉。尽管已经开发出一系列有前景的方法用于分析三维医学图像，但数据限制与实际需求依然存在。目前可公开使用的三维数据集仍非常有限。其中规模较大的包括英国生物样本库(UK Biobank，约100,000份MRI扫描)与癌症影像档案库TCIA(约50,000项研究)。尽管组织病理学中会常规收集三维样本，但三维成像并未成为标准操作，因此缺乏可公开访问的三维组织病理数据集。此外，标准化问题仍然突出，主要源于病理图像采集过程中的可变性。仪器设定、染色方法及机构间操作流程的差异会引入批次效应，而这些问题在训练数据有限的情况下会进一步加剧。

训练高精度的人工智能模型需要大量数据：卷积神经网络(CNN)通常在拥有约10,000张标注图像时表现良好，但Transformer模型则需要数量级更高的数据量。尽管诸如MIMIC-CXR(含377,000张影像)和CheXpert Plus(约226,000张正位放射图像，配有放射报告与患者元数据)等数据集十分重要，但其规模仍远小于ImageNet(约1,400万张图像)。数据完整性与偏差仍是关键挑战。

图5.3.2展示了不同主流医学语言模型与图像模型的训练数据词元体量，并与通用文本与图像模型进行对比。例如，GatorTron是一款面向电子健康档案中非结构化患者信息提取的大型临床语言模型，其训练语料包含820亿个词元；相比之下，Llama 3的训练量达到15万亿个词元，约为前者的182倍。在图像模型方面，RadImageNet是一个开放的放射学深度学习研究数据集，包含相当于1,600万个图像词元；而OpenAI早期图像生成器DALL-E的训练量约为60亿词元，相当于前者的375倍。

医学与非医学语言和图像模型训练所用数据集的词元数量

资料来源：RAISE Health, 2025 | 图表：2025年人工智能指数报告

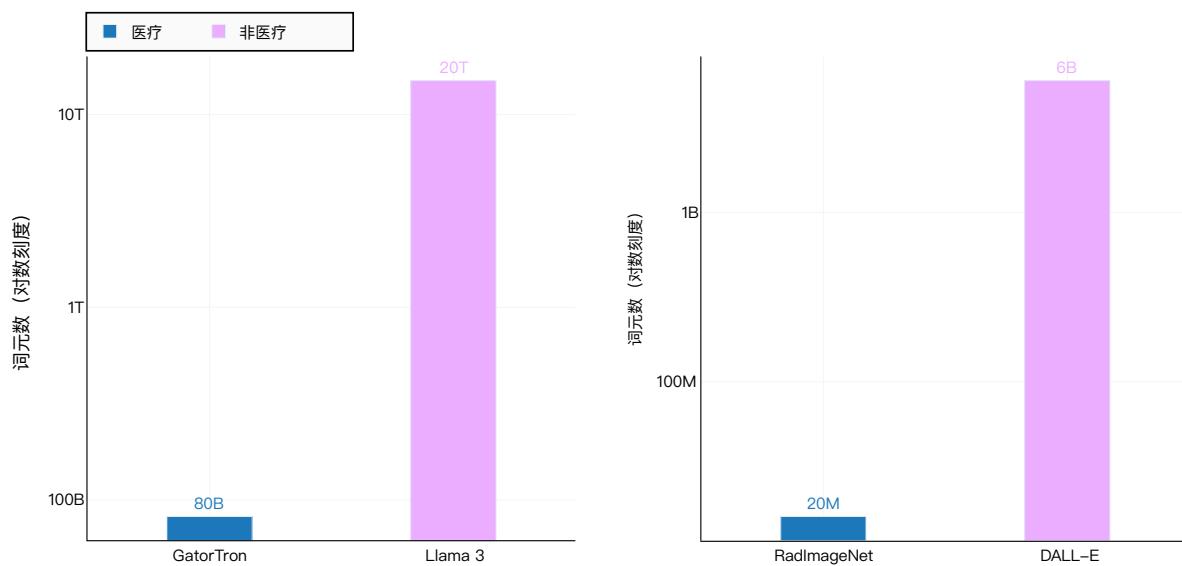


图 5.3.2

纵向影像数据对于建模疾病进展具有重要意义，但目前仍明显不足。例如 ADNI 项目（阿尔茨海默病神经影像计划）涵盖约 2,000 名参与者、研究跨度超过 15 年，是该类研究的典型代表。然而，可扩展的多模态纵向数据集仍非常罕见。填补这些空白，需要以下措施的结合：注重隐私保护的数据共享策略（如联邦学习 (federal learning)）、合成数据生成技术以及

更完善的注释策略。为了训练和验证稳健的医学影像人工智能模型，必须构建更大规模、更全面、多队列来源的训练数据集。提高高质量、带标签数据的可用性，有望提升模型性能。同时，改善验证实践也将增强对模型的信心，促进其更顺利地应用于临床实践中。

先进建模方法

图 5.3.3 展示了主流的临床医学影像建模方法、每种方法下的重要模型发布，以及各自面临的主要挑战。

医学影像建模方法与代表性人工智能模型

资料来源：2025年人工智能指数

建模方法	代表性模型发布	优势	挑战
扩散模型	1. RoentGen (2022) 2. RNA-CDM (2023) 3. XReal (2024)	生成用于训练、隐私保护与病理特异性增强的合成医学图像，在稳定性与多样性方面优于 GAN 模型	数据集偏差、幻觉性伪影、诊断不确定性
大型视觉-语言模型 (LVLMs)	1. CheXagent (2024) 2. Merlin (2024) 3. Med-Gemini (2024) 4. PathChat (2024) 5. TITAN (2024) 6. PRISM (2025) 7. BiomedParse (2025)	融合医学图像与文本，实现更优的诊断、分割与报告自动生成，扩展多模态能力	数据稀缺、对低资源环境的泛化能力不足、计算需求高
纯二维视觉基础模型	1. CTransPath (2022) 2. Virchow (2024) 3. UNI (2024) 4. MedSAM(2024)	可用于泛癌检测、生物标志物预测与图像分割，减轻注释负担	领域泛化能力弱、跨模态适应能力有限
多尺度 / 切片级模型	1. HIPT (2022) 2. MEGT (2023) 3. MG-Trans (2023) 4. HIGT (2023) 5. Prov-GigaPath (2024)	通过分层 Transformer 与图神经网络增强整片切片图像分析，提升诊断精度与可解释性	扩展性、计算效率与数据集变异性方面存在挑战

图 5.3.3

第五章：科学与医学

5.3 临床诊疗：影像学领域

近年来，基础模型在医学影像领域的应用显著增长。图 5.3.4 按照医学专科对代表性模型进行了分类。值得注意的是，病理学方向的新发布模型数量显著增加，成为增长最为集中的领域之一。

医学各专科领域与代表性人工智能模型

资料来源：2025年人工智能指数

医学专科	代表性发布模型
超声心动图学	1. EchoCLIP (2024)
肿瘤学	1. MUSK (2025)
眼科学	1. RETFound (2023) 2. VisionFM (2024)
病理学	1. CTransPath (2022) 2. CHIEF (2024) 3. Prov-GigaPath (2024) 4. PathChat (2024) 5. TITAN (2024) 6. Virchow (2024) 7. UNI (2024)
放射科	1. RoentGen (2022) 2. CheXagent (2024) 3. Merlin (2024) 4. PRISM (2025)

图 5.3.4

5.4 临床诊疗：非影像领域

临床知识

本节探讨大语言模型及近期人工智能模型在关键医学知识基准测试中的表现。

MedQA

评估人工智能模型的临床知识表现，需衡量其医学专业水平，尤其是可应用于临床场景的知识。

MedQA 是一项于 2020 年推出的综合性数据集，源自职业医学考试，包含超过 60,000 道临床问题，旨在挑战医生水

平。人工智能在 MedQA 基准测试上的表现已显著提升。微软与 OpenAI 的研究团队近期对模型 o1 进行了测试，其取得了 96.0% 的最新最优得分，相较 2023 年创下的记录提高了 5.8 个百分点（图 5.4.1）。自 2022 年底以来，该基准测试的表现已累计提升 28.4 个百分点。如第二章所述的其他通用知识基准测试一样，MedQA 可能正接近饱和状态，这意味着有必要开发更具挑战性的评估体系。

MedQA：测试准确率

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

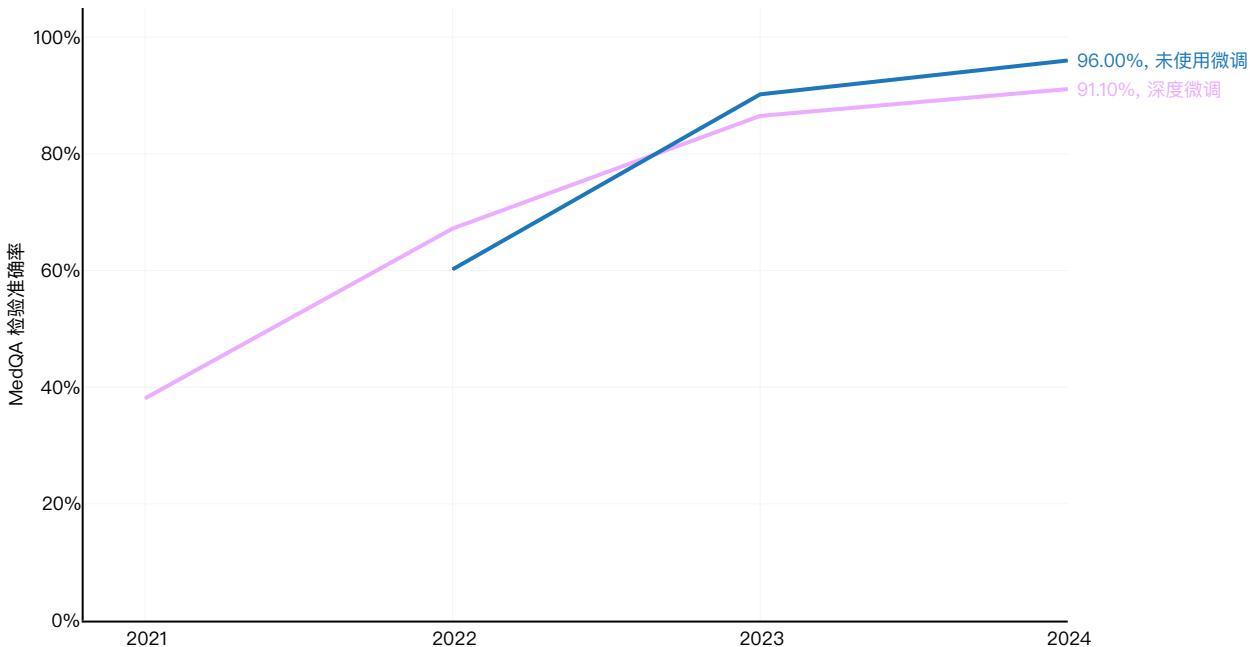


图 5.4.1

第五章：科学与医学

5.4 临床诊疗：非影像领域

重点：

人工智能医生和成本效益考虑

有研究者指出，评估医学领域的大语言模型不应仅依赖 MedQA，应采用涵盖更广泛医学子领域的基准测试。虽然 MedQA 具有一定价值，但若单独使用，可能无法反映真实临床应用中的复杂性。相比之下，采用多项基准可带来更强的临床相关性与更稳健的模型表现评估。

2024 年，加州大学圣克鲁兹分校、爱丁堡大学与美国国立卫生研究院合作开展了一项更具广度的人工智能医学系统测试研究。该研究评估了五个领先的大语言模型，包括最新开发的 o1 模型（具备“链式推理”功能）。其余模型包括 GPT-3.5、Llama 3-8B、GPT-4，以及专门构建的医学模型 Meditron-70B。这些模型在 19 个医学数据集上进行了测试，任务涵盖概念识别、文本摘要、基于知识的问答、临床决策支持与医学计算等内容。图 5.4.2 展示了五个大语言模型在所有数据集上的平均表现。结果显示，大语言模型的临床知识能力正持续提升，尤其是如 o1 这类具备实时推理能力的新模型。尽管进展显著，但挑战依然存在，包括幻觉问题与多语种表现不一致等。

此前的研究成果（已在去年的人工智能指数中引用）表明，像 Medprompt 这样的提示策略可以在无需额外微调的情况下，显著提升大语言模型在医学基准测试中的表现。OpenAI 最新发布的 o1 模型借鉴了上述策略中的一些洞见，通过在生成最终回答前引入运行时推理机制，从而增强了模型性能。研究人员发现，即便不使用专门的提示工程技术，o1 在医学任务中的表现依然优于结合 Medprompt 的 GPT-4 系列模型。然而，他们的分析也强调了 o1 所面临的准确率与成本之间的权衡：尽管其在 MedQA 基准上的得分比使用 Medprompt 的 GPT-4 Turbo 高出 5.8 个百分点，但在计算资源方面，o1 的成本却大约高出 1.5 倍。如图 5.4.3 所示，该基准测试中的成本与精度呈明显权衡关系。这一现象为临床场景生成式人工智能功能部署的医疗从业者提出了关键考量：必须在性能提升与计算成本之间寻求平衡。

选定的大语言模型在医学数据集上的性能表现

资料来源：Xie 等，2024 | 图表：2025 年人工智能指数报告

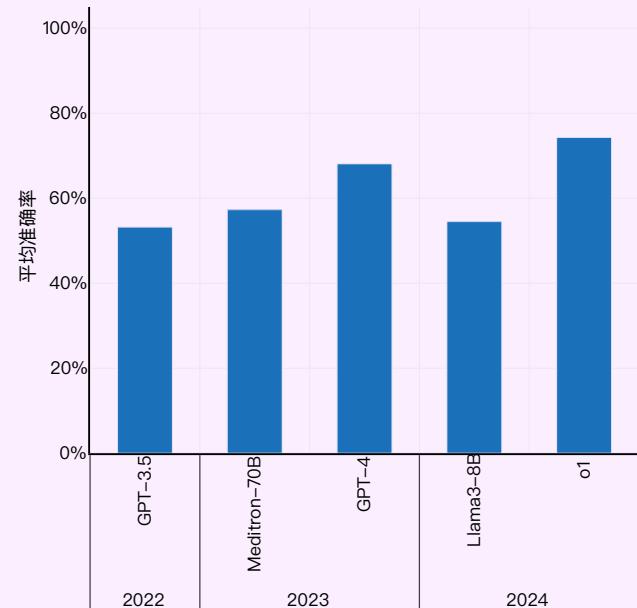


图 5.4.2

增强型帕累托前沿：准确率与成本

资料来源：Nori 等，2024

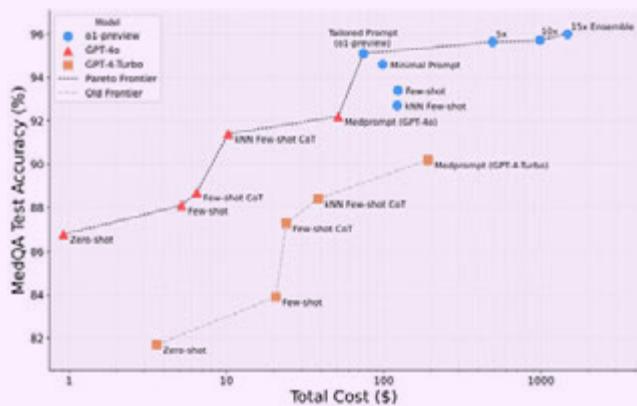


图 5.4.3

第五章：科学与医学

5.4 临床诊疗：非影像领域

医疗大语言模型性能评估

概述

近年来，对大语言模型在医疗任务中表现的评估兴趣急剧上升。在 PubMed 数据库中搜索“大语言模型”一词，共检索到 1,566 篇论文，其中仅在 2024 年就发表了 1,210 篇（图 5.4.4）。

2019–2024 年PubMed 收录的有关大语言模型的论文数量

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

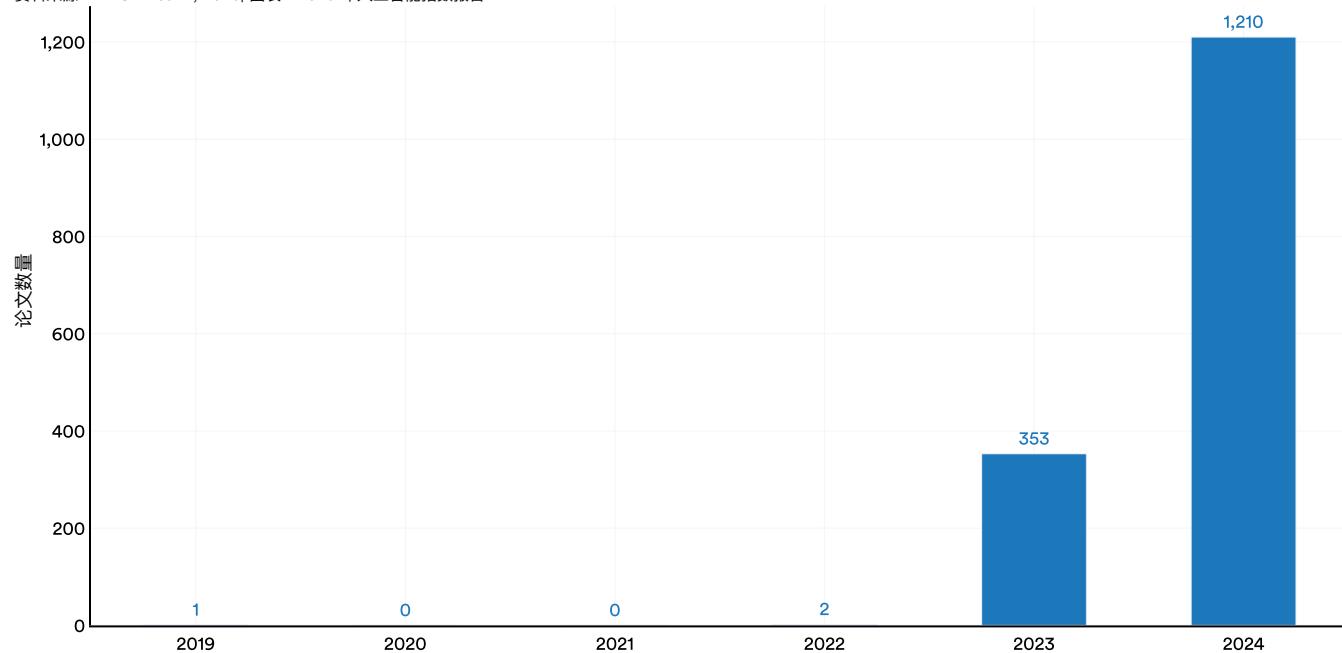


图 5.4.4

第五章：科学与医学

5.4 临床诊疗：非影像领域

2024 年初开展的一项系统性综述识别出超过 500 篇论文，内容聚焦于评估自然语言处理(NLP)在医疗任务中的性能表现，且重点集中于医疗决策支持领域(图 5.4.5)。大多数评

估医疗 NLP 系统性能的研究，集中于两类任务：医学知识增强任务(419 篇)和诊断辅助任务(178 篇)。

医疗任务、NLP 与 NLU 任务，以及 519 篇研究中的评估维度

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

	准确性	全面性	事实性	鲁棒性	公平性、偏见与有害性评估	部署指标	校准与不确定性
任务类别	增强医学知识	222	91	44	33	16	10
	进行诊断	100	38	11	11	14	4
	向患者传达信息	88	68	32	22	18	3
	提出治疗建议	47	22	9	8	3	1
	与患者沟通	35	29	8	15	22	1
	医疗协调与规划	36	24	5	5	7	1
	分诊患者	24	7	5	2	8	0
	开展文献综述	18	7	3	2	2	0
	综合研究数据	16	7	2	3	2	0
	生成医学报告	8	8	2	0	3	0
	进行医学研究	8	7	3	3	3	0
	提供异步护理	8	5	3	3	1	1
	管理临床知识	5	5	1	1	0	0
	临床笔记记录	6	2	1	1	0	0
	生成转诊建议	3	0	0	0	0	0
	优化手术操作	3	3	1	1	0	0
	生物医学数据挖掘	2	0	0	0	0	0
	生成账单编码	1	0	0	0	0	0
	开具处方	1	0	0	0	0	0
NLP/NLU 任务	问答任务 (*)	398	194	71	61	54	14
	文本分类 (*)	29	10	6	5	10	2
	信息抽取 (*)	29	12	8	5	4	6
	摘要生成 (*)	29	21	7	3	8	0
	对话交互 (*)	6	6	1	1	5	1
	翻译 (*)	5	1	2	2	1	0

图 5.4.5⁴

4、其中标有星号 (*) 的任务为自然语言处理 (NLP) 或自然语言理解 (NLU) 任务。

第五章：科学与医学

5.4 临床诊疗：非影像领域

大语言模型在临床诊断推理中的应用

诊断错误导致大量患者受到伤害，众多机构正在探索将人工智能作为改进诊断流程的工具。

重点：

大语言模型对诊断推理的影响

2024 年，一项单盲、随机对照实验测试了在处理复杂临床病例时，GPT-4 的辅助效果与传统医疗资源之间的对比。这项研究共涉及 50 位拥有美国执照的执业医生，评估的核心问题是：人工智能辅助的决策制定是否能提升医生的诊断准确性与效率。研究结果显示，在使用 GPT-4 辅助的情况下，医生的整体表现并未显著优于仅依赖传统工具的医生群体。事实上，使用人工智能辅助的医生诊断准确率为 76%，而使用传统工具的医生为 74%，仅有微弱提升。然而，在一项二次分析中，研究者发现单独使用 GPT-4 模型的表现超过了所有医生群体，其在诊断推理得分上达到了 92%，比不使用人工智能的医生群体高出 16 个百分点（图 5.4.6）。尽管人工智能模型在独立任务中表现出色，将其集成进实际临床工作流程中却面临挑战。在时间效率方面，各组医生完成病例的时间无显著差异，这表明引入大语言模型的临床工作流仍未带来效率优势。

即便人工智能模型在独立测试中表现优异，仅仅让医生使用大型语言模型，并不足以提升他们的诊疗表现。这一现象也出现在其他人工智能与人类协作场景中，提示我们：要实现模型性能与临床专业人员之间的真正协同，需重新设计工作流程、改进用户培训与人机交互界面。

大语言模型在临床诊断方面的性能表现

资料来源：Goh 等, 2024 | 图表：2025 年人工智能指数报告

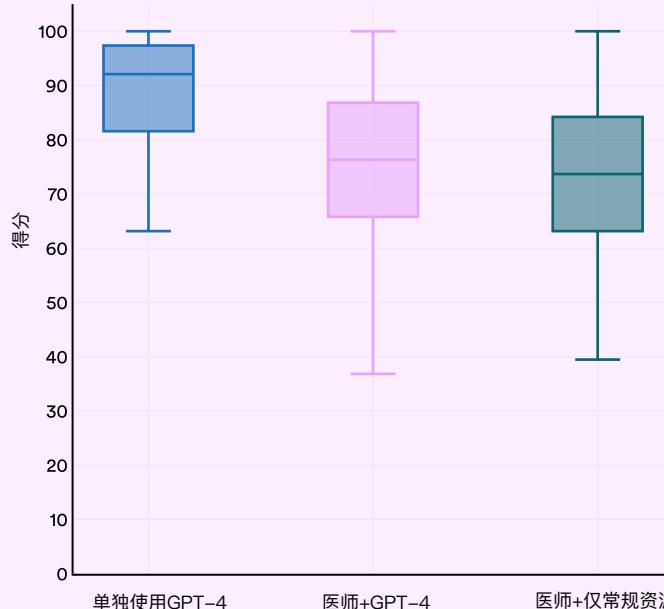


图 5.4.6

管理性推理和患者护理决策

除了诊断本身，医生还必须处理治疗方案选择、风险与获益权衡以及患者偏好等多维度决策任务，统称为“管理性

推理（management reasoning）”。研究人员正在探索大型语言模型是否能够提升这些复杂、情境依赖的推理能力。

重点：

GPT-4 辅助临床管理任务的效果评估

2024–2025 年间，一项前瞻性、随机、对照试验评估了 GPT-4 在复杂临床管理决策中的辅助效果。研究共涉及 92 名执业医生，其中一半使用 GPT-4 辅助并结合标准资源，另一半仅依赖传统医学参考资料。结果显示：使用 GPT-4 的医生在任务表现上平均领先对照组约 6.5 个百分点（见图 5.4.7）。有趣的是，仅使用 GPT-4 的表现与 GPT-4 辅助医生组相当，表明在某些定义明确的场景中，接近自主的人工智能管理支持可能具有可行性。然而，引入人工智能辅助也带来权衡：使用 GPT-4 的医生在每个案例上耗时略长。研究人员将此归因于医生在决策过程中进行更深入的思考与分析。总体而言，生成式人工智能可以在临床决策中带来实质性改进，但其影响可能更多体现在决策质量上，而非单纯提升效率。

大语言模型在临床诊断方面的性能表现

资料来源：Goh 等, 2024 | 图表：2025 年人工智能指数报告

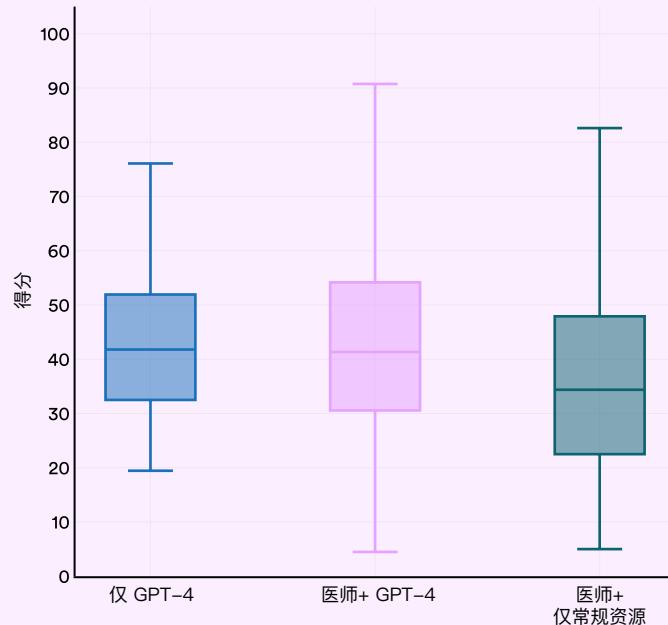


图 5.4.7

第五章：科学与医学

5.4 临床诊疗：非影像领域

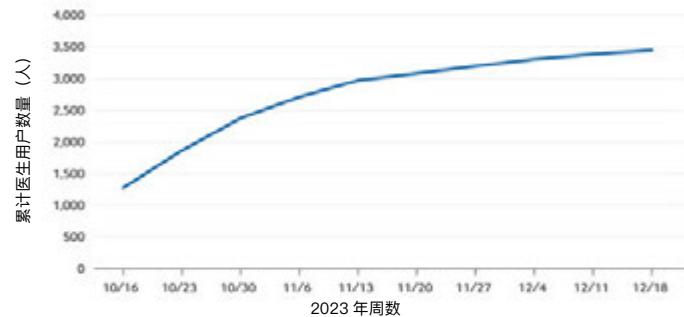
环境型人工智能语音助手

临床文档工作长期以来是医生负担与倦怠的主要来源之一。环境型语音记录技术正迅速演化，并将大语言模型整合进医生与患者交流的处理流程中。第一项相关研究发表在《NEJM Catalyst》期刊，描述了 2023 年底环境型人工智能语音助手技术在 Kaiser Permanente Northern California 的部署。这项技术在试点结束前已被数千名临床医生采用（图 5.4.8）。紧随其后的是第二项研究，发表于《JAMIA》期刊，介绍了该技术在 Intermountain Health 的试点经验。值得注意的是，这两项研究所评估的都是早期版本的技术，这些版本尚未完全实现自动化或与电子健康档案系统（EHR）集成。

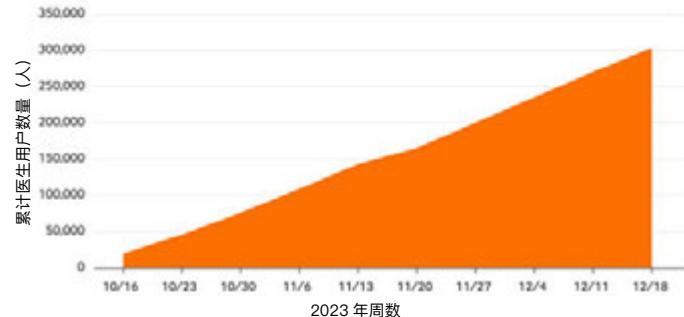
2023 年 10 月 16 日至 12 月 24 日环境型人工智能语音助手工具的累计使用情况

在 2023 年 10 月 16 日上线至 12 月 24 日期间，共有 3,442 名独立注册医生与员工用户启用该人工智能语音助手（见 Panel A）；期间累计记录 303,266 次医生 – 患者交流（见 Panel B），这些交流中均启用了语音助手功能，且每次交流持续时间不少于 2 分钟。

Panel A 为累计启用人工智能语音助手的独立医生人数



Panel B 为累计人工智能语音助手服务次数



资料来源：Tierney 等, 2024

图 5.4.8

第五章：科学与医学

5.4 临床诊疗：非影像领域

斯坦福大学的研究人员开展了一项两阶段研究，用于评估环境型人工智能语音助手技术的效果。该研究在以往工作的基础上，测试了一种全流程集成、自动化的人工智能医疗文书系统。结果显示，该系统在客观指标（如文书记录所需时间）与主观指标（如医生的工作体验）方面均取得了明显改善。技术采纳情况良好，在参与的医生中平均采纳率达到了 55%。人工智能

语音助手带来了显著的效率提升，每条记录平均节省医生约 30 秒，整体电子健康档案（EHR）操作时间每日减少约 20 分钟（图 5.4.9）。此外，医生普遍报告称工作负担与职业倦怠显著下降，平均降幅分别为 35% 和 26%。这些结果表明，由人工智能驱动的语音助手技术能够切实改善医生的工作流程与福祉，不仅节省了时间，还缓解了繁重的行政负担。

人工智能记录对医生使用电子健康档案（EHR）的影响

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

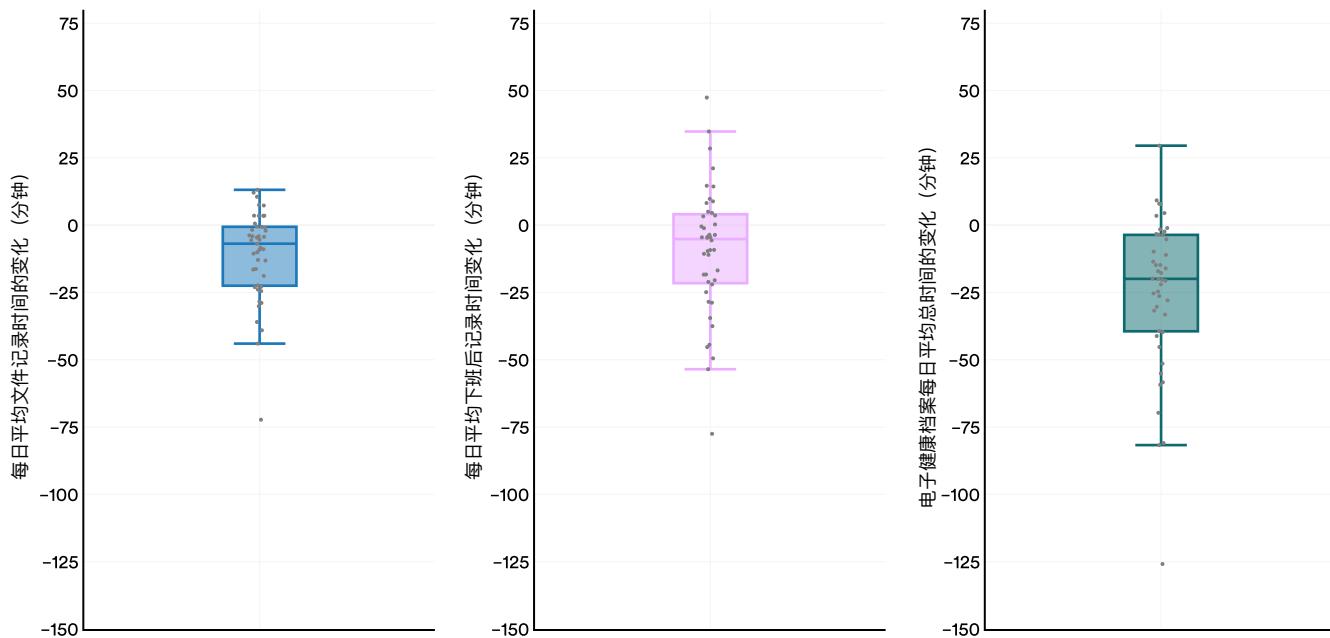


图 5.4.9

据报道，到 2024 年，对环境型语音助手技术的投资将近 3 亿美元。虽然当前的技术应用主要集中在临床文档撰写这一起点，但研究与产业界的乐观人士预期，该类技术将在未来

实现全面部署，覆盖门诊与住院场景，最终实现对医嘱下达、账单与编码、实时临床决策支持等流程的自动化辅助。

第五章：科学与医学

5.4 临床诊疗：非影像领域

部署、实施与撤除

FDA 对人工智能医疗设备的批准

人工智能在临床环境中的应用在过去十年中呈指数级增长，尤其体现在美国食品药品监督管理局(FDA)批准的人工智能医疗设备数量激增。

FDA 于 1995 年首次批准一款具备人工智能功能的医疗设备。此后将近 20 年的时间里，每年的批准数量一直保持在个位数。直到 2015 年，当年共有 6 款人工智能医疗设备获得批准。自此之后，年度批准数量开始迅猛增长，并在 2023 年达到峰值——223 件(图 5.4.10)。

1995–2023年FDA批准的人工智能医疗设备数量

资料来源：FDA, 2024 | 图表：2025 年人工智能指数报告

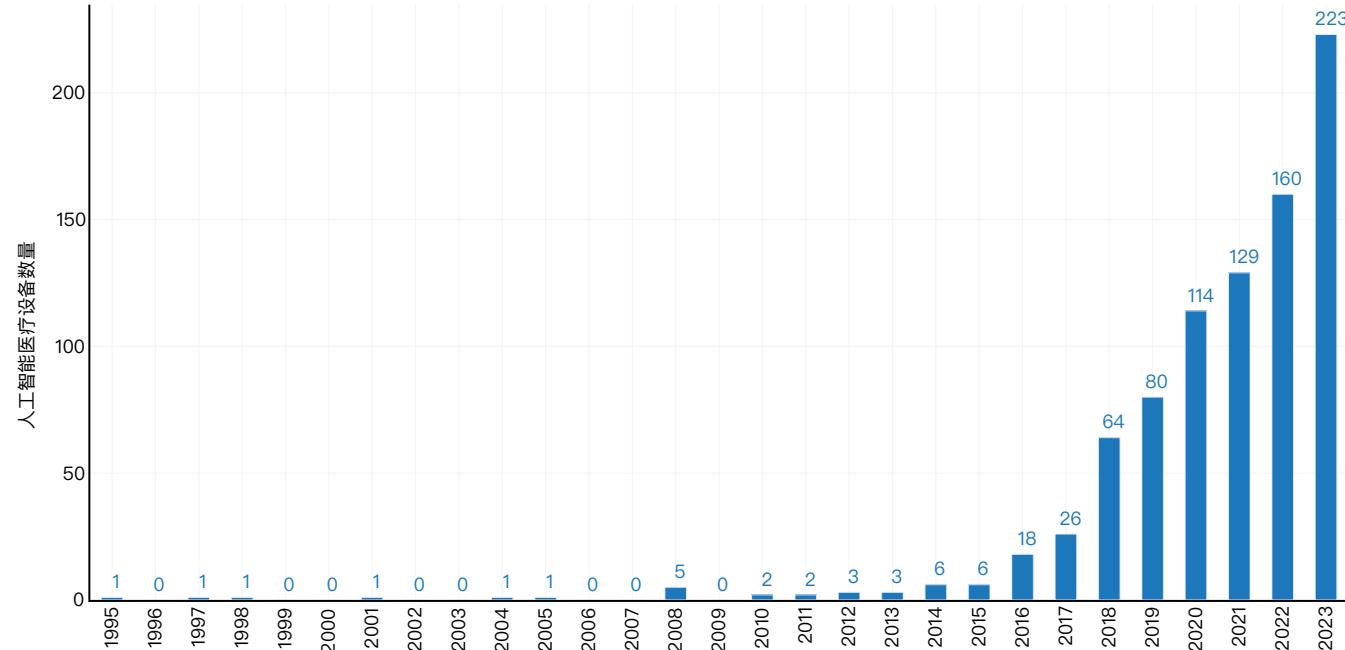


图 5.4.10

成功应用案例：斯坦福医疗系统 (Stanford Health Care)

在临床实践中成功生成式人工智能功能部署模型，需依托一套确保其公平性、实用性与可靠性的系统框架。斯坦福医疗系统在评估与生成式人工智能功能部署工具时，采用其内部开发的 FURM 框架(公平(Fair)、实用(Useful)、可靠(Reliable)、

可量化(Measurable))。在已评估的 6 个人工智能应用案例中，有 2 个已实现成功落地部署：(1) 外周动脉疾病(PAD)筛查；(2) 住院患者的文档记录与编码改进。本节将详细说明外周动脉疾病筛查的应用。

第五章：科学与医学

5.4 临床诊疗：非影像领域

外周动脉疾病筛查

外周动脉疾病是一种慢性血管性疾病，早期常被忽视，从而导致严重并发症，如肢体严重缺血甚至截肢。为提升早期检测与干预能力，斯坦福医疗系统开发并部署了一套具备人工智能功能的 PAD 分类模型，用于优化筛查流程并改善患者治疗。

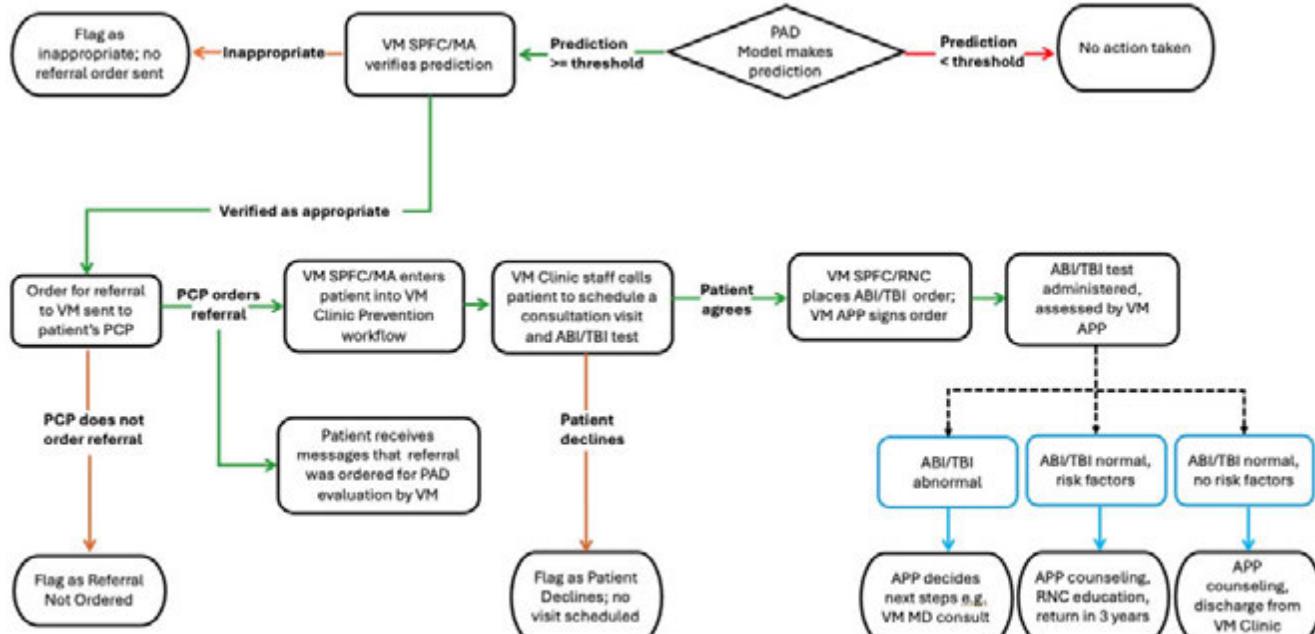
这项筛查工具的主要目标是：在初级保健人群中实现更早期的诊断，以便在病情恶化前采取医疗或手术干预。通过识别高风险患者，该模型还能帮助优化资源分配，确保最需要

干预的人群能够立即获得随访与治疗。

为实现与临床工作流程的无缝集成，该人工智能工具被设计为可自动评估外周动脉疾病（PAD）风险，并对高风险个体进行标记提示，以便进一步评估。一旦病情被确认，患者将转诊至血管外科接受会诊。图 5.4.11 展示了将 PAD 筛查整合进临床工作流程的建议模型与操作路径，包括风险评估、转诊流程以及患者随访等关键环节。

将PAD筛查整合入临床实践的建议模型与工作流程

资料来源：Callahan 等, 2024



ABI/TBI – ankle/toe brachial index; APP – advanced practice provider; MA – medical assistant; RNC – registered nurse coordinator; SPFC – specialty patient flow coordinator; VM – vascular medicine

图 5.4.11

该人工智能工具在完成试点阶段后，进入第 2 阶段（Stage 2），并已在斯坦福医疗系统全面部署。模型预计每年将影响约 1,400 名患者。除去显著的临床效益外，该项目还证明具备财务可持续性，可在无外部资金支持下正常运行。

通过提升早期发现率、降低严重并发症风险，并改善患者预后，该人工智能驱动的策略正逐步重塑外周动脉疾病的标准化治疗路径。

健康社会决定因素

大语言模型与基于人工智能的临床决策支持（CDS）系统正在推动医学实践的变革，但在不同专科的采用程度存在差异。有些专科快速拥抱这些工具，而另一些则持谨慎态度。本节回顾了相关研究与创新，重点强调证据基础的重要性。其中一个核心方向是社会健康决定因素（Social Determinants of Health, SDoh），如社会经济地位与生活环境等。2024年，人工智能的进步集中应用于SDoh领域，以改善患者护理与促进健康公平。

多个模型在放射治疗任务中识别 SDoh 的表现

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

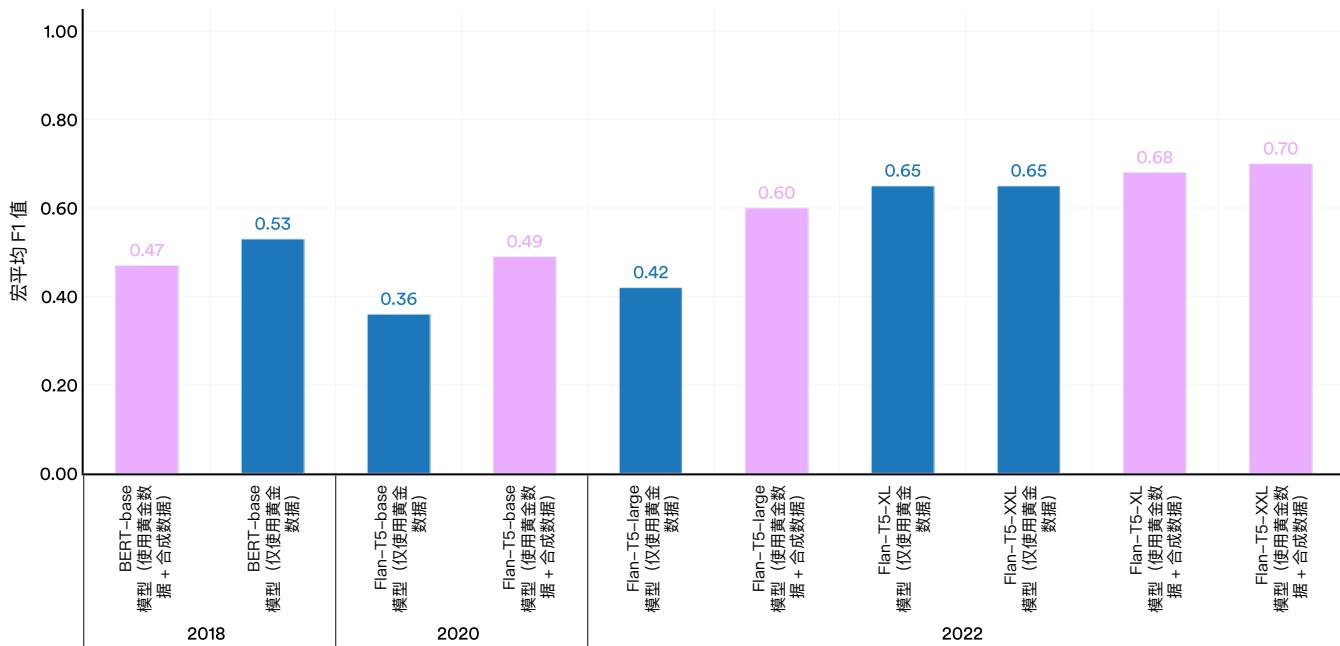


图 5.4.12

从电子健康档案中提取 SDoh 数据能帮助医生识别患者的社会需求（如住房不稳定或食物短缺）。这类研究凸显了大

从电子健康档案和临床记录中提取 SDoh

经过微调的多标签分类器（如 Flan-T5 XL）在识别临床笔记中的 SDoh 信息方面，表现优于 ChatGPT 系列模型，且对人口学描述的敏感性更低。这类模型在引入种族、族裔或性别变量时展现出更低的偏差。图 5.4.12 展示了多个模型在放射治疗（RT）测试集上识别 SDoh 的表现。新一代模型（如 Flan-T5-XXL），在加入合成与标注数据（SDoh 标签句）后，性能表现最佳。总体来看，模型规模扩大与数据融合优化显著提升了 SDoh 识别能力。

语言模型提升 SDoh 记录质量、资源配置效率与健康公平性的潜力，同时也强调了减少偏差与强化合成数据方法的重要性。

第五章：科学与医学

5.4 临床诊疗：非影像领域

医疗领域的人工智能应用与 SDoH 的整合

图 5.4.13 重点介绍了各个医学专科领域，并说明了人工智能如何将健康的社会决定因素纳入各个领域。

专科领域	最新研究	整合描述
肿瘤学	stasy等, 2024	人工智能工具被用于将 SDoH 纳入癌症治疗计划中，例如考虑患者就医便利性与社会支持情况，帮助肿瘤科医生制定个性化、可行的治疗策略
心脏病学	Snowdon等, 2023 Quer等, 2024	心脏病人工智能模型已开始融合 SDoH，用于提升对高血压、心力衰竭等疾病的风险评估精度，优化治疗管理。
精神病学	Stade等, 2024	大语言模型被用于分析社区层面的 SDoH 数据，有助于识别社会风险因子集中的地区，从而优先部署心理健康干预资源。

图 5.4.13

合成数据

合成数据正通过提升隐私保护型分析、临床建模与人工智能训练，彻底改变医疗健康领域。它优化了工作流程，能够模拟罕见病例，并支持以人工智能为驱动的创新实践。然而，正如本年度人工智能指数第一章所指出的，关于其可扩展性的担忧，促使我们在采用过程中必须保持谨慎。

临床风险预测

近期研究验证了合成数据在隐私保护临床风险预测中的价值。一项近期研究验证了合成数据在隐私保护临床风险预测

中的有效性。研究人员利用 ADGAN、PATEGAN 与 DPGAN 三种生成模型，在英国生物样本库(UK Biobank)中，对曾经吸烟者(ever-smokers)群体的肺癌风险进行了建模。⁵ 图 5.4.14 展示了主成分分析(PCA)特征值的对比结果，表明 ADGAN 与 PATEGAN 生成的数据分布与真实数据高度一致，从而能够支持可靠的聚类与特征选择。这些研究结果表明，合成数据集在不依赖真实且可识别的患者信息的前提下，能够保留统计特征的真实性，支持探索性分析，并可用于开发预测模型。

主成分分析

资料来源资料来源：Qian 等, 2024

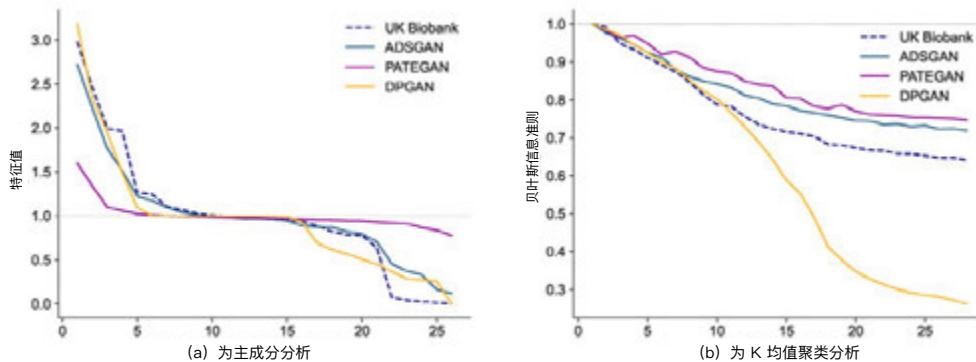


图 5.4.14

5、常吸烟者是指一生中至少吸过 100 支烟的人。

第五章：科学与医学

5.4 临床诊疗：非影像领域

药物发现

《自然》期刊近期发表的一项研究，提出了一种生成式人工智能方法，用于在药物研发中实现体外配方优化与粒子工程建模。该方法通过受关键质量属性 (critical quality attributes, CQAs) 指导的图像生成器，生成可用于分析的数字药物配方，从而减少对大规模物理实验的依赖。研究团队通过预测微晶纤维素 (MCC) 在口服片剂中的渗流阈值 (percolation threshold)，验证了该方法的有效性。图 5.4.15 将真实片剂体积的曲折率计算结果 (绿色方块) 与人工智能合成体积的结果 (红色圆圈) 进行了对比。⁶ 两者高度一致的结果表明：合成数据在模拟药物特性方面展现出巨大潜力，能够提升人工智能驱动的药物发现效率与建模能力。

数据生成平台

平台对于展示、标准化和自动化合生成数据的创建是必不可少的。最新发表的研究表明，通过其提出的合成表格神经生成器 (STNG) 框架，大规模合成数据生成与验证不仅可行，还能显著增强医学领域人工智能应用的效果。图 5.4.16 通过对比真实与合成心脏病数据集的曲线下面积 (AUC) 值，评估了不同合成数据生成方法的有效性。在多数情况下，真实数据集与合成数据集之间存在高度吻合，这证明合成数据能够精准建模复杂的健康状况。合成数据生成方法的进步可在提升数据保真度的同时，有效降低隐私风险。

基于人工智能生成合成结构的渗滤阈值预测与验证

资料来源：Hornick 等, 2024

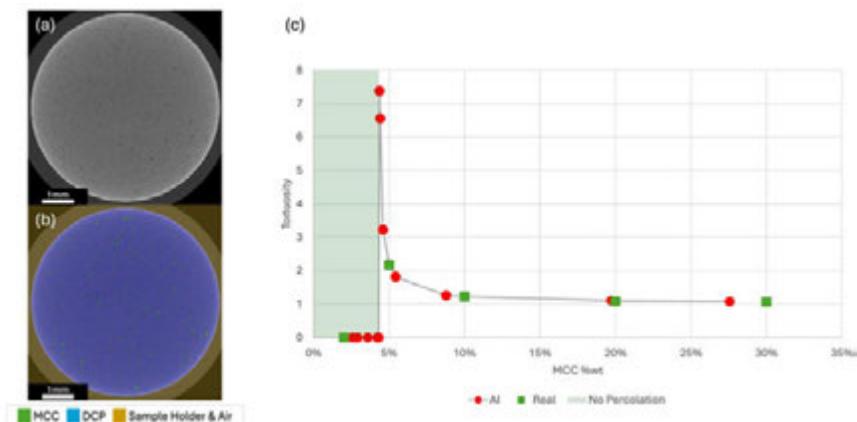


图 5.4.15

用于评估合成心脏病数据集的曲线下面积

资料来源：Rashidi 等, 2024

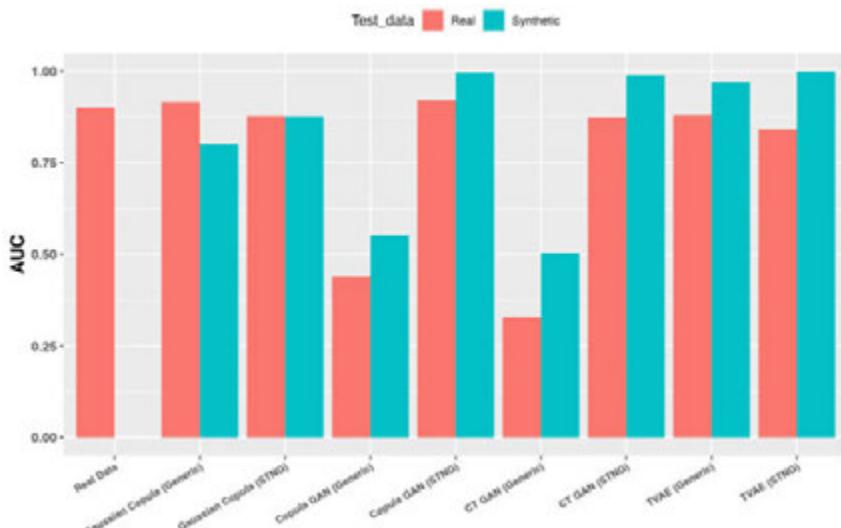


图 5.4.16

6、曲折度是指与两点之间可能的最短直线距离相比，路径的曲折程度。

电子健康档案系统

将人工智能集成进电子健康档案(EHR)系统，可通过简化管理流程、增强临床决策支持与改善患者护理质量，从而缓解医疗系统负担。目前，EHR市场由几家主要厂商主导，包括：Epic、Oracle Health(前身Cerner)、Meditech与TruBridge(前身为CPSI)。这些厂商的人工智能工具因其市场占有率，有望在其生态系统中得到广泛应用。截至2021年，美国医院中对任何类型EHR系统的采用率已接近90%，其中通过认证的EHR系统的采用率约为80%。

一项由美国医院协会(American Hospital Association, AHA)于2023年开展的IT调查发现，大多数在其电子健康档案(EHR)系统中使用机器学习或预测模型的医院，主要依赖某一主流厂商提供的住院服务解决方案(图5.4.17)。其中，Epic、Cerner和Meditech的采用率最高。Epic、Cerner和CPSI所服务的医院主要采用厂商自行开发的模型，而Meditech及其他厂商的用户则更常使用第三方或医院自建的解决方案(见图5.4.18)。

预测模型在主要住院电子健康档案供应商中的应用

资料来源：AHA调查，2024|图表：2025年人工智能指数报告

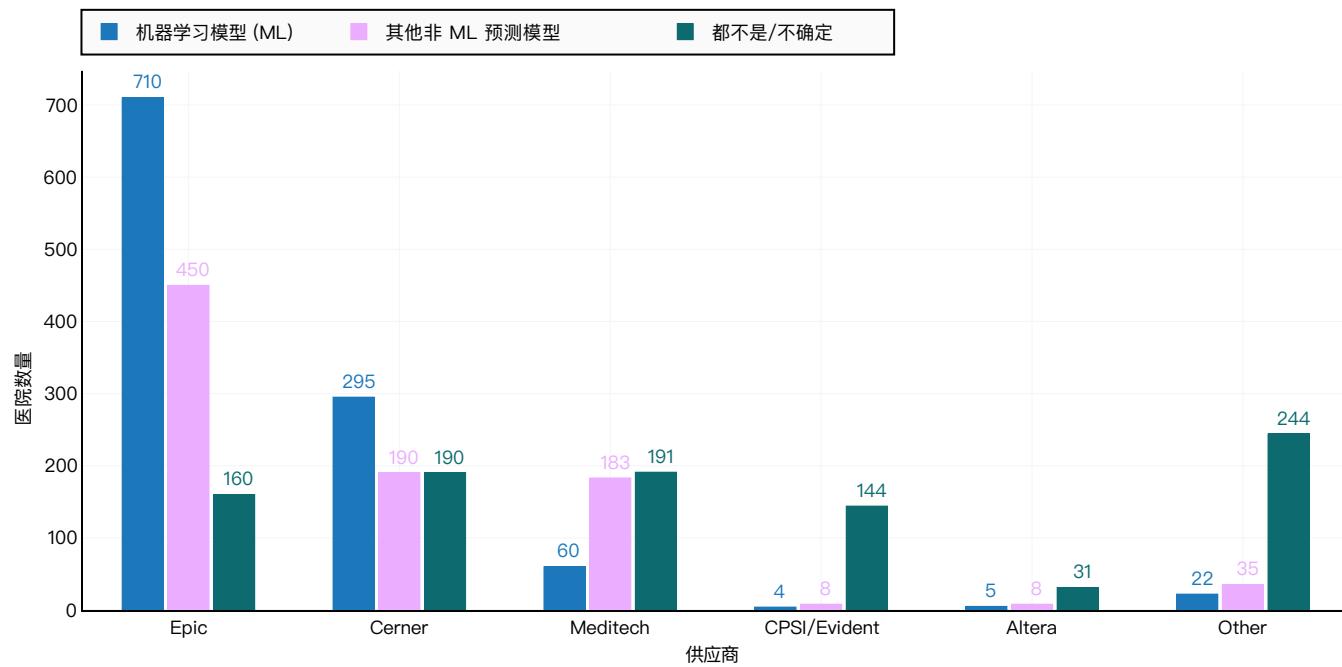


图 5.4.17

第五章：科学与医学

5.4 临床诊疗：非影像领域

电子健康档案（EHR）供应商的预测模型开发来源

资料来源：AHA 调查，2024 | 图表：2025 年人工智能指数报告

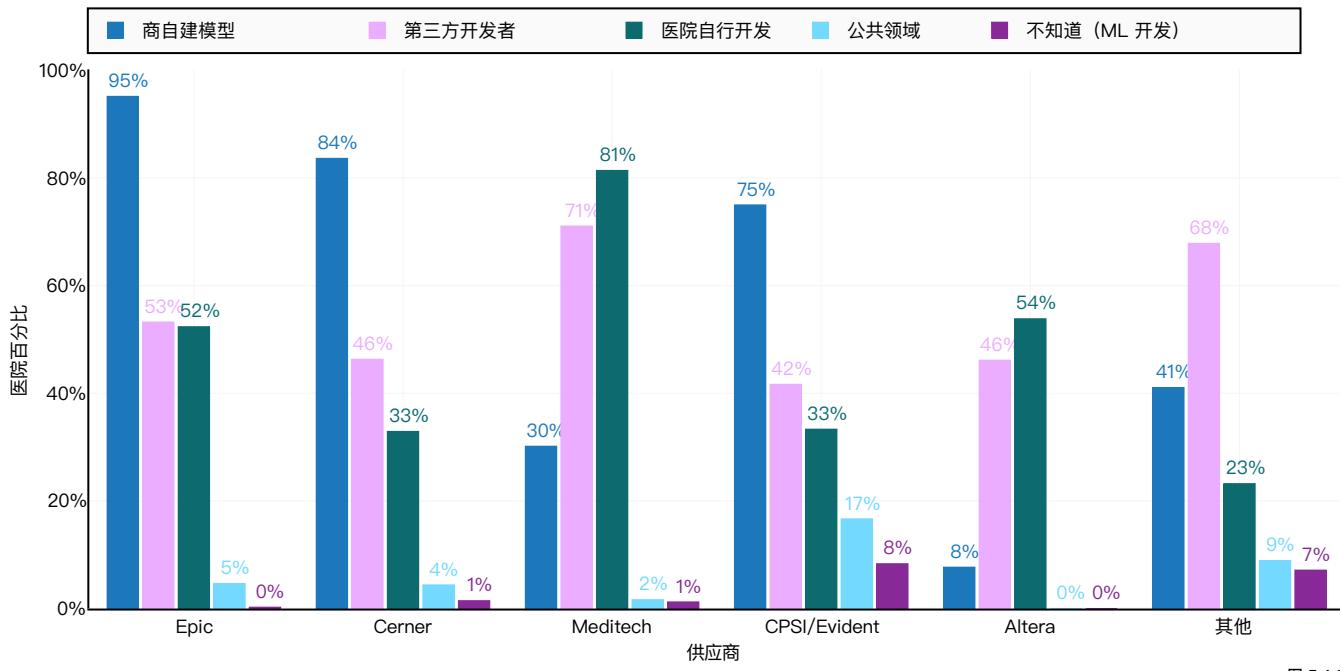


图 5.4.18

将人工智能系统集成至电子健康档案（EHR）平台，有望简化临床工作流程，同时改善医疗服务提供者与患者的整体体验。然而，目前仍不确定此类配备人工智能的健康信息技术是否真正能使资源匮乏群体从中受益。这些群体通常在技术采纳方面面临更高的壁垒。例如，生活在农村地区的群体往往受到网络带宽受限、医疗信息技术基础设施薄弱，以及 EHR 系统功能有限等结构性条件的制约。而这些因素正是

实现人工智能驱动医疗系统的基础支撑条件。此外，仍需进一步评估现有人工智能工具是否在功能基础薄弱的 EHR 系统中同样适用。目前，许多医疗环境仍依赖于功能较为简化的 EHR 平台。因此，若想真正实现人工智能在医疗中的公平部署，就必须正视技术适配性与基础条件不均所带来的结构性差异。

第五章：科学与医学

5.4 临床诊疗：非影像领域

临床决策支持

人工智能正在彻底改变医疗从业者的疾病诊断、预测和管理模式，且日益强调通过临床试验对人工智能系统进行严格评估。临床决策支持系统(CDS)中人工智能技术的发展演进，体

现了从新冠疫情期间的被动干预，逐步转向基于数据的主动性临床决策，相关临床试验数量也逐年递增。如图 5.4.19 所示，涉及人工智能技术的临床试验数量正呈现稳定增长态势。

2014–2024年提及人工智能的临床试验数量

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

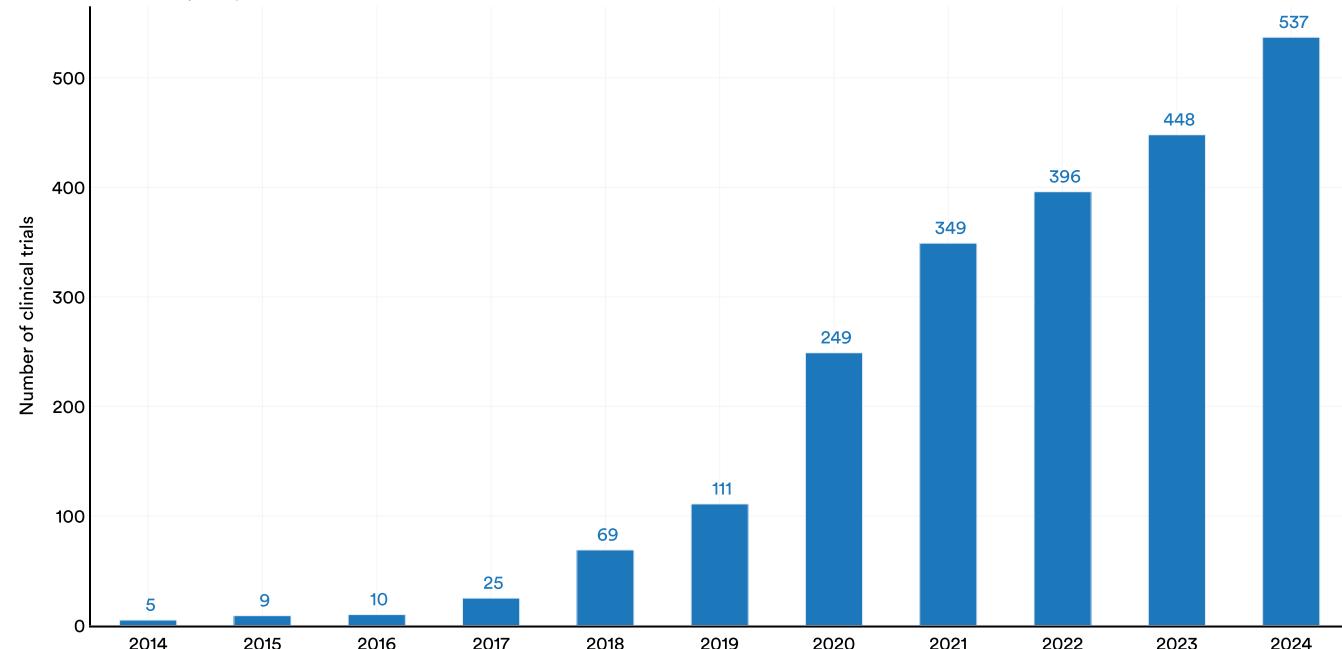


图 5.4.19

第五章：科学与医学

5.4 临床诊疗：非影像领域

新冠肺炎疫情加速了人工智能在分诊、资源调配和预后预测领域的应用，凸显了该技术在实时临床决策支持系统(CDS)中的潜力。后疫情时代，人工智能的应用范畴已从应急响应扩展到慢性病管理、诊疗流程优化及工作流整合。例如《CERTAIN研究》证明，人工智能驱动的实时手术辅助可显著提高胃肠手术的诊断准确率。截至2023年，人工智能在CDS中的应用已

延伸至用药安全和工作流优化领域，典型案例如《药学实践中的用药错误预防研究》，该研究利用人工智能实现了实时药物差错监测。全球范围内人工智能驱动的临床试验数量激增，2024年中国(105项试验)、美国(97项)和意大利(42项)位列前三(图5.4.20)。

2021—2024年各地区包含人工智能内容的临床试验数量

资料来源：RAISE Health, 2025 | 图表：2025年人工智能指数报告

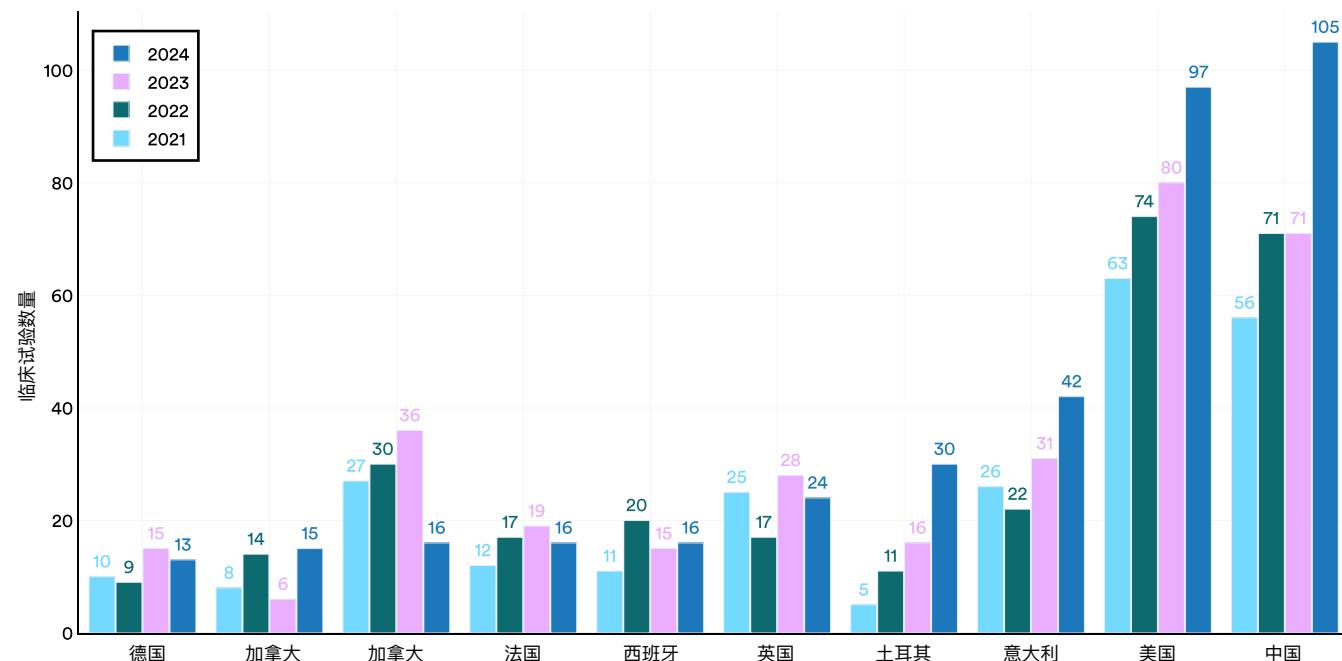


图 5.4.20

第五章：科学与医学

5.5 伦理考量

如前述章节所讨论，人工智能在医学研究和临床诊疗中的日益深入应用，既带来希望也伴随挑战。人工智能系统在训练过程中严重依赖大量数据，而这些数据的收集、使用和共享——特别是在医疗等高风险领域——可能引发多重伦理问题。

5.5 伦理考量 元分析

本节中，人工智能指数团队通过对数千项医学伦理研究进行元分析 (meta-review)，系统梳理了该领域的發展现状。图 5.5.1 展示了研究团队采用的方法学框架。数据显示，过去五年间，医疗人工智能伦理问题的关注度逐年攀升，相关出版物数量在 2020 至 2024 年间增长达四倍 (图 5.5.2)。

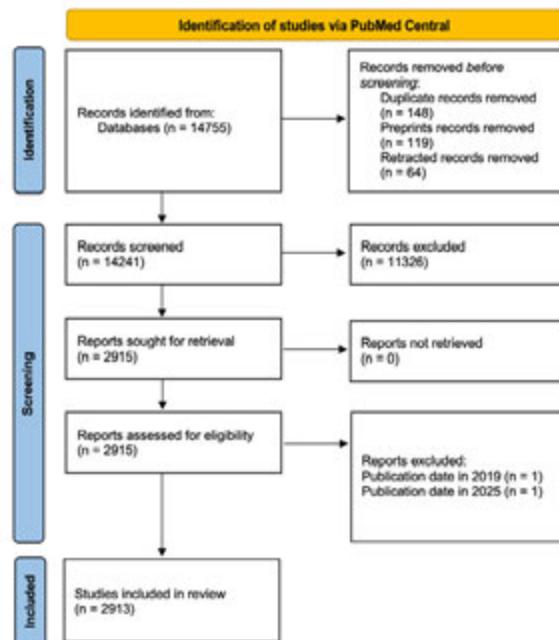


图 5.5.1

2020–2024年人工智能伦理医学成果统计数

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

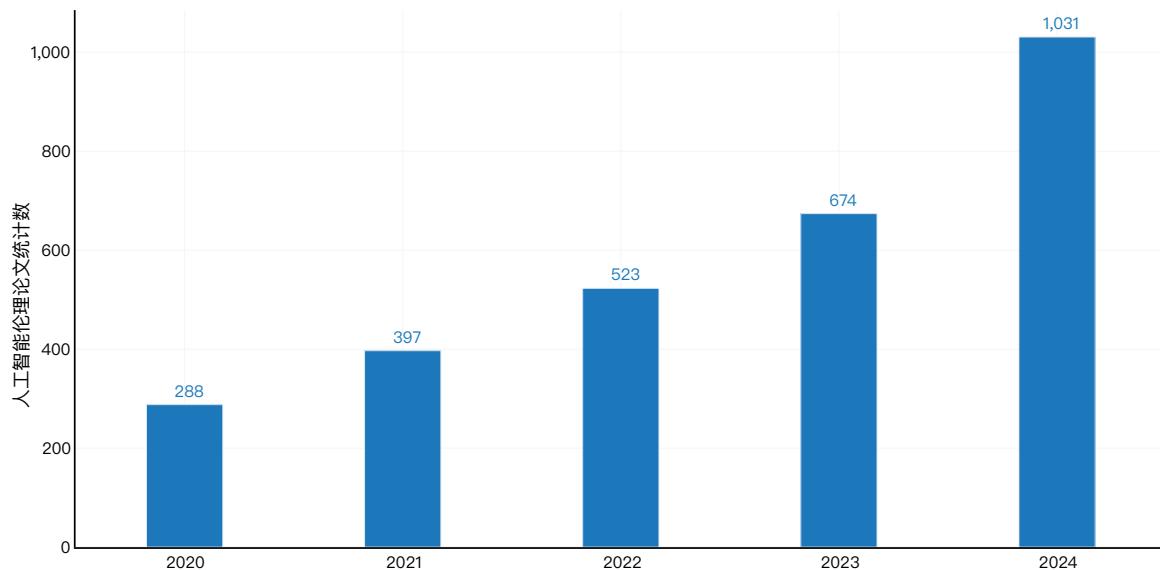


图 5.5.2

第五章：科学与医学

5.5 伦理考量

医学伦理论文中人工智能应用的关注焦点随时间推移不断演变。图 5.5.3 展示了 2020 至 2024 年 AI 医学论文中讨论的伦理议题分布。2024 年数据显示，偏见和隐私成为最受关注的伦理问题，公平性次之。值得注意的是，2020 年隐私问题的讨论热度曾显著高于偏见问题，但这一趋势在后续研究中发生了明显转变。

2020–2024年医学人工智能伦理论文中讨论最多的十大伦理问题

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

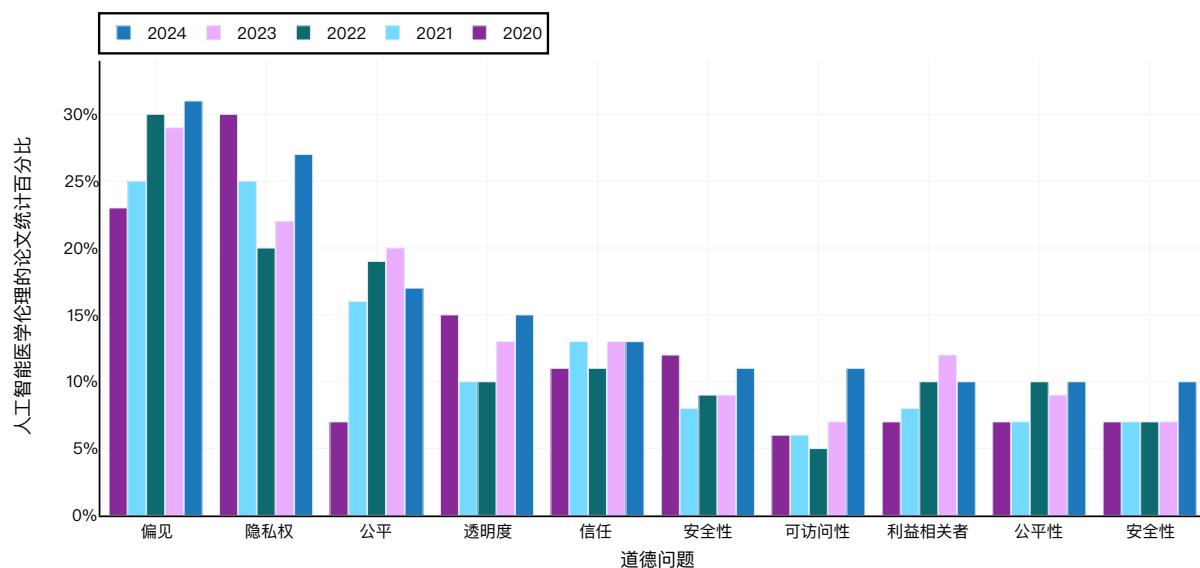


图 5.5.3

在人工智能工具方面，医学伦理文献对 OpenAI 的 GPT 系列（如 ChatGPT）给予了高度关注（图 5.5.4）。这一现象反映出过去几年间，大型语言模型在医学伦理领域引发的关注度正在持续扩大。

2020–2024年医疗人工智能伦理论文中讨论的人工智能工具

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

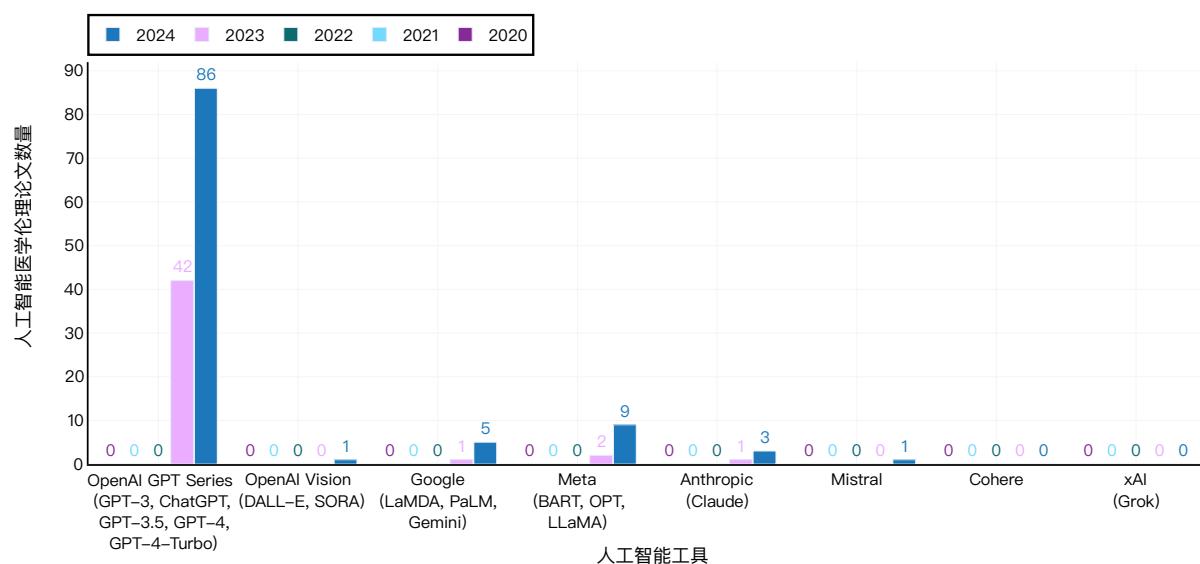


图 5.5.4

第五章：科学与医学

5.5 伦理考量

图 5.5.5 与图 5.5.6 分别按财年展示了美国国立卫生研究院（NIH）资助医学人工智能伦理项目的立项数量与资金总额。数据显示，2023 至 2024 年间，立项数量从 25 项激增至 337

项（图 5.5.5）；同期资助总额更从 1600 万美元飙升至 2.76 亿美元，短短一年内实现近 17 倍的增长。

2020–2024财年NIH对医疗人工智能伦理研究资助的数量

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

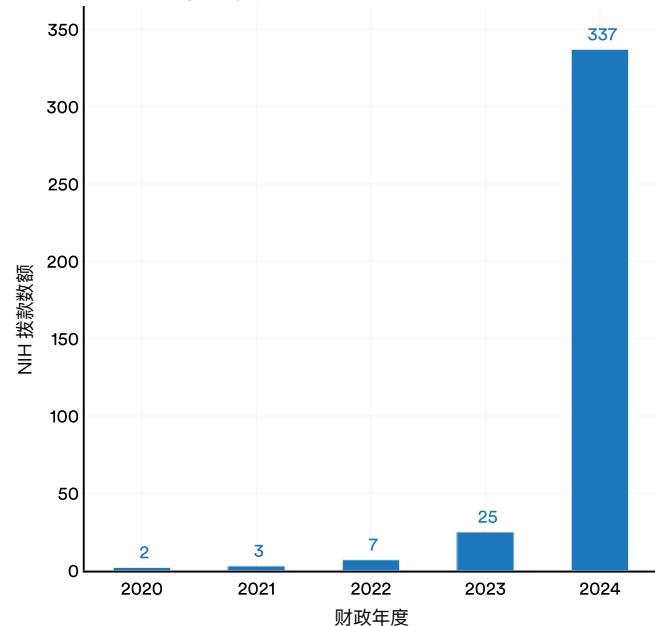


图 5.5.5

2020–2024财年NIH对医疗人工智能伦理的研究资助金额

资料来源：RAISE Health, 2025 | 图表：2025 年人工智能指数报告

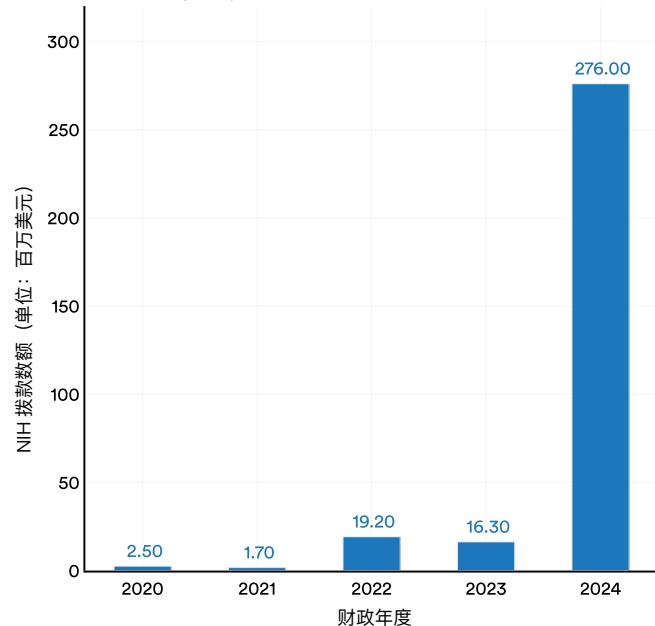


图 5.5.6

第五章：科学与医学

5.6 科学领域的人工智能基础模型

本年度，多个基础模型在科学各领域中相继问世。一些模型是在大型语言模型的基础上，针对特定领域的文献进行微调；另一些则从头开始使用专门的数据（如时间序列或气象数据）进行训练。随后，这些基础模型被进一步微调，用于具体的科学任务或应用场景。

5.6 科学领域的人工智能基础模型

重点：

标志性的模型发布

人工智能推动了物理、化学、地球科学等多个科学领域的显著进展。下表汇总了这些领域中一些最具标志性的模型发布情况，以及用于追踪这些进展的新资源。本节分析是

人工智能指数项目的初步尝试，未来将继续拓展并深化对人工智能驱动的科学进展在更广泛学科中的覆盖。

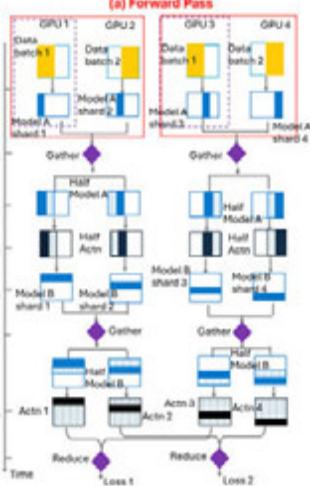
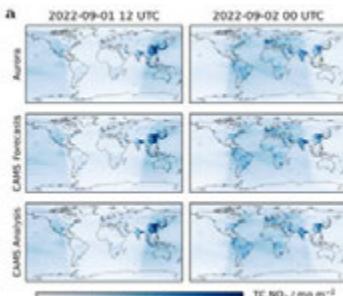
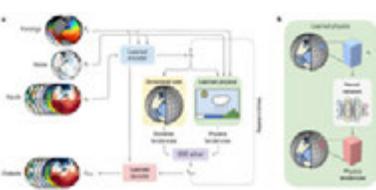
发布日期	模型名称	领域	技术意义	Image
2024年2月6日	Crystal <u>大语言模型</u>	材料科学	研究人员在 LLaMA-2 70B 模型基础上，对编码为文本的原子级数据进行微调，用于生成稳定材料，其亚稳态生成率几乎是领先扩散模型的两倍（49% 对比 28%），同时仍具备物理合理性。该方法支持无条件生成、结构补全与文本引导设计等灵活应用，并通过扩大模型规模增强对称性的感知	<p>图 5.6.1 资料来源：Gruver 等, 2024</p>
2024年2月14日	LlaSMol	化学	为应对大型语言模型在化学任务中的低效表现，研究人员引入 SMolInstruct —— 一个包含超过 300 万个样本、覆盖 14 项任务的高质量数据集；并基于该数据集开发了 LlaSMol 模型系列。其中，基于 Mistral 的 LlaSMol 在多个任务中大幅超越 GPT-4 与 Claude 3 Opus，并在仅调整 0.58% 参数的前提下，接近于任务专用模型的表现，显示出面向特定领域的指令微调的强大能力。	<p>图 5.6.2 资料来源：Yu 等, 2024</p>

第五章：科学与医学

5.6 科学领域的人工智能基础模型

重点：

标志性模型发布（续）

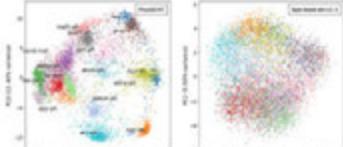
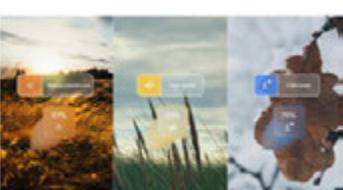
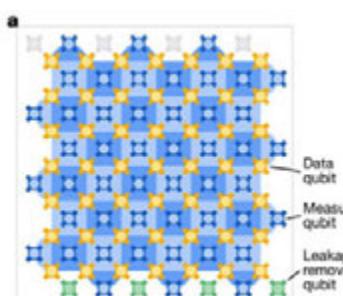
2024年 4月23日	ORBIT	地球科学	<p>橡树岭国家实验室发布了 ORBIT 模型，这是迄今为止气候科学领域规模最大的人工智能模型，拥有 1130 亿参数，规模比先前模型大出 1000 倍。该模型采用新型并行计算技术进行训练，并在 Frontier 超级计算机上测试，其持续性能最高达到 1.6 exaFLOPS，标志着人工智能驱动地球系统预测的新高度。</p>	 <p>图 5.6.3 资料来源：Wang 等, 2024</p>
2024年 5月20日	Aurora	地球科学	<p>Aurora 是一种大规模基础模型，训练数据涵盖逾 100 万个小时的地球系统记录。它可提供空气质量、海浪状况、气旋轨迹及高分辨率天气等领域的最先进预测能力，在运算成本极低的情况下仍超越传统系统，并可用极少资源跨领域微调，是向普惠型人工智能地球系统预测迈出的重要一步。</p>	 <p>图 5.6.4 资料来源：Bodnar 等, 2024</p>
2024年 7月22日	NeuralGCM	天气预报	<p>该研究推出了 NeuralGCM，一种将可微分的基于物理的求解器与机器学习组件结合的混合模型，能够同时模拟天气和气候。在短期和中期预测中，该模型的表现与当前领先的机器学习模型及物理模型相当甚至更优，能准确追踪长期气候指标，捕捉热带气旋等复杂现象，同时实现大幅计算节约。</p>	 <p>图 5.6.5 资料来源：Kochkov 等, 2024</p>

第五章：科学与医学

5.6 科学领域的人工智能基础模型

重点：

标志性模型发布（续）

2024年8月18日	PhysBERT	物理学	<p>由于其专业术语与复杂概念，物理学文本对自然语言处理极具挑战。PhysBERT 是首个专为物理学设计的文本嵌入模型，显著超越通用模型在物理任务中的表现。该模型在 120 万篇 arXiv 论文上训练，并通过监督数据微调，在信息检索与子领域适配任务中显著提升效果。</p>	 <p>图 5.6.6 资料来源：Hellert 等, 2024</p>
2024年9月16日	FireSat	火灾预测	<p>谷歌发布的 FireSat 是一个基于卫星的森林火灾检测系统，利用人工智能在起火 20 分钟内识别最小为 5×5 米的火点。系统通过分析实时影像和环境数据实现此功能。该项目由谷歌与 Earth Fire Alliance、Muon Space 联合开发，不仅强化了灾害响应能力，也推动了全球野火研究的进展。</p>	 <p>图 5.6.7 资料来源：谷歌, 2024</p>
2024年12月4日	GenCast	天气预报	<p>Google DeepMind 推出的 GenCast 是一种人工智能驱动的天气模型，基于扩散方法，提供极高精度的 15 天天气预测，在几乎所有指标上都优于现有的 ENS 等传统系统。GenCast 可在数分钟内生成预测结果，广泛适用于灾害响应、可再生能源与农业等领域。</p>	 <p>图 5.6.8 资料来源：谷歌, 2024</p>
2024年12月9日	AlphaQubit	量子计算	<p>2024年底，Google DeepMind 与 Google Quantum AI 联合发布了 AlphaQubit——一款具备最先进量子纠错能力的 AI 解码器。随后推出的 Willow 是首款实现在表面码阈值以下进行指指数级错误抑制与纠正的量子芯片，标志着该领域的一项重大突破。Willow 还完成了一项基准任务，仅耗时五分钟，而该任务在全球最快的超级计算机上则需逾十垓年 (septillion years)，远超过宇宙年龄。</p>	 <p>图 5.6.9 资料来源：谷歌, 2024</p>

第六章： 政策



第六章：政策

概述	325
章节要点	326
6.1 2024 年全球人工智能政策要闻	327
6.2 人工智能与政策制定	336
全球人工智能立法记录	336
概述	336
按地理区域划分	337
重点：全球人工智能立法深度观察	338
美国立法记录	339
联邦层面	339
州级层面	340
重点：美国州级人工智能立法案例概览	342
重点：有关反深度伪造技术的政策制定	343
全球立法讨论中的人工智能提及频次	345
概述	345
美国国会委员会提及情况	348
美国监管条例	349
概述	349
按机构分类	349
重点：美国联邦法规深度观察	351
6.3 人工智能公共投资	352
人工智能公共投资总额	353
按机构与职能领域划分的人工智能公共支出分布	360
重点：美国人工智能科研拨款分析	362

获取公共数据

第六章： 政策

概述

人工智能技术的快速发展已引发全球政策制定者的高度关注，各国相继出台人工智能相关政策。近年来，以美国和欧盟为代表的多个国家及政治实体相继推出重要监管法规。最新动态显示，多国政府已宣布对人工智能基础设施进行大规模投资。这一政策浪潮反映出国际社会对人工智能“双轨治理”共识的不断增强，即规范风险的同时释放变革潜力。

本章节系统梳理全球人工智能治理格局：首先呈现 2024 年关键政策事件时间轴，继而分析全球及美国立法进展，量化统计立法讨论中的人工智能议题热度，并解读美国监管机构的人工智能治理路径。最终以美国人工智能公共投资分析作结，其中大部分数据由人工智能指数团队独立采集。

第六章：

政策

章节要点

1. 美国各州正引领人工智能立法进程，而联邦层面进展相对迟缓。2016 年，只有一项州级人工智能相关法律获得通过，而到 2023 年，增加到 49 项。仅在去年，这一数字就翻了一番多，达到 131 项。虽然联邦层面的人工智能法案提案也有所增加，但通过的数量仍然很少。

2. 世界各国政府都加大人工智能基础设施投资力度。加拿大宣布了一项 24 亿美元的人工智能基础设施一揽子计划，而中国设立 475 亿美元半导体产业基金。法国承诺为人工智能基础设施投资 1170 亿美元，印度承诺投资 12.5 亿美元，而沙特阿拉伯的“超越计划”则包括对人工智能的 1000 亿美元投资。

3. 在全球范围内，人工智能在立法程序中的提及率不断上升。在 75 个国家中，2024 年立法程序中提及人工智能的次数增加了 21.3%，从 2023 年的 1557 次增加到 1889 次。自 2016 年以来，人工智能被提及的总数增长了 9 倍多。在全球范围内，人工智能安全研究机构加速扩张与协同合作。

4. 2024 年，各国相继成立国际人工智能安全研究机构。首批机构由美国和英国于 2023 年 11 月首届人工智能安全峰会闭幕后率先设立。随着 2024 年 5 月首尔人工智能峰会的召开，日本、法国、德国、意大利、新加坡、韩国、澳大利亚、加拿大及欧盟等国家和地区也相继承诺成立相关机构。

5. 美国人工智能相关联邦法规数量激增。2024 年，美国出台了 59 项人工智能相关法规，是 2023 年 25 项的两倍多。这些法规来自 42 个机构，是 2023 年出台法规的 21 个机构的两倍。

6. 美国多州加强深度伪造监管立法。2024 年之前，只有加利福尼亚、密歇根、华盛顿、得克萨斯和明尼苏达五个州颁布了法律，对选举中的深度伪造行为进行监管。2024 年，俄勒冈州、新墨西哥州和纽约州等 15 个州出台了类似措施。此外，到 2024 年，已有 24 个州通过了针对深度伪造的法规。

第六章：政策与管理

6.1 2024 年全球人工智能政策要闻

本章节节概述了由人工智能指数指导委员会遴选的 2024 年全球范围内最重要的人工智能相关政策事件。

6.1 2024 年全球人工智能政策要闻

2024 年 2 月 21 日

新加坡计划 5 年内投资 10 亿美元发展人工智能

(时任) 副总理兼财政部长黄循财在 2024 年 2 月 16 日的预算报告中宣布，政府将在未来五年内拨款超过 10 亿美元，用于支持人工智能计算、人才培养和行业发展。



来源：《海峡时报》，2024

2024 年 3 月 11 日

阿布扎比成立千亿美元人工智能投资公司

2024 年 3 月，阿布扎比成立了国有投资基金 MGX Fund Management Limited，专注于人工智能技术领域，目标管理资产规模达 1000 亿美元。此举与阿联酋将自身定位为全球人工智能技术创新引领者的战略目标相一致。



资料来源：彭博社，2024

2024 年 3 月 13 日

欧洲议会通过《人工智能法案》

欧盟议会在达成初步协议三个月后，正式通过了具有里程碑意义的《人工智能法案》。该法案是全球首个全面人工智能监管框架，包含诸多规定，如透明度与报告义务、基于风险的监管机制，以及禁止社交评分、人为操控与基于“敏感特征”的生物识别分类等用途。大部分条款将在两年实施期后，于 2026 年生效。由于该法案采取更为严格的监管手段，相较于其他地区更具限制性，引发了业界广泛关注与争议。



资料来源：时间，2023

第六章：政策与管理

6.1 2024 年全球人工智能政策要闻

2024 年 3 月 15 日

印度撤销强制审批新人工智能模型的计划

在发布要求科技公司推出新人工智能模型前须获政府批准的行政指导不到一个月后，印度因面临企业家和投资者的强烈反对，于近日发布了修订后的企业自律指导方针。新规要求企业对用户明示其人工智能模型是否未经充分测试或存在可靠性问题。印度电子信息技术部仍强调，人工智能模型不得破坏选举公正性，或助长偏见与歧视。



资料来源: TechCrunch , 2024

2024 年 3 月 17 日

印度启动印度人工智能使命计划 (IndiaAI Mission), 投资 12.5 亿美元

2024 年 3 月，印度正式启动印度人工智能使命计划以强化其人工智能生态系统。这项总投资达 12.5 亿美元的倡议将通过公私合作模式实现三大目标：建设超 1 万块 GPU 的算力基础设施、开发国家非个人数据平台、扶持本土人工智能模型与深度科技初创企业。该计划同时强调构建伦理人工智能治理框架，并通过在非中心城市扩展人工智能实验室来促进技术普惠发展。



资料来源《自然》 2024

2024 年 3 月 20 日

法国政府因谷歌使用受版权保护的信息对其处以 2.5 亿欧元罚款

法国竞争监管机构“竞争管理局”(Autorité de la Concurrence)对谷歌公司处以 2.5 亿欧元罚款，原因是该公司在未通知媒体机构的情况下，使用法国新闻内容训练其人工智能聊天机器人 Bard(现更名为 Gemini)。监管机构指出，这一行为违反了欧盟知识产权规则，并导致新闻出版商和通讯社无法就内容使用进行公平定价协商。谷歌已接受处罚决定，并提出一系列整改措施以解决内容抓取问题。



资料来源来源: NBC 新闻 2024

第六章：政策与管理

6.1 2024 年全球人工智能政策要闻

2024年 3 月 21 日

联合国大会通过促进“安全、可靠、可信”的人工智能的决议

在 120 多个成员国的支持下，联合国大会通过了一项由美国主导的具有“历史性”的决议（尽管该决议在法律上并不具有约束力），旨在促进“安全、可信任”的人工智能系统的使用。大会呼吁各方确保人工智能系统的使用应遵循人权法律，并承认这些系统在加速实现联合国可持续发展目标方面可能发挥的作用。该决议获得包括中国在内的 120 多个国家支持，并经联合国 193 个成员国一致通过，无需投票表决。



资料来源：联合国新闻，2024

2024 年 4 月 7 日

加拿大承诺投资 24 亿加元来确保本国的在人工智能领域优势地位

加拿大 2024 年联邦预算案提出了一揽子 24 亿加元的措施，在全球人工智能开发和采用竞争日益激烈的背景下，“确保加拿大在人工智能方面的优势”。资金将用于一系列举措，包括增强研究人员和开发人员的能力与基础设施、扶持人工智能初创企业、帮助中小企业通过人工智能提升生产率、支持受人工智能影响的工人，以及成立新的“加拿大人工智能安全研究所”。



资料来源：国际中心 治理创新，2024

2024 年 5 月 11 日

英国人工智能安全研究所推出评估人工智能模型安全性的开源工具

研究机构发布了一套名为 Inspect 的工具集，旨在评估人工智能模型在多个领域的能力，包括核心知识、推理能力以及自主能力。该研究所表示，这是首次由政府支持机构牵头推出的人工智能安全测试平台，并以开源许可形式公开发布，以惠及业界、研究机构和学术界。



资料来源：TechCrunch，2024

第六章：政策与管理

6.1 2024 年全球人工智能政策要闻

2024 年 5 月 21 日

英国和韩国在首尔联合举办人工智能安全峰会

在人工智能首尔峰会期间，与会国家分享了它们依据《布莱奇利宣言》所采取的安全措施。该宣言是前一年在英国人工智能安全峰会期间签署的，强调人工智能的伦理与负责任开发。在英国峰会成果的基础上，各国随后陆续设立或宣布成立人工智能安全研究机构。在首尔，相关国家更进一步，签署了一份意向书，旨在建立一个由各机构组成的协作网络，强调推进全球合作以提升人工智能安全的重要性。



资料来源：战略与国际研究中心，2024

2024 年 5 月 27 日

中国成立有史以来规模最大的国家支持投资基金，用于推动半导体产业发展中

中国启动了一项规模达 475 亿美元的基金，用以提升半导体产能。该基金的设立标志着中国“国家集成电路产业投资基金”（简称“大基金”）第三期的启动。自 2014 年以来，该基金已通过关键投资支持国内两大芯片制造龙头企业的发展。此次举措是在美国对诸如半导体等关键技术加强出口管制的背景下提出的，这些技术是训练人工智能系统所依赖的 GPU 等硬件组件的基础。



资料来源：路透社，2024

2024 年 5 月 28 日

欧盟委员会成立人工智能办公室

在《人工智能法案》提出三年后，欧盟委员会公布了其核心执行机构——人工智能办公室（AI Office）。该办公室将在法案实施过程中发挥关键作用，具体包括：执行通用人工智能模型的监管标准、协调实践准则的制定、并对违反行为实施处罚。该机构拥有逾 140 名员工，由五个部门组成，分别负责不同的人工智能相关目标，包括通过人工智能促进社会福祉以及推动人工智能与机器人领域的卓越发展。



资料来源：战略与国际研究中心，2024

第六章：政策与管理

6.1 2024 年全球人工智能政策要闻

2024 年 6 月 26 日

美国国家标准与技术研究院(NIST)发布框架，帮助机构识别与缓解生成式人工智能风险

美国国家标准与技术研究院 (NIST) 发布了一项自愿性框架，旨在协助各类组织识别生成式人工智能所带来的独特风险，并就缓解这些风险提出了一系列建议措施。该框架是在 2023 年发布的《NIST 人工智能风险管理框架》的基础上延伸而来。建议措施包括：确定组织的人工智能风险容忍度与相应的风险管理需求、明确管理人工智能风险的责任分工、以及定期邀请非开发人员的专家参与风险评估与更新过程。该框架发布前，NIST 还发布了一份关于对抗性机器学习的文档，系统阐述了攻击类型分类、此类攻击的潜在影响，以及相关的缓解策略。



资料来源：FedScoop, 2024

2024 年 7 月 25 日

美国国务院发布《人工智能与人权风险管理指南》

美国国务院设计了《人工智能与人权风险管理档案》，为各国政府、企业与民间社会提供指导，以便将人工智能风险管理与人权保障相结合。该档案基于 NIST 的人工智能风险管理框架，提出四项关键功能——治理、映射、评估与管理，用于评估并缓解从偏见到监控滥用等各类人工智能相关风险。通过连接人工智能治理与人权保护，该档案为全球负责任地开发与部署人工智能提供了一种可适用的工具。



资料来源：美国国务院, 2024

2024 年 8 月 2 日

英国撤销 13 亿英镑技术与人工智能基础设施承

英国工党政府取消了前任政府所承诺的 13 亿英镑技术与人工智能项目资金，称这些承诺“资金不足”。这些项目原于 2023 年宣布，包括为“人工智能研究资源”(AI Research Resource) 提供 5 亿英镑，以资助算力基础设施，以及为爱丁堡大学建造百亿亿次级超算提供 8 亿英镑资金。



资料来源：英国广播公司, 2024

第六章：政策与管理

6.1 2024 年全球人工智能政策要闻

2024 年 9 月 13 日

美国白宫成立人工智能数据中心基础设施工作组

白宫召集了一场会议，邀请联邦官员与科技企业高管，共同讨论保障强大数据中心基础设施的能源来源问题，而此类基础设施对于支撑人工智能模型的运行至关重要。出席会议的企业包括 OpenAI、Anthropic、亚马逊云服务(AWS)、英伟达与谷歌母公司 Alphabet。白宫在新闻稿中强调，推进人工智能在美国的发展对国家安全至关重要，并有助于确保人工智能系统的安全性、保障性与可信赖性。新成立的人工智能数据中心基础设施工作组将负责识别建设机会，并与相关机构合作，优先发展人工智能数据中心。

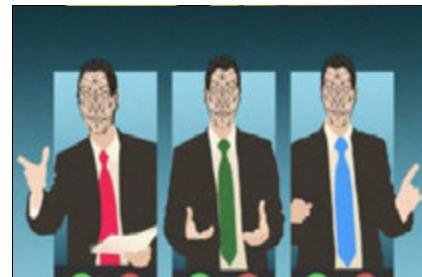


资料来源：FedScoop, 2024

2024 年 9 月 17 日

加州州长签署三项关于人工智能与选举传播的法案

在 2024 年旧金山市长选举前夕，加州州长 Gavin Newsom 宣布签署三项旨在打击深度伪造选举内容的法案。分别为 AB 2655、AB 2839 与 AB 2355，这些法案要求大型在线平台在特定时期内移除或标注数字伪造的选举内容；延长禁止传播误导性人工智能生成选举内容的时间窗口；并强制要求所有使用人工智能生成或修改内容的选举广告必须附带适当的披露说明。



资料来源：《华尔街日报》，2024

2024 年 9 月 22 日

联合国通过《全球数字契约》以确保包容与安全的数字未来

在“未来峰会”，联合国成员国通过了《全球数字契约》(Global Digital Compact)，旨在为所有人构建一个包容、开放、可持续、公正、安全与保障的数字未来。该契约强调以下目标：弥合数字鸿沟、扩大数字经济带来的普惠成果、营造一个尊重人权的数字空间、推动公平的数据治理、并加强人工智能的国际治理。契约所依据的原则以国际法与人权为核心，力图通过数字技术的力量，加速实现可持续发展目标。



资料来源：联合国，2024

第六章：政策与管理

6.1 2024 年全球人工智能政策要闻

2024 年 9 月 29 日

加州州长否决扩张性人工智能立法

加州州长 Gavin Newsom 否决了一项加州人工智能安全法案。该法案原本有望在全国范围内为人工智能监管设立先例，因加州是众多领先人工智能企业的所在地。法案拟要求对前沿人工智能模型在公开展示前进行安全测试，并赋予州总检察长就人工智能相关危害起诉企业的权力。支持者认为该法案是在确保人工智能安全与问责方面迈出的必要一步，而批评者则认为该法案过于限制，可能抑制人工智能的发展，尤其是不利于开源权重的人工智能生态。鉴于加州作为世界第五大经济体的地位，该法案的影响或将超出州界，类似“布鲁塞尔效应”，在国家乃至国际层面塑造人工智能治理方向。Newsom 为其否决辩护称，该法案施加了过高的标准。

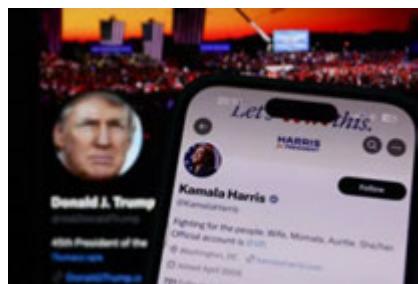


资料来源：《金融时报》，2024

2024 年 10 月 2 日

美国法官叫停加州新人工智能法案，因涉及卡玛拉·哈里斯的深度伪造视频

加州一项新人工智能法律在签署仅两周后，便被一名联邦法官暂时中止。在裁决中，法官 Judge Mendez 指出，该法律对“有害”影像的定义模糊，可能对宪法保护的言论自由构成威胁。此前该法律被用于起诉一位 X 平台用户，该用户曾发布一段涉及副总统卡玛拉·哈里斯的深度伪造视频。



资料来源：《洛杉矶时报》，2024

2024 年 11 月 8 日

沙特阿拉伯宣布“超越项目”

2024 年 11 月，沙特阿拉伯宣布启动“超越计划”(Project Transcendence)，这是一项价值 1000 亿美元的人工智能倡议，旨在将该国建设成为全球科技中心。该计划由沙特主权财富基金——公共投资基金(Public Investment Fund, PIF)牵头执行，并与谷歌母公司 Alphabet 达成合作协议，计划投资 50 亿至 100 亿美元，用于开发阿拉伯语人工智能模型。此举与沙特“2030 愿景”高度契合，该愿景致力于推动经济多元化，摆脱对石油的依赖，并将人工智能建设为未来发展的重要支柱。



资料来源：《电信评论》，2024

第六章：政策与管理

6.1 2024 年全球人工智能政策要闻

2024 年 11 月 14 日

欧盟人工智能办公室发布《通用人工智能实践准则》初稿

欧洲人工智能办公室发布了《通用人工智能实践准则》四份初稿中的第一份。该准则由四个独立专家工作组制定，重点关注以下领域：透明度与版权、风险识别与评估、风险缓解以及内部治理。一旦正式定稿，该准则将作为《人工智能法案》的补充机制，使人工智能模型提供方在最终标准发布前，得以通过该准则证明其合规性。



资料来源：欧洲联盟，2024

2024 年 11 月 25 日

美国与全球合作伙伴共同启动国际人工智能安全网络

2024 年 11 月，美国商务部与国务院在旧金山共同主办了“国际人工智能安全研究机构网络”(International Network of AI Safety Institutes) 的首次会议。该倡议旨在加强全球在安全人工智能创新方面的协调合作，重点包括：管理合成内容风险、测试基础模型、以及对先进人工智能系统进行风险评估。美国担任首任轮值主席国，初始成员包括：澳大利亚、加拿大、欧盟、法国、日本、肯尼亚、大韩民国、新加坡以及英国。该网络已获得超过 1100 万美元的全球研究资助承诺，以支持其后续工作。



资料来源：美联社，2024

2024 年 12 月 2 日

美国加强对华半导体制造设备和软件的出口管制

美国商务部工业与安全局宣布了新的出口管制措施，进一步限制中国制造先进半导体的能力。新规包括：对 24 类半导体制造设备、3 类软件工具的出口限制，以及额外的限制措施。美国商务部长强调，此类措施对维护美国国家安全至关重要。



资料来源：CNBC，2024

第六章：政策与管理

6.1 2024 年全球人工智能政策要闻

2024 年 12 月 19 日

联合国安理会讨论人工智能在冲突中的应用并呼吁建立全球框架

2024 年 12 月 19 日，联合国安理会召开会议，专门讨论人工智能在军事环境中所带来的挑战。联合国秘书长安东尼奥 · 古特雷斯强调，人工智能的迅猛发展正超越现有治理框架的适应能力，可能会削弱人类对武器系统的控制能力。他呼吁建立“国际护栏”(international guardrails)，以确保人工智能的安全与包容性应用。此次讨论的背景，是关于乌克兰战争中广泛使用自主无人机和机器人武器系统的持续报道。



资料来源：《伯克利政治评论》，2016

6.2 人工智能与政策制定

全球人工智能立法记录

概述

人工智能指数分析了 2016 年至 2024 年间 114 个国家包含“人工智能”一词的立法。¹其中，39 个国家已制定至少一项与人工智能相关的法律(图 6.2.1)。²总体来看，这些国家共通过了 204 项与人工智能相关的法律。图 6.2.2 展示了自 2016 年以来

每年颁布的与人工智能相关的法律数量。2024 年通过的人工智能相关法律共计 40 项，高于 2023 年的 30 项，使其成为继 2022 年之后立法数量第二高的一年。自 2016 年以来，人工智能相关法律的通过数量从 1 项增长至 40 项。

2016–2024 年各国通过的人工智能相关法律数量分布

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

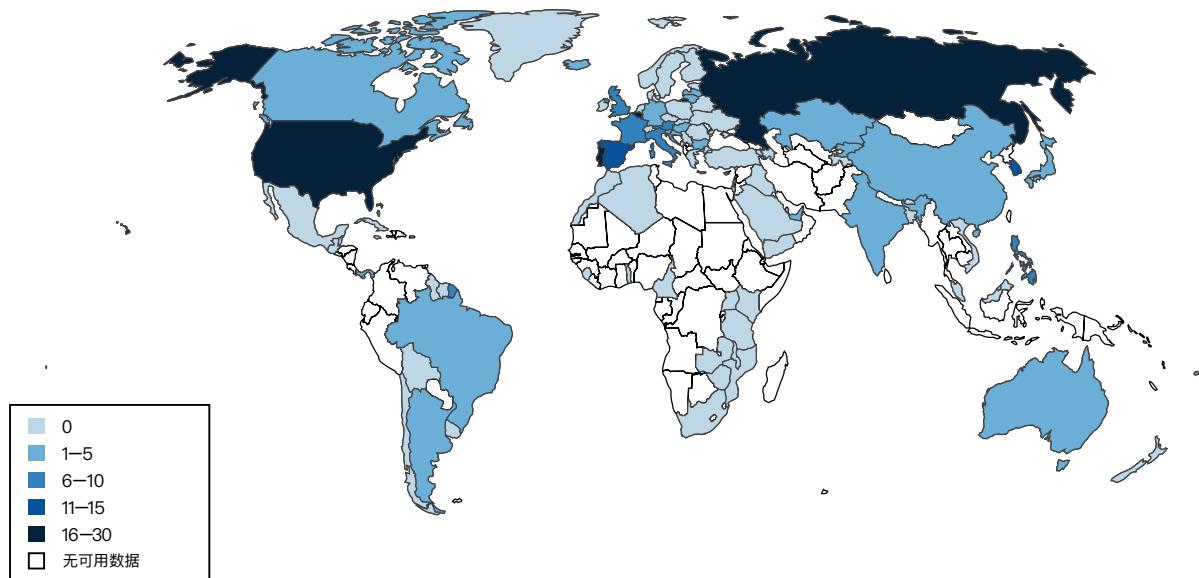


图 6.2.1

1、由于提出的大型法案可能包括多个与人工智能有关的章节，因此该分析可能低估了实际通过的法律数量。例如，《国防授权法案》是作为一个单一的综合法案提出的，但其中包括一系列较小的法案，这些法案最初是单独提出的，后来合并为一个单一的综合。

2、由于某些国家的立法数据库访问受限，人工智能指数在本年度缩减了分析国家样本，因此本年度所报告的人工智能相关法律数量可能低于往年报告。此外，香港与澳门尽管并非被正式承认为主权国家，亦被纳入统计，总体分析涵盖 116 个国家和地区。

第六章：政策与管理

6.2 人工智能与政策制定

2016–2024 年 116 个选定地区通过人工智能相关法律的数量

资料来源：人工智能指数，2025 | 图表：2025 年人工智能指数报告

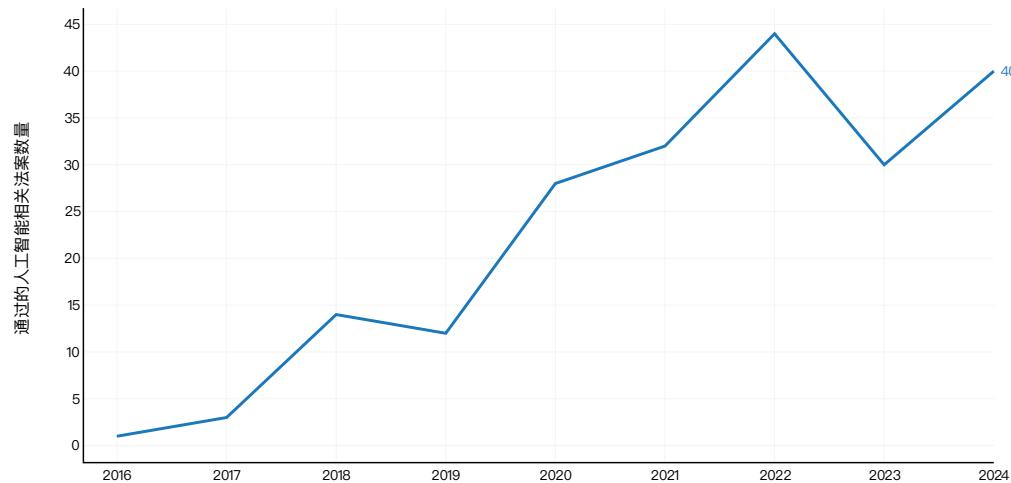


图 6.2.2

按地理区域划分

图 6.2.3 展示了 2024 年全球前 15 大地区颁布的人工智能相关法律数量。俄罗斯以 7 部法律居首，比利时和葡萄牙各以 5 部法律紧随其后。图 6.2.4 显示了自 2016 年以来全球颁布的人

工智能相关法律总数，美国以 27 部法律位居榜首，葡萄牙和俄罗斯各以 20 部法律并列第二。³

2024 年选定地区通过人工智能相关法律数量

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

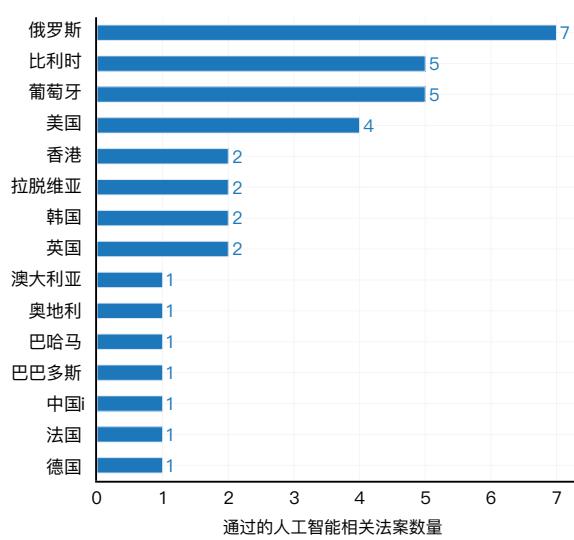


图 6.2.3

2016–2024 年部分地区通过人工智能相关法律数量（总和）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

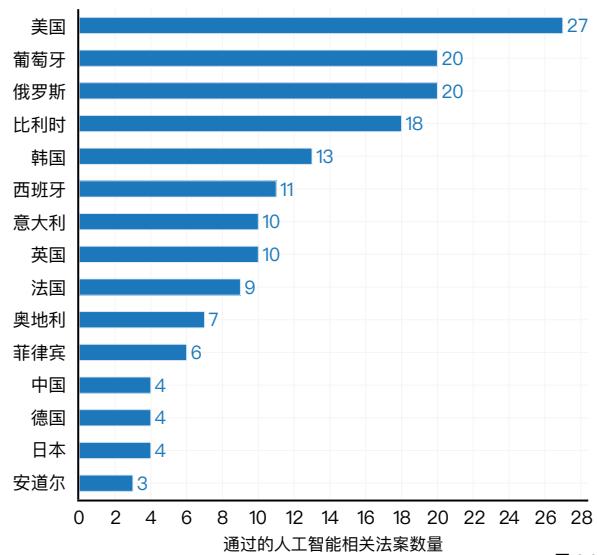


图 6.2.4

3、为简明起见，图 6.2.3 和图 6.2.4 显示了按计数排名前 15 位的地理区域的数据。完整的国家级总数将 2025 年夏季更新的 Global AI Vibrancy Tool 中提供。如需立即获取，请联系人工智能指数团队。

重点：

全球人工智能立法深度观察

本节详细分析 2024 年通过的部分人工智能相关法律。图 6.2.5 选取了五个国家的法律案例，涵盖人工智能治理的多个领域。

国家	法案名称	内容摘要
奥地利	《联邦法律修正案：修订〈通讯监管局法〉与〈2021年电信法〉》	本法案设立“人工智能服务中心”，负责支持、建议并协调在媒体、电信与邮政领域的人工智能治理。该法设立一个人工智能咨询委员会，负责监测人工智能发展趋势、向政府提供政策建议，并协助塑造国家人工智能政策。服务中心需运营一个人工智能项目信息门户，特别是涵盖公共资金资助项目，并就人工智能监管、网络安全与合规性提供指导。为支持相关活动，法案每年拨款70万欧元，后续将依据通胀调整。
比利时	《关于设立人工智能指导委员会的皇家法令》	本法案设立联邦人工智能指导委员会，为政府提供人工智能政策建议，并作为人工智能治理的主要联络机制。委员会成员由各部委及公共机构代表组成，定期召开会议，以协调比利时的人工智能政策实施。
法国	《2021年10月25日第2021–1382号法：关于数字时代文化作品访问的规范与保护》（2024年由第2024–449号法修订） ⁴	本法案设立视听与数字传播监管局（ARCOM），通过合并原“视听高级委员会”（CSA）与“网络作品传播与版权保护高级机构”（HADOPI）而成。该法强化了打击网络盗版的措施，并加强对数字平台的监管，以保障文化内容的数字访问权。修订案扩大了ARCOM的职权，授权其使用人工智能工具监管数字平台，特别是在版权侵权识别与打击盗版方面。
拉脱维亚	《竞选宣传法修正案》	本修正案规范政治广告中人工智能的使用，要求在付费竞选材料中对人工智能生成内容进行明确披露。同时，禁止在选举活动中使用带有虚假或匿名社交媒体账户的自动化系统。
俄罗斯	《关于修订〈个人数据联邦法〉及〈在俄罗斯联邦主体——联邦城市莫斯科中建立促进人工智能技术发展与实施的特定规制试验法〉，以及对〈个人数据法〉第6条与第10条的修正案》	本法案确立了匿名化个人数据处理与共享的框架，以支持政府职能中的人工智能发展。法案对人工智能驱动的决策制定进行规范，设定了生物识别数据的安全标准，并限制对敏感人工智能相关数据集的外国访问权限。

图 6.2.5

4. 2024 年通过的第 2024–449 号法律对原 2021 年颁布的第 2021–1382 号法律进行了修订，扩展其适用范围至人工智能领域，并正式授权 ARCOM 运用人工智能技术。

第六章：政策与管理

6.2 人工智能与政策制定

美国立法记录

联邦层面

图 6.2.6 展示了美国国会在 2016 至 2024 年间通过与提出的人工智能相关法案数量，并揭示了提案数量的显著增长。

⁵ 2023 年，美国国会提出了 171 项人工智能相关法案，而 2024 年该数字增至 221 项，自 2022 年以来几乎增长了三倍。然而，

在所有提出的人工智能法案中，实际通过的比例仍然较低。此类立法活跃趋势，反映了政策制定者对人工智能技术（尤其是生成式人工智能）能力与公众关注度上升的回应。⁶

2016–2024 年美国国会提出的与人工智能相关的法案数量及通过的法律数量

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

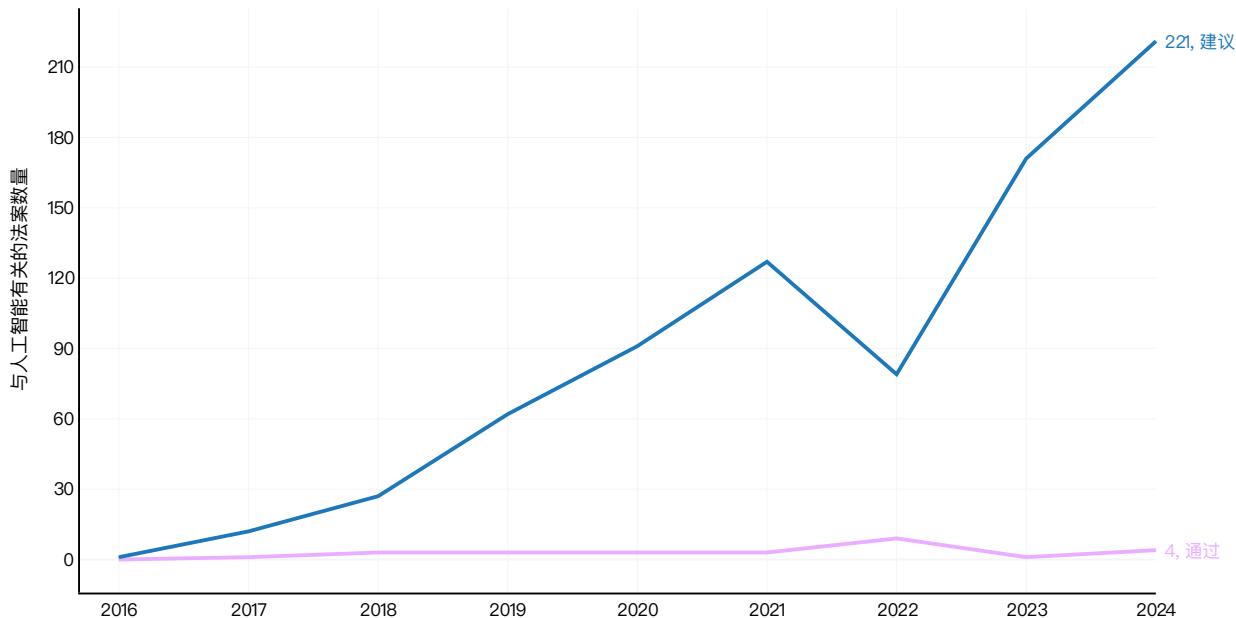


图 6.2.6

5、图中“通过”法案指已在参众两院获得通过。

6、本节仅涵盖国会立法，不含行政部门（如特朗普总统的“Stargate”声明）及联邦监管机构（如 FTC 对人工智能生成评论与社交媒体机器人的新规）制定的政策。

第六章：政策与管理

6.2 人工智能与政策制定

州级层面

人工智能指数亦追踪了美国各州在人工智能领域的立法情况。图6.2.7展示了2024年各州通过的人工智能相关法案数量，加利福尼亚州以22项居首，其后为犹他州（12项）与马里兰州（8项）。图6.2.8展示了2016至2024年间各州累计通过的人工智能立法数量，加州同样居首（42项），其次为马里兰州、弗吉尼亚州与犹他州，三州均为17项。

2024年美国部分州通过的人工智能相关法案数量

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

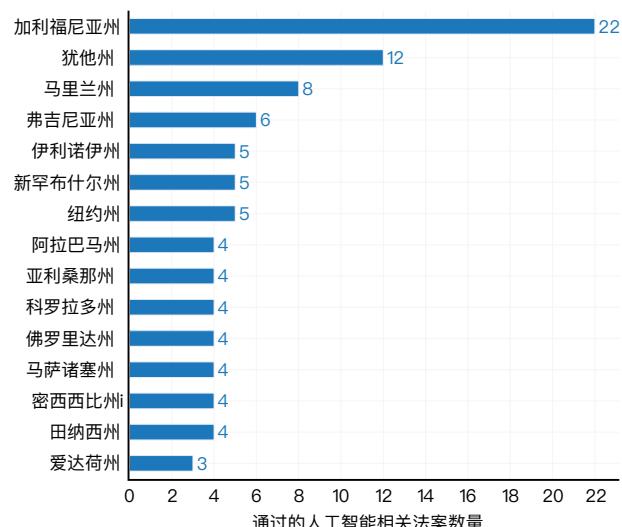


图 6.2.7

2016–2024年美国各州通过的人工智能相关法案数量（总和）

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

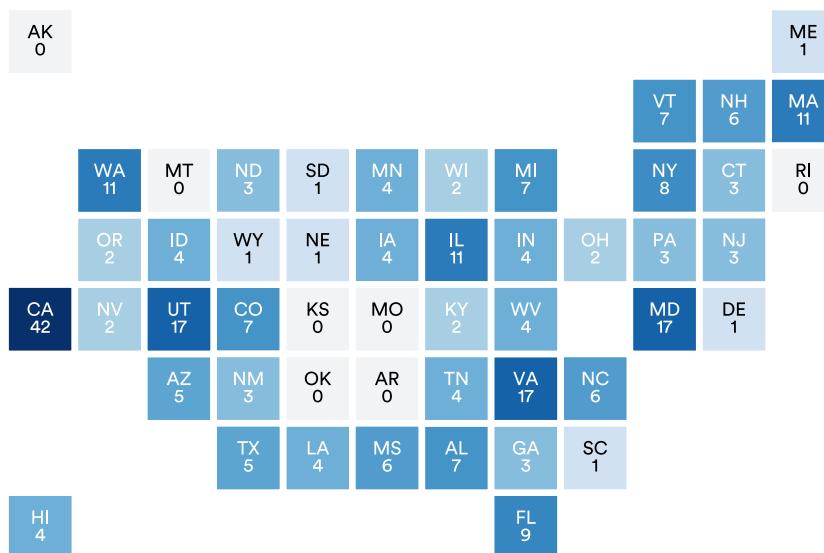


图 6.2.8

第六章：政策与管理

6.2 人工智能与政策制定

自2016年以来，美国州一级通过的人工智能相关法律数量显著上升。2016年仅通过1项，至2023年增长至49项。而在过去一年内，这一数字翻倍以上，达到了131项（图6.2.9）。

2016–2024 年美国各州通过的人工智能相关法案数量

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

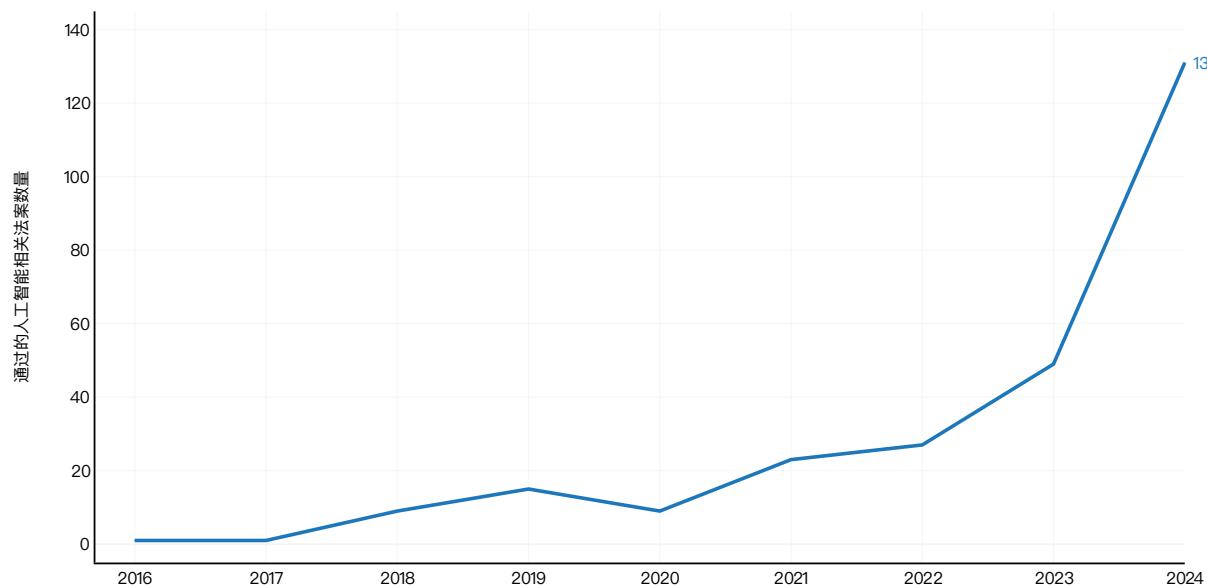


图 6.2.9

重点：

美国州级人工智能立法案例概览

本节精选了2024年美国多个州通过的人工智能相关法律案例，涵盖加利福尼亚和纽约等人工智能企业聚集地，也包括阿拉巴马与科罗拉多等在产业中地位相对较小的州，反映出州一级在人工智能治理方面的多样化关注焦点（图6.2.10）。

州	法案名称	内容摘要
阿拉巴马州	《有关选举；规定散布重大误导性媒体为犯罪行为》	本法案禁止在选举前90天内散布人工智能生成的、旨在误导选民或损害候选人的欺骗性媒体。若无明确免责声明，违规者可被判处轻罪；重复违规者可被判重罪。新闻报道、讽刺内容及带有免责声明的媒体可获豁免，受害方可提起民事诉讼。
加利福尼亚州	《加州人工智能透明法案》	该法案要求大型人工智能提供方免费提供人工智能检测工具，并在人工智能生成内容中标示清晰且永久的来源说明。每次违反将被罚款5000美元，由总检察长或地方执法机关执行。
科罗拉多州	《人工智能消费者保护法案》 ⁷	本法案对高风险人工智能系统提出透明、公正的使用要求。开发与部署方必须防止算法歧视，保障用户对人工智能决策的申诉权，并进行定期影响评估。
马萨诸塞州	《关于满足马萨诸塞州未来信息技术需求的法案》	法案拨款12.6亿美元，用于全州的信息技术、网络安全与宽带基础设施现代化建设。其中2500万美元用于将人工智能与机器学习技术整合进州政府系统，以提高自动化、效率与安全性。
纽约	《修订〈普通商业法〉关于社交媒体服务条款披露要求的法案》	本法案要求社交媒体公司以明确且可访问的方式公开其各平台的服务条款，并向州总检察长提交条款报告。若不合规，将面临惩罚。

图 6.2.10

重点：

有关反深度伪造技术的政策制定

美国各州在通过立法打击深度伪造技术（deepfake）方面表现尤为积极。深度伪造是指通过人工智能生成的合成媒体，能够篡改或替换视频、音频或图像中的人物形象，通常会产生逼真但具有欺骗性的内容。如今年人工智能指数第三章所讨论的，深度伪造可能被用于操纵选举结果，或生成不雅图像。非营利组织“Public Citizen”维护着一个追踪人工智能深度伪造法规的数据库，涵盖选举相关滥用和私密图像滥用两方面内容。图6.2.11展示了美国各州随时间推移通过的反深度伪造法律数量，包括与选举和私密图像相关的法规。⁸

图6.2.12重点呈现了各州通过选举相关人工智能深度伪造法规

的时间节点：在2024年之前，加利福尼亚、华盛顿、得克萨斯、密歇根和明尼苏达五个州已通过此类法律；2024年，又有包括俄勒冈、新墨西哥和纽约在内的12个州出台了类似法规。

针对私密深度伪造的州级法规远比针对选举滥用的法规更为普遍。共有25个州颁布了覆盖所有个人的法律，另有5个州通过了仅适用于未成年人的法规（图6.2.13）。怀俄明和俄亥俄是目前仅有的两个尚未实施任何形式私密深度伪造法规的州。

2019—2024 年美国各州通过的反深度伪造法律数量统计

资料来源：Public Citizen, 2025 | 图表：2025 年人工智能指数报告

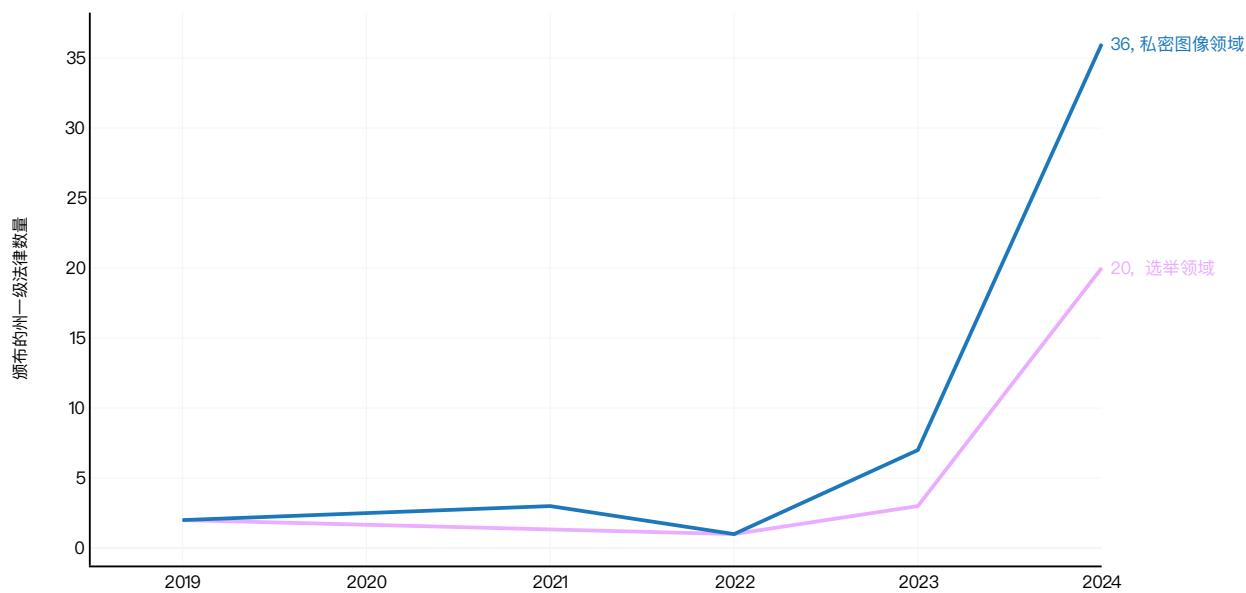


图 6.2.11

8、由于部分州的反深度伪造法律生效日期未完全核实，图 6.2.11 仅统计已确认通过时间的法案。

第六章：政策与管理

6.2 人工智能与政策制定

重点：

有关反深度伪造技术的政策制定（续）

截至 2024 年，美国各州针对人工智能生成的选举类深度伪造内容的州级法规及其状态

资料来源：Public Citizen, 2025 | 图表：2025 年人工智能指数报告

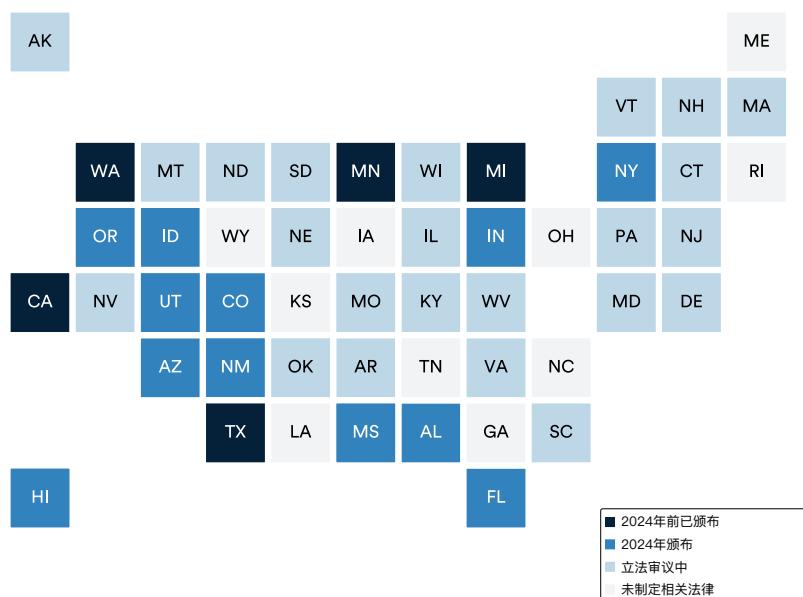


图 6.2.12

截至 2024 年，美国各州针对人工智能生成的私密影像深度伪造内容的州级法规及其状态

资料来源：Public Citizen, 2025 | 图表：2025 年人工智能指数报告

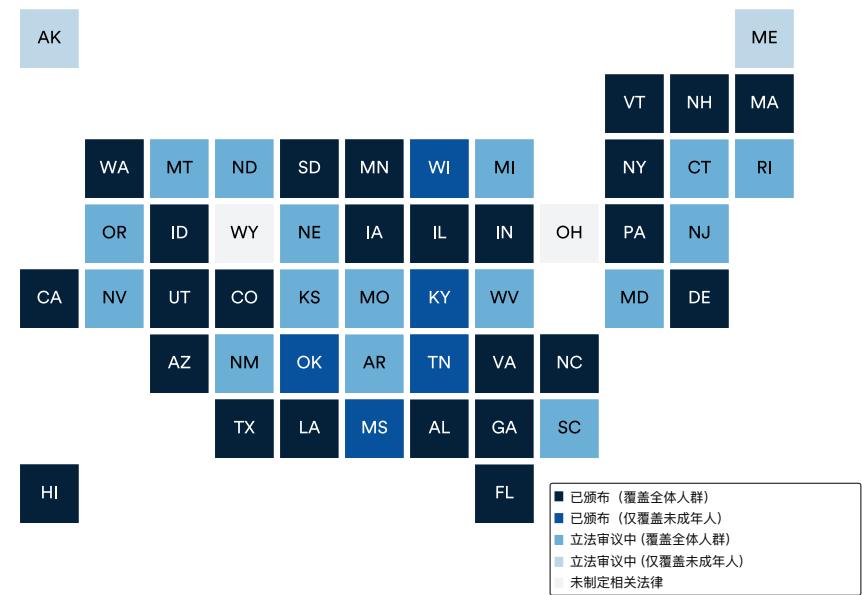


图 6.2.13

第六章：政策与管理

6.2 人工智能与政策制定

全球立法讨论中的人工智能提及频次

立法关注度的另一项指标是政府及议会会议记录中提及“人工智能”的次数。人工智能指数分析了2016年至2024年间73个国家和地区的立法会议记录，统计包含关键词“人工智能”的会议场次。⁹

概述

图6.2.14展示了2016至2024年间，全球立法程序中“人工智能”一词的年提及次数。2023年为1557次，2024年增长至1889次，增幅达21.3%。自2016年起，该数字增长超过九倍。

2016—2024年 75个地区立法会议中人工智能提及频次统计

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

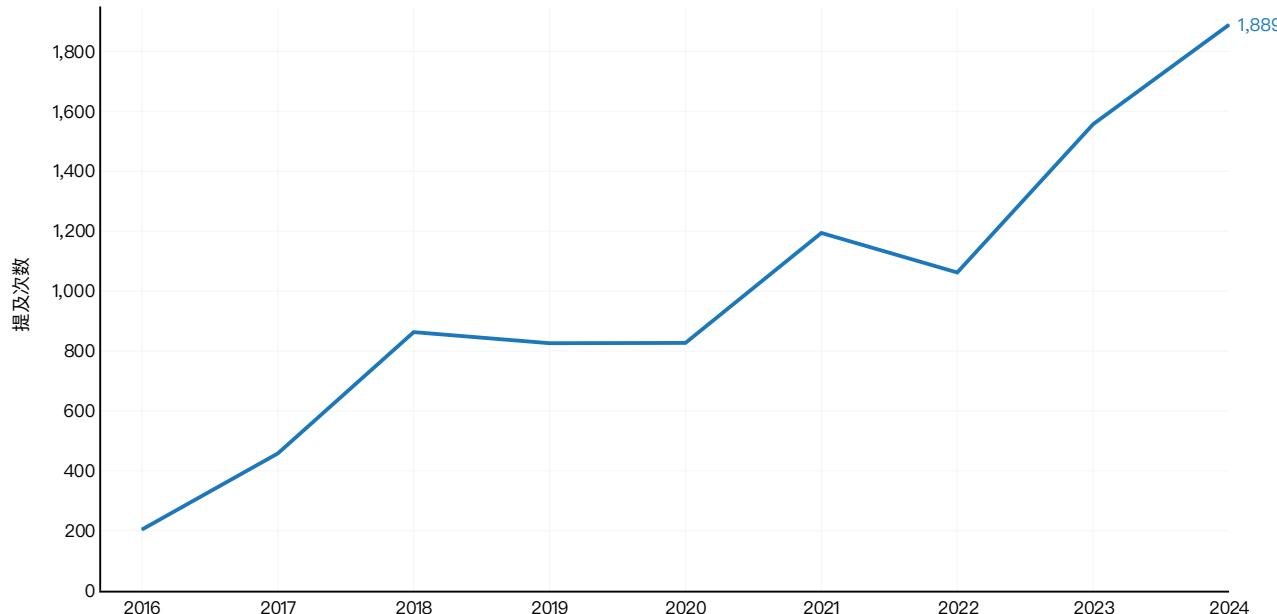


图 6.2.14

9、被分析国家的完整名单见附录。人工智能指数研究小组的目标是审查世界各地的政府和议会程序，但并非所有国家都有可公开访问的数据库。今年，指数略微调整了跟踪方法，导致与以往的总数略有不同。更具体地说，提及次数按会议计算，因此在同一立法会议上多次提及人工智能指数算作一次。完整的方法详见附录。此外，人工智能指数还跟踪了澳门和香港的提及情况。虽然它们不是正式的国家，但它们的提及也包括在图 6.2.14 的统计中。该指数总共跟踪了 75 个人地理区域的人工智能提及情况。

第六章：政策与管理

6.2 人工智能与政策制定

2024年在立法程序中提及人工智能最多的国家为西班牙（314次），其后为爱尔兰（145次）和澳大利亚（123次）。在被分析的75个国家和地区中，有57个在至少一次立法程序中提及了人工智能。

2024年各立法程序中提及人工智能的次数

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

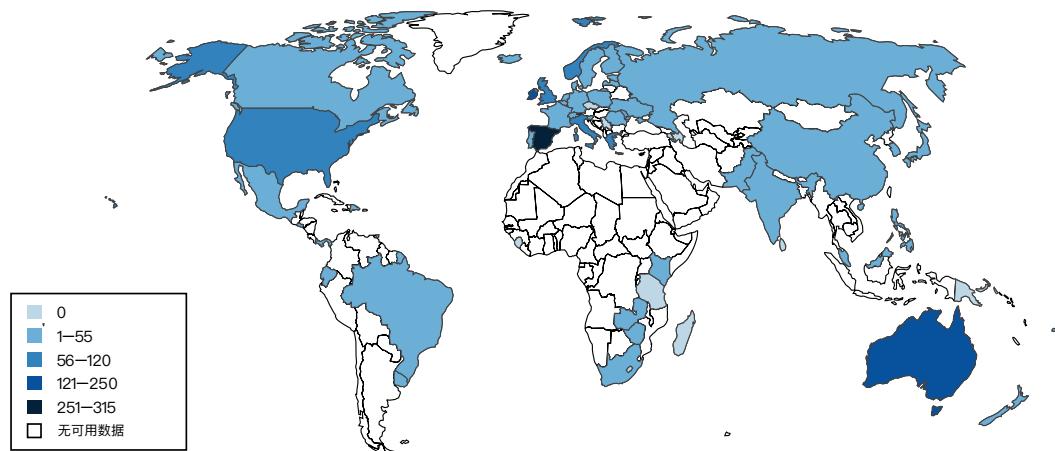


图 6.2.15

将 2016 年至 2024 年的立法提及次数汇总后，出现了一种类似的趋势（图 6.2.16）。西班牙以 1200 次位列第一，紧随其后的是英国（710 次）和爱尔兰（659 次）。

2016–2024年各立法程序中提及人工智能的次数（总和）

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

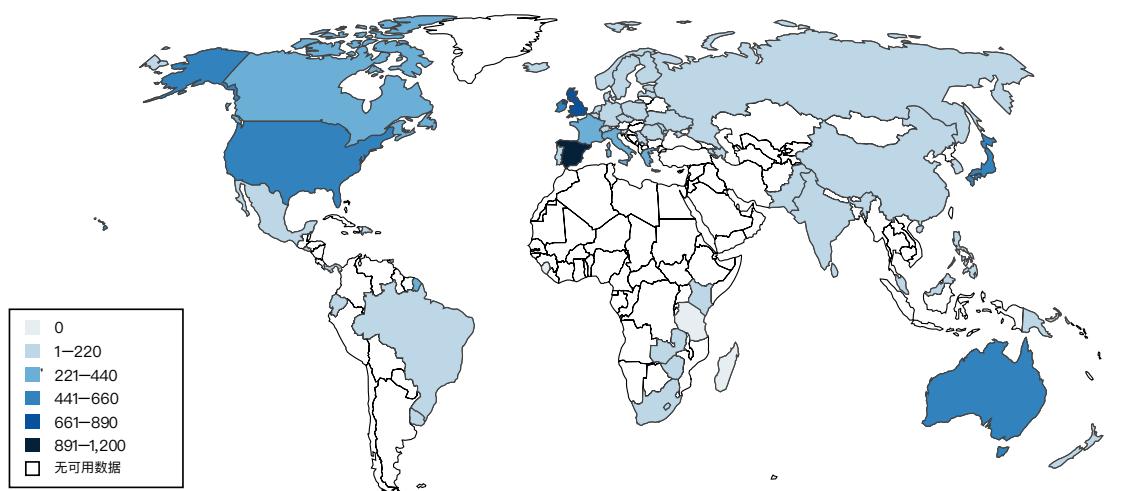


图 6.2.16

第六章：政策与管理

6.2 人工智能与政策制定

图6.2.17基于部分国家的数据，对比了人工智能在议会讨论中被提及的频率与通过的人工智能相关法律数量。总体而言，议会中对人工智能的高关注度与人工智能立法数量呈正相

关。然而，部分国家如比利时、葡萄牙和俄罗斯，在讨论频率与实际立法之间存在明显偏离，表明议会关注不一定直接转化为立法成果。

2016–2024年各立法程序中提及人工智能的次数与人工智能相关法案通过数量的对比

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

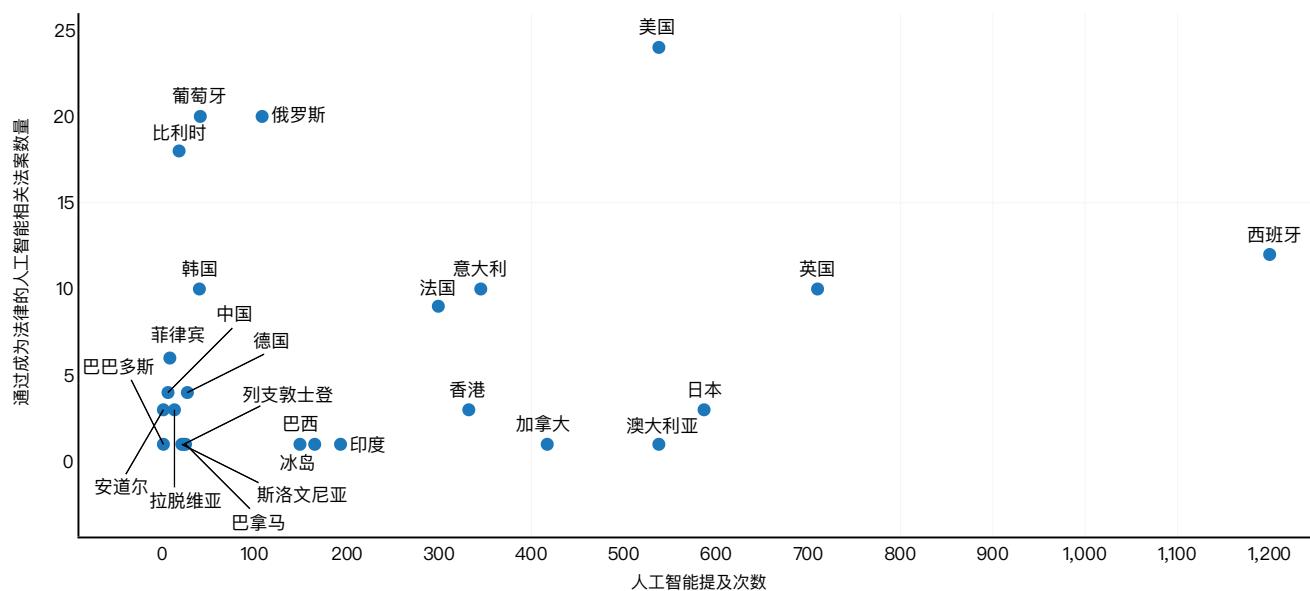


图 6.2.17

第六章：政策与管理

6.2 人工智能与政策制定

美国国会委员会提及情况

美国众议院与参议院各委员会在其报告中对人工智能的提及情况，是衡量国会对人工智能关注程度的另一指标。这些委员会通常负责立法、政策事务、调查与内部事务。图6.2.18记录了2001年至2024年间，美国各届国会委员会报告中提及人工智能的频次。

第118届国会（2023–2024）提及次数达136次，创历史新高，较第117届增长了83.8%。

2001–2024年美国各届国会会期委员会报告中的人工智能提及次数统计

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

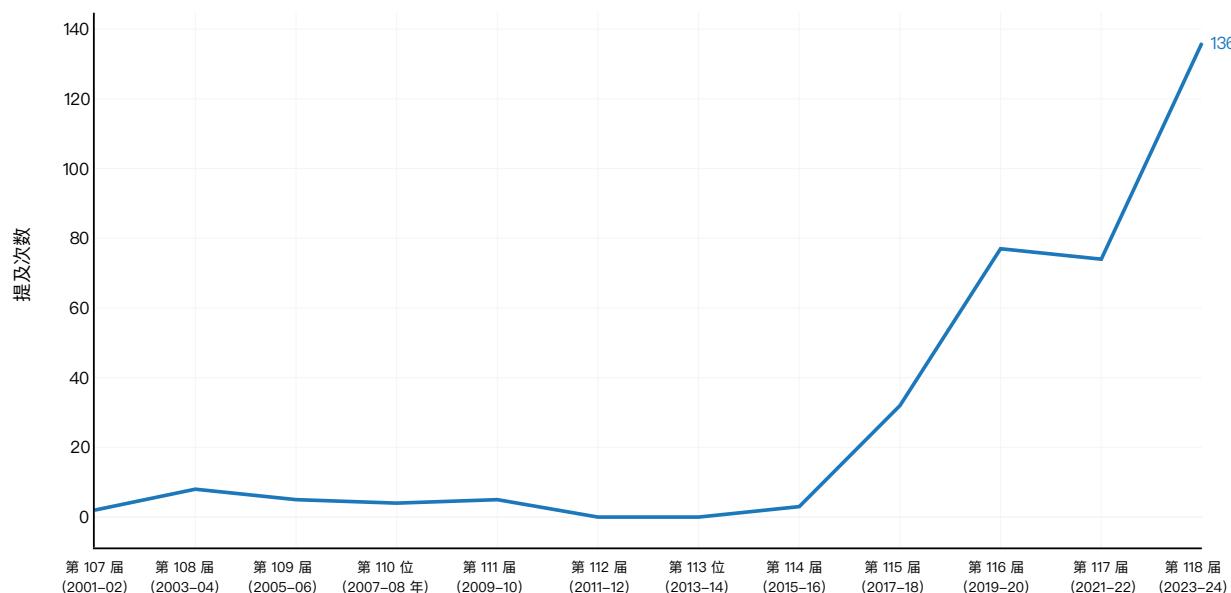


图 6.2.18

第六章：政策与管理

6.2 人工智能与政策制定

美国监管条例

人工智能的兴起引发了监管机构的高度关注——这些联邦机构负责监管经济中的特定领域，并指导法律的执行。本节探讨的是美国的人工智能监管情况。与确立国家法律框架的立法不同，监管条例是由行政当局制定的详细指令，用以执行立法条文。在美国，具有代表性的监管机构包括：环境保护署（EPA）、食品药品监督管理局（FDA）、联邦通信委员会（FCC）。由于立法的具体内容通常通过监管行动得以体现，理解人工智能监管格局是深入理解人工智能政策制定的重要组成部分。本节分析了2016年至2024年间由美国监管机构颁布的人工智能相关监管条例，涵盖监管总数、主题、适用范围、监管意图与发起机构等多个方面。为收集相关数据，人工智能

指数研究团队在美国联邦公报（Federal Register）中进行了关键词“artificial intelligence”的检索。联邦公报是涵盖美国政府几乎所有部门文件的综合性数据库，汇集了来自超过436个联邦机构的信息。

概述

过去六年中，人工智能相关监管数量急剧上升。尤其在过去一年中，增长趋势尤为显著（图6.2.19）。2024年，美国共发布59项人工智能相关监管条例，远高于2023年的25项，实现了超过两倍的增长。

2016–2024 年间美国人工智能相关监管条例数量

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

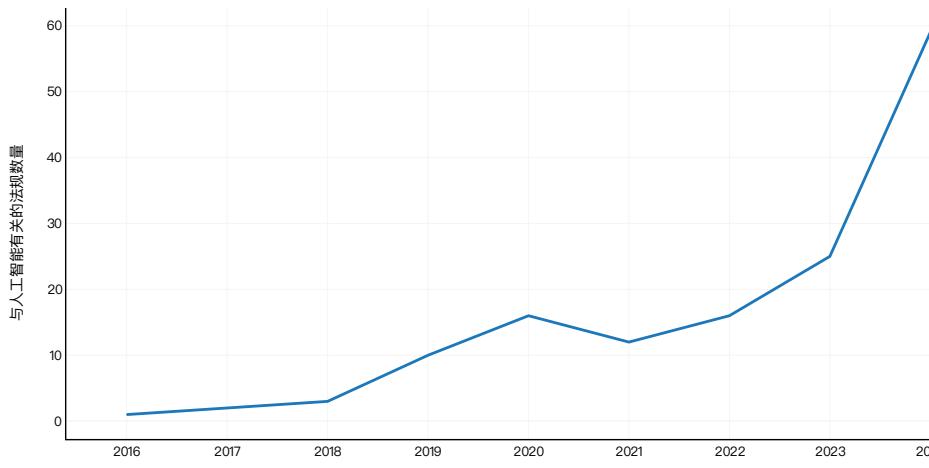


图 6.2.19

按机构分类

图6.2.20展示了自2016年以来，美国不同联邦监管机构所发布的人工智能相关监管条例数量。¹⁰ 2024年，美国卫生与公众服务部发布了最多的人工智能监管条例（共14项），其次是

医疗保险与医疗补助服务中心（7项）和商务部（7项）。人工智能监管条例的来源机构数也达到了历史最高——共计42个机构，远高于2023年的21个与2022年的17个。这一趋势反映出：人工智能正受到美国越来越多监管机构的广泛关注。

10、监管条例可由多个机构共同发起，因此图6.2.20中的机构总数与图6.2.19中的总条例数不完全一致。图6.2.20中的“机构”术语遵循联邦公报的标准用法。

第六章：政策与管理

6.2 人工智能与政策制定

2016—2024 年美国各联邦机构发布的人工智能相关监管条例数量

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

	2016	2017	2018	2019	2020	2021	2022	2023	2024
Agency									
Agency for International Development									1
Census Bureau								1	
Centers for Medicare and Medicaid Services	1	2	4	5	4	3	7		
Children and Families Administration			1	1	1				
Commerce Department			1	1	3	5	7		
Commodity Futures Trading Commission								1	
Comptroller of the Currency						1		2	
Consumer Financial Protection Bureau		1				1	1	4	
Copyright Office, Library of Congress		1					1	1	
Defense Acquisition Regulations System								1	
Defense Department								1	
Education Department			2					2	
Employee Benefits Security Administration						1			
Employment and Training Administration					1				
Energy Department							1	1	
Environmental Protection Agency								2	
Executive Office for Immigration Review								1	
Executive Office of the President	2	6	5	2	3	5	5	5	
Federal Aviation Administration								1	
Federal Communications Commission					1			2	
Federal Deposit Insurance Corporation								1	
Federal Election Commission								1	
Federal Energy Regulatory Commission								1	
Federal Housing Finance Agency								1	
Federal Railroad Administration							1	1	
Federal Reserve System								1	
Federal Trade Commission								2	
Financial Crimes Enforcement Network								1	
Food and Drug Administration				1	1	2	1	2	
Health and Human Services Department	1	2	5	5	5	4	14		
Homeland Security Department					3		1	2	
Housing and Urban Development Department					1		1	1	
Industry and Security Bureau						3	4	4	
Investment Security Office				1				1	
Justice Department								1	
Labor Department					1	1	1	1	
Library of Congress			1					1	
National Credit Union Administration								1	
National Highway Traffic Safety Administration								1	
National Oceanic and Atmospheric Administration								1	
National Science Foundation							1		
Nuclear Regulatory Commission						1			
Occupational Safety and Health Administration							1		
Office of the Inspector General			2	1	1				
Office of the Secretary			1	3	1	1	6		
Patent and Trademark Office			1					1	
Personnel Management Office								1	
Public Health Service					1	1			
Securities and Exchange Commission						1	2	3	
Small Business Administration								1	
Transportation Department							1	4	
Treasury Department					1	1		4	

图 6.2.20

重点：

美国联邦法规深度观察

本节重点介绍了2024年美国联邦政府通过的部分以“规则”与“行政令”形式发布的人工智能相关监管条例（图6.2.21）。

机构	法规名称	内容摘要
总统行政办公室	《防止“关注国家”获取美国敏感个人数据和与美国政府相关的数据》(Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern)	本行政命令将“关注国家”(countries of concern)利用人工智能构成的国家安全威胁列为重大风险。命令特别警告称，外国敌对势力可能借助大量敏感个人数据和美国政府相关数据，训练人工智能算法以执行间谍活动、网络行动和舆论操控。为应对此类风险，命令提出一系列数据保护措施，包括限制或禁止与这些国家的数据交易行为，并强化网络基础设施安全防护。
工业与安全局	《外国生产直接产品规则的补充与对先进计算与半导体制造产品的出口控制强化》(Foreign-Produced Direct Product Rule Additions, and Refinements to Control for Advanced Computing and Semiconductor Manufacturing Items)	本规则修订《美国出口管理条例》，收紧对半导体制造设备和超级计算机出口的管控，特别是对中国的相关限制。规则新增了对半导体制造的进一步限制，更新现有条款，并引入“红旗”(Red Flags)制度，以识别潜在的非法出口风险。此举旨在遏制中国规避原有限制的行为，并限制其开发可对美国国家安全构成威胁的高级计算与人工智能系统的能力。
消费者金融保护局	《2024-06号消费者金融保护通告：用于招聘、晋升和其他就业决策的背景档案与算法评分》(Consumer Financial Protection Circular 2024-06: Background Dossiers and Algorithmic Scores for Hiring, Promotion, and Other Employment Decisions)	本通告明确规定，雇主在作出就业决策时，不得在未经合规的前提下依赖背景档案、算法评分或第三方报告。该规定重申了《公平信用报告法》(Fair Credit Reporting Act, FCRA)下的核心义务，特别是对人工智能驱动系统而言，如获取员工同意后方可调用消费者报告等条款。通过该通告，监管机构对在招聘与用人中使用算法评分设定了明确的边界。
联邦选举委员会	《关于竞选权威虚假陈述的解释性规则》(Fraudulent Misrepresentation of Campaign Authority)	该解释性规则针对人工智能生成内容日益增多的背景，对《联邦选举法案》(Federal Election Campaign Act, FECA)进行了补充说明。规则重申 FECA 具有“技术中立性”，监管重点并非特指人工智能误用，而是聚焦于个人或组织是否参与了与选举有关的虚假陈述行为。
财政部投资安全办公室	《关于美国在“关注国家”中对某些国家安全技术与产品投资的条款》(Provisions Pertaining to U.S. Investments in Certain National Security Technologies and Products in Countries of Concern)	本最终规则落实了第 14105 号总统行政命令，要求美国公民在与“关注国家”中从事敏感技术领域的实体进行交易时，必须向财政部进行通报，并在某些情况下禁止交易。该命令发布于 2023 年，所涉高风险技术领域包括人工智能、半导体与量子计算。美国政府认为，对这些领域的投资可能提升敌对国家对美国家安全构成威胁的能力。

图 6.2.20

第六章：政策与管理

6.3 人工智能公共投资

随着人工智能持续推动医疗、交通、国防等关键领域的创新，公共资金已成为各国实现其人工智能战略的关键支柱。了解各国政府在人工智能研究与开发上的投入，对于理解更广阔的地缘政治人工智能格局至关重要，但对这些投资进行追踪仍面临重大挑战。尽管国家预算可能会列出与人工智能相关的支出项目，但这些预算拨款并不总能直接反映为实际支出。此外，人工智能投资往往被嵌套在更广义的科学或技术倡议中，导致精准识别人工智能专属资金配置变得较为困难。

为解决这一问题，人工智能指数使用自然语言处理（NLP）技术分析了各国公开发布的招标与合同文件，从中识别与人工智能相关的政府支出。¹² 这种对招标文件的分析方式，能更直接反映投资趋势，也更好展现政府随时间推移如何配置资源。由于人工智能指数仅分析了公开披露合同和招标数据的国家，一些国家因此未被纳入分析范畴。¹³ 本节还单独分析了美国在人工智能领域的科研拨款支出总额。

人工智能指数报告特别提醒：基于本节公布的政府支出数据进行跨国直接比较需谨慎。尽管本次分析涵盖了多国政府合同数据，但仅包含美国联邦层面的科研资助支出数据。这种数据不对称性源于从欧盟、中国等其他国家和地区获取可比性资

助数据的复杂性和难度。以美国为例，政府人工智能指数支出中科研资助占据重要比重。2023年，据人工智能指数估算，美国政府人工智能指数相关公共合同金额约8.3亿美元，而同期人工智能指数相关科研资助高达45亿美元。鉴于当前跨国数据可获性和一致性的局限，进行各国人工智能指数公共支出的比较分析仍为时过早。本项研究旨在为建立更全面的全球数据覆盖迈出第一步。人工智能指数报告将持续推进这项工作，并欢迎有意提升数据范围和质量的研究人员、机构及政府开展合作。

11、本节的分析工作由 Lapo Santarasci 主持。

12、附录中详细介绍了这一分析方法背后的全部方法。报告的滞后性可能导致 2024 年的数据不完整，最新的分析是 2023 年底的数据。

13、由于数据获取限制，本次分析未涵盖欧盟（整体层面）和中国等主要政府 AI 项目资助地区。人工智能指数承诺将在未来版本中扩大研究范围，将这些地区及其他区域纳入统计。

第六章：政策与管理

6.3 人工智能公共投资

人工智能公共投资总额

图6.3.1总结了各国人工智能相关合同数量及其价值的核心数据。¹⁴ 从2013年至2023年，美国在人工智能公共投资方面位居全球首位，共签订2678份独立人工智能合同，总金额约为52亿美元（图6.3.1和图6.3.2）。在欧洲，英国、德国和

法国的人工智能合同总价值最高，合计占欧洲公共人工智能投资的56%。在欧洲，英国、德国和法国的人工智能合同总价值最高，合计占欧洲公共人工智能投资的56%。

2013–2023 年部分国家人工智能相关公共合同支出总额

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

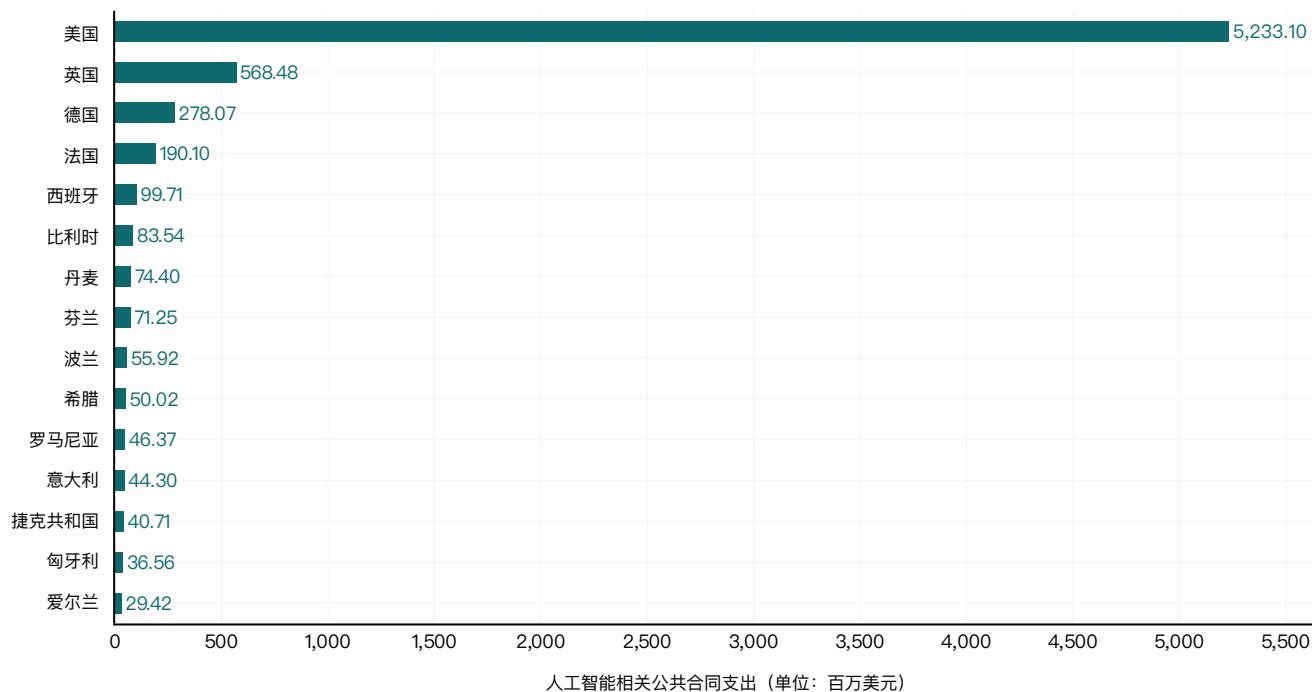


图 6.3.1

14、所提供的结果和数字受特定匹配投标样本缺失值比率的影响：NAICS 代码为 0.16%，美元价值为 26.8%。值得注意的是，样本中不包括北爱尔兰的投标书，因为其办事处不提供 API 服务或批量下载选项以进行大规模数据收集。

第六章：政策与管理

6.3 人工智能公共投资

2013–2023 年部分国家人工智能相关合同总数

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

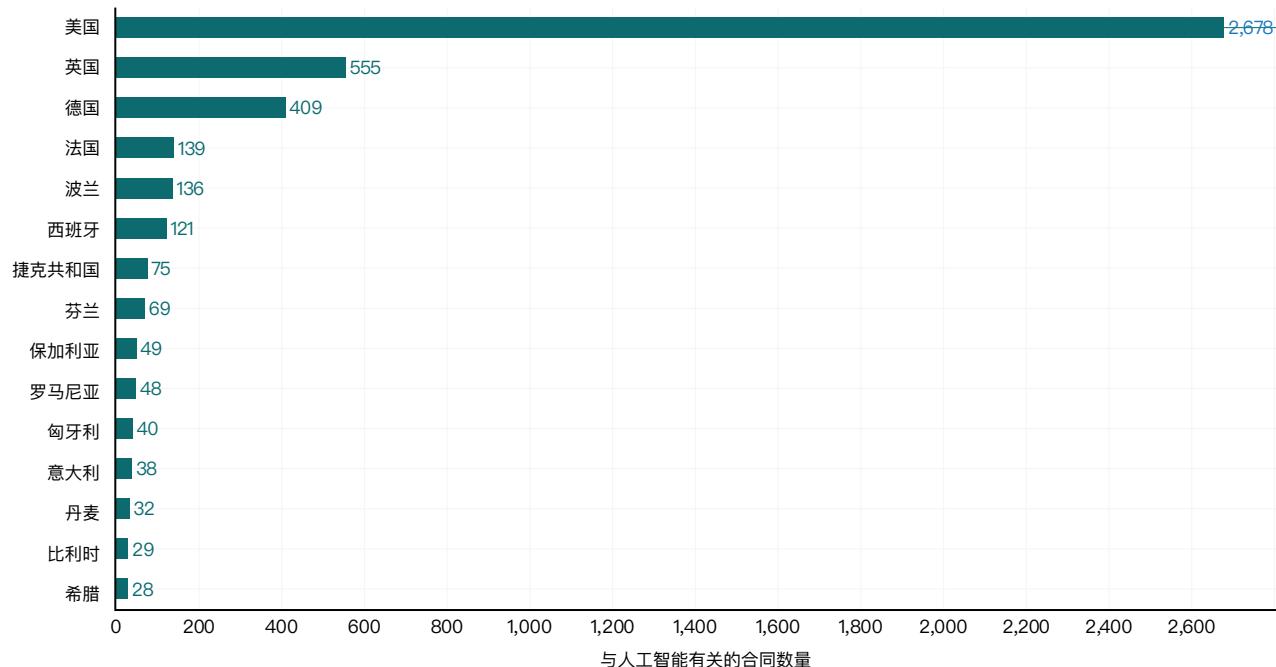


图 6.3.2

2013–2023 年部分国家人工智能相关公共合同的中位数金额

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

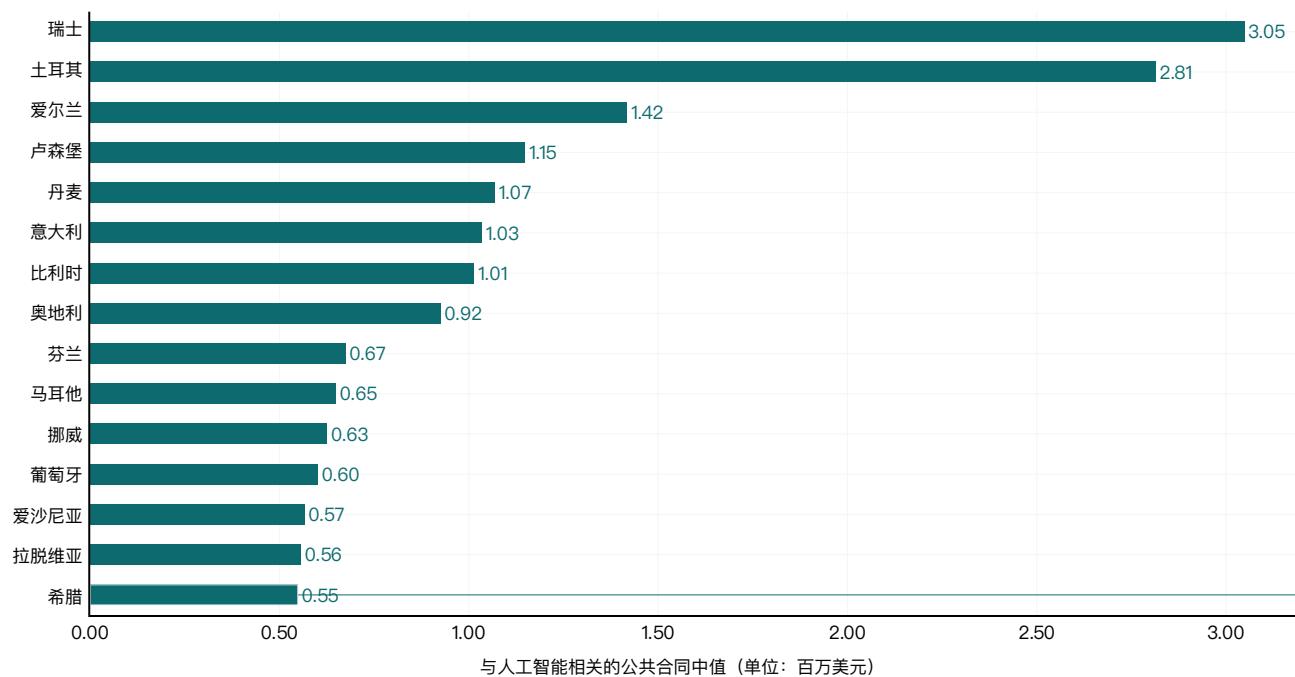


图 6.3.3

第六章：政策与管理

6.3 人工智能公共投资

过去十年间，哪些国家在人工智能领域的投入最高？美国以每 10 万居民投入 158 万美元位居榜首，芬兰(130 万美元)和丹麦(130 万美元)紧随其后(图 6.3.4)。

2013–2023 年，部分国家每 10 万居民在人工智能相关公共合同支出金额（合计）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

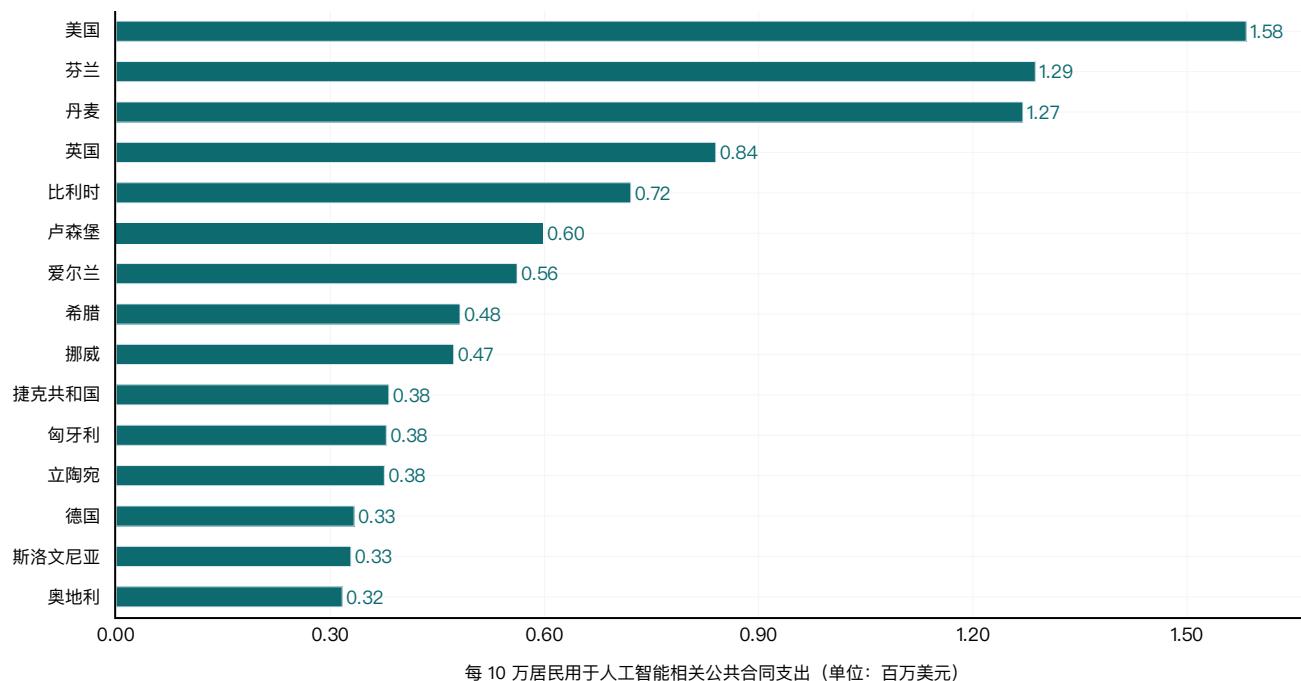


图 6.3.4

第六章：政策与管理

6.3 人工智能公共投资

图 6.3.5 展示了 2023 年人工智能领域的公共投资情况。美国的人工智能合同支出为 8.31 亿美元，远超其他国家，英国以 2.63 亿美元位列第二。虽然德国、西班牙、英国依然是欧洲

投资大户，但罗马尼亚、希腊、匈牙利和波兰等以往排名靠后的国家也进入了前十。这一变化表明，人工智能资金在欧洲的分布趋于更加平衡。

2023 年部分国家用于人工智能相关公共支出合同金额

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

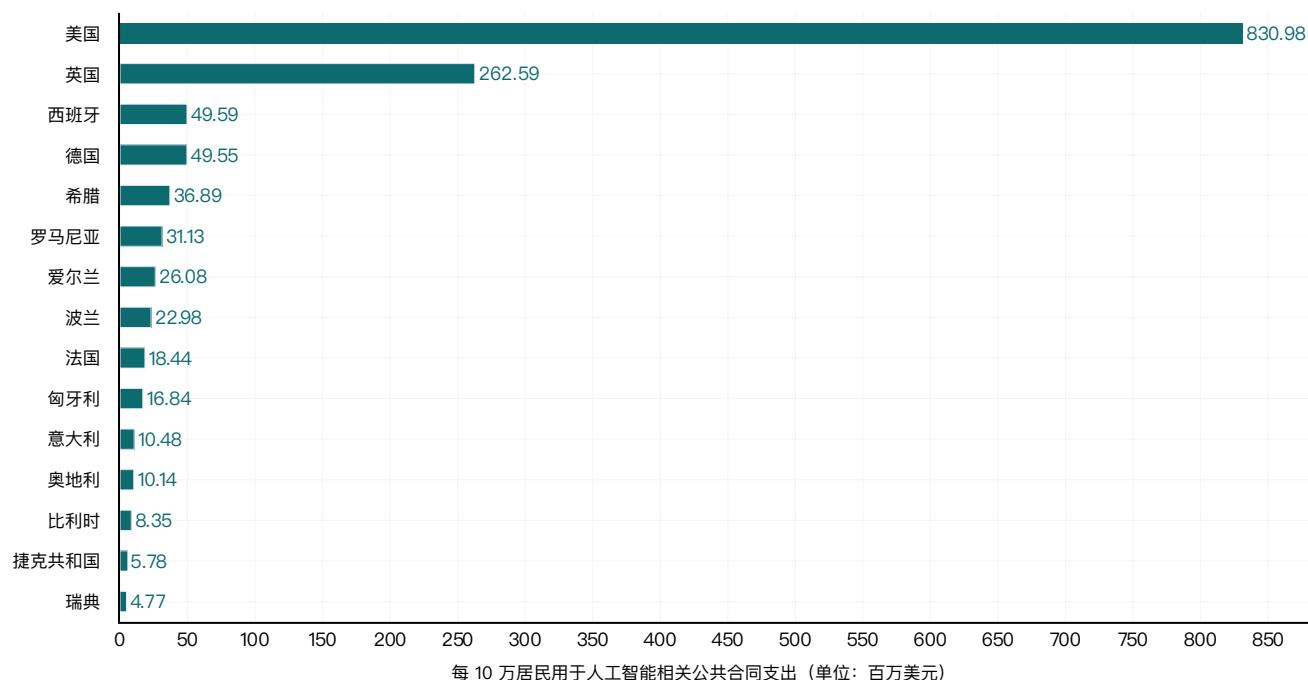


图 6.3.5

第六章：政策与管理

6.3 人工智能公共投资

图6.3.6展示了美国与欧洲这两个人工智能投资重点地区在过去十年间的公共人工智能投资趋势。数据显示，两地区的人工智能相关支出均显著增长。尤其值得注意的是，2023年欧洲的人工智能投资总额较2013年增长约67倍，而美国则增长约15倍。欧洲在2017年和2019年经历了特别显著的跃升：2017年同

比增长400%，2019年再次上涨200%。值得一提的是，2019年也是全球范围内国家人工智能战略发布数量达到峰值的一年。这一持续上升趋势清晰地说明了各国政府在财政上对人工智能的重视程度与投入意愿。

2013–2023年美国和欧洲人工智能相关公共合同支出金额

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

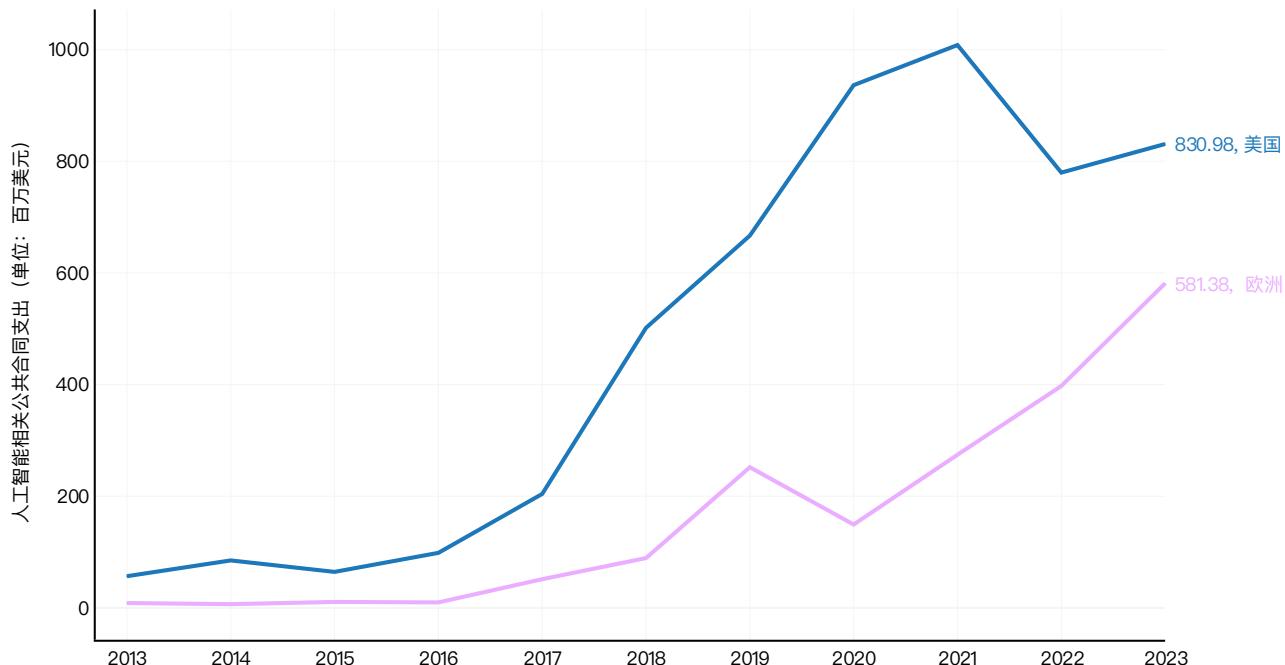


图 6.3.6

第六章：政策与管理

6.3 人工智能公共投资

图6.3.7呈现了美国与欧洲在人工智能公共支出方面的差额变化。数据显示，这一差距在2020年前持续扩大，但过去三年

开始逐步缩小，表明欧洲国家正在逐步赶上美国在人工智能相关公共支出方面的步伐。

2013–2023年美国和欧洲在人工智能相关公共合同支出差异

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

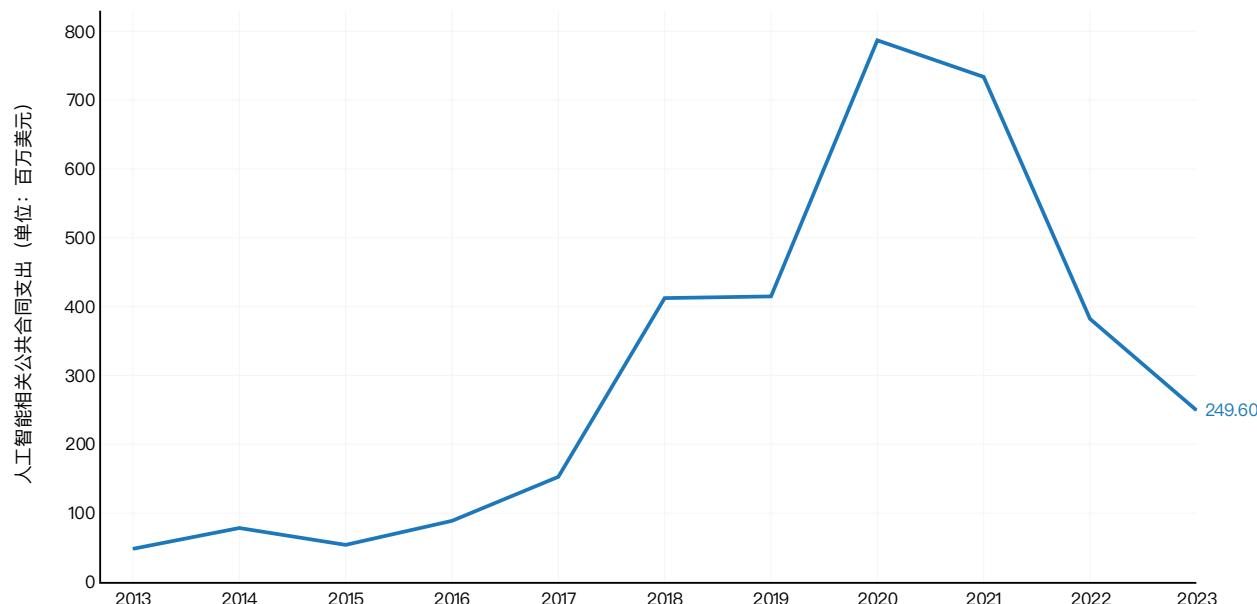


图 6.3.7

第六章：政策与管理

6.3 人工智能公共投资

图6.3.8展示了2013年至2023年欧洲五大主要国家——比利时、法国、德国、西班牙和英国的人工智能相关公共支出数据。显示投资呈现稳定增长趋势，并伴有周期性峰值。其中，德国在2018年11月发布国家人工智能战略后，于2019年实现显著增长。英国则在2021年和2023年出现两次人工智能公共投资

的急剧上升，这与其人工智能委员会提出的国家战略密切相关——该独立专家委员会成立于2019年，旨在为政府提供建议并对人工智能生态系统进行高层指导。相比之下，比利时、法国和西班牙的增长幅度较为温和但保持稳定态势。

2013–2023 年欧洲五大主要国家用于人工智能相关公共合同支出金额

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

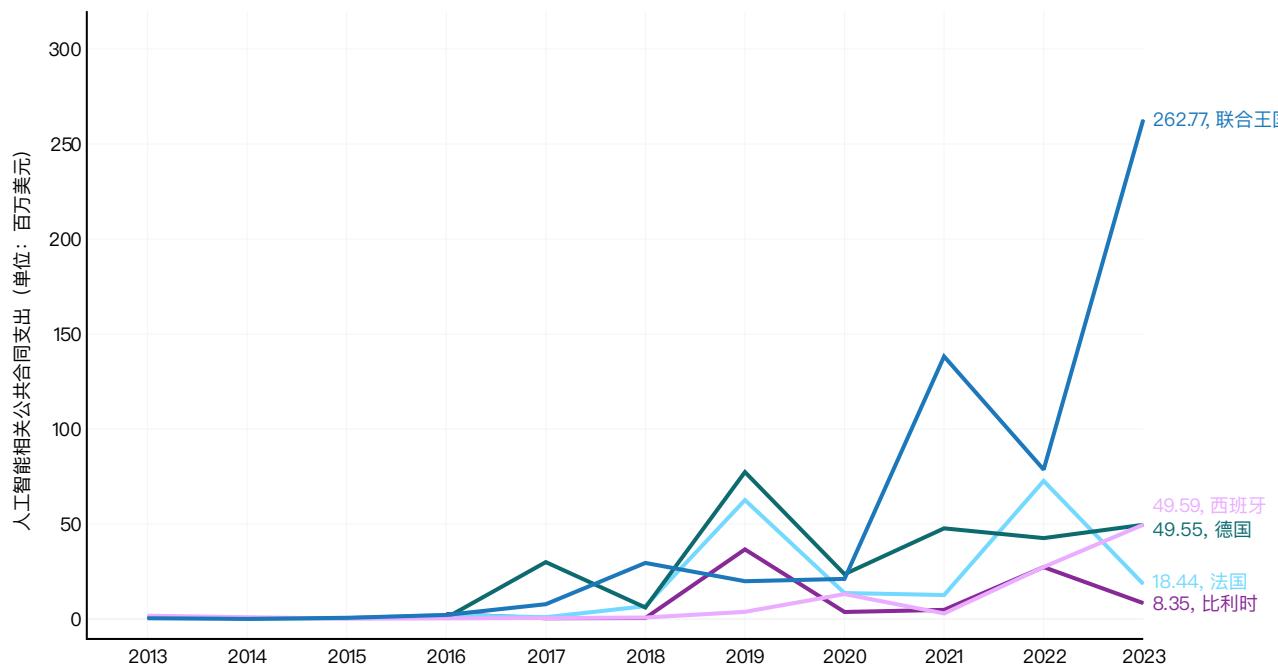


图 6.3.8

按机构与职能领域划分的人工智能公共支出分布

美国与欧洲在人工智能公共招标投资的分布格局存在显著差异，这种差异源于双方在战略优先级与制度结构上的不同。如图6.3.9所示，自2013年以来，美国绝大多数人工智能相关合同由国防部签订，这与该机构在美国技术创新体系中长期扮演的核心角色相一致。2023年，美国人工智能相关公共合同中，国防部占比为75.04%。退伍军人事务部位列第

二，占比为6.83%；财政部位居第三，占比为5.34%。

退伍军人事务部在人工智能领域的投资主要集中在医疗和康复相关应用，包括基于人工智能的辅助诊断、机器人义肢研发以及心理健康支持系统的构建等方向。这些应用反映出该部门在推动智能医疗服务方面的持续投入。

2013–2023 年按机构分类的人工智能相关科研拨款的公共支出（占总额的百分比）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

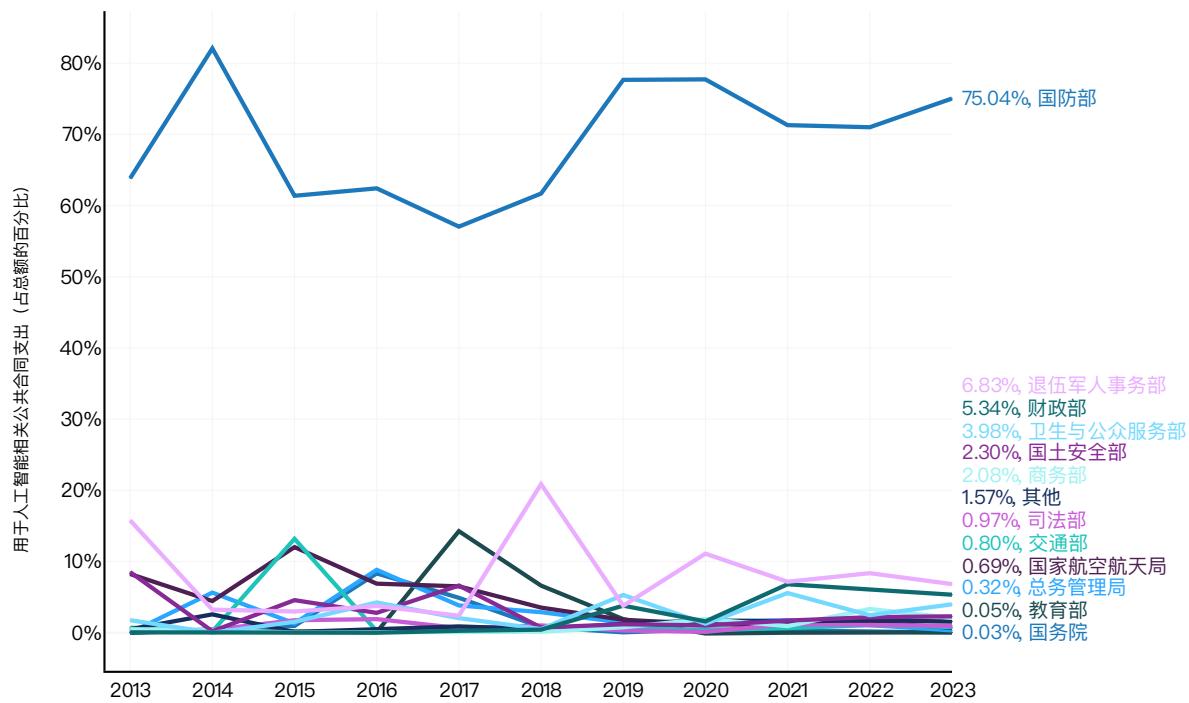


图 6.3.9

第六章：政策与管理

6.3 人工智能公共投资

在欧洲，通过公共招标实现的人工智能投资呈现出与美国显著不同的模式。由于缺乏类似美国那样集中汇总的财政支出数据，人工智能指数采用了按主要职能类别对资助实体进行分类的方法，以分析欧洲的人工智能公共投资结构。如图6.3.10所示，欧洲的人工智能投资在不同职能类别之间分布更加均衡。

其中，2023年排名前三的资助领域为：一般公共服务、教育、医疗健康，三者合计占据了当年欧洲人工智能相关公共投资总额的约84%。在同一年，国防相关支出仅占全部人工智能公共招标投资的0.84%，这一比例与美国形成了鲜明对比。在美国，国防是人工智能资金的绝对主要接收方。

2013–2023年，按资金提供机构分类，欧洲各国政府在人工智能相关公共合同支出（占总支出的百分比）

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

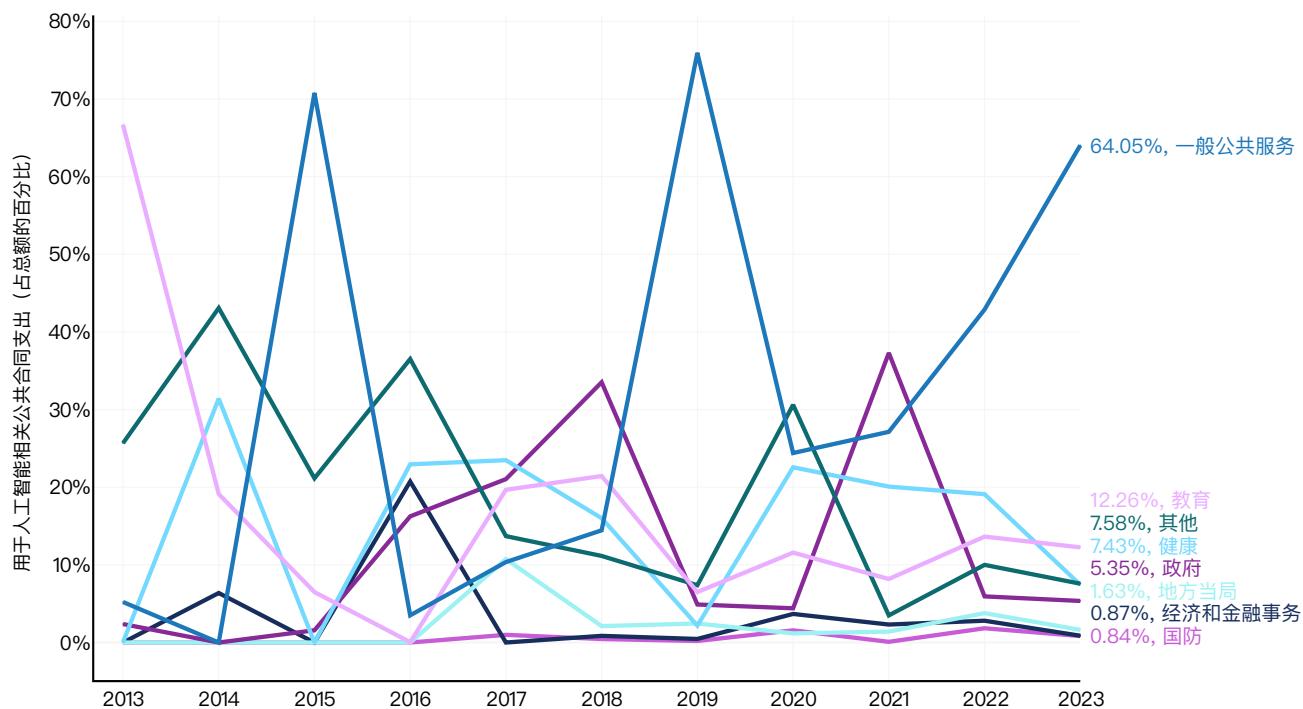


图 6.3.10

第六章：政策与管理

6.3 人工智能公共投资

重点：

美国人工智能科研拨款分析

公共拨款是政府支持人工智能项目和相关计划的重要渠道之一。通过此类拨款，公共机构可以直接投资于多种人工智能应用项目，例如，提升 X 光血管造影(X-ray angiography)的解读效果，构建用于自动化土壤监测的人工智能驱动无人机系统。开发可解释的机器学习工具等。政府可通过向国家科学基金会(NSF)或卫生与公众服务部(HHS，其包括国家卫生研究院(NIH))等机构提供研究拨款，以支持专注于人工智能的科研项目。在本节中，人工智能指数对美国政府在人工智能相关专项中的拨款数据进行了分析。与前文一致，人工智能指数同样采用自然语言处理(NLP)方法来识别人工智能相关拨款项目。¹⁵

图 6.3.11 展示了 2013 年至 2023 年美国人工智能相关拨款支出的汇总数据。在此期间，美国联邦政府累计向人工智能相关研究项目拨款总额约 197 亿美元。

2013–2023 年美国人工智能相关拨款金额统计

资料来源：2025 年人工智能指数 | 表：2025 年人工智能指数报告

资助金额统计	价值
资助项目数量	18,399
总金额（百万美元）	19,748.44
中位数（单位：千美元）	247.53
平均金额（单位：千美元）	1,073.34
人均资助额（千美元/10万人）	5,967.69

图 6.3.11

图 6.3.12 展示了美国人工智能拨款随时间的变化趋势。

2013 年至 2023 年间，美国人工智能科研拨款总额增长近 19 倍，从最初的 2.3 亿美元上升至 45 亿美元。在 2014 年至 2020 年期间，拨款年均增长率约为 40%。这一下拨款规模的迅速扩张，与人工智能技术在过去十年间的持续演进密切相关。特别是在深度学习、自然语言处理以及计算机视觉等核心领域取得关键性进展的背景下，公共部门对于人工智能在具体应用场景中的部署需求不断上升，进而推动了政府对相关研究项目的持续加码投入。

2013–2023 年按机构分类的人工智能相关科研拨款的公共支出（占总额的百分比）

资料来源：2025 年人工智能指数 | 图表：2025 年人工智能指数报告

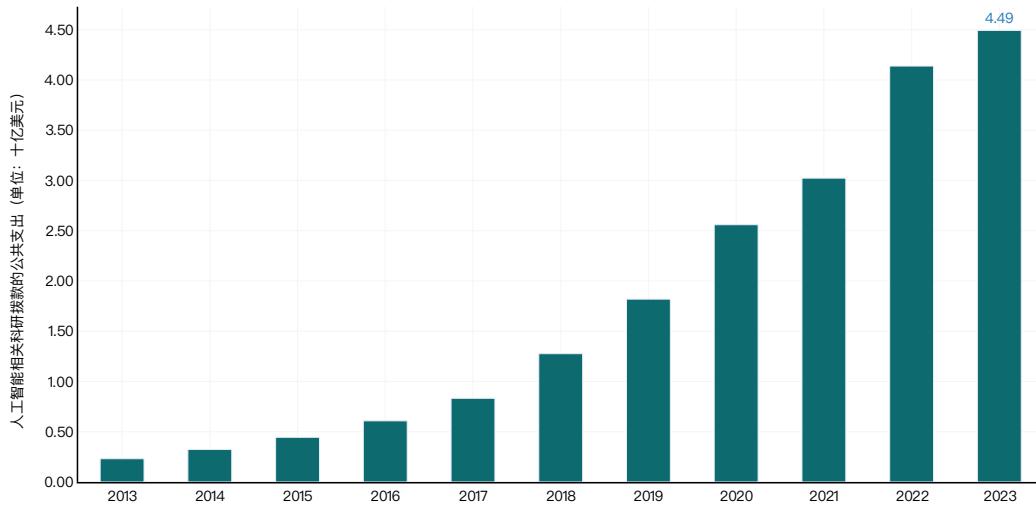


图 6.3.12

15、这种方法的全部原理见附录。

重点：

美国人工智能科研资助分析（续）

图6.3.13展示了2013年至2023年间，美国人工智能相关拨款在各资助机构之间的分布情况。其中，卫生与公众服务部所获得的拨款比例最高，达43.6%，其次为国家科学基金会，占27.9%，商务部位列第三，占5.4%。

2013–2023年按机构分类的人工智能相关科研拨款的公共支出（占总额的百分比）

资料来源：2025年人工智能指数 | 图表：2025年人工智能指数报告

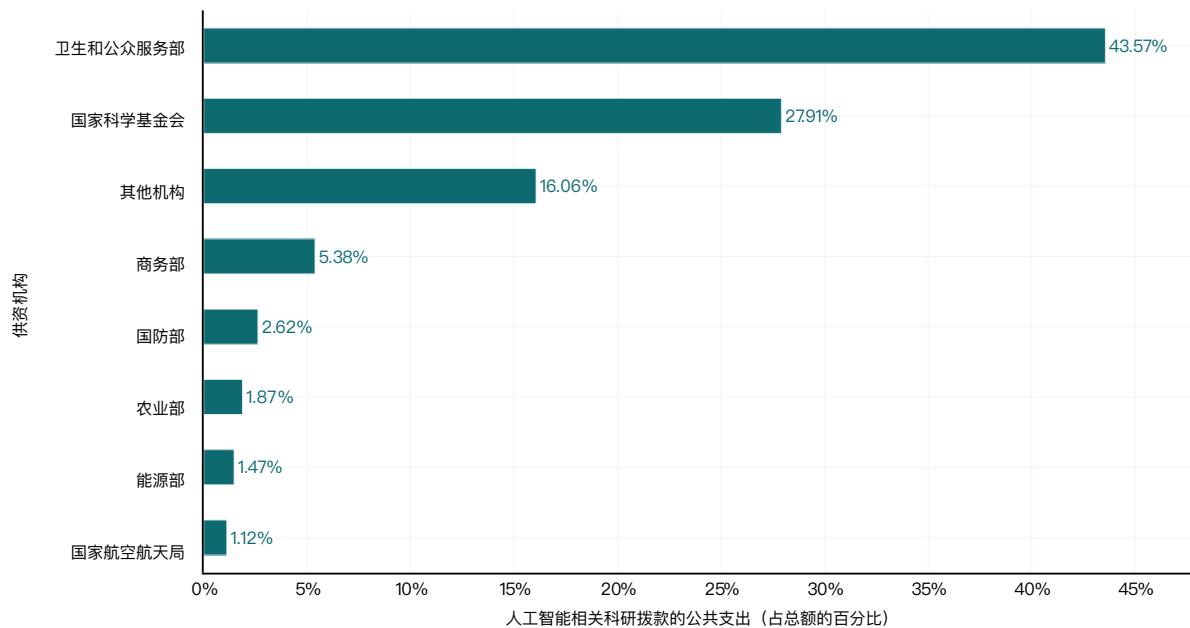


图 6.3.13

第七章： 教育



第七章：教育

概述	366
章节要点	367
7.1 背景介绍	368
7.2 K–12 阶段的计算机科学与人工智能教育	369
美国	369
基础计算机科学教育	369
高阶计算机科学教育	373
教育标准与政策指导	376
教师视角	377
全球现状	379
普及情况	379
政策指导	380
7.3 高等教育阶段的计算机科学与人工智能教育	382
学位授予情况	382
美国	382
全球	388
指导政策	392
7.4 展望未来	393

获取公共数据

第七章： 教育

概述

人工智能通过生成式人工智能对工作的影响——提升效率、实现任务自动化——已进入公众视野，同时也推动了教育领域个性化学习的创新。然而这项技术虽前景广阔，却暗藏风险：从生成虚假输出的“幻觉问题”，到强化社会偏见、削弱批判性思维。随着人工智能教育市场规模预计将大幅增长，技术滥用引发的伦理问题日益凸显——人工智能工具曾错误指控边缘化学生作弊的事件，正警示着负责任的技术开发与部署的紧迫性。

应对这些挑战需要技术素养与批判性审视社会影响的双重能力。培养 AI 专业人才必须从 K-12 阶段基础教育与高等教育着手，确保学生成为负责任的使用者与开发者。人工智能教育不能孤立存在——必须与更广泛的计算机科学 (CS) 教育体系协同发展。本章节将审视全球人工智能与计算机科学教育发展现状、教育机会差异，以及塑造人工智能教育角色的政策框架。

本章节由卡普尔基金会、计算机科学教师协会 (CSTA)、公共利益技术大学联盟 (PIT-UN) 与人工智能指数联合撰写。卡普尔基金会致力于种族平等与技术创新的交叉领域，通过构建公平包容的计算教育路径、推进减轻技术危害并促进机会平等的政策、部署资本支持负责任且符合伦理的技术方案来实现使命。计算机科学教师协会作为全球会员组织，通过团结、支持并赋能教育工作者来提升计算机科学教育的质量、可及性与包容性。公共利益技术大学联盟 (PIT-UN) 则促进高校间合作，共同建设公共利益技术领域，培养心系公共利益的新一代技术人才。

第七章：

教育

章节要点

1. 美国高中计算机科学(Computer Science, CS)课程的普及率与选修人数较上一学年略有提升，但教育差距依然存在。学生的参与情况因州、种族和民族、学校规模、地理位置、收入、性别和残疾而异。

2. 美国计算机科学的教师希望传授人工智能，但认为自己不具备这样的能力。尽管 81% 的计算机科学教师认同应将人工智能应用及人工智能基础知识纳入计算机科学基础课程体系，但仅有不足半数的高中计算机科学教师认为自己具备开展人工智能教学的专业能力。

3. 全世界有三分之二的国家提供或计划提供 K-12 阶段的计算机科学教育。自 2019 年以来，这一比例翻了一番，其中非洲和拉丁美洲国家进展最为显著。然而，由于学校缺乏电力供应，非洲国家学生获得计算机科学教育的机会最少。

4. 2022 年至 2023 年期间，美国获得人工智能硕士学位的毕业生人数几乎翻了一番。尽管人工智能在学士学位和博士学位中的关注度提升将较为缓慢，但硕士学位的激增可能预示着所有学位层次都将出现这一发展趋势。

5. 在培养信息、技术和通信技术(ICT)毕业生方面，美国持续处于全球领先地位。西班牙、巴西和英国紧随美国之后，成为各层次毕业生最多的国家，而土耳其则是男女比例最均衡的国家。

7.1 背景介绍

要深入把握当前人工智能教育的发展态势，必须对教育领域的人工智能应用、人工智能素养教育以及人工智能专业教育这三者作出明确界定（图 7.1.1）。教育领域的人工智能应用主要指人工智能技术在教学过程中的实际运用；人工智能素养教育则侧重于培养对人工智能技术的基本认知能力，包括理解其

运作机制、掌握使用方法以及认识潜在风险；而人工智能专业教育不仅包含上述素养要求，更致力于培养学生开发人工智能系统所需的专业技术能力，例如支撑人工智能技术的数据分析能力、数据偏差的识别与修正等关键技能。本章节所采用的数据指标主要针对人工智能专业教育这一范畴。

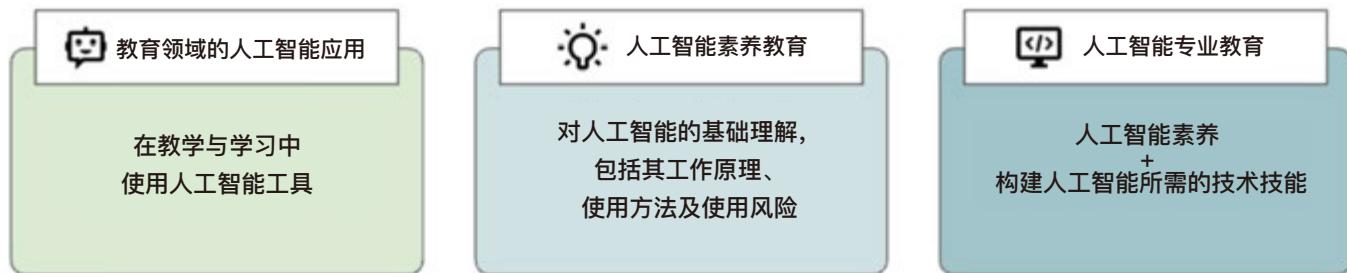


图 7.1.1

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

全球在构建一个具备韧性与多样性的劳动力队伍方面面临严峻挑战，尤其是在基础设施、资源获取、课程参与等方面存在巨大差异，而这些差异进一步加剧了K-12阶段学生在迈向科技驱动未来中的不平等起点。尽管由于数据收集与指标设定缺乏标准化，问题的实际规模难以精准估算，本节仍聚焦于计算技术教育最早期的环节，结合现有的全球数据审视当前K-12阶段的计算机科学及人工智能教育现状。

7.2 K-12 阶段的计算机科学与人工智能教育¹

美国

在探讨美国人工智能教育的普及程度与教育质量之前，有必要先回顾其在计算机科学教育方面的历史演进。自2016年奥巴马总统发起“全民计算机科学教育”倡议以来，联邦政府投入数十亿美元，旨在确保所有K-12学生皆具备计算机科学学习机会，以培养他们成为数字经济的创造者与科技社会中的负责任公民。该项拨款重点支持专业教师培训、教学资源优化及区域合作机制建设，以扩大计算机科学教育的覆盖面。国家科学基金会(NSF)也主导开发与推行两门新课程——《探索计算机科学》和《AP计算机科学原理》，以吸引更广泛的学生群体参与计算教育。同时，科技产业界与慈善机构亦共同出资，推动全国性项目，使数以百万计的学生得以接触计算机科学教育。

基础计算机科学教育

过去十年中，教育倡导者敦促政策制定者通过立法改善计算机科学教育的普及。这一系列努力已取得初步成效：2017—2018学年，仅有35%的美国高中开设计算机科学课程；而到了2023—2024学年，该比例已上升至60%。然而，国家层面的整体数据仍可能掩盖不同州之间的差异。例如，阿肯色州与马里兰州的所有高中（100%）皆开设计算机科学课程，而蒙大拿州的覆盖率为31%（图7.2.1）

2024年各州公立高中开设计算机科学基础课程的比例（占该州高中总数的百分比）

资料来源：Code.org, CSTA, and ECEP Alliance, 2024 | 图表：2025年人工智能指数报告

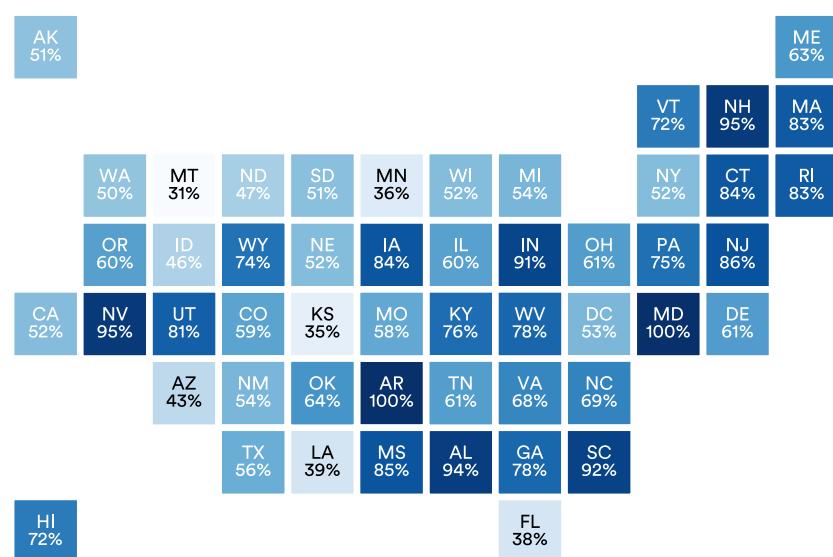


图 7.2.1

1. 由于人工智能传统上属于计算机科学的分支领域，本章节在缺乏人工智能专项数据时引用计算机科学教育数据。

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

在公平接受计算机科学教育方面仍存在显著差距，部分学生群体被边缘化。2023–2024 学年，部分学生群体的课程覆盖仍显不足：包括符合免费或减价午餐(FRL)资格的学生、小型

学校学生、城市与乡村地区的学 生，以及美洲原住民学生（图 7.2.2 至 7.2.5）。

2024 年按规模分类的开设计算机科学基础课程的学校

资料来源：Code.org, CSTA, and ECEP Alliance, 2024 | 图表：2025 年人工智能指数报告

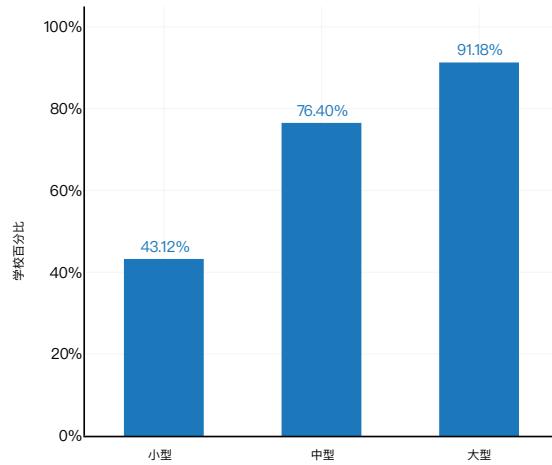


图 7.2.2

2024 年按地理区域划分的开设计算机科学基础课程的学校

资料来源：Code.org, CSTA, and ECEP Alliance, 2024 | 图表：2025 年人工智能指数报告

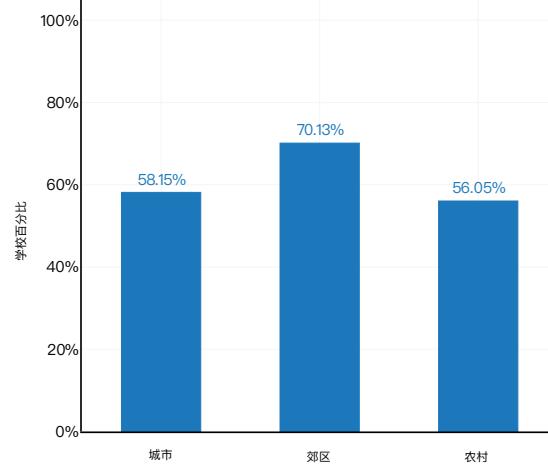


图 7.2.3

2024 年按免费和减免午餐学生人数分类的开设计算机科学基础课程的学校

资料来源：Code.org, CSTA, and ECEP Alliance, 2024 | 图表：2025 年人工智能指数报告

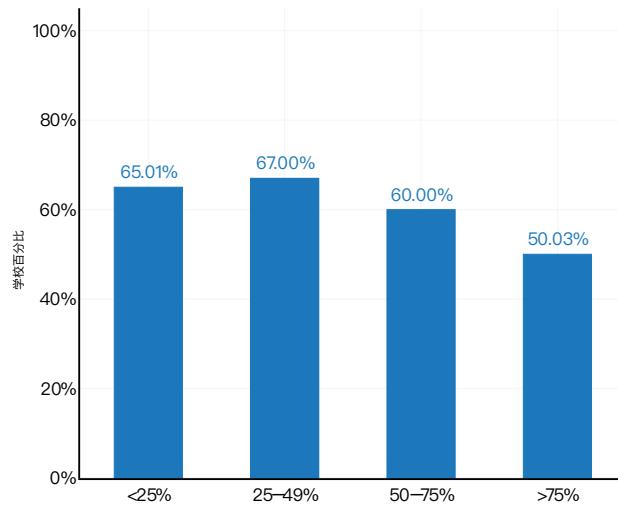


图 7.2.4

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

2024 年按种族 / 族裔分类的学习计算机科学基础课程的情况

资料来源 : Code.org, CSTA, and ECEP Alliance, 2024 | 图表: 2025 年人工智能指数报告

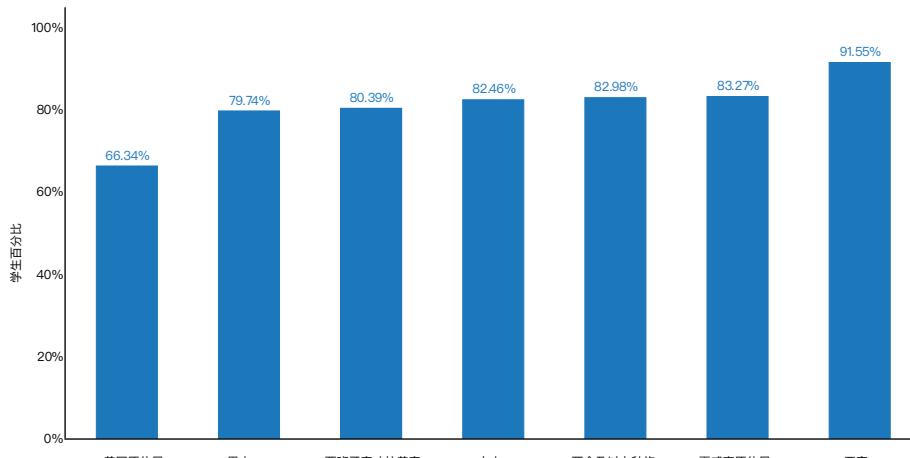


图 7.2.5

此外,根据 41 个州的学生参与数据,计算机科学课程的实际选修情况亦显不足。2020–2021 学年,仅有 5.1% 的高中生参与了计算机科学课程;至 2023–2024 学年,该比例仅微幅上升至 6.4%。各州之间的参与差异同样显著——例如南卡罗来纳州有 26% 的高中生参与计算机科学课程,而佛罗里达州、亚利桑那州与爱达荷州的参与率仅为 2% (图 7.2.6)

2024 年高中计算机科学专业招生人数占学生总数的比例

资料来源 :: Code.org, CSTA, and ECEP Alliance, 2024 | 图表: 2025 年人工智能指数报告

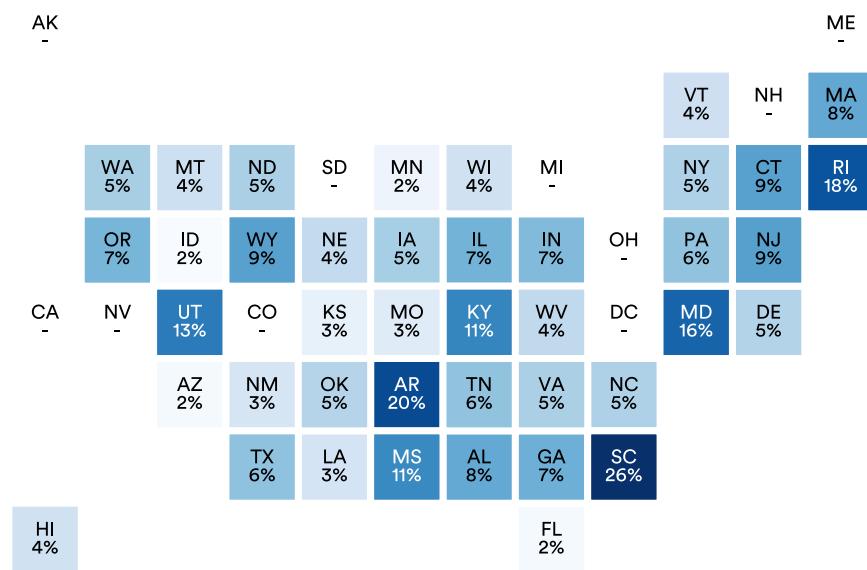


图 7.2.6

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

按种族和族裔划分的计算机科学课程参与数据显示，推广工作的确在一定程度上提高了非洲裔、美洲原住民 / 阿拉斯加原住民以及白人学生的参与度，在全国范围内接近甚至超过其人口占比（图 7.2.7）。然而，数据不完整（尤其缺少九个州的数据）也提醒我们应谨慎解读整体趋势。女性学生在计算机科学课

程中的参与度明显低于其在 K-12 阶段人口中的比例。此外，拉丁裔与太平洋岛裔学生、持有个性化教育计划（IEP）学生、FRL 学生以及英语学习者在全国范围内均呈现参与不足的趋势（图 7.2.7 与图 7.2.8）。

2024 年按种族 / 族裔分类计算机科学公立高中入学率与全国人口统计数据对比

资料来源：Code.org, CSTA, and ECEP Alliance, 2024 | 图表：2025 年人工智能指数报告

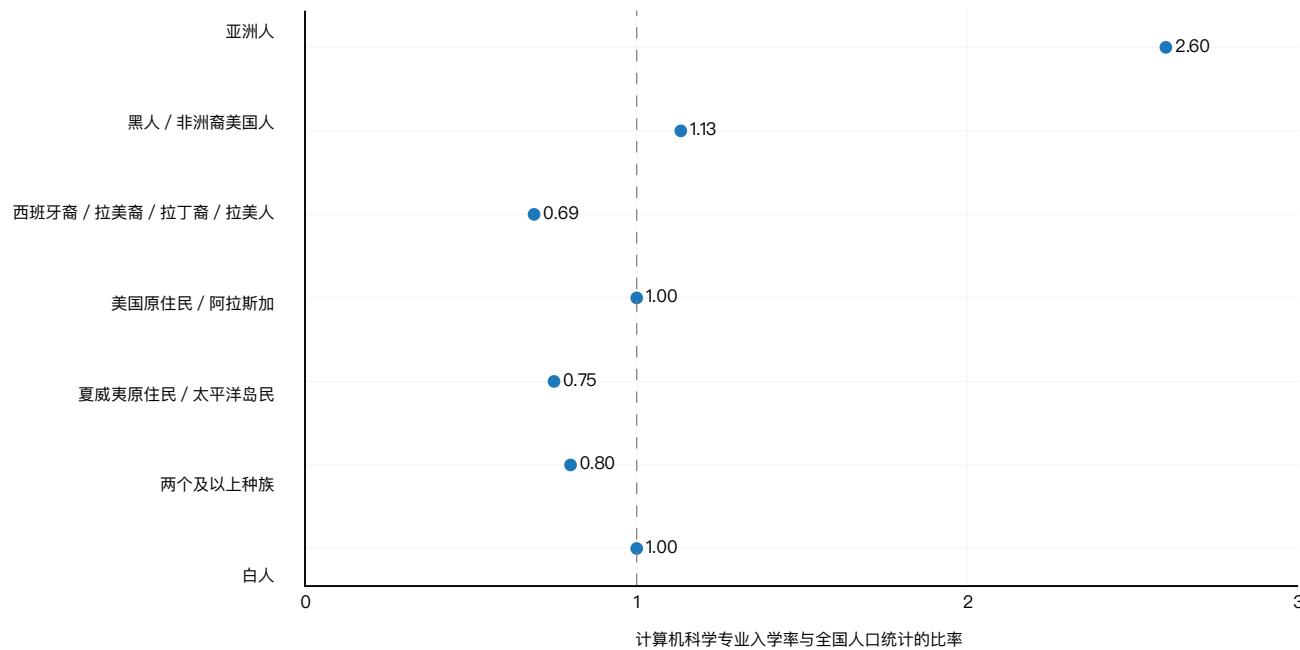


图 7.2.7

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

2024 年按子群体划分的计算机科学公立高中入学率与全国人口统计数据对比

资料来源：Code.org, CSTA, and ECEP Alliance, 2024 | 图表：2025 年人工智能指数报告

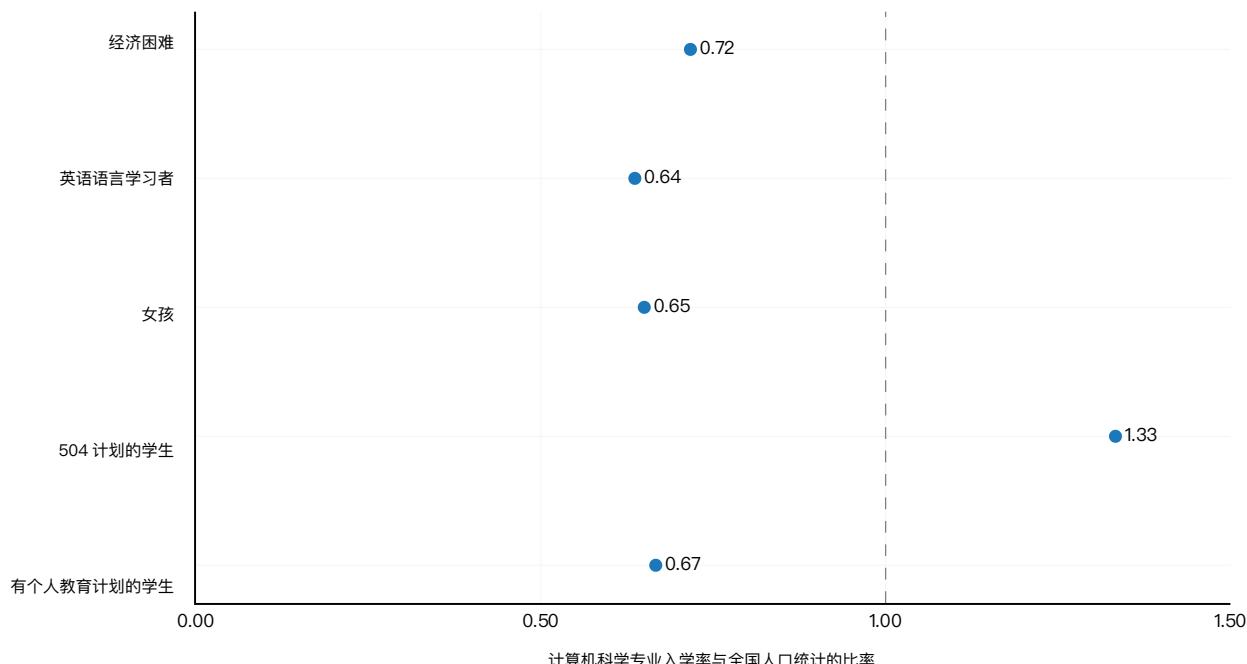


图 7.2.8²

高阶计算机科学教育

为提升学生在人工智能领域的能力，仅提供基础课程远远不够，还需确保其能接触更高阶的课程内容。尽管当前的 AP 计算机科学 A(AP CS A)课程并未明确覆盖人工智能内容，但《AP 计算机科学原理》(AP CS Principles, AP CS P)课程已开始纳入相关主题。因此，AP CS P 具备潜力，可让更广泛的学生群体初步接触人工智能相关知识。尽管参加 AP CS 考试的学生总人

数呈现持续增长(图 7.2.9)，但就整体学生人口的种族与族裔构成而言，各群体参与情况仍存在失衡(图 7.2.10 和图 7.2.11)。具体而言，亚裔学生、白人男生以及多种族学生在 AP CS 考试中比例显著偏高，而其他学生群体的参与程度则明显不足(图 7.2.12)。

²、504 计划为《1973 年康复法案》第 504 节所规定，保障残障学生在教育环境中的平等机会；而 IEP（个别化教育计划）则依据《残障人士教育法案》，为有特殊需要的学生量身定制教育方案，具有法律效力。

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

2007–2023 年参加 AP 计算机科学考试的人数

资料来源：Code.org, CSTA, and ECEP Alliance, 2024 | 图表：2025 年人工智能指数报告

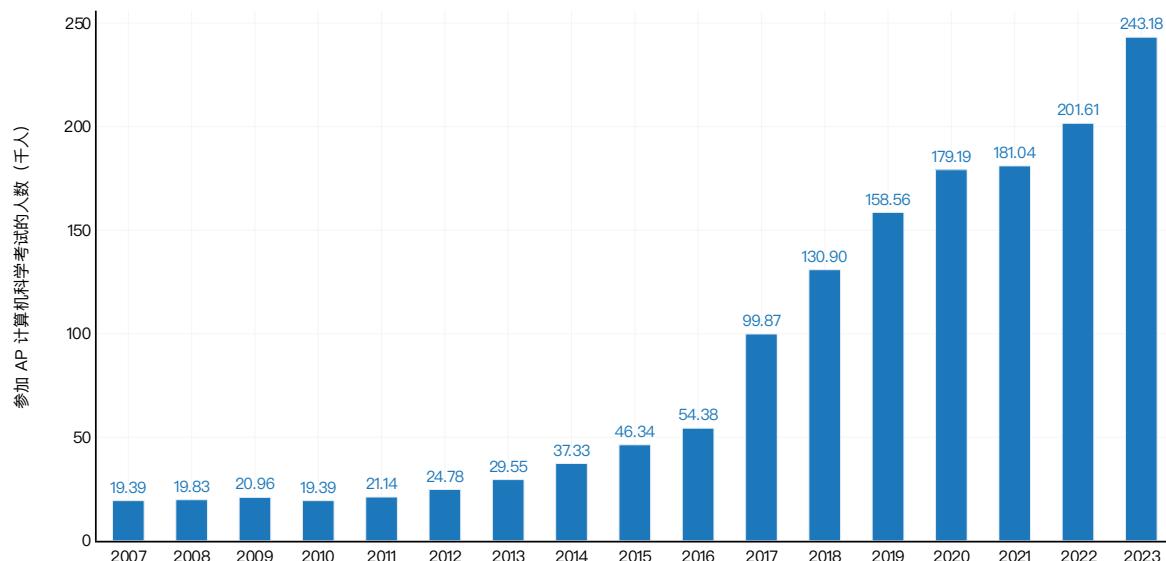


图 7.2.9

2007–2023 年按种族 / 族裔分类的 AP 计算机科学考试参加情况

资料来源：Code.org, CSTA, and ECEP Alliance, 2024 | 图表：2025 年人工智能指数报告

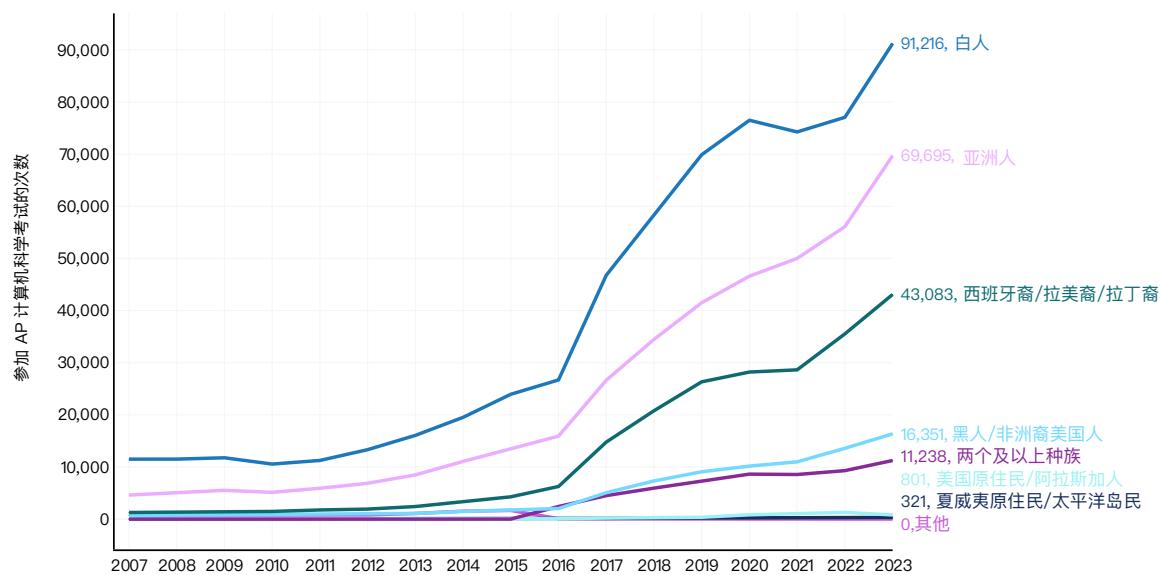


图 7.2.10

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

2007-2023 年按种族 / 族裔分类的参加 AP 计算机科学考试的学生人数（占答卷学生总人数的百分比）

资料来源：Code.org, CSTA, and ECEP Alliance, 2024 | 图表：2025 年人工智能指数报告

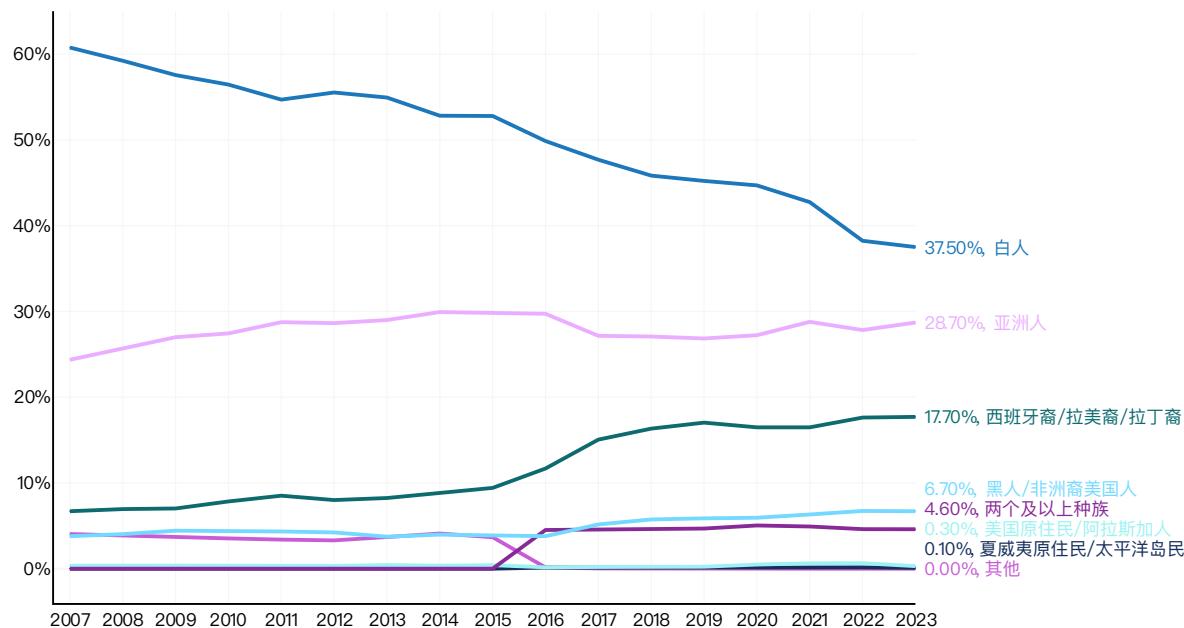


图 7.2.11

2023 年按种族 / 族裔分类的 AP 计算机科学考试参与率与全国人口统计对比情况

资料来源：Code.org, CSTA, and ECEP Alliance, 2024 | 图表：2025 年人工智能指数报告

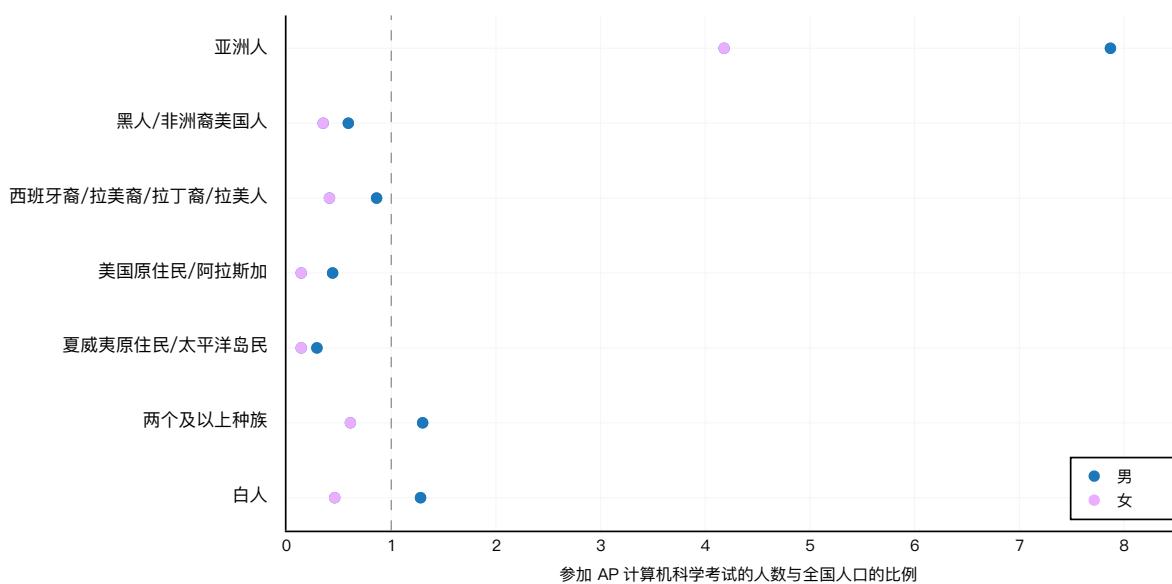


图 7.2.12

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

教育标准与政策指导

迄今为止，美国联邦层面的政策指导主要聚焦于“教育中的人工智能”（AI in education），而非“人工智能教育”（AI education）。2023 年与 2024 年，美国教育部教育技术办公室发布了一系列有关教育中人工智能应用的报告。其中一份面向教育技术开发者，另外两份则面向教育工作者、教育管理者以及政策制定者。最新的一份报告于 2024 年 10 月发布，旨在为 K-12 阶段学校提供关于人工智能安全与有效实施的政策建议。

截至 2025 年 1 月，美国已有 26 个州发布了有关“教育中人工智能应用”的指导性文件。尽管计算机科学与人工智能教育在内容上存在大量重叠，且教师在实际教学中常常交叉涵盖两者，但 K-12 阶段的计算机科学课程标准中涉及人工智能的内容极为有限。由计算机科学教师协会 (CSTA) 于 2017 年发布

的 K-12 阶段计算机科学课程标准中，仅在高中高年级设有两条标准明确要求学生具备人工智能相关知识。然而，该标准仍对人工智能教育的基础知识与技能提供了支撑，涵盖感知、数据结构与算法等主题。美国在 K-12 阶段所采纳的计算机科学课程标准，平均覆盖了 CSTA 标准中相同子概念的 97%，显示出国家层面的教学内容高度一致性。在 44 个已采纳 K-12 CS 标准的州中，有 33 个州设定了与人工智能相关的课程标准。这些标准通常内容简略，符合 CSTA 框架，且主要集中于高中阶段（图 7.2.13）。³ 其中，科罗拉多州（2024 年）、佛罗里达州（2024 年）、俄亥俄州（2022 年）与弗吉尼亚州（2024 年）最近颁布了覆盖 K-12 全阶段、内容更为详尽的人工智能课程标准；阿肯色州则为高中阶段的人工智能与机器学习课程制定了独立标准。

美国各州采用人工智能特定的 K-12 阶段阶段计算机科学标准

资料来源：CSTA and IACE, 2024 | 图表：2025 年人工智能指数报告

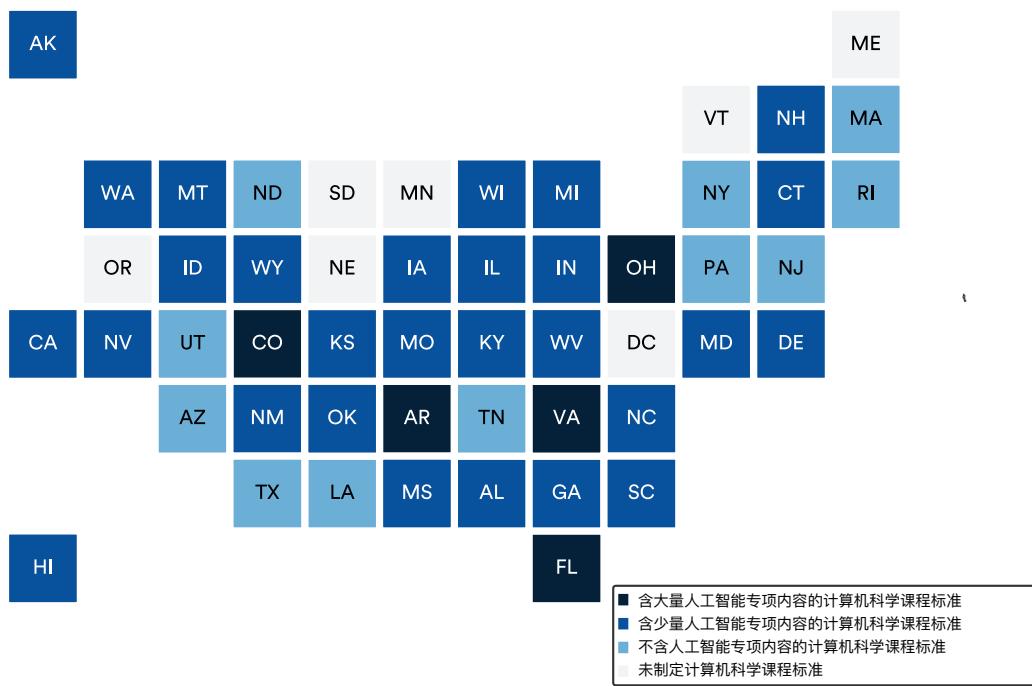


图 7.2.13

3、本项目由美国国家科学基金会（NSF）资助，资助号为 2311746。本材料中表述的任何观点、发现、结论或建议均为作者个人观点，不代表美国国家科学基金会的观点。

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

教师视角

为了探讨计算机科学教师在人工智能教育中的认知与实践，计算机科学教师现状调查收集了全国 2,901 名学前教育至 12 年级计算机科学教师的数据（其中 33% 的受访者为小学教师，36% 为初中教师，51% 为高中教师）。⁴⁵

随着人工智能教育在未来劳动力发展中愈发重要，评估现有教师群体的准备状况变得尤为关键。尽管 81% 的计算机科学教师认为人工智能应被纳入计算机科学基础教育体系中，但仅有 46% 的高中教师、44% 的初中教师与 34% 的小学教师表示自己具备传授人工智能的能力（图 7.2.14）。

当被问及实际教授的内容时，超过三分之二的初中与高中计算机科学教师表示尽管课程标准中并未明确规定，他们仍在课堂中主动涵盖 AI 内容，而小学教师中这一比例为 65%（图 7.2.15）。除此之外，更多教师表示他们涵盖了与人工智能相关的各类内容，如算法、计算系统、计算思维以及编程。

按年级分类，认为自己有能力传授人工智能的教师百分比

资料来源：计算机科学教师现状调查，2024 | 图表：2025 年人工智能指数报告

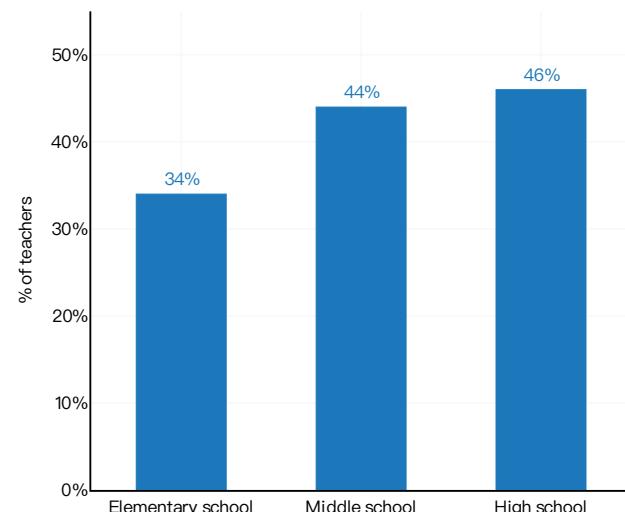


图 7.2.14

按年级划分的计算机科学课堂中教授的人工智能概念

资料来源：计算机科学教师现状调查，2024 | 图表：2025 年人工智能指数报告

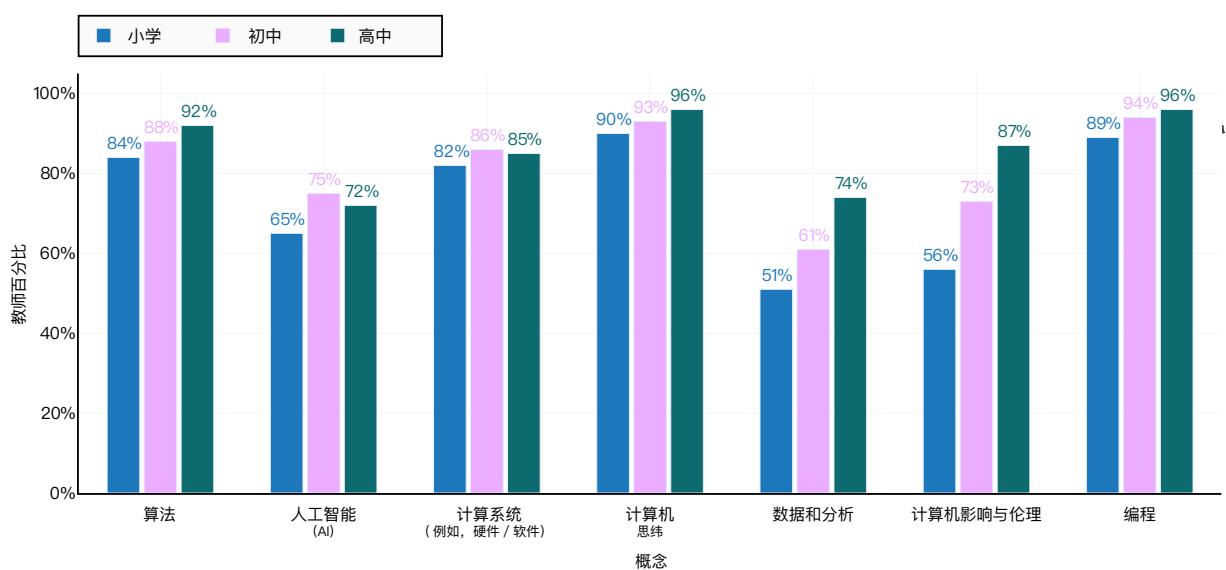


图 7.2.15

4、本项目由美国国家科学基金会（NSF）资助，资助号为 2118453。本材料中表述的任何观点、发现、结论或建议均为作者个人观点，不代表国家科学基金会观点。由于某些问题允许受访者选择多个选项，因此调查回答的总数可能不是 100%。

5、图中的百分比总和不等于 100%，因为如果受访者教授的超过一个，则可以选择多个选项。

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

在2,245名在课堂中传授人工智能内容的教师中，大多数每门课花费在该主题上的时间不足五小时。小学教师所投入时间最少，70%仅教授1至2小时（图7.2.16）。

按年级分类学生在计算机科学课堂上学习人工智能所花费的时间

资料来源：计算机科学教师现状调查，2024 | 图表：2025 年人工智能指数报告

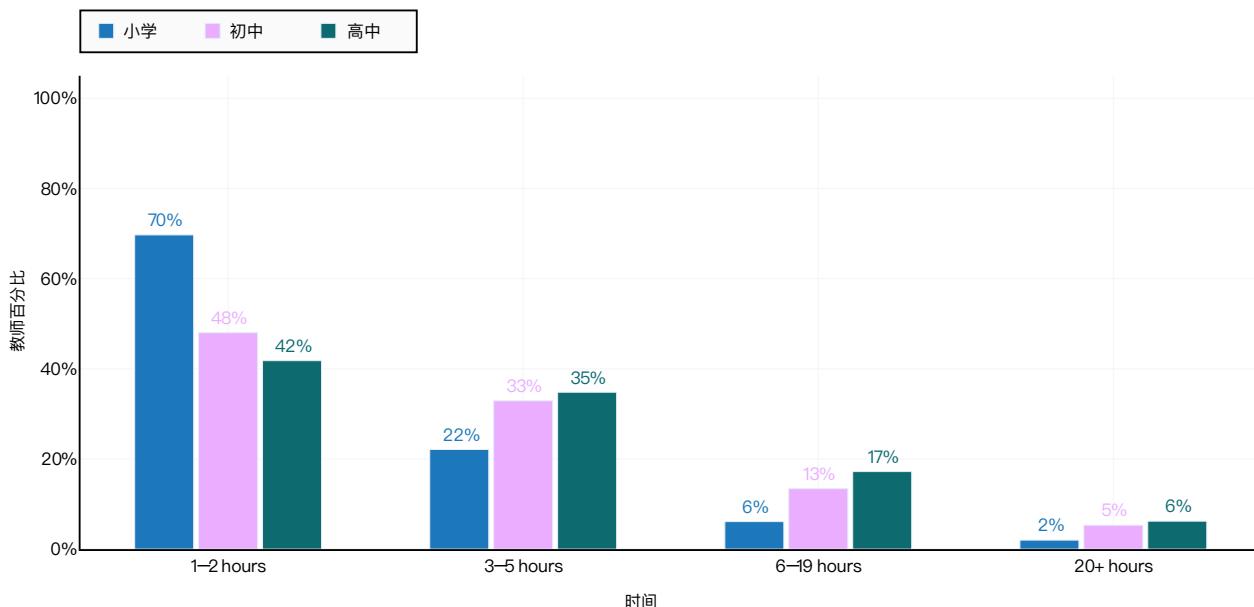


图 7.2.16

在谈及人工智能在课堂中的主要益处时，教师普遍认为其有助于提升教学效率、实现因材施教、改善学生的学业支持机制，并为学生的未来做好准备。然而，当被问及潜在风险时，教师最担忧的问题则包括人工智能滥用（尤其涉及学术诚信）、技术对学生学习与参与度的削弱、对人工智能的过度依赖、人工智能生成虚假信息与复制偏见的风险，以及学生隐私等伦理问题。

为了使学生能够负责任地使用人工智能，教师群体本身也需要接受技能提升。2024 年一项针对 364 位计算机科学教师的调查显示，88% 的受访者表示急需获得更多人工智能相关的专业发展资源。具体而言，教师认为他们亟需提升自身在人工智能领域的素养，包括了解人工智能的工作机制、使用方式与伦理影响。

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

全球现状

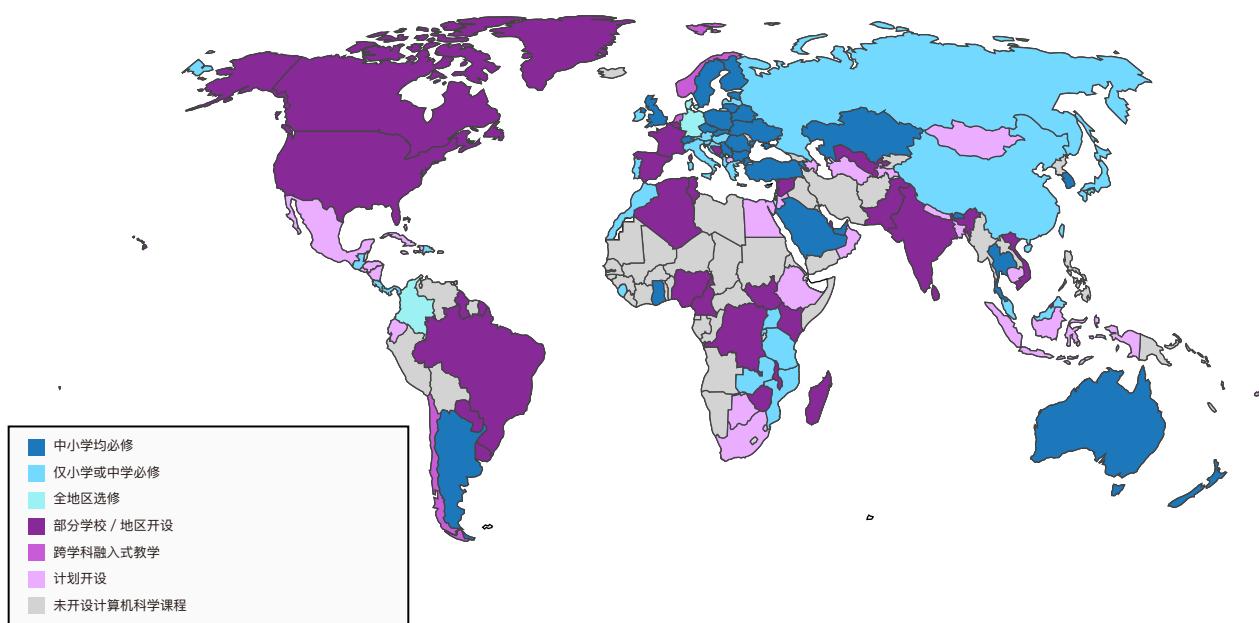
截至目前，明确在国家课程中纳入人工智能教育的国家仍属少数(如加纳、韩国与荷兰)；多数国家虽在教育战略层面强调人工智能教育的重要性，但尚未提出具体的实施方案。由于历史上人工智能教育往往被纳入计算机科学 (CS) 或信息与通信技术 (ICT) 教育体系之中，故在本节分析中，以 CS 和 / 或 ICT 教育的普及程度作为人工智能教育发展的替代指标。然而，与追踪美国计算机科学教育发展所面临的挑战类似，在解读全球教育指标时也需保持审慎态度，因为计算机科学 (CS) 与信息通信技术 (ICT) 教育常被混同为数字素养或计算机素养教育。⁶

普及情况

截至 2024 年，全球约有三分之二的国家已实施或计划实施计算机科学教育(图 7.2.17)。其中，约 30% 的国家将计算机科学教育设为小学和 / 或中学阶段的必修课程，且欧洲是实施此类政策国家数量最多的地区。过去五年内，全球各大洲在推广计算机科学教育方面均取得了不同程度的进展，非洲与拉丁美洲的增长尤为显著(图 7.2.18)。尽管如此，非洲国家的学生依然是全球范围内最难获得计算机科学教育机会的群体。造成这一现象的主要原因在于基础设施匮乏：截至 2023 年，撒哈拉以南非洲地区仅有 34% 的小学具备电力供应，这一现实不仅制约了学生的计算机基础技能培养，更进一步阻碍了计算机科学与人工智能课程的实施。

2024 年按国家分类的计算机科学教育普及情况

资料来源：Raspberry Pi 计算教育研究中心，2024| 图表：2025 年人工智能指数报告



6. 数字素养指“运用信息通信技术查找、评估、创建和传递信息的能力，需兼具认知与技术技能”；计算机素养则指“对计算机及生产力软件等程序的通用操作能力”。

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

全球范围内，由于缺乏标准化的数据收集机制，追踪人工智能教育进展面临挑战。语言障碍及各国实施情况更新滞后，进一步加大了跨国精准监测的难度。

2019 年与 2024 年按大洲分类的接受计算机科学教育的变化

资料来源：Raspberry Pi 计算教育研究中心，2024 | 图表：2025 年人工智能指数报告

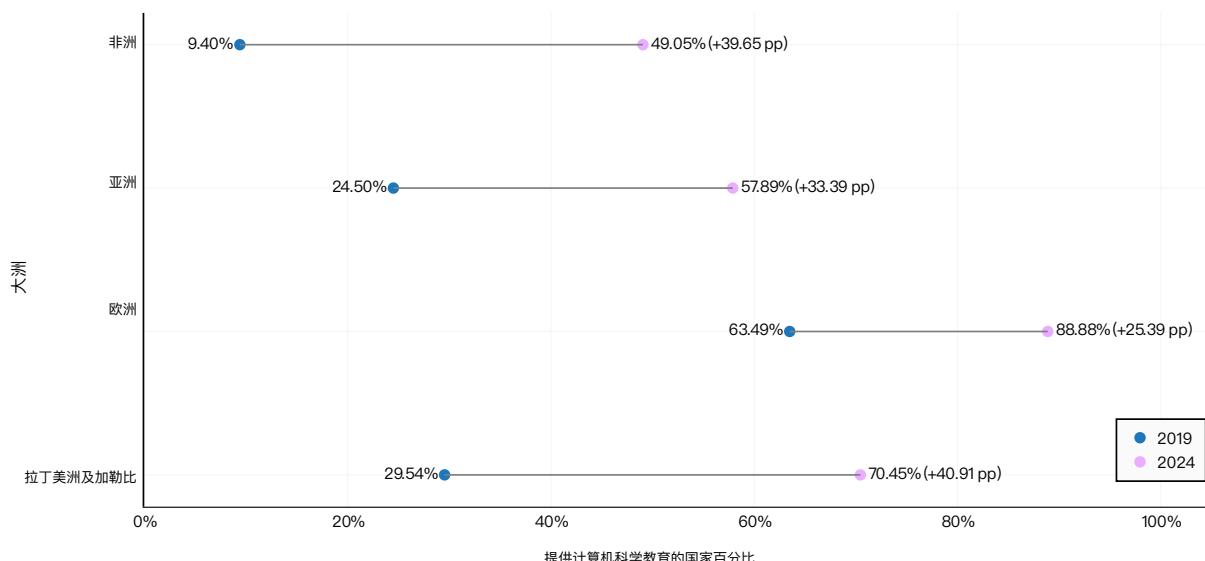


图 7.2.18

政策指导

在全球范围内，各国在发展“人工智能教育标准”方面的推进速度显著落后于“教育中人工智能应用”的政策制定。截止 2024 年 11 月，共有 10 个国家发布了人工智能教育相关的指导文件，分别为：澳大利亚、比利时、加拿大、日本、新西兰、韩国、乌克兰、英国、美国与乌拉圭。这一发展轨迹并不令人意外，各国就人工智能在教育领域的政策与指南制定已展开长达十年的探讨。早在 2015 年，联合国教科文组织的成员国便在全球

层面承诺推动科技发展，以确保“包容与公平的优质教育，并促进全民终身学习机会”（即可持续发展目标 4）。随后，教科文组织于 2019 年发布《北京共识：人工智能与教育》，旨在提出具体建议，引导各国在 2030 年前实现全民公平获取优质教育的目标（详见《教育 2030 议程》）。在这一框架中，有四项关于 K-12 阶段教育政策与执行层面的指导意见明确涉及人工智能相关内容。

第七章：教育

7.2 K-12 阶段的计算机科学与人工智能教育

与 AI4K12 倡议相似，国际组织也正积极构建人工智能教育课程框架，以供各国采纳与本地化应用。AI4K12 提出的“五大核心理念”(Five Big Ideas in AI) 成为当前 K-12 阶段人工智能教育的重要内容组织框架(图 7.2.19)。2023 年，联合国教科文组织发布了面向学生与教师的人工智能能力框架。学生框架涵盖四大核心能力：以人为本的价值观、人工智能伦理、人工智能技术与应用，以及人工智能系统设计。在每一项能力中，学生需经历“理解、应用、创造”的认知递进过程。在欧盟层面，多数成员国采用 DigComp 2.2 框架，作为提升公民数字能力的重要指南，并结合计算机科学学习目标推动学生数字素养的全面发展。DigComp 2.2 的最新版内容已包括关于与人工智能互动所需知识、技能与态度的建议，尽管其并未明确纳入教授公民构建人工智能系统的相关指导。

AI4K12 指导方针围绕人工智能领域的五大理念展开

资料来源：AI4K12, 2024

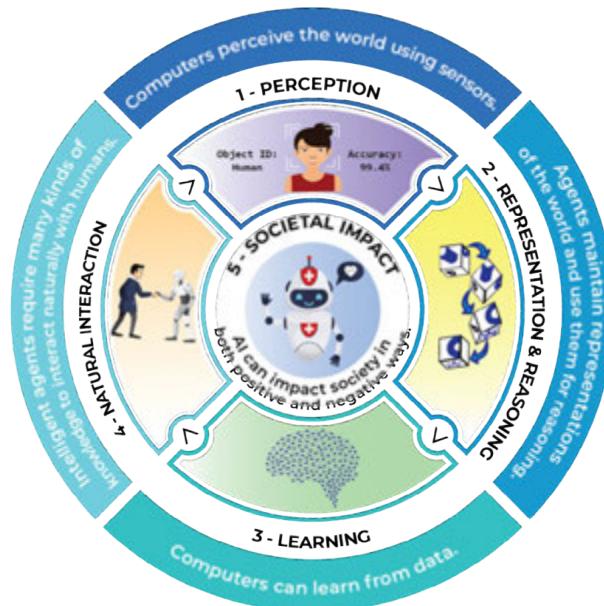


图 7.2.19⁷

7、注：图 7.2.19 介绍了“五大核心”，这五个核心理念分别如下：1. 感知 (Perception)，计算机通过传感器感知世界。2. 表征与推理 (Representation and Reasoning)，智能体维护世界的表征，并利用这些表征进行推理。3. 学习 (Learning)，计算机可以从数据中学习。4. 自然交互 (Natural Interaction)，智能体需要多种知识以自然地与人类互动。5. 社会影响 (Societal Impact)，人工智能可能以积极或消极的方式影响社会。

第七章：教育

7.3 高等教育阶段的计算机科学与人工智能教育

人工智能将在美国劳动力结构和经济未来中扮演的角色尚未完全明晰，但其影响预计极为深远。当前，技术类从业者已在美国经济中占据重要地位，全国已有 960 万人在各行业担任技术岗位。尽管自动化所引发的就业替代问题令人担忧，但对人工智能相关岗位的需求——如数据库管理与数据基础设施解决方案——预计将持续增长。因此，全球必须共同致力于确保高等教育机构具备培训未来劳动力的能力，并进一步扩展计算技术人才培养体系。

7.3 高等教育阶段的计算机科学与人工智能教育

学位授予情况

美国

本节中关于美国高等教育阶段计算机科学与人工智能教育趋势的数据，主要来源于国家教育统计中心(NCES)。值得注意的是，学科分类标准(CIP)由该中心隶属于美国教育部制定，用于统一划分各类学术项目。自 2016 年起，人工智能相关课程被纳入 CIP 编码 11.0102，涵盖“以符号推理、知识表示和仿真为核心的课程，重点在于通过计算机和软件模拟人类学习与推理的过程与能力，以及人类运动控制与动作的计算建模。相关课程内容包括计算理论、控制论、人因工程、自然语言处理，以及工程技术及具体应用领域中的相关知识”。

尽管过去十年来计算机科学副学士学位获得者数量大致保持稳定，部分社区学院已率先探索人工智能教育，开设了相关证书课程，以及人工智能与相关领域的副学士与学士学位项目(图

7.3.2)。代表性机构包括：马里科帕社区学院系统(Maricopa)、休斯敦社区学院(Houston Community College)、迈阿密 - 戴德学院(Miami Dade College)以及湾区社区学院联盟中的若干成员。

过去十年间，计算领域学士学位的毕业人数增长了 22% (图 7.3.1)。2023 年，计算机科学学士毕业生人数最多的五所高校依次为：Western Governors University、加州大学伯克利分校、南新罕布什尔大学、德州大学达拉斯分校以及密歇根大学。⁸ 尽管人工智能在学士层面的增长尚需时间体现（因其学制通常为四年），人工智能领域的快速扩张已在硕士阶段显现出来：2022 至 2023 年间计算机科学硕士毕业生数量增长了 26%，过去十年累计增长达 83%。

8、Western Governors University 和南新罕布什尔大学主要是在线教育机构。

第七章：教育

7.3 高等教育阶段的计算机科学与人工智能教育

2013–2023 年美国高等教育阶段计算机科学毕业生人数

资料来源：国家教育统计中心高等教育综合数据系统，2013–2023 年 | 图表：2025 年人工智能指数报告

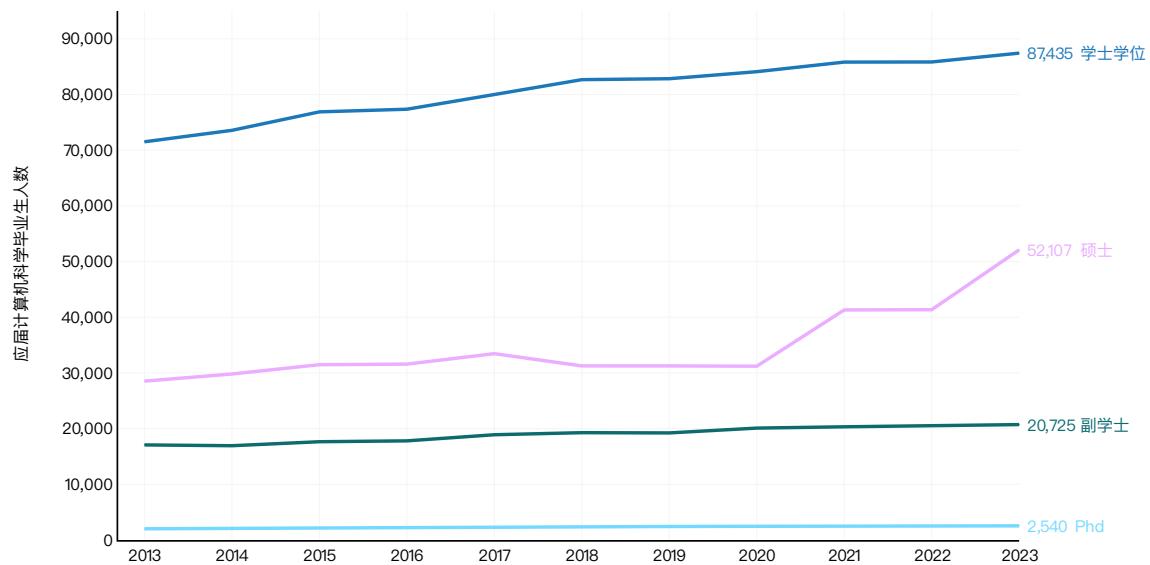


图 7.3.1

尽管女性整体高等教育毕业率高于男性，但学位完成率数据显示女性在计算机科学领域中仍处于代表性不足的状态（图 7.3.2）。

2023 年按性别分类的美国高等教育阶段计算机科学毕业生

资料来源：国家教育统计中心高等教育综合数据系统，2013–2023 年 | 图表：2025 年人工智能指数报告

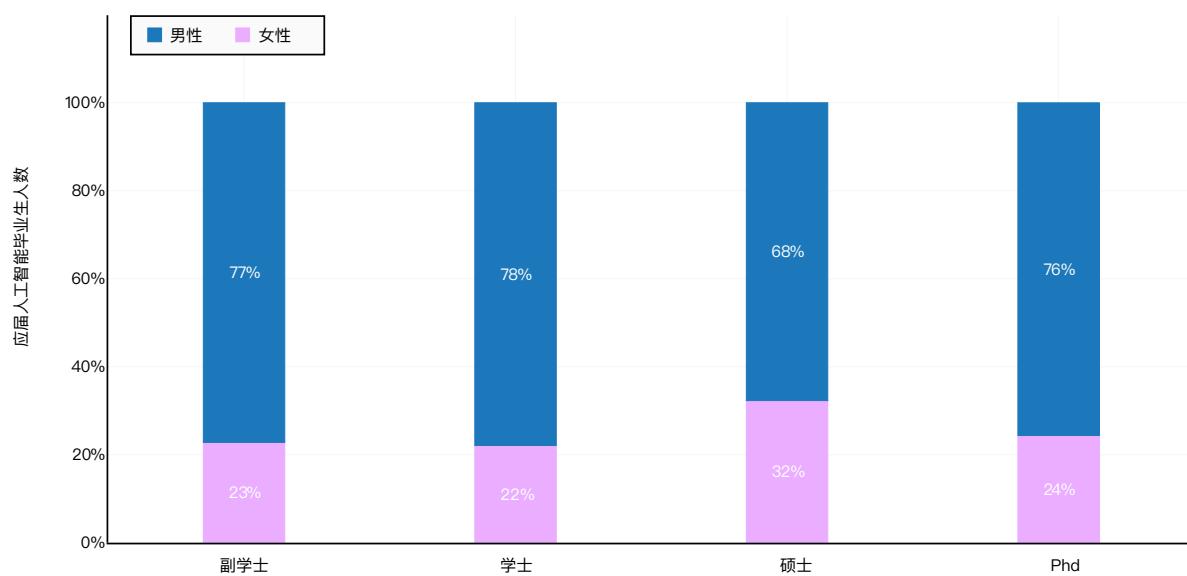


图 7.3.2

第七章：教育

7.3 高等教育阶段的计算机科学与人工智能教育

非洲裔学生在计算机相关学科中所占比例为：本科阶段8%，硕士阶段8%，博士阶段7%（图7.3.3）。拉丁裔学生在本科阶段占比为13%，在硕士阶段为8%，在博士阶段则下降至4%。相较之下，白人学生在本科阶段的占比为46%，在博

士阶段超过半数（52%）。亚裔学生在高等教育阶段的计算机专业中则表现出显著的过度代表趋势，其在本科、硕士与博士阶段的占比分别为23%、28%与17%。

2023年按种族 / 族裔分类的美国计算机科学与所有高等教育毕业生的对比（仅美国居民）

资料来源：国家教育统计中心中学后教育综合数据系统，2013–2023年 | 图表：2025年人工智能指数报告

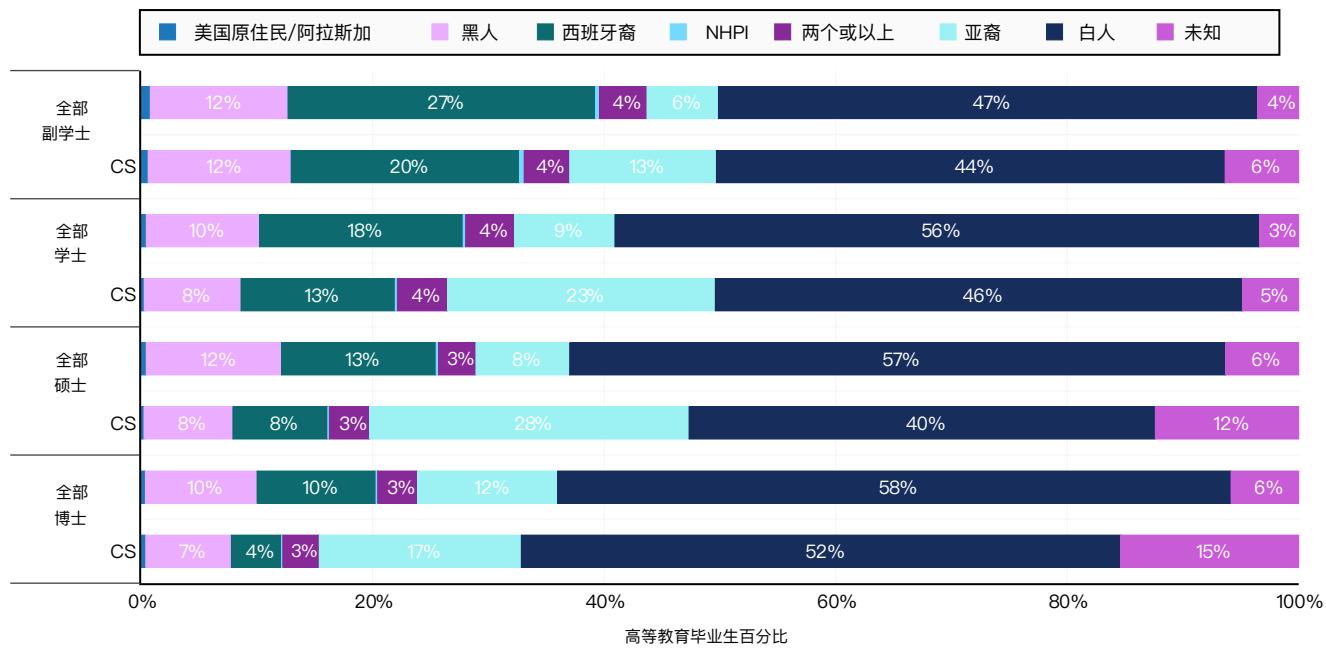


图 7.3.3

计算机相关研究生项目中的学生大多数来自美国以外国家，这一比例在过去几年中持续上升。2023年，非本国居民占计算机硕士学位毕业生的67%，博士学位毕业生的60%。在2022年至2023年间，国际计算机科学硕士生人数增长超过两倍，从15,811人增至34,850人（数据来源：IPEDS）。其中，来自印度与中国的学生成为了该研究生群体的绝大多数，

分别占国际计算机科学硕士生总数的93%（共95,130人）和国际CS博士生总数的60%（共13,070人）（图7.3.4与图7.3.5）。与此同时，开设人工智能专属学士学位课程的美国高校数量在2022年至2023年间几乎翻倍；提供人工智能专属硕士项目的院校数量亦呈现出显著增长（图7.3.6）。

第七章：教育

7.3 高等教育阶段的计算机科学与人工智能教育

2022 年美国高校国际计算机科学硕士在读人数

资料来源：美国国家科学委员会；美国国家科学基金会，2023| 图表：2025 年人工智能指数报告

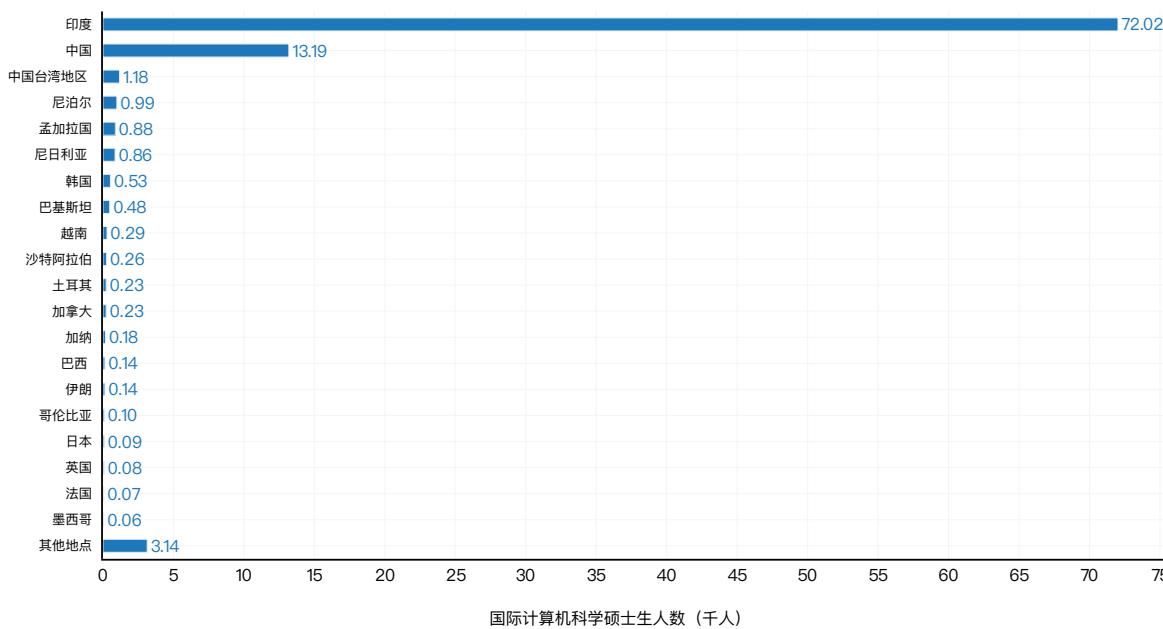


图 7.3.4

2022 年美国高校国际计算机科学博士在读人数

资料来源：美国国家科学委员会；美国国家科学基金会，2023| 图表：2025 年人工智能指数报告

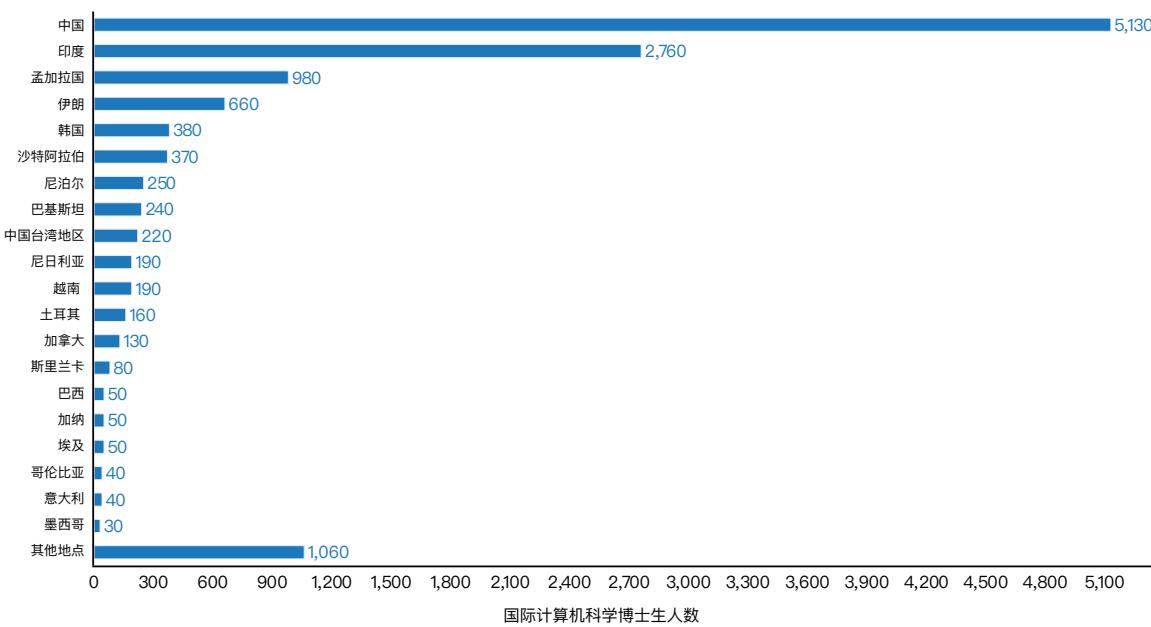


图 7.3.5

第七章：教育

7.3 高等教育阶段的计算机科学与人工智能教育

2013–2023 年美国提供人工智能学士和硕士学位院校数量

资料来源：国家教育统计中心中学后教育综合数据系统，2013–2023 年 | 图表：2025 年人工智能指数报告

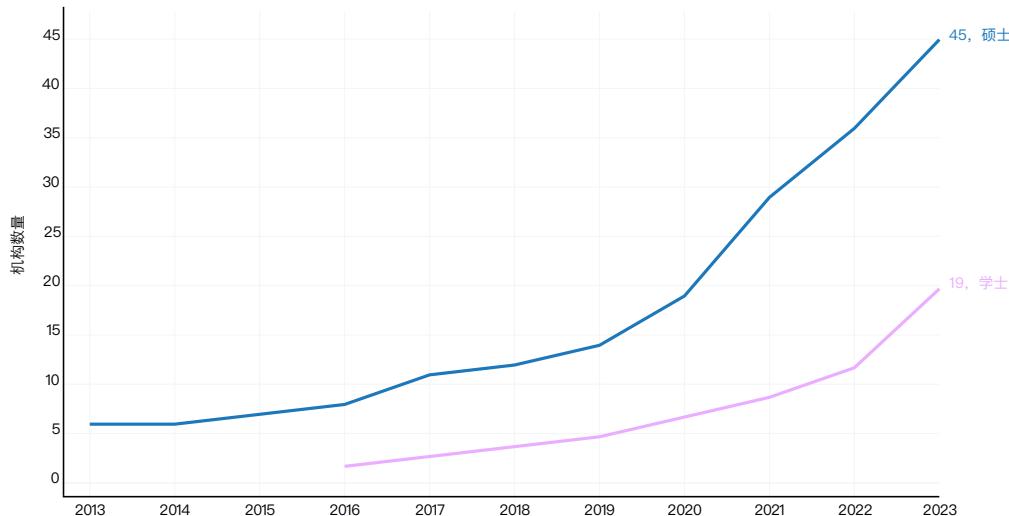


图 7.3.6

2022年至2023年间，获得人工智能硕士学位的学生人数出现显著增长（图7.3.7）。卡耐基梅隆大学作为当年人工智能专业毕业生人数最多的高校，其毕业人数在一年内实现翻番；

与此同时，宾夕法尼亚州立大学则于2022年迎来了首届人工智能专业的毕业生（图7.3.8）。在此之前，卡耐基梅隆大学曾长期是为数不多开设人工智能专属学位项目的高校之一。

2013–2023 年美国人工智能学士和硕士应届毕业生

资料来源：国家教育统计中心中学后教育综合数据系统，2013–2023 年 | 图表：2025 年人工智能指数报告

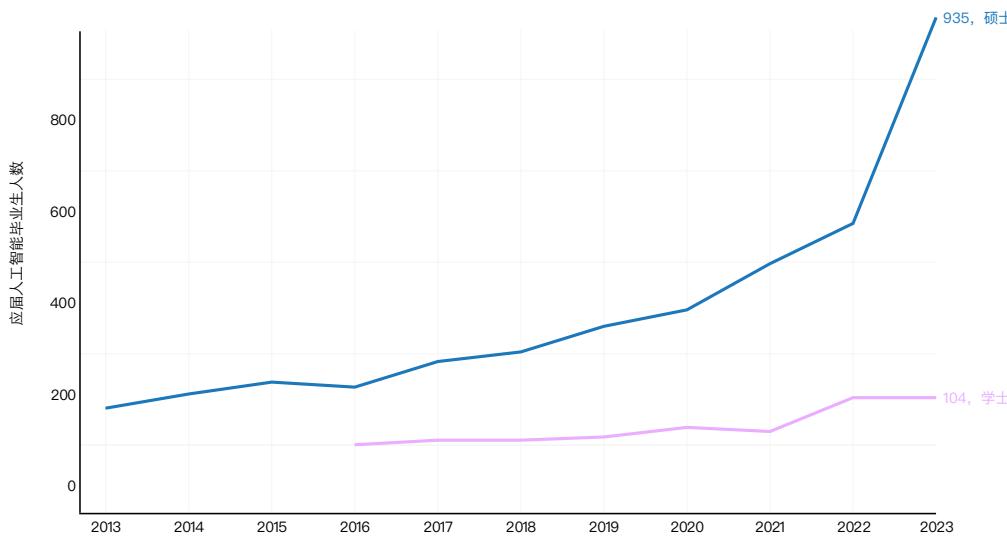


图 7.3.7

第七章：教育

7.3 高等教育阶段的计算机科学与人工智能教育

按学位类型分类的 2023 年人工智能专业毕业生最多的高等教育机构⁹

资料来源：国家教育统计中心的《中学后教育综合数据系统》，2023 年。

人工智能本科毕业生	
卡内基梅隆大学	32
富勒塞尔大学	19
威斯康星康考迪亚大学	16
高等技术大学	10
宾夕法尼亚州立大学主校区	7
人工智能硕士课程毕业生	
卡内基梅隆大学	178
宾夕法尼亚大学	98
北德克萨斯大学	76
东北大学	55
圣何塞州立大学	52
人工智能博士课程毕业生	
卡内基梅隆大学	28
国会技术大学	4
匹兹堡大学匹兹堡校区	1

图 7.3.8

9、本列表仅包括使用人工智能专业 CIP 代码大学，而不是使用一般 CS 代码的大。然而，全球许多学习人工智能的学生很可能就读于更广泛的 CS。

第七章：教育

7.3 高等教育阶段的计算机科学与人工智能教育

全球

在全球范围内，尚无任何单一数据集能够对所有国家的人工智能或计算机科学高等教育状况进行统一且标准化的统计。然而，经济合作与发展组织（OECD）已整合了其成员国及若干非成员国的相关数据，用以提供一定程度上的比较基础。¹⁰ 其中，教育统计的国际标准分类（ISCED）被用于进行跨国对比，该分类体系亦为OECD评估全球教育进展的基准。信息与通信技术（ICT）涵盖“信息学、信息与通信技术及计算机科学”等研究领域，其核心内容涵盖一系列新兴技术，涉

及数字信息的处理与传输，包括计算机、计算机网络（如互联网）、微电子学、多媒体、软件及程序设计等方面。

在该数据集中，美国依然是信息与通信技术相关领域的全球领导者，其在副学士、学士、硕士及博士层级的毕业生数量均高于其他任何国家（见图7.3.9至图7.3.12）。尤其值得注意的是，在副学士、硕士与博士层级，美国的毕业生人数超过第二位国家的两倍；在学士层级，美国的毕业生人数亦接近第二位国家的两倍。

2022年按国家分类的信息和通信技术短周期高等教育应届毕业生人数

资料来源：经合组织，2022 | 图表：2025年人工智能指数报告

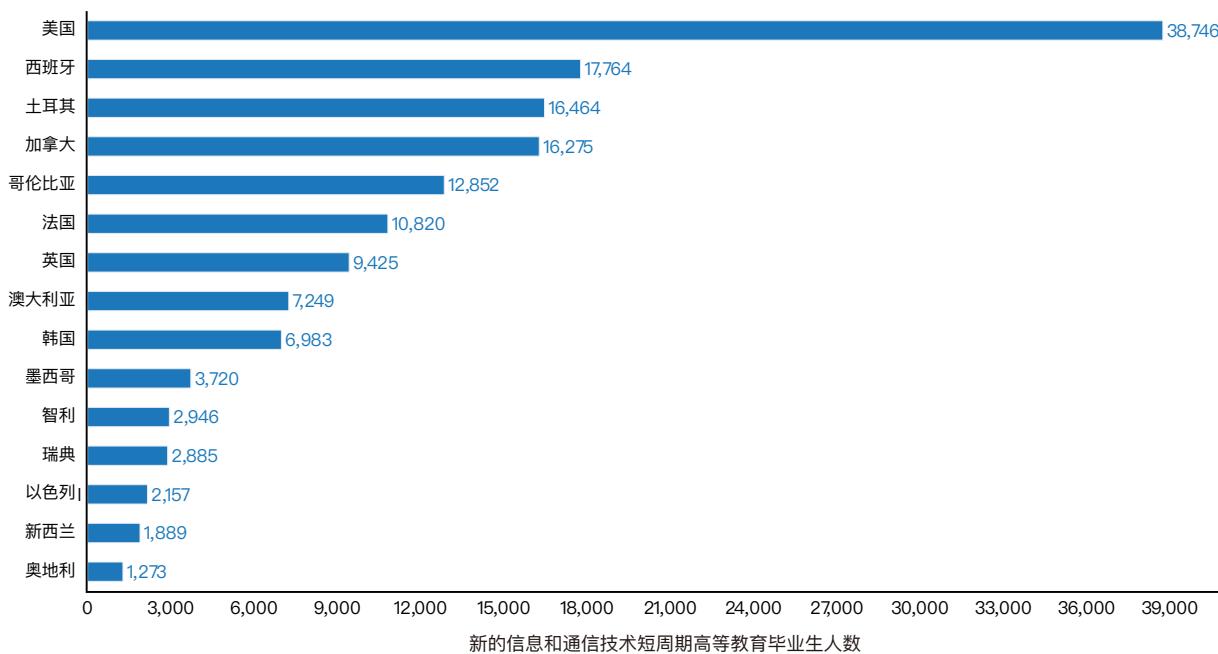


图 7.3.9

10、虽然该数据集提供了对一些国家的深入了解，但它忽略了一些拥有大量 ICT 毕业生的国家。印度、中国和非洲国家被排除在外，这凸显了全球标准化数据收集的必要性，以确保将那些在计算机教育方面进行了大量投资并在全球大多数国家占很大比例的国家包括在内。在收集和报告全球教育数据方面也存在很大的滞后性；因此，可获得数据的最近年份是 2022 年。

第七章：教育

7.3 高等教育阶段的计算机科学与人工智能教育

2022 年按国家分类的新增信息与通信技术本科毕业生

资料来源：经合组织，2022 | 图表：2025 年人工智能指数报告

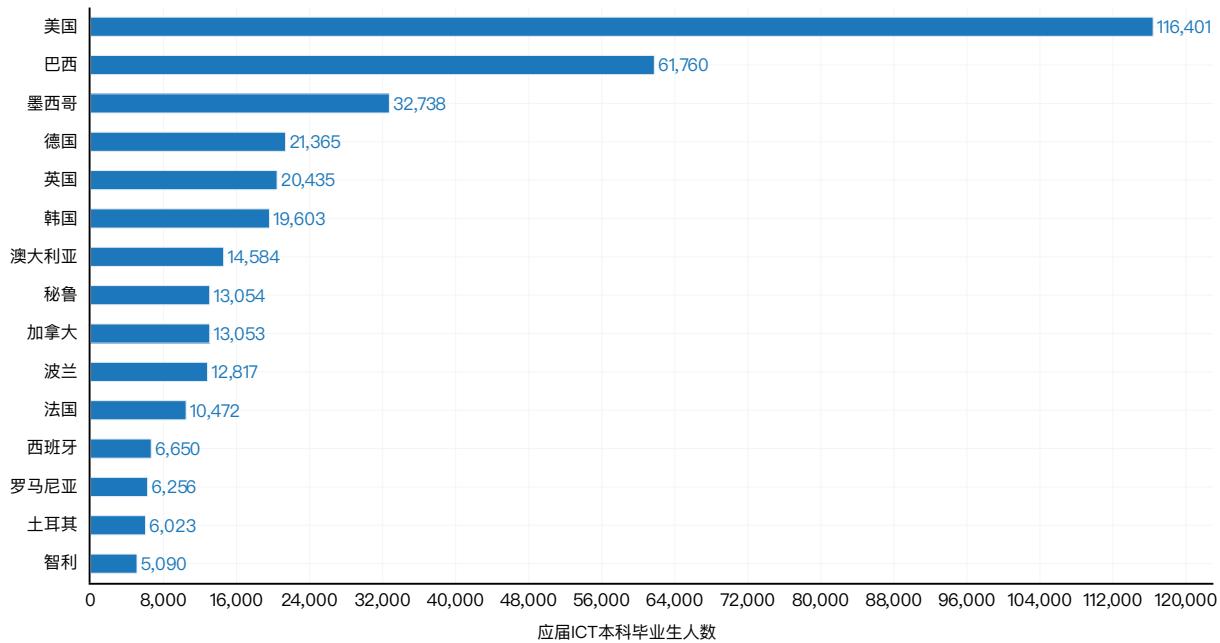


图 7.3.10

2022 年按国家分类的新增信息与通信技术硕士毕业生人数

资料来源：经合组织，2022 | 图表：2025 年人工智能指数报告

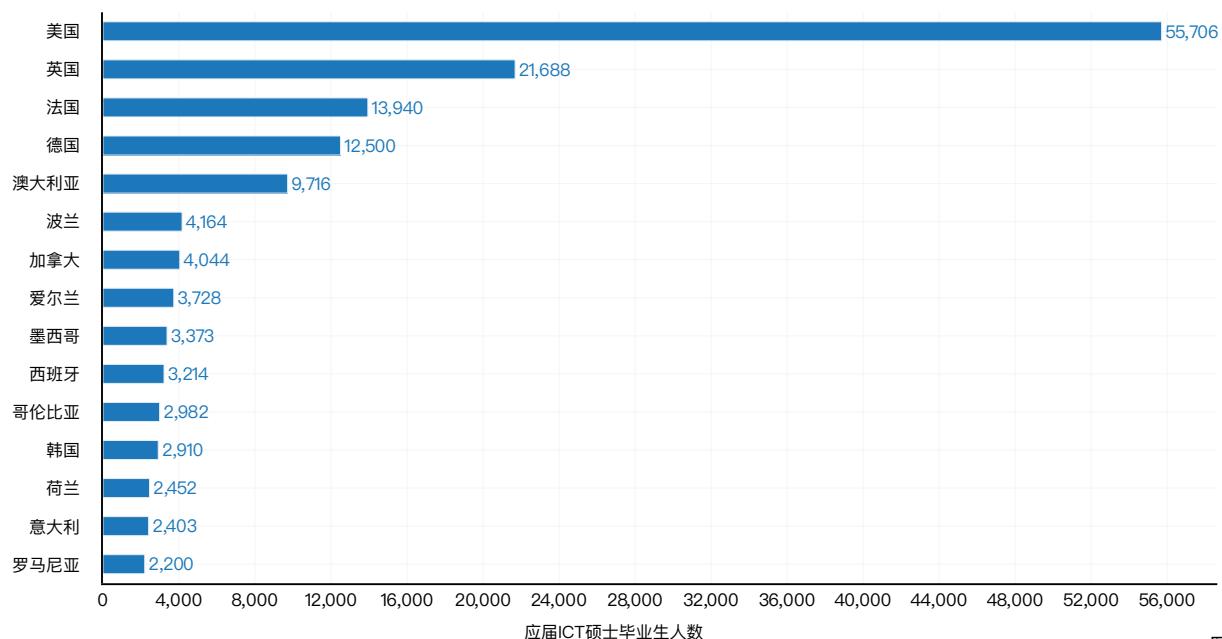


图 7.3.11

第七章：教育

7.3 高等教育阶段的计算机科学与人工智能教育

2022 年按国家分类的新晋的信息与通信技术博士毕业生

资料来源：经合组织，2022 | 图表：2025 年人工智能指数据报告

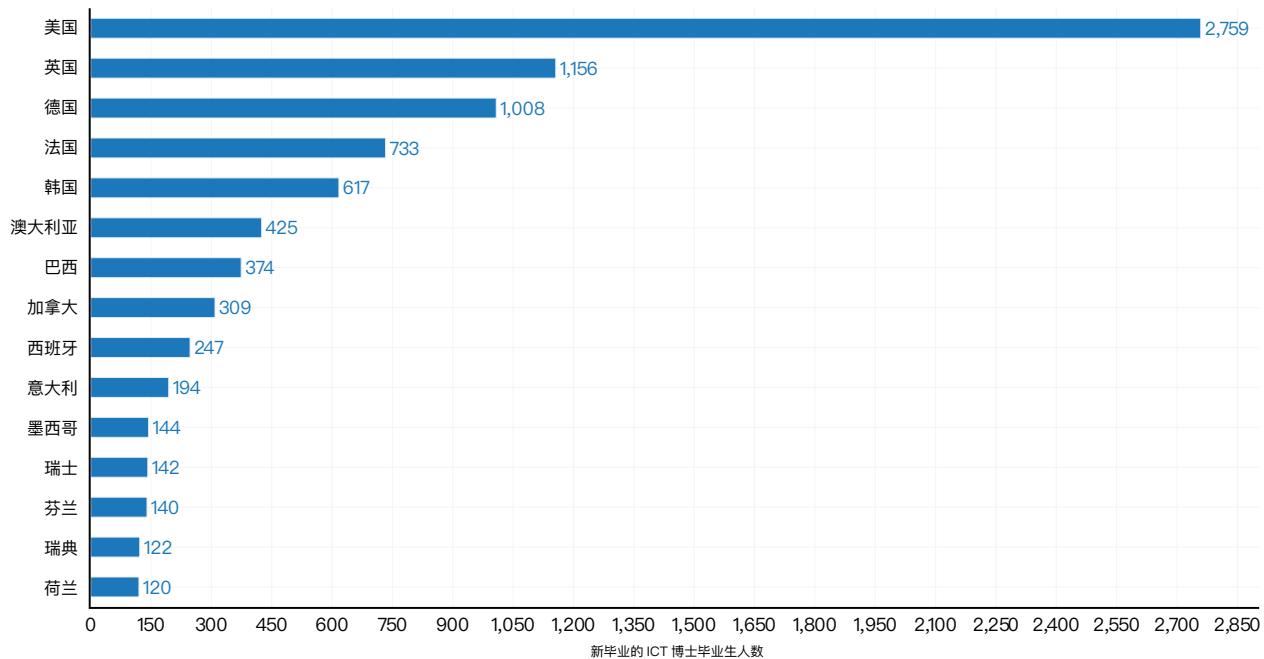


图 7.3.12

性别平衡仍是全球人工智能相关学科中的一项重大挑战（图7.3.13）。在全球范围内，女性在信息与通信技术相关高等教育毕业生中所占比例平均约为四分之一，在副学士、学士及博士层级皆为如此。相较之下，硕士阶段女性占比略高，约

接近三分之一。在各国之中，土耳其在性别平衡方面表现尤为突出，该国女性在副学士、学士、硕士及博士层级的毕业生中均至少占据一半以上的比例。

第七章：教育

7.3 高等教育阶段的计算机科学与人工智能教育

2022 年按国家分类的 ICT 高等教育应届毕业生中女性所占百分比

资料来源：经合组织，2022 | 图表：2025 年人工智能指数报告

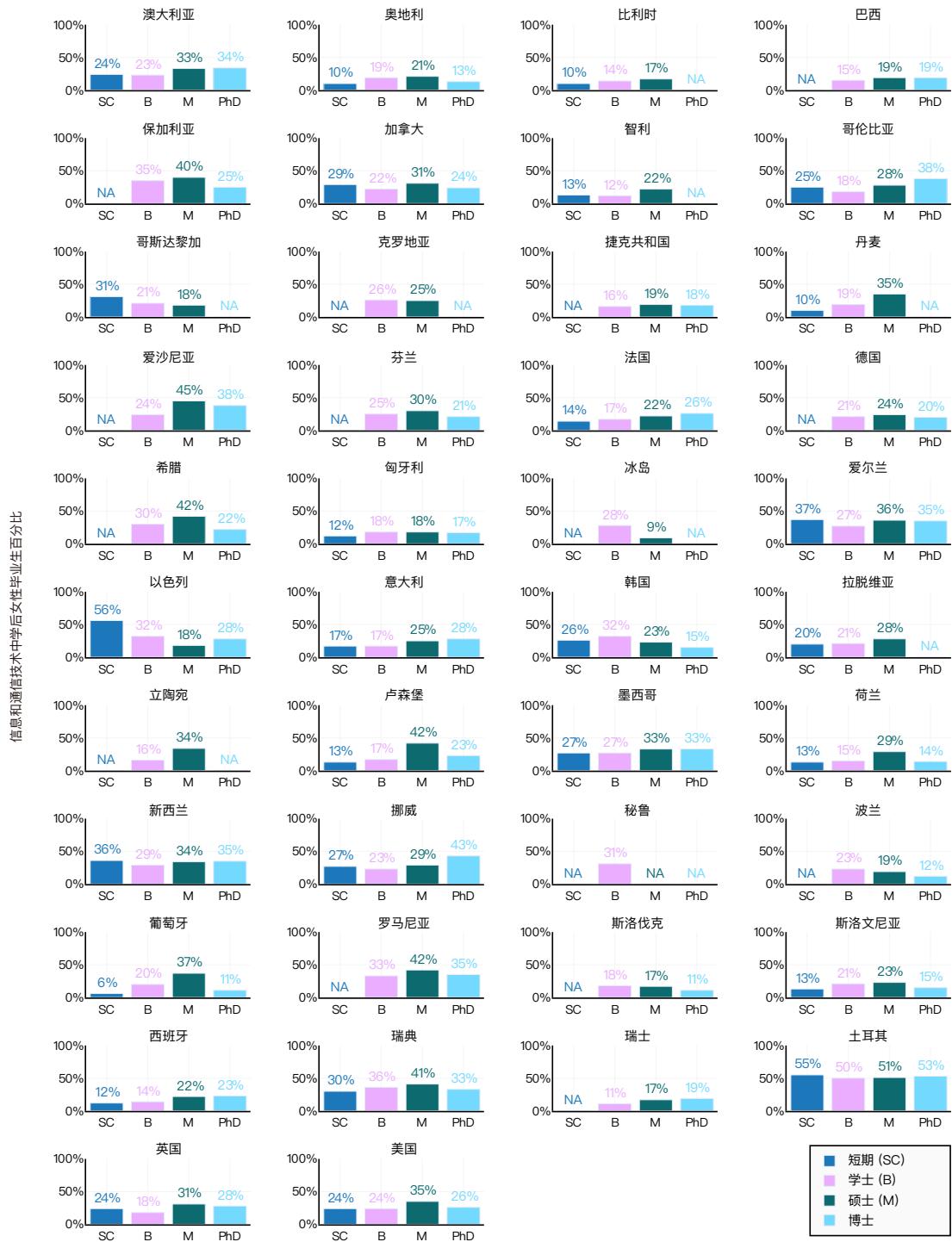


图 7.3.13

指导政策

目前，大学层面的人工智能政策和指导主要集中于学生在作业中使用人工智能的行为规范，而针对人工智能教育本身的指导通常由各个院系（主要为计算类院系）在内部制定。

人工智能在高校校园中的使用已极为普遍，无论是学生还是教师群体都高度依赖：有 86% 的学生在学习中使用人工智能，61% 的教师在教学中使用人工智能。然而，围绕人工智能使用的相关指导方针在高校内部仍缺乏清晰性与标准化。至 2025 年初，仅有 39% 的高等教育机构制定了与人工智能相关的“可接受使用政策”，尽管这一比例较 2024 年增长了 16 个百分点。在学生规模超过一万人的大型高校中，制定相关政策的比例显著高于学生人数不足五千人的小型高校。尽管人工智

能对教学与学习政策的影响最为显著，但实际上，几乎所有类型的高校政策都受到人工智能技术的波及，包括技术采购流程（如是否可使用校方资源采购人工智能工具）、知识产权与版权法律的遵守、是否允许使用人工智能创建恶意软件或病毒等。在网络安全、数据隐私、在线教学以及数据分析等诸多政策领域，人工智能的应用与管控均引发了系统性的连锁反应。

除了 2019 年《北京共识：人工智能与教育》为 K-12 阶段提供的指导之外，联合国教科文组织还发布了适用于 K-12 及高等教育的综合性指导政策，意在借助人工智能技术，推动实现“教育 2030 议程”中所设定的全球目标。该报告中提出了五项专门针对高等教育阶段人工智能教育的执行与政策建议。

7.4 展望未来

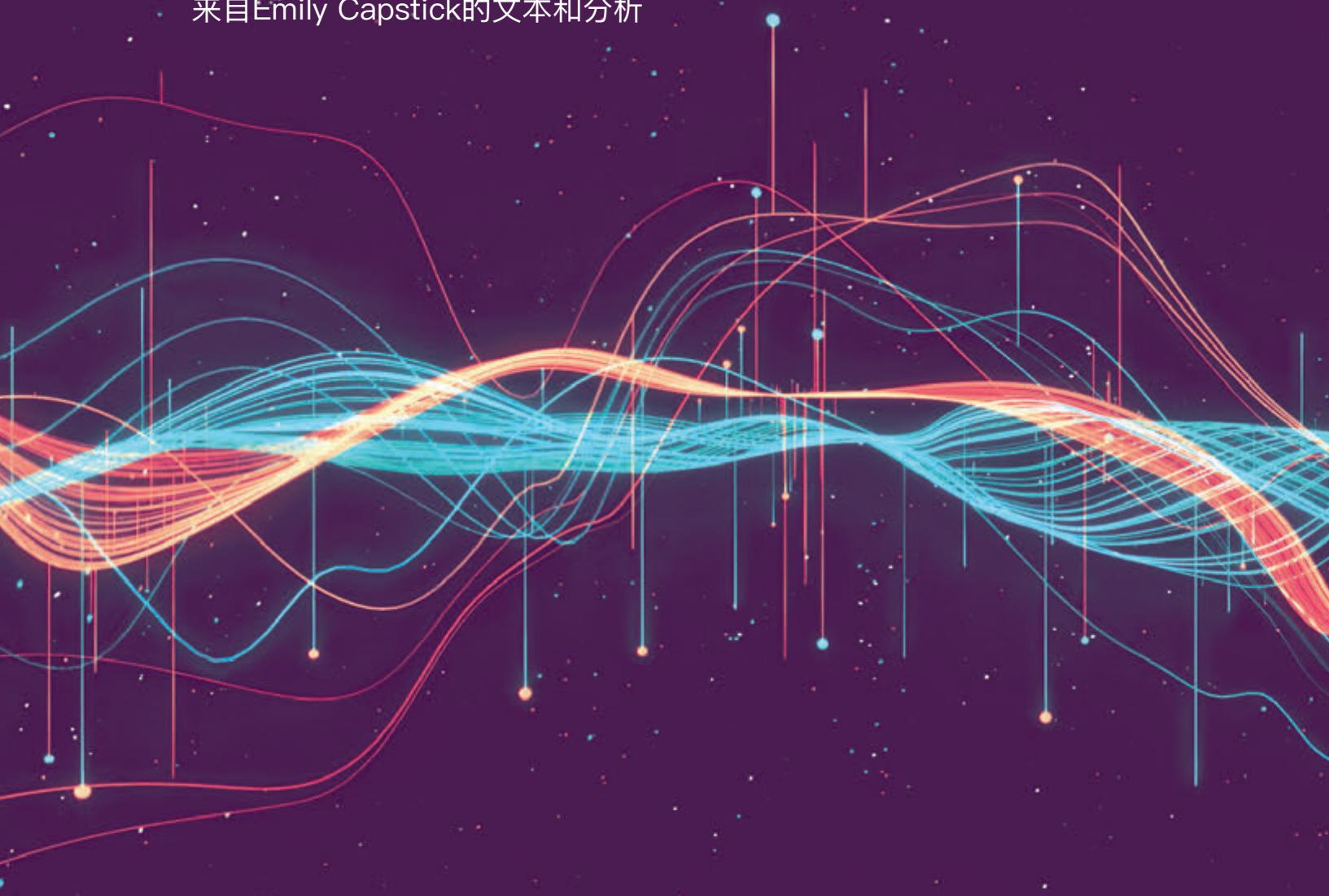
公平的人工智能教育生态系统的有意识设计，将对未来技术创新的负责任发展与部署发挥至关重要的作用。当前人工智能得以迅速扩展的制度环境，已导致一系列不良后果，包括操纵国家政治进程的虚假信息与误导性信息活动、人工智能赋能武器的发展，以及对受版权保护的知识产权的侵犯。这些现象凸显出：在人工智能构建路径上亟需优先确立一种更加稳健与负责任的方式。为实现这一目标，必须重新设想人工智能教育的整体方案，将人工智能能力视作学生通往技术驱动未来所不可或缺的核心素养。其中，这些能力不仅应涵盖技术开发本身，也应包括一种伦理视角，即能够主动识别、分析并质询人工

智能所带来的社会影响。当前已有以计算机科学为基础的基础设施、政策框架与实施机制，为人工智能教育的系统性整合提供了现实路径。然而，随着人工智能技术的持续演进，教育体系的转型已刻不容缓。唯有如此，才能确保未来技术的构建者充分意识到人工智能可能带来的风险，并具备有效减缓其负面影响的能力。全球范围内的高等教育机构应持续推进人工智能教育通道的构建，并在此过程中进行阶段性成效监测。同时，有必要制定相应政策以扩大课程获取的公平性，并实施切实可行的策略，提升教师专业能力，促进学生的广泛参与，从而实现人工智能核心素养在教育系统中的公平普及与能力构建。

第八章:

公众观点

来自Emily Capstick的文本和分析



第八章：公众观点

概述	396
章节要点	397
8.1 公众观点	399
全球公众观点	399
人工智能产品和服务	399
人工智能与就业	405
人工智能与民生	407
重点：自动驾驶汽车	409
8.2 美国政策制定者的观点	410

获取公共数据

第八章： 公共观点

概述

随着人工智能日益深入社会各个层面，理解公众对于这一技术的态度正变得愈发重要。有关人们如何看待人工智能的洞察，不仅有助于预测其可能引发的社会影响，也能揭示不同国家与人口群体在采纳与接受程度上的差异。初步数据显示，公众对于人工智能的焦虑情绪正在上升，且某些地区的悲观情绪远高于其他地区。随着技术的持续演进，这一趋势是否将延续，仍有待进一步观察。

本章节将从全球、国家、人口统计和族群等多个维度探讨公众对人工智能的看法。所用数据来自多项研究来源，包括益普索公司持续追踪全球人工智能态度的长期调查、美国汽车协会关于自动驾驶车辆的民意调查，以及针对美国地方政策制定者对人工智能观点的最新研究。

第八章： 公共观点

章节要点

1. 全球对人工智能产品和服务的态度持谨慎乐观的态度。在益普索(益普索)2022–2024年持续追踪的26个国家中，18国的受访者对人工智能产品与服务“利大于弊”的认同比例呈现上升趋势。在全球范围内，认为人工智能产品和服务利大于弊的个人比例从2022年的52%上升到2024年的55%。

2. 人工智能对日常生活影响的预期认知度持续攀升。在世界各地，三分之二的人现在认为，人工智能驱动的产品与服务将在未来3至5年内显著改变日常生活——这一比例较2022年上升6个百分点。除马来西亚、波兰和印度外，其余国家自2022年以来该认知度均有所提升，其中加拿大(增长17%)与德国(增长15%)涨幅最为显著。

3. 对人工智能公司伦理行为的怀疑正在增加，同时对人工智能公平性的信任正在下降。在全球范围内，人们对人工智能公司保护个人数据的信心从2023年的50%下降到2024年的47%。同样，与相比，如今相信人工智能系统不偏不倚、不受歧视的人也越来越少。

4. 人工智能乐观程度的地区差异依然存在。2023年人工智能指数首次指出，人工智能乐观程度的地区差异依然存在。在中国(83%)、印度尼西亚(80%)和泰国(77%)等国家，绝大多数人认为人工智能驱动的产品和服务利大于弊，而在加拿大(40%)、美国(39%)和荷兰(36%)，只有少数人持这种观点。

5. 美国人自动驾驶汽车仍持不信任态度。根据美国汽车协会(American Automobile Association, AAA)最新调查数据显示，61%的美国民众对自动驾驶汽车存在恐惧心理，仅有13%的受访者表示信任该技术。尽管这一恐惧比例较2023年68%的峰值有所下降，但仍高于2021年54%的水平。

6. 美国地方政策制定者普遍支持对人工智能实施监管。2023年，美国73.7%的地方政策制定者(涵盖镇、市、县三级政府)支持对人工智能实施监管，较2022年的55.7%显著提升。民主党人士的支持率(79.2%)明显高于共和党人士(55.5%)，但两党支持率相较2022年均呈现显著增长。

第八章： 公共观点

章节要点（续）

7. **此前对人工智能持最强烈怀疑态度的国家中，对人工智能的乐观态度呈现急剧上升趋势。**在全球范围内，人们对人工智能产品和服务的乐观态度有所提高，其中以前最持怀疑态度的国家乐观态度增幅最大。2022年，英国(38%)、德国(37%)、美国(35%)、加拿大(32%)和法国(31%)是最不倾向于认为人工智能利大于弊的国家。此后，这些国家对人工智能的乐观态度分别提升了8%、10%、4%、8%和10%。

8. **劳动者预期人工智能将重塑就业结构，但其对岗位被替代的担忧程度相对较低。**在全球范围内，60%的受访者认为人工智能将在未来五年内改变个人的工作方式。然而，有一小部分受访者(36%)认为，人工智能将在未来五年内取代他们的工作。

9. **美国地方政府制定者在人工智能政策优先事项上存在明显分歧。**美国地方政府决策者虽普遍支持人工智能监管，但在具体政策优先事项上存在显著差异。支持率最高的政策包括更严格的数据隐私法规(80.4%)、失业人员再培训计划(76.2%)以及人工智能应用监管条例(72.5%)。然而，针对执法部门人脸识别禁令(34.2%)、工资下降补贴(32.9%)和全民基本收入(24.6%)等政策的支持率则显著降低。

10. **人工智能被视为提升效率的工具和娱乐体验的助推器，但其经济影响仍存疑虑。**全球对人工智能影响的看法各不相同。55%的人认为人工智能将节省时间，51%的人预计它将提供更好的娱乐选择，但对其健康或经济效益有信心的人较少。只有38%的人认为人工智能将提升医疗健康水平，36%的人认为人工智能将改善国民经济，31%的人认为人工智能将对就业市场产生积极影响，37%的人认为人工智能将提升自己的工作效率。

8.1 公众观点

全球公众观点

本节通过益普索在 2022 年、2023 年与 2024 年所开展的调查，探讨全球公众对人工智能的看法差异。结果表明，不同国家与人口群体间对人工智能的认知与态度存在显著差异。

人工智能产品和服务

2024 年，益普索开展了一项关于全球公众对人工智能态度的调查。该调查通过对 32 个国家 23,685 名成年人进行访谈完成。¹ 图 8.1.1 展示了受访者对特定陈述表达同意的比例。2022 年至 2024 年间，公众对人工智能的认知总体保持相对

稳定。2024 年，67% 的受访者表示对人工智能有良好理解，66% 预期人工智能将在未来三到五年内深刻改变其日常生活。认为人工智能驱动的产品与服务利大于弊的人群比例从 2022 年的 52% 小幅上升至 2024 年的 55%。

然而，图 8.1.1 也揭示了人们日益增长的担忧。过去一年中，相信人工智能企业会保护其个人数据的受访者比例下降了 3 个百分点，认为人工智能不会对任何群体产生歧视或偏见的受访者比例下降了 2 个百分点。

2022–2024 年全球公众对使用人工智能的产品和服务的看法（占总数的百分比）

资料来源：益普索，2022–2024 年 | 图表：2025 年人工智能指数报告

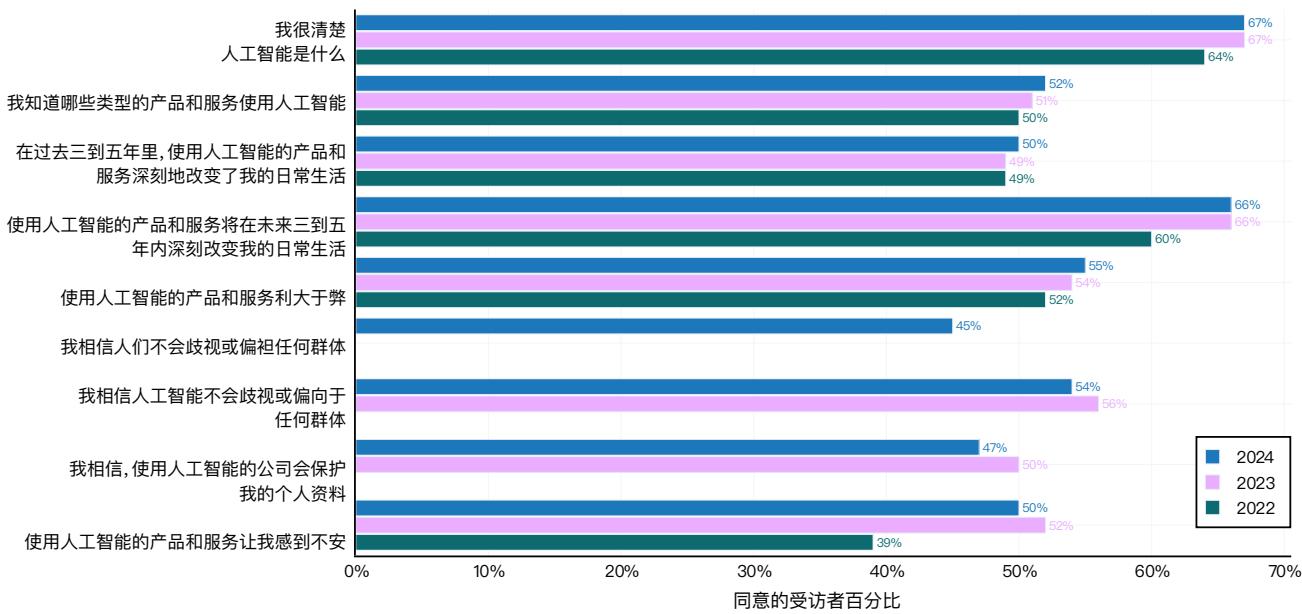


图 8.1.1

1、有关调查方法的详细信息，请参见附录。调查时间为 2024 年 4 月至 5 月。

第八章：公共观点

8.1 公众观点

根据益普索调查，不同国家之间对人工智能利弊的认知差异显著。总体而言，亚洲与拉丁美洲的受访者更倾向于认为人工智能利大于弊，例如中国（83%）、墨西哥（70%）与印度（62%）的受访者对人工智能持正面看法。相较之下，欧洲及英语国家则更为谨慎，如英国（46%）、澳大利亚（44%）、加拿大（40%）与美国（39%）的受访者认为人工智能利大于弊的比例明显偏低。

值得注意的是，曾在 2022 年持较强怀疑态度的国家中，公众情绪正逐步改善。益普索对 26 个国家的 2022 年与 2024 年数据对比显示，其中 18 个国家中，对人工智能持正面态度的比例有所提升。2022 年，法国（31%）、加拿大（32%）、美国（35%）、德国（37%）、澳大利亚（37%）与英国（38%）是最不看好人工智能的国家之一，而到了 2024 年，这些国家的相关比例均有所上升。

2022–2024 年各国公众认为人工智能利大于弊的比例

资料来源：益普索，2024 年 | 图表：2025 年人工智能指数报告

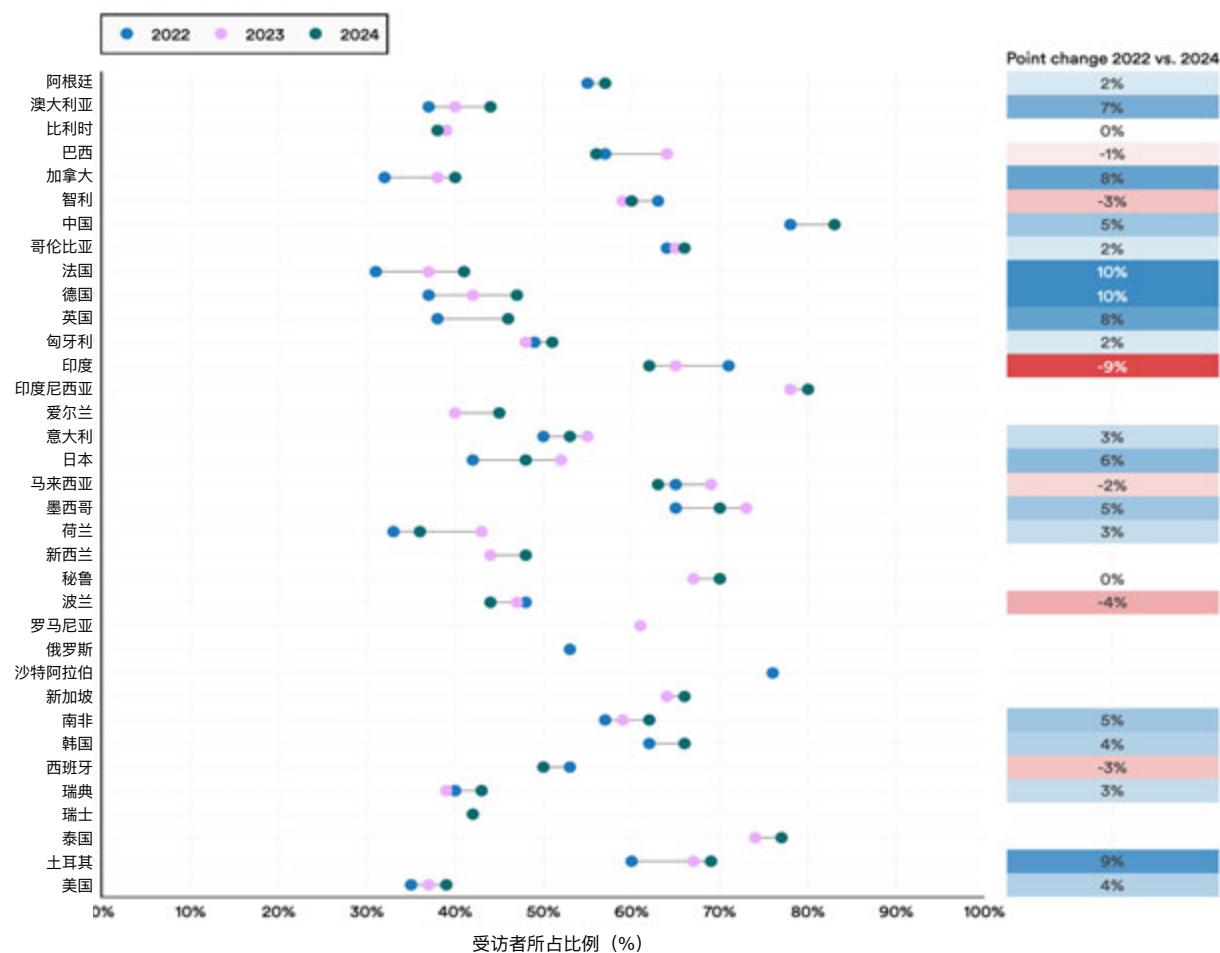


图 8.1.2

第八章：公共观点

8.1 公众观点

图 8.1.3 进一步展示了各国公众对人工智能产品与服务的认知、信任与情绪反应。在所有国家中，中国受访者对人工智能的认知程度、信任水平与兴奋感均为最高：81% 的人知道哪些产品或服务使用了人工智能，80% 对这些产品表示兴奋，76% 信任人工智能不会对群体产生偏见，86% 预期人工

智能将在三至五年内深刻改变其生活。与之形成对比的是，仅有 58% 的美国受访者相信人工智能将在未来三至五年内显著影响其生活，仅 34% 对人工智能产品表示兴奋。

对个人数据隐私的担忧在日本与加拿大最为显著，而对人工智能歧视性的担忧则在瑞典与比利时最高。

2024 年各国公众对人工智能产品的认知与态度

资料来源：益普索，2024 年 | 图表：2025 年人工智能指数报告

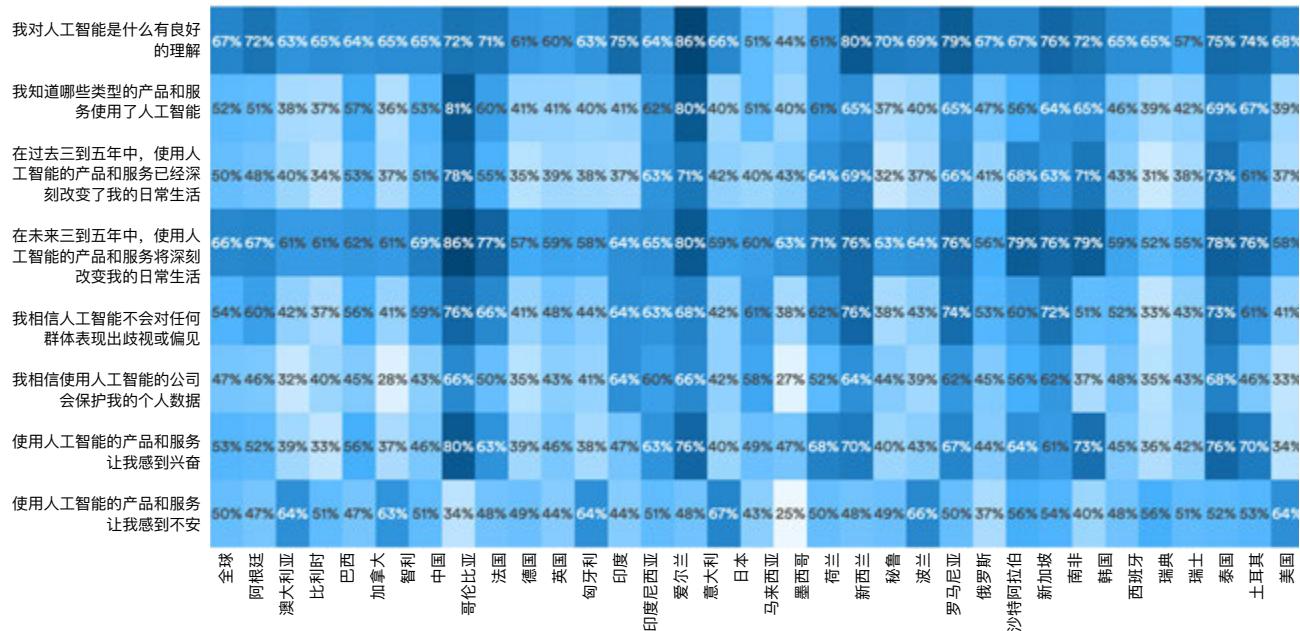


图 8.1.3

第八章：公共观点

8.1 公众观点

图 8.1.4 揭示了不同国家受访者对于人工智能的“兴奋感”与“焦虑感”之间的关系。结果显示，英语国家（如英国、美国、加拿大、澳大利亚与新西兰）对人工智能的焦虑感显著高于其他

国家，而兴奋感却普遍偏低。相对而言，亚洲国家。如中国、韩国与印度尼西亚的受访者则表现出较高的兴奋度与较低的焦虑水平。日本是该地区中对人工智能最为谨慎的例外。

2024 年各国公众对人工智能产品的“兴奋”与“焦虑”交叉态度分布图

资料来源：益普索，2024 | 图表：2025 年人工智能指数报告

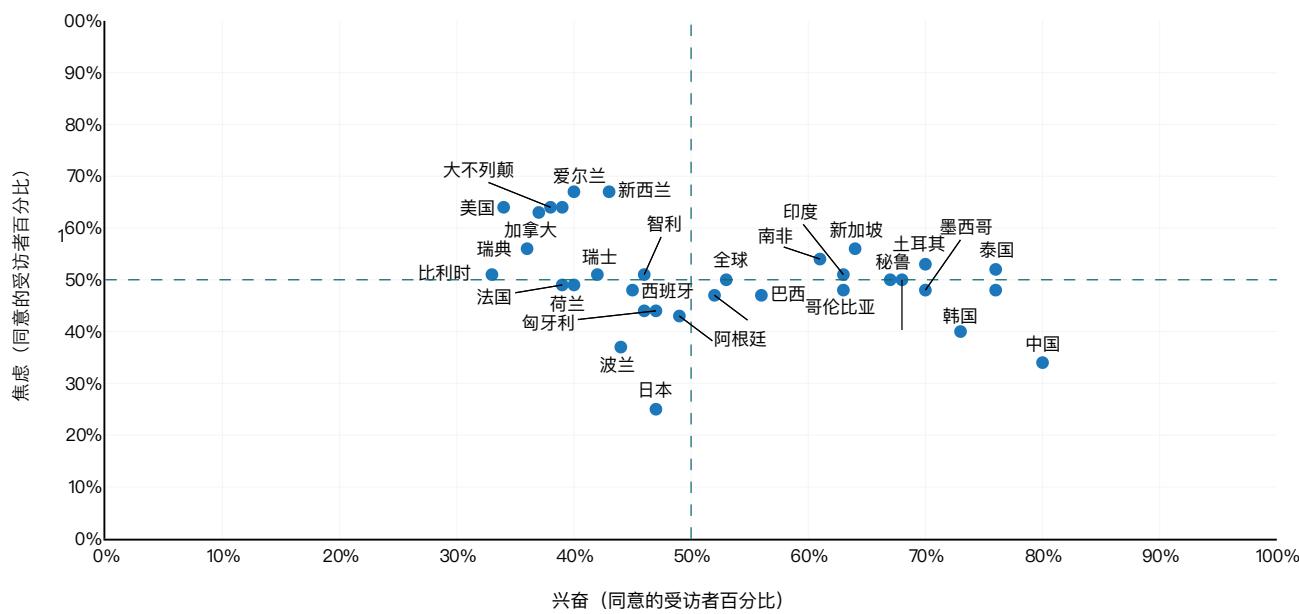


图 8.1.4

第八章：公共观点

8.1 公众观点

益普索连续两年对大多数国家进行调查，使得跨年度对比成为可能。图 8.1.5 显示了各项人工智能相关问题的年度变化趋势。整体来看，人工智能指数观察到对使用人工智能的担忧略有上升，正面回答平均下降了 0.6%。这主要是由于对使用人工智能公司是否会保护个人数据的信任度下降了 3%，以及人工智能是否不会歧视或偏袒任何群体信任度下降了 2%。²

巴西和马来西亚对人工智能的认知度、信任度和热情度降幅最

为显著。在这两个国家，负面趋势主要源于信任人工智能企业能保护其个人数据的受访者比例急剧下降。

南非和爱尔兰则在人工智能认知度、信任度和热情度方面呈现最显著的平均增幅。爱尔兰的积极趋势似乎源于良好的用户体验——该国自认日常生活深受人工智能产品服务深刻影响的受访者比例增幅高居全球之首。

2023–2024 年全球公众对人工智能的态度变化

资料来源：益普索，2022–2024 年 | 图表：2025 年人工智能指数报告

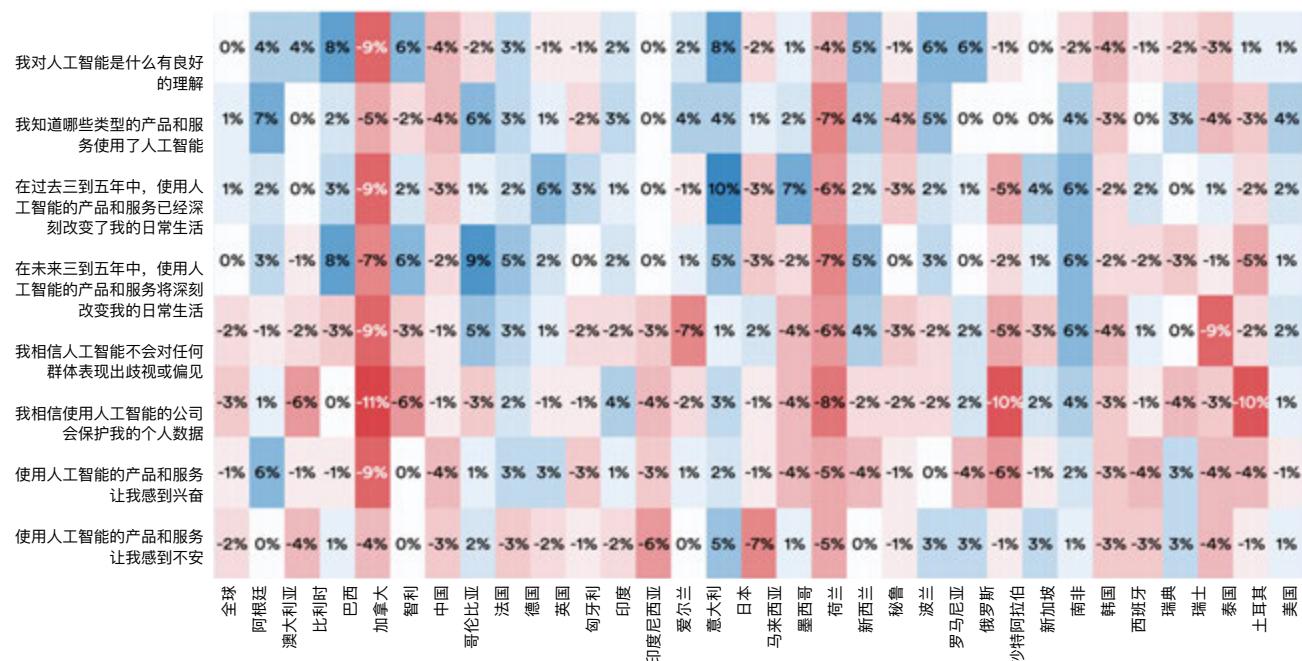


图 8.1.5

2、对“使用人工智能的产品和服务让我感到紧张”这一问题的全球平均回答不包括在内因为这是唯一一个正分数会产生规范性负面结果的问题。

第八章：公共观点

8.1 公众观点

图 8.1.6 对比了 2022 年和 2024 年益普索（益普索）调查数据，重点呈现了自 ChatGPT 推出以来的公众态度变化。全球范围内，认为人工智能驱动的产品和服务将在未来 3–5 年内深

刻改变日常生活的观点比例上升了 6%。除印度、马来西亚和波兰外，所有国家这一认知度自 2022 年以来均有提升，其中加拿大(17%)和德国(15%)涨幅最为显著。

2022 年和 2024 年公众对“人工智能将深刻改变生活”的认知变化比较

资料来源：益普索，2022–2024 年 | 图表：2025 年人工智能指数报告

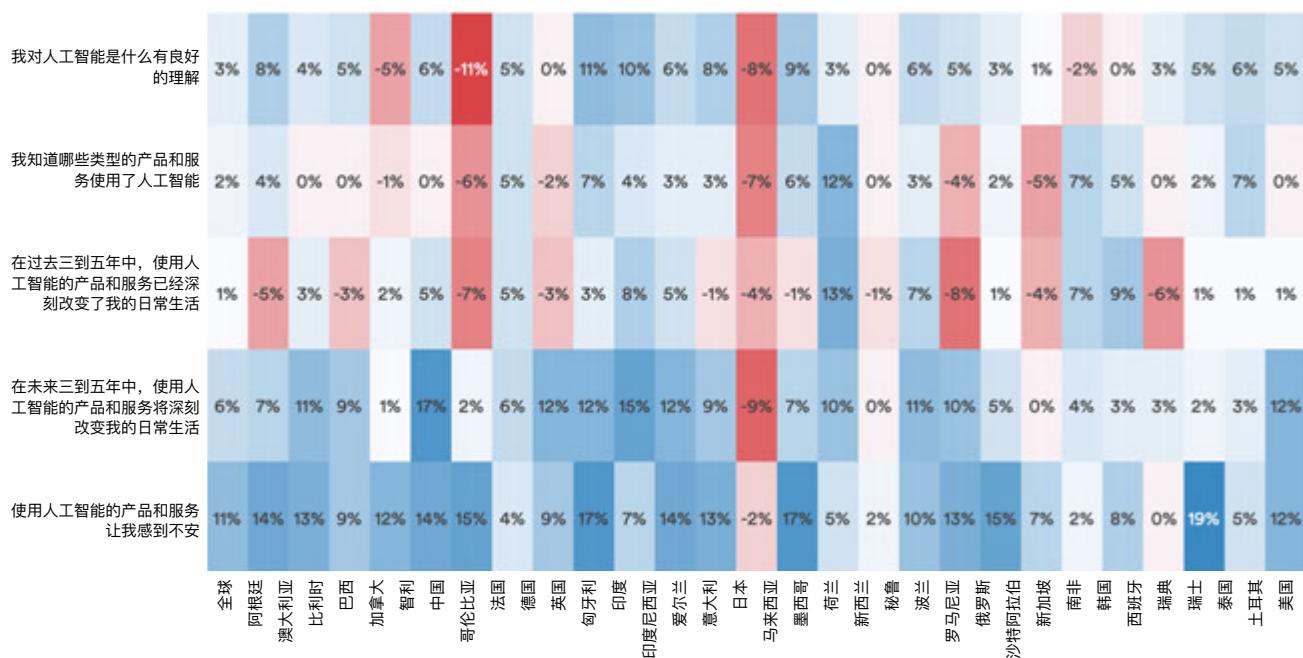


图 8.1.6

人工智能与就业

2024 年益普索调查新增了若干问题，探讨公众对人工智能影响当前工作的看法。图 8.1.7 展示了全球公众对人工智能将如何改变工作方式或取代现有岗位的预期。总体而言，60% 的受访者认为人工智能“可能”在未来五年内改变他们的工作方式，另有 36% 的受访者认为人工智能“可能”在同一时间内取代其当前职位，即每三人中就有一人持此观点。由于 2023

年版本未区分“非常可能”与“有些可能”，跨年比较存在一定局限性。但将 2024 年数据按总“可能性”加总后与 2023 年进行对比，整体情绪变化不大。2023 年，有 57% 的受访者认为人工智能将改变工作方式，36% 认为其可能取代当前工作，与 2024 年几乎一致。

2024 年全球对人工智能对当前工作影响的看法

资料来源：益普索，2024 | 图表：2025 年人工智能指数报告

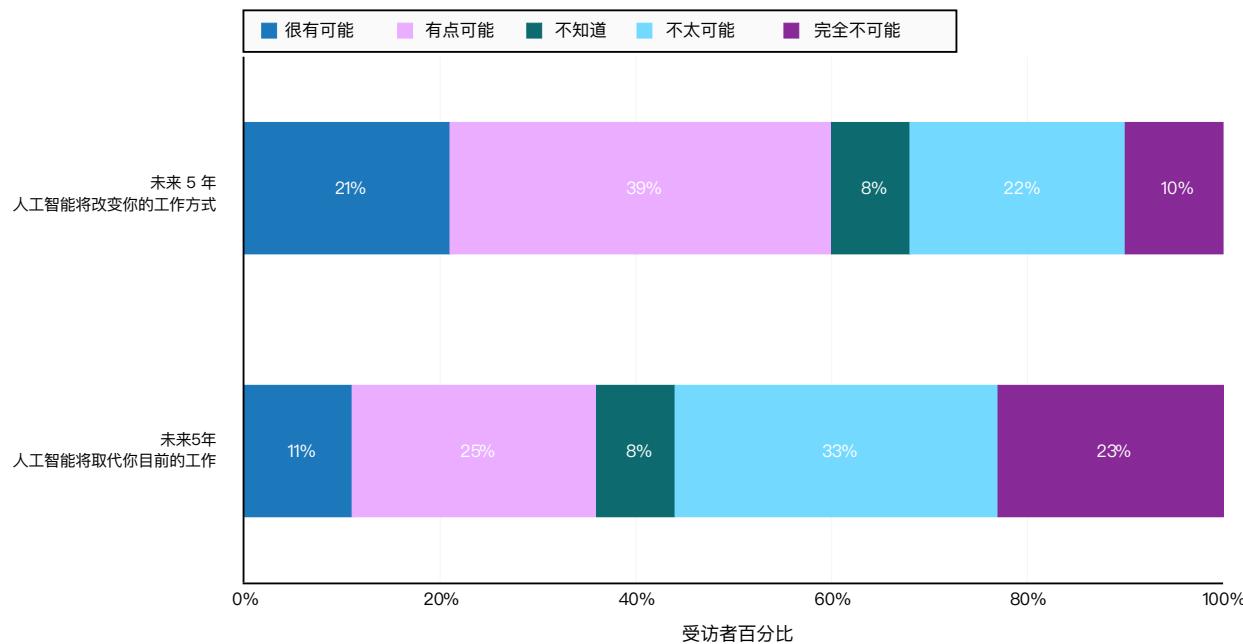


图 8.1.7

第八章：公共观点

8.1 公众观点

关于人工智能是否将在未来五年内改变人们当前工作的看法，不同代际群体之间存在显著差异（见图 8.1.8）。年轻一代（如 Z 世代与千禧一代）比年长群体（如 X 世代与婴儿潮一代）更倾向于认为人工智能将会改变他们的工作方式。具体而言，2024 年有 67% 的 Z 世代认同人工智能可能影响其当前工作，而婴儿潮一代中这一比例仅为 49%。

从 2023 年到 2024 年，所有代际群体中认为人工智能将改变工作方式的比例均有所上升。值得注意的是，在新增的 3% 受访者中，千禧一代与婴儿潮一代的增幅最大，或许表明跨代对人工智能影响力的意识正在增强。

2023 年与 2024 年全球公众对“人工智能将在未来五年改变当前工作方式”的认同比例（按代际划分）

资料来源：益普索，2024 年 | 图表：2025 年人工智能指数报告

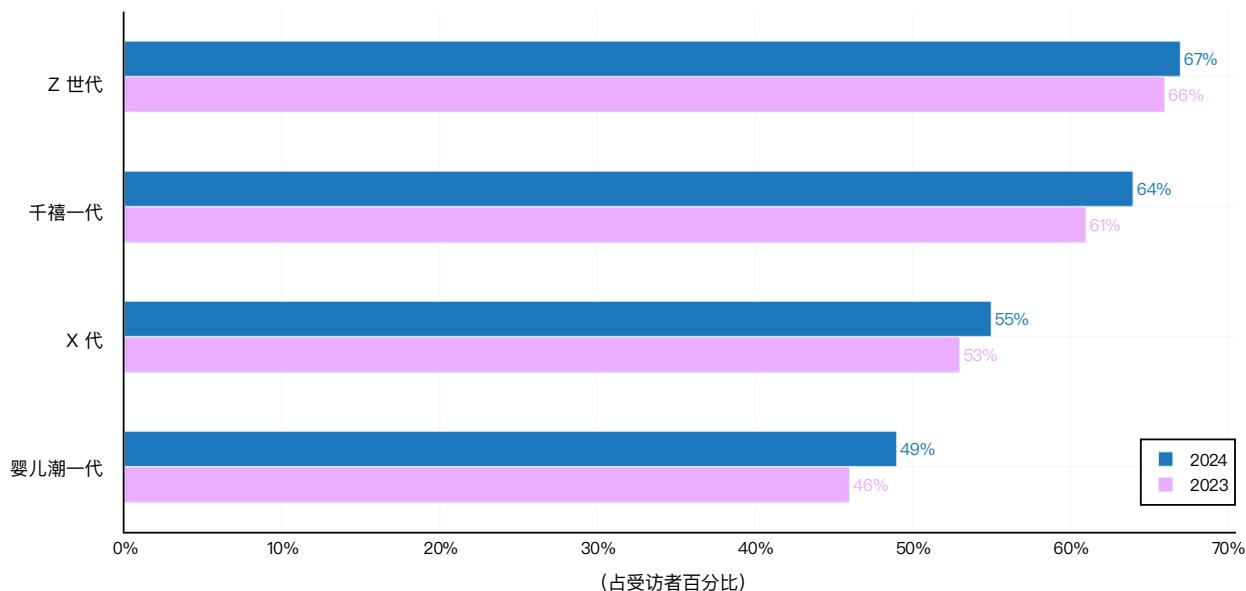


图 8.1.8

人工智能与民生

益普索调查还探讨了受访者在人工智能将对生活各个方面产生影响问题上的看法，包括经济、娱乐和健康等领域。

根据图 8.1.9，55% 的全球受访者认为人工智能将减少完成任务所需的时间，51% 的人认为人工智能将改善其娱乐选择。而在经济与就业市场方面，受访者则表现出更为谨慎的态度，分别仅有 36% 与 31% 的人认为人工智能将在这些领域产生积极影响。

图 8.1.9 还显示，不同国家中对人工智能“将改善本国经济”抱有信心的比例差异显著。亚洲国家最为乐观，其中中国有 72% 的受访者表达正面预期，其次是印尼(54%)。相比之下，

在荷兰、美国、比利时、瑞典与加拿大，认为人工智能将改善经济的比例均低于 25%。

此外，在每一个国家中，若受访者对人工智能改善经济抱有乐观态度，他们在其他领域的评价也往往更为积极。例如，那些认为人工智能将改善本国经济的群体也普遍相信人工智能能够节省时间并改善健康。

全球平均来看，有 38% 的受访者认为人工智能将改善健康。其中，墨西哥受访者的乐观程度最高，达 56%；而日本则为最低，仅有 19%。

2024 年全球公众对人工智能在经济、娱乐、健康等方面影响的认知

资料来源：益普索，2024 年 | 图表：2025 年人工智能指数报告

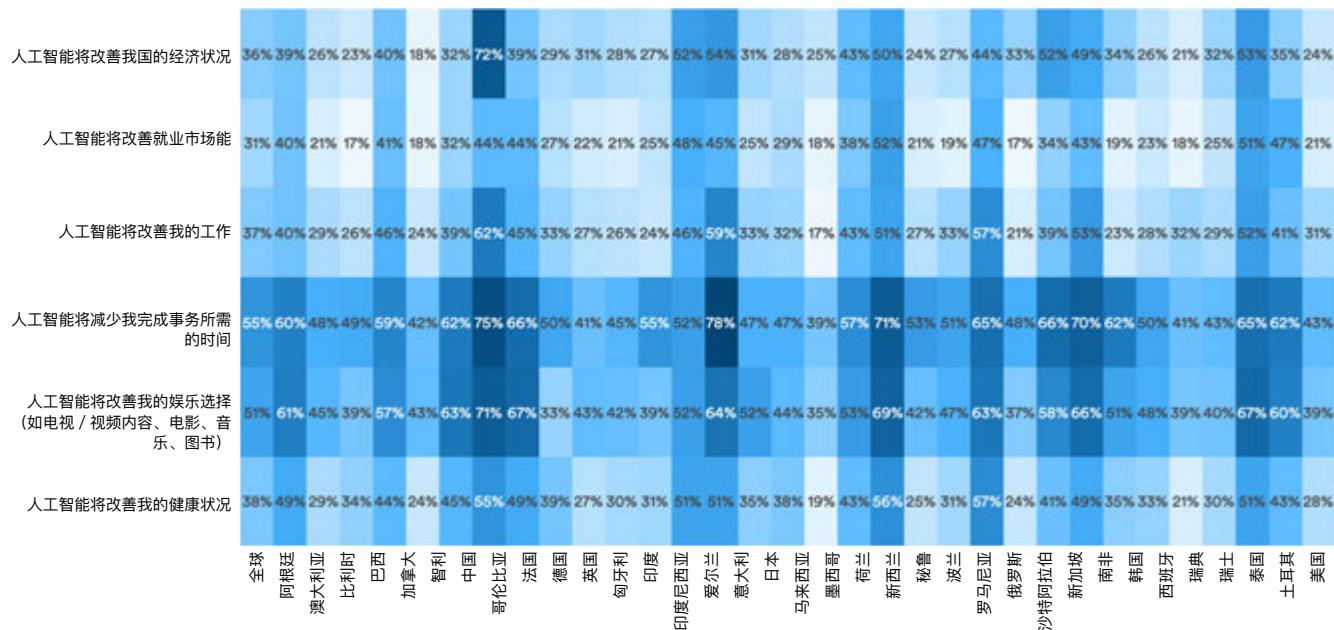


图 8.1.9

第八章：公共观点

8.1 公众观点

图 8.1.10 与图 8.1.11 对前文数据进行了相关性分析，旨在探究不同问题反馈之间的关联程度。研究发现，受访者对于“人工智能将改善就业市场”的认同度与其认为“人工智能将使自身工作受益”的信念呈现高度相关性。部分国家如波兰在这两方面的乐观程度均较低，分别仅有 17% 和 21% 的受访者表示认同。相比之下，中国受访者态度更为积极，44% 认为人

工智能将促进就业市场发展，62% 相信人工智能会改善自身工作状况。

同样地，在那些受访者认为人工智能将缩短任务完成时间的国家，民众也更倾向于认为人工智能会改善他们的个人工作状况。

2024 年全球对人工智能改善就业市场潜力与个人工作的看法

资料来源：益普索，2024 年 | 图表：2025 年人工智能指数报告

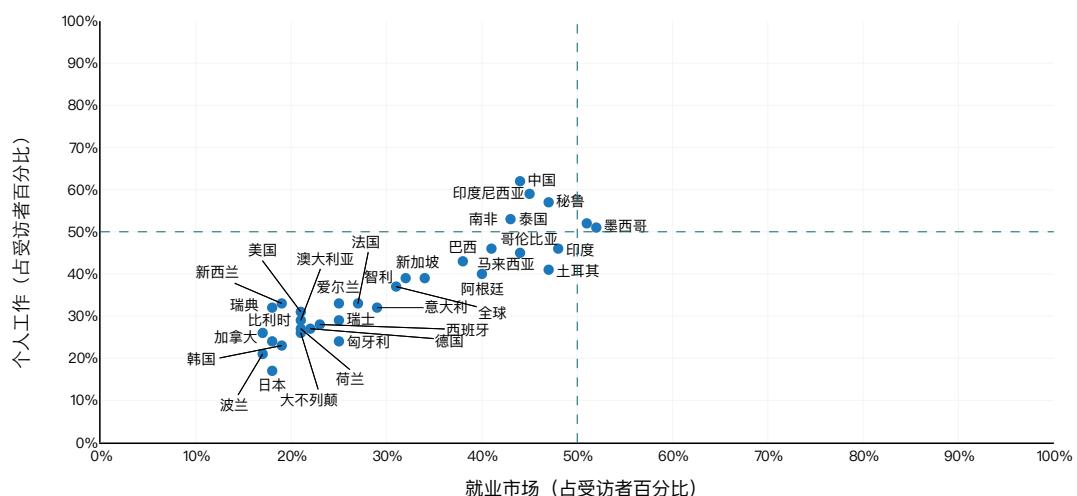


图 8.1.10

2024 年全球对人工智能改善完成工作时间与个人工作的潜力的看法

资料来源：益普索，2024 年 | 图表：2025 年人工智能指数报告

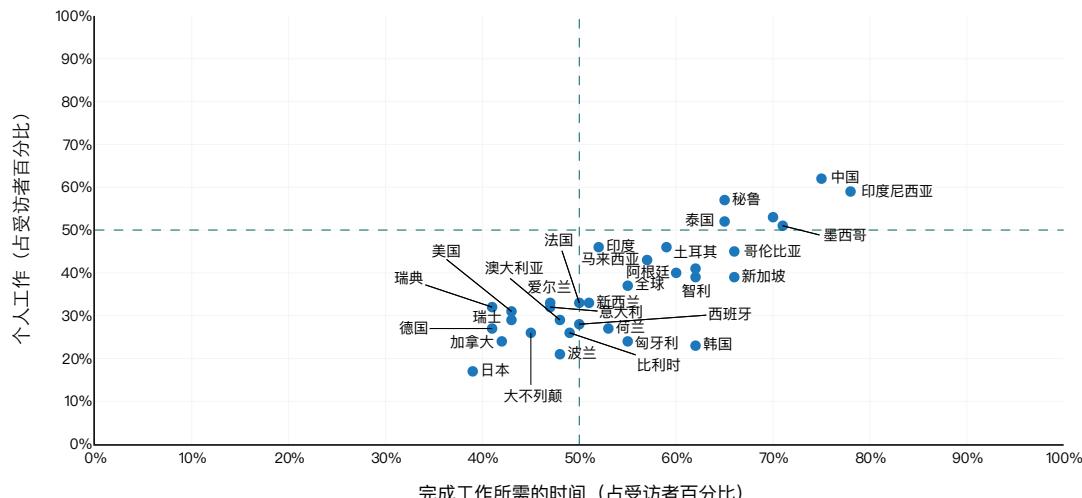


图 8.1.11

第八章：公共观点

8.1 公众观点

重点：

自动驾驶汽车

正如第二章技术性能中所讨论的，自驾车技术在能力与实际部署方面均取得显著进展。随着 Waymo 与 Zoox 等公司的发展，理解公众对自动驾驶技术的态度变得愈加关键。

美国汽车协会（AAA）每年都会就自驾车公众认知开展调查。最新调查于 2025 年 1 月进行，覆盖约 97% 的美国家庭。图 8.1.12 显示，尽管自动驾驶车辆正在逐步进入公共道路，但仍有一半以上的美国人表示对该技术感到恐惧。仅有 13% 的受访

者表示信任自动驾驶汽车。尽管“恐惧感”较 2023 年的高峰（68%）略有下降，但仍高于 2021 年的 54%。支持这一观点，较 2022 年的 55.7% 显著上升。ChatGPT 的发布似乎是推动政策制定者转向支持监管的关键因素。民主党人对人工智能监管的支持率（79.2%）高于共和党人（55.5%），但 2022 年后两党在此问题上的支持率均呈现明显提升。

2021–2025 年美国驾驶员对自动驾驶汽车的态度

资料来源：AAA, 2025 | 图表：2025 年人工智能指数报告

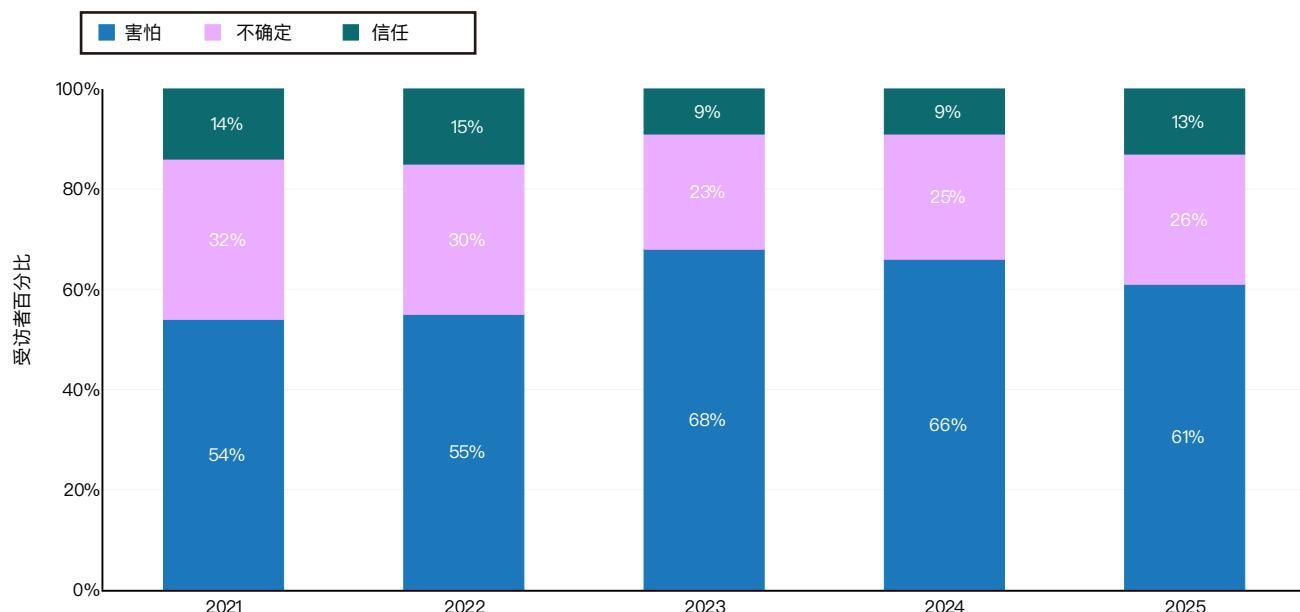


图 8.1.12

8.2 美国政策制定者的态度

在评估公众对人工智能的态度时，仅考察大众舆情是不够的，还必须关注关键利益相关方的看法，尤其是政策制定者，因为他们在人工智能监管与政策制定中扮演着核心角色。2022年与2023年，由乌普萨拉大学、牛津大学、哈佛大学和雪城大学联合组成的研究团队开展了一项涵盖美国地方政策制定者（包括市政、镇区与县级层面）的调研。这项研究共收集了约1,000名地方官员的反馈，其时间跨度恰好涵盖ChatGPT发

布前后，从而使得研究者能够比较政策制定者态度的变化。图8.2.1显示了地方政策制定者对“政府应监管人工智能”这一陈述的认同程度。2023年，73.7%的美国地方政策制定者支持这一观点，较2022年的55.7%显著上升。ChatGPT的发布似乎是推动政策制定者转向支持监管的关键因素。民主党人对人工智能监管的支持率（79.2%）高于共和党人（55.5%），但2022年后两党在此问题上的支持率均呈现明显提升。

按政党和年份分类的美国地方官员对政府监管人工智能的支持程度

资料来源：Hatz等，2025 | 图表：2025年人工智能指数报告

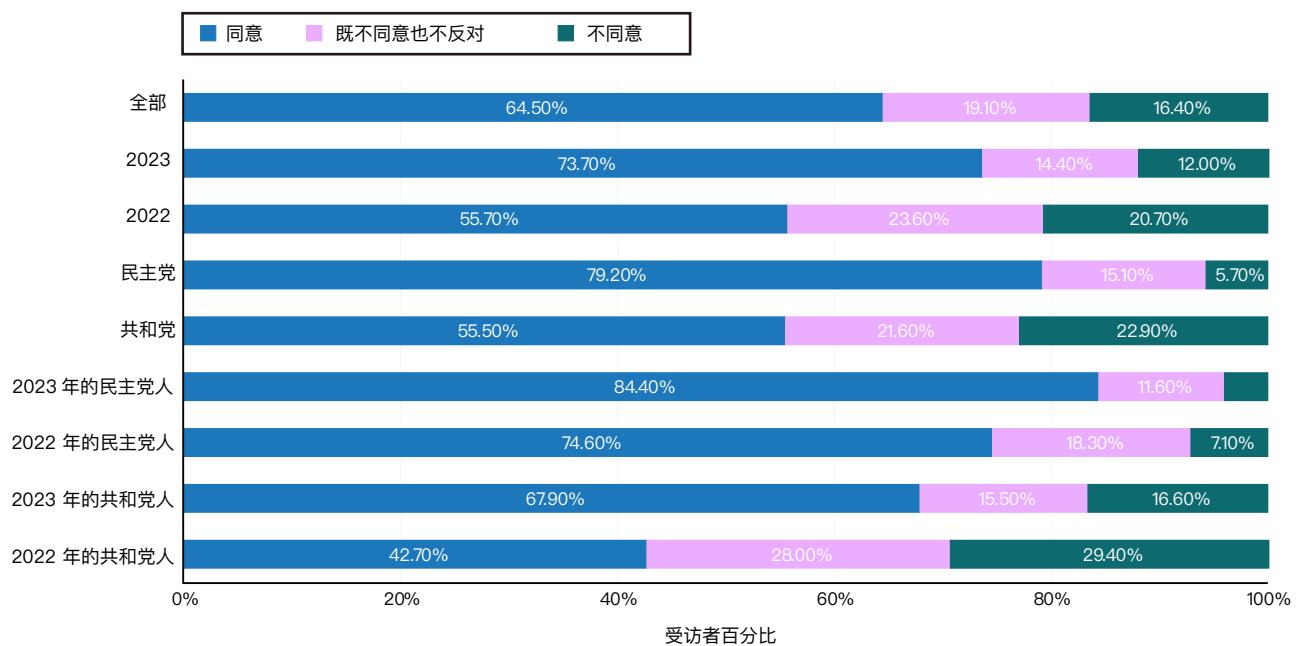


图 8.2.1

第八章：公共观点

8.2 美国政策制定者的意见

鉴于大多数地方政策制定者支持某种形式的人工智能监管，他们具体倾向于哪些政策？数据显示（图 8.2.2），支持率最高的是加强数据隐私监管（80.4%）。此外，76.2% 的受访者支持针对失业人员的再培训计划，72.5% 支持实施人工智能应用

监管条例。相比之下，对再分配措施的支持率明显较低：仅 33.9% 支持通过工资补贴抵消收入下降，24.6% 支持全民基本收入制度。

美国地方官认为在 2025–2050 年间具有积极意义的人工智能政策选项

资料来源：Hatz 等，2025 | 图表：2025 年人工智能指数报告

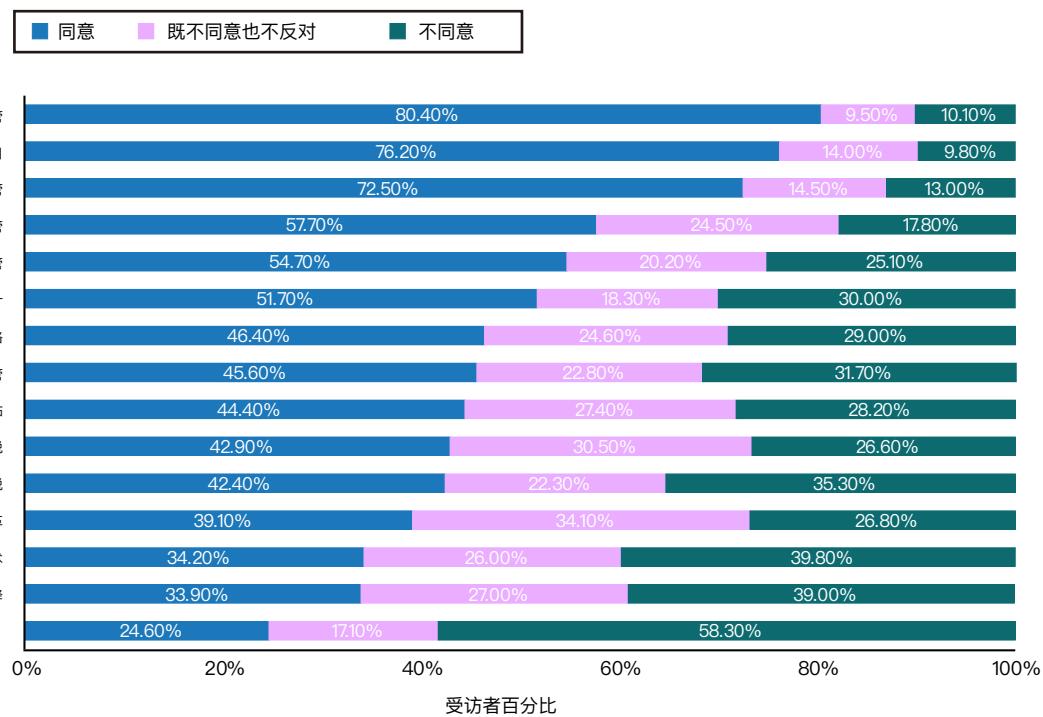


图 8.2.2

第八章：公共观点

8.2 美国政策制定者的意见

尽管多数地方官员支持人工智能监管，但只有少数人认为自己需要在短期内采取行动（见图 8.2.3）。2023 年，仅有 34.3% 的受访者认为其将在未来数年内需要做出相关决策，而

56.5% 认为不太可能。但这一比例已较 2022 年有所上升：从 32.2% 升至 36.6%。这一变化反映了诸如 ChatGPT 等重要人工智能发展成果对政策制定者态度的影响。

按党派和年份分类的美国地方官员制定人工智能相关政策的决定的可能性

资料来源：Hatz 等，2025 | 图表：2025 年人工智能指数报告

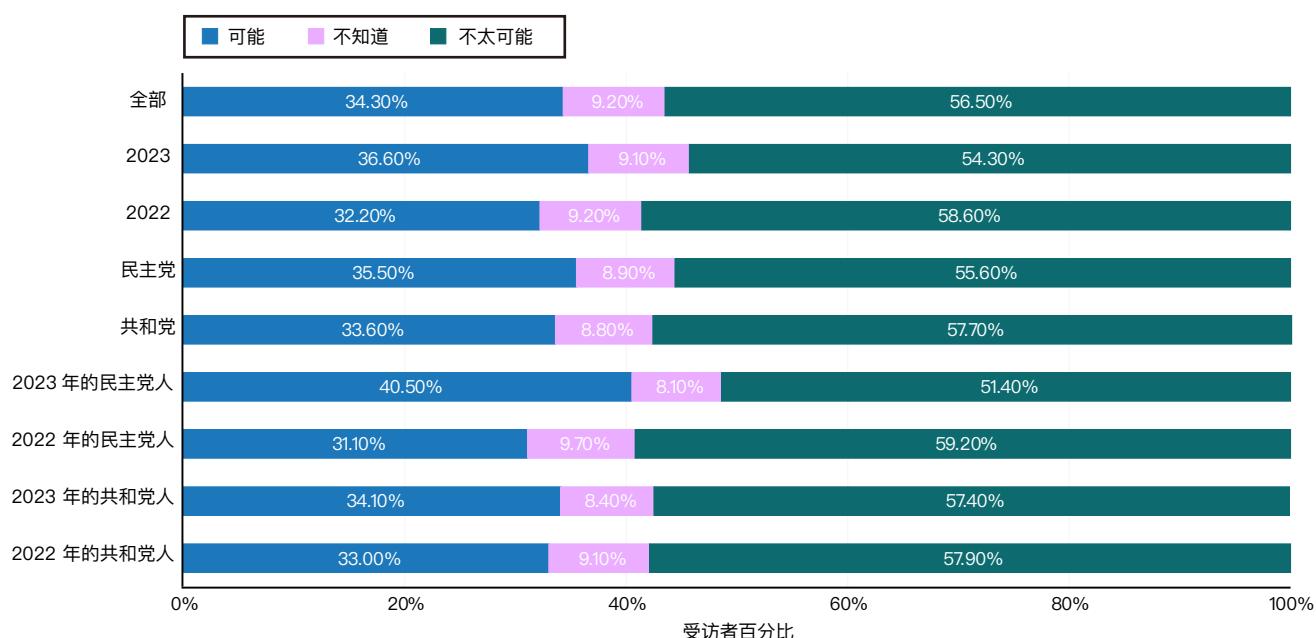


图 8.2.3

第八章：公共观点

8.2 美国政策制定者的意见

仅有 29.8% 的当地选举产生的官认为自己具备足够的信息来制定人工智能政策(图 8.2.4)。虽然从 2022 到 2023 年，民主党与共和党官员的信心水平均有所提升，但整体水平仍相对较低。

按政党和年份分类的美国地方官员对制定人工智能政策具备充足信息的认同程度

资料来源：Hatz 等，2025| 图表：2025 年人工智能指数报告

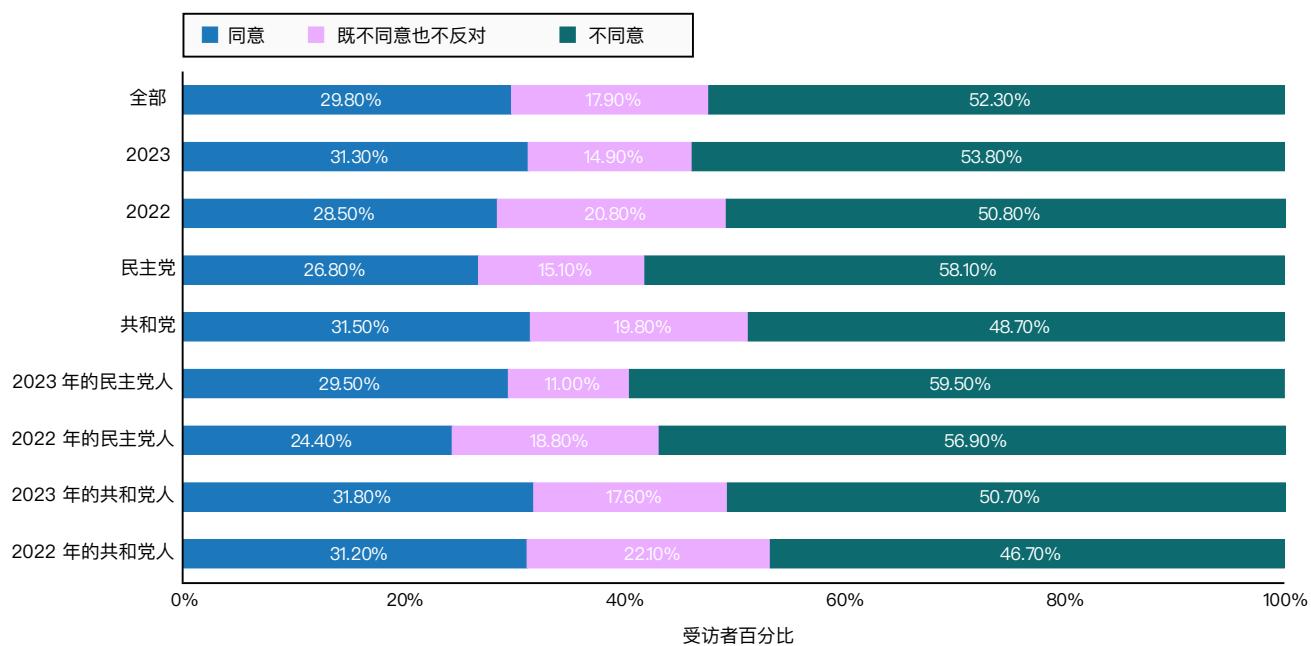


图 8.2.4

附录



附录

第一章	研究与开发	416
第二章	技术性能	420
第三章	负责任的人工智能	427
第四章	经济	431
第五章	科学与医学	441
第六章	政策与治理	451
第七章	教育	454
第八章	公共观点	455

附录

第一章：研究与开发

第一章：研究与开发

致谢

人工智能指数谨此致谢：Angelo Salatino 对人工智能论文分类的贡献，Ben Cottier 领导的机器学习推理成本分析，Lapo Santarlasci 领导的人工智能专利分析，Andrew Shi 领导的人工智能模型环境影响分析。

人工智能论文发表分析

在这项分析中，人工智能指数使用 OpenAlex 作为主要据源，这是一个开放式学术数据库，拥有超过 2.6 亿份研究论文统计。OpenAlex 使用自己的知识组织系统对论文进行分类，该系统被称为 OpenAlex Topics——这是一个包含约 4500 个主题的分类系统，将 Scopus 代码 和 CWTS 分类 结合在一起。该系统使用深度学习模型，考虑了标题、摘要、期刊名称和引文网络进行分类。为了更精确地识别人工智能相关主题，人工智能指数分析了 OpenAlex 识别的计算机科学论文统计，并使用计算机科学本体和 CSO 分类器完善了分类。

计算机科学本体 (CSO) 是一个大规模的、自动生成的研究领域本体，它是利用 Klink-2 算法从 1600 万个论文统计得出的。它的特点是具有数千个子主题的分层结构，可以精确率地将特定术语映射到更广泛的研究领域。与 OpenAlex、Scopus 和 Web of Science 等通用学术数据库相比，CSO 提供了更详细的 和更精细的研究领域表述。, 它已被广泛用于 学术数据探索、分析、建模以及专家鉴定和推荐。本分析中使用的 3.4.1 版包括计算机科学领域的约 15,000 个主题和 166,000 个关系。该版本于 2025 年 1 月 17 日发布，引入了 150 多个人工智能方面的新研究课题，使人工智能相关的课题总数达到

2369 个，仅人工智能领域就有 12620 个层次关系。为了分析研究趋势，人工智能指数使用了 CSO 分类器——一种基于 CSO 主题自动对研究论文进行分类的无监督方法。分类器采用三阶段流程处理论文标题和摘要：语法模块检测 CSO 主题的直接提及；语义模块使用词嵌入来识别相关概念；后处理模块合并结果，过滤掉不相关的主题，并添加更广泛的类别，以实现更精细的分类。在本次分析中，人工智能指数扩展了 CSO 分类器，特别人工智能及其子主题。自首次发布以来，该分类器因其多功能性获得了越来越多的关注。例如，Springer Nature 使用它对论文集书籍进行常规分类，提高元数据质量。除了 学术出版 之外，它还被成功应用于研究软件、YouTube 视频、新闻稿、招聘广告和 IT 博物馆藏品的分类。

准确地将研究论文归类为会议论文集或期刊论文对本分析至关重要。OpenAlex 的元数据字段——类型、交叉引用类型和来源类型——有时会发生冲突。为了解决这些不一致问题，人工智能指数将 OpenAlex 记录映射到 DBLP，这是一个领先的计算机科学论文统计数据库。DBLP 以其高质量的元数据而著称，它通过严格的半自动化整理过程不断增加新的论文统计，目前已为 360 万篇会议论文和 300 万篇期刊论文编制了索引。OpenAlex 和 DBLP 之间的初始匹配是使用 DOI 进行的。对于其余未匹配的论文，人工智能索引使用标题和出版年份的组合。为了简化这一过程，人工智能指数建立了一个标题索引，以优化搜索并确保跨数据集的高效映射。

人工智能论文统计根据多个参数进行汇总，以提供综合分析。论文统计考虑到最新版本的出版日期，按年份对论文进行了分

附录

第一章：研究与开发

组。此外人工智能指数团队还按照作者所属的地理区域或世界银行地区进行分组。这就意味着，如果一篇论文由来自不同国家的研究人员共同撰写，那么这篇论文可能会被多次统计，每个国家都会被统计一次。如果作者所属单位缺失，这些论文统计将被映射为“未知”。此外，如果有作者所属单位，则通过作者所属单位将部门与论文统计联系起来，这可能导致一篇出版物被多个部门统计。如有引文数据，则包括引文计数；没有引文数据的出版物归类为“未知”。

百强论文统计分析

人工智能指数通过收集和分析 OpenAlex、谷歌学术（Google Scholar）和 Semantic Scholar 等多个来源的引文数据，对有影响力的人工智能论文发表统计进行了全面分析。最初从 OpenAlex 收集的是每个出版年被引用最多的 150 篇论文，经过仔细审核后，该名单被细化为 100 篇论文统计。

该方法将论文统计归属于作者所属单位所代表的所有国家和地区，这意味着一篇论文可以被多次统计。例如，一篇由美国和中国人员共同撰写的论文在每个国家都会被统计一次。这种方法可能会导致汇总统计中的总数重叠。无论是在期刊、会议还是 arXiv 这样的资料库中，发表年份都以最新版本为准。为保持准确率，对组织隶属关系进行了核实和标准化，根据总部所在地分配国家。

人工智能论文发表统计 100 强的完整名单可在这里查阅。

人工智能专利分析

人工智能指数采用混合分类法，将基于关键词的文本分析与基于分类代码的识别相结合，识别与人工智能相关的专利。

专利文献数据来自 PATSTAT Global，这是一个由欧洲专利局（EPO）发布的综合数据库。¹ 专利根据最早记录的授权出版物的出版机构归属于各个国家。

使用 deep-translator 工具、谷歌翻译引擎和 Meta NLLB-200 机器翻译模型翻译了最初以英语以外的语言出版的专利摘要和标题。翻译后，使用自然语言处理（NLP）技术对专利文本进行处理。这些技术包括删除停顿词和特殊字符、保留关键语法类别的语篇（POS）标记、小写转换、词法化以及 <NUM> 标记替换数字量。

与人工智能相关的专利是通过使用正则表达式（regex）搜索标题和摘要中的相关术语来确定的。人工智能专用关键词词典是通过一个结构化的多步骤过程开发的，其中纳入了人工智能模型生成的关键词，并使用已建立的人工智能词典（如 Yamashita 等（2021）中的词典）进行了扩展，还通过基于 Word2Vec 的同义词识别进行了完善。使用 BERTopic 主题建模和基于 DeBERTA 的零镜头分类进行了进一步验证，并采用人工检查以减少误报。

除了基于关键词的分类外，使用 国际专利分类（IPC）和 合作专利分类（CPC）代码识别了与人工智能相关的专利。通过人工智能模型分析、基于 regex 的搜索和先前的研究（包括来自 Pairolo 等的分类（2023）和 WIPO（2024）），编制了一份人工智能相关代码的整理清单。最终的数据集是通过合并两种方法的结果构建的，同时兼顾了覆盖率和准确率。

Epoch 标志性模型分析

人工智能预测研究小组 Epoch AI 维护着一个 数据集其中包括具有里程碑意义的人工智能和机器学习模型，以及关于这些

¹ 尽管采用了这种汇总程序，但在同一 DOCDB 族中的申请具有相同最早申请日的边缘情况下，偶尔会出现重复。人工智能指数在进行分析时，会删除与汇总变量（如按年份计算）有关的重复值。

附录

第一章：研究与开发

模型的创建者和论文统计的相关信息，如作者列表、引用次数、完成的人工智能任务类型以及在训练中使用的计算量。这些论文作者的国籍对人工智能地缘政治预测具有重要影响。随着各种研究机构和技术公司开始生产先进的机器学习，人工智能发展的全球分布可能会转移或集中在某些地方，进而影响地缘政治格局，因为人工智能有望在不久的将来成为经济和军事实力的重要组成部分。

为了追踪各国在地标性论文统计上的人工智能研究贡献分布情况，大纪元数据集按照以下方法进行编码：

1. 数据集截取于 2025 年 3 月。其中包括有关里程碑式模型的论文，这些论文是根据“计算趋势”数据集文档中所述的重要性、相关性和独特性等纳入标准筛选出来的。
2. 作者所属国家以论文中注明的所属单位为准。对于国际组织，除非注明更具体的地点，否则作者归属于该组织的总部所在国。
3. 所有具有里程碑意义的论文统计都会在时间段内（如每月或每年）进行汇总，并对各国的贡献进行汇编，以确定每个国家在每个时间段内对具有里程碑意义的人工智能研究的贡献程度。
4. 对不同国家的贡献进行长期比较，以确定任何趋势。

训练成本分析

为了创建成本估算数据集，Epoch 数据库过滤了大规模机器学习时代 2 发布的模型，这些模型在发布时的训练计算量排名前 10 位。这样就筛选出了最大规模的机器学习。Transformer 模型被添加到这组模型中，以进一步了解背景情况。

对于所选的机器学习，其训练时间、类型、数量和硬件使用率均根据适用情况从出版物、新闻稿或技术报告中确定。这些模型使用的计算硬件的云租赁价格是从云供应商网站的在线历史档案中收集的。³

训练成本是根据硬件类型、数量和时间估算得出的，方法是将每小时的云费（培训时）⁴乘以硬件小时数。然而，有些开发人员购买了硬件而不是租用云计算，而且云计算价格因供应商和租用承诺而异，因此开发人员的真实成本可能会有所不同。

在估算这些模型的训练成本时遇到了各种挑战。通常情况下，开发人员不会透露培训时间或所使用的硬件。在其他情况下，硬件的云计算价格也无法获得。关于训练成本趋势的调查，在 Epoch AI 的另一份报告中有更详尽的说明。

人工智能会议参会规模

人工智能指数联系了 2024 年各种人工智能学术会议的组织者，请他们提供有关总出席人数的信息。对于在网上公布了总出席人数的会议，人工智能指数采用了这些报告的总人数，而没有联系会议组织者。

GitHub

确定人工智能项目

GitHub 与哈佛商学院、微软研究院和微软人工智能公益实验室的研究人员合作，按照 Gonzalez, Zimmerman, and Nagappan (2020) 和 Dohmke, Iansiti, and Richards (2023) 的方法识别公共人工智能资源库，分别使用与 AI/ML 生成式人工智能相关的主题标签，以及通过雪球取样确定的其他相关关

2. 根据 Compute Trends Across Three Eras of Machine Learning (Epoch, 2022 年)，选定的截止日期为 2015 年 9 月 1 日。

3. 历史价格来自亚马逊网络服务、微软 Azure 和谷歌云平台价格目录的存档快照，可通过 Internet Archive Wayback Machine 查看。

4. 所选的租用率是模型开发者使用硬件和云供应商的最新公布价格，按三年承诺租用率计算，减去培训时间和公布日期后的两个月。如果无法提供该价格，则使用最类似的价格 —— 要么是不同日期的相同硬件和供应商，要么是不同云供应商的相同硬件。如果没有三年期承诺租用率，根据给定云供应商的经验平均折扣从其他租用率中推算。如果没有准确的硬件类型（如 Nvidia A100 SXM4 40GB），则使用通用类型（如 Nvidia A100）。

附录

第一章：研究与开发

关键词，如“机器学习”、“深度学习”和“人工智能”。GitHub 通过依赖 Python 的 PyTorch、TensorFlow、OpenAI、Transformers、XGBoost、scikit-learn 和 SciPy 库的资源库进一步扩充了数据集。

将人工智能项目映射到地理区域

公共人工智能项目通过 IP 地址地理定位映射到地理区域，以确定项目所有者每年的模式位置。每个项目所有者在与 GitHub 交互时都会根据其 IP 地址分配一个位置。如果项目所有者在一年内更换了位置，那么项目的位置将由这一年中每天采样的所有者的模式位置决定。此外，即使项目所有者没有执行任何活动，项目所有者的最后已知位置也会按日结转。如果项目所有人在美国境内开展了活动，然后有六天没有开展活动，那么该项目所有人在那七天中将被视为在美国境内。

环境影响分析

人工智能指数使用（2019 年）提出的计算器估算了训练语言和视觉模型的碳排放量。分析的重点是培训阶段的排放，不包括硬件生产、闲置基础设施和部署排放。研究考察了四类模型：工业语言模型、学术语言模型、工业愿景模型和学术愿景模型。

计算器的准确率已根据公布的排放值进行了验证。计算器的输入包括硬件类型、图形处理器小时数、提供商和计算区域。对于较新的硬件，如 H100 图形处理器（2022 年发布），计算中使用 A100 SXM4 80GB 作为替代。提供商的选择基于已知的合作伙伴关系（例如，谷歌模型使用 GCP，OpenAI 使用 Azure），而计算区域则由团队所在地决定。

特别考虑了在定制硬件上训练的模型，如 BLOOM 在法国使用

的 Jean Zay 超级计算机。在这些情况下，私人基础设施的计算纳入了碳效率（千克 / 千瓦时）和抵消百分比。

该研究共评估了 50 个模型：34 个行业语言模型（2018–24 年）、8 个行业愿景模型（2019–23 年）、4 个学术语言模型（2020–23 年）和 4 个学术愿景模型（2011–22 年），并选择了在各自领域具有特别影响力的模型。

附录

第二章：技术性能

第二章：技术性能

致谢

人工智能指数谨此致谢：Andrew Shi（负责生成 Midjourney 和 Pika 视频制作样本）以及 Armin Hamrah（负责梳理人工智能重大技术进展时间线）所作出的贡献。

比较基准

在本章节中，人工智能指数对比较基准进行了报告，承认比较基准在跟踪人工智能技术进步方面的重要性。作为标准做法，指数从 leaderboard、公共资料库（如 Papers With Code 和 RankedAGI）以及公司论文、博客文章和产品发布中获取比较基准分数。指数运行的前提是各公司报告的分数是准确和真实的。本节中的比较基准分数是截至 2025 年 2 月中旬的最新分数。不过，自人工智能指数发布以来，可能会更新的机型发布，从而超过当前的最先进分数。

1.ARC–AGI 通用人工智能：有关 ARC–AGI 的数据取自 2025 年 2 月 ARC–AGI 论文和 OpenAI 视频。要了解有关 ARC–AGI 的更多信息，请阅读原文。

2.Arena–Hard–Auto：Arena–Hard–Auto 的数据来自 2025 年 2 月的 LMSYS leaderboard。要了解有关 Arena–Hard–Auto 的更多信息，请阅读原文。

3.Bench2Drive：有关 Bench2Drive 的数据摘自 2025 年 2 月 Bench2Drive 论文。要了解有关 Bench2Drive 的更多信息，请阅读原文。

4.伯克利函数调用：有关伯克利函数调用的数据取自 2025 年 2 月伯克利函数调用 leaderboard。要了解有关伯克利函数调用的更多信息，请阅读原著。

5.BigCodeBench：有关 BigCodeBench 的数据取自 2025 年 2 月 BigCodeBench leaderboard。欲了解有关 BigCodeBench 的更多信息，请阅读原文。

6.Chatbot Arena：有关 Chatbot Aren 的数据取自 2025 年 2 月 Chatbot Arena Leaderboard。要了解有关 Chatbot Arena 的更多信息，请阅读原文。

7.FrontierMath：有关 FrontierMath 的数据来自 FrontierMath 论文和 OpenAI 视频。要了解有关 FrontierMath 的更多信息，请阅读原文。视觉效果还辅以 OpenAI 的 o3 模型的比较基准数据，这些数据来自 YouTube 视频，该视频宣布于 2025 年 12 月推出该模型。

8.GAIA：有关 GAIA 的数据取自 2025 年 2 月 GAIA leaderboard。要了解有关 GAIA 的更多信息，请阅读原文。

9.GPQA：有关 GPQA 的数据来自 2025 年 2GPQA 论文和 OpenAI 视频。要了解有关 GPQA 的更多信息，请阅读原文。

10.GSM8K：有关 GSM8K 的数据取自 2025 年 2 月的 GSM8K Papers With Code leaderboard。要了解有关 GSM8K 的更多信息，请阅读原文。

11.HELMET：有关 HELMET(如何有效、彻底地评估长语境模型)的数据摘自 2025 年 2 月的 HELMET 论文。要了解有关 HELMET 的更多信息，请阅读原文。

12.HLE：有关 HLE 的数据摘自 2025 年 2 月 HLE 论文。要了解有关 HLE 的更多信息，请阅读原文。

13.HumanEval：有关 HumanEval 的数据来自 2025 年 2 月的 HumanEval Papers With Code leaderboard。要了解有关 HumanEval 的更多信息，请阅读原文。

14.LRS2：Oxford–BBC Lip Reading Sentences

附录

第二章：技术性能

15.2 (LRS2) 的数据取自 2025 年 2 月 LRS2 Papers With Code leaderboard。要了解有关 LRS2 的更多信息,请阅读原文。

16.MATH: 有关 MATH 的数据取自 2025 年 2 月 MATH Papers With Code leaderboard 和 o3-mini 模型发布会。要了解有关 MATH 的更多信息,请阅读原文。

17.MixEval: 有关 MixEval 的数据取自 2025 年 2 月的 MixEval leaderboard。要了解有关 MixEval 的更多信息,请阅读原文。

18.MLU: 关于 MLU 的数据取自 2025 年 2 月 MMLU Papers With Code leaderboard 要了解更多关于 MLU 的信息,请阅读原文。

19.MMLU-Pro: 有关 MMLU-Pro 的数据取自 2025 年 2 月 MMLU-Pro leaderboard。要了解有关 MMLU-Pro 的更多信息,请阅读原文。

20.MMMU: 有关 MMMU 的数据取自 2025 年 2 月的 MMMU leaderboard。要了解有关 MMMU 的更多信息,请阅读原文。

21.MTEB: 有关海量文本嵌入比较基准 (MTEB) 的数据取自 2025 年 2 月 MTEB leaderboard。要了解有关 MTEB 的更多信息,请阅读原文。

22.MVBench: 有关 MVBench 的数据取自 2025 年 2 月 MVBench leaderboard 要了解有关 MVBench 的更多信息,请阅读原文。

23.PlanBench: 有关 PlanBench 的数据摘自 2025 年 2 月 PlanBench paper。要了解有关 PlanBench 的更多信息,请阅读原文。

24.RE-Bench: RE-Bench 的数据取自 2025 年 2 月 RE-Bench paper。要了解有关 RE-Bench 的更多信息,请阅读原文。

25.RLBench: 关于 RLBench 的数据来自 2025 年 2 月的 RLBench Papers With Code leaderboard。要了解有关 RLBench 的更多信息,请阅读原文。

关 RLBench 的更多信息,请阅读原文。

26. 标尺有关 Ruler 的数据取自 2025 年 2 月的 Ruler 资源库。要了解有关 Ruler 的更多信息,请阅读原文。

27.SWE-bench: 有关 SWE-bench 的数据取自 2025 年 2 月的 SWE-benchleaderboard。要了解有关 SWE-bench 的更多信息,请阅读原文。

28.VAB: VisualAgentBench (VAB) 上的数据取自 2025 年 2 月的 VAB leaderboard。要了解有关 VAB 的更多信息,请阅读原文。

29.VCR: 有关 VCR 的数据取自 2025 年 2 月 VCRleaderboard。要了解有关 VCR 的更多信息,请阅读原文。

30.WildBench: 有关 WildBench 的数据取自 2025 年 2 月 WildBench leaderboard 要了解有关 WildBench 的更多信息,请阅读原文。

附录

第二章：技术性能

引用作品

Akter, S. N., Yu, Z., Muhamed, A., Ou, T., Bäuerle, A., Cabrera, Á. A., Dholakia, K., Xiong, C., & Neubig, G. (2023). *An In-Depth Look at Gemini's Language Abilities* (arXiv:2312.11444). arXiv. <https://doi.org/10.48550/arXiv.2312.11444>

Bairi, R., Sonwane, A., Kanade, A., C, V. D., Iyer, A., Parthasarathy, S., Rajamani, S., Ashok, B., & Shet, S. (2023). *CodePlan: Repository-Level Coding Using LLMs and Planning* (arXiv:2309.12499). arXiv. <https://doi.org/10.48550/arXiv.2309.12499>

Bauza, M., Chen, J. E., Dalibard, V., Gileadi, N., Hafner, R., Martins, M. F., Moore, J., Pevciciute, R., Laurens, A., Rao, D., Zambelli, M., Riedmiller, M., Scholz, J., Bousmalis, K., Nori, F., & Heess, N. (2024). *DemoStart: Demonstration-Led Auto-Curriculum Applied to Sim-to-Real With Multi-fingered Robots* (arXiv:2409.06613). arXiv. <https://doi.org/10.48550/arXiv.2409.06613>

Bommasani, R., Kapoor, S., Klyman, K., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2024). “Considerations for Governing Open Foundation Models.” *Science*, 386(6718), 151–53. <https://doi.org/10.1126/science.adp1848>

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., ... & Zitkovich, B. (2023). *RT-2: Vision-Language Action Models Transfer Web Knowledge to Robotic Control.* (arXiv:2307.15818). arXiv. <https://arxiv.org/abs/2307.15818>

Budagam, D., Kumar, A., Khoshnoodi, M., KJ, S., Jain, V., & Chadha, A. (2024). *Hierarchical Prompting Taxonomy: A Universal Evaluation Framework for Large Language Models Aligned With Human Cognitive Principles* (arXiv:2406.12644; Version 4). arXiv. <https://doi.org/10.48550/arXiv.2406.12644>

Cao, Z., Long, M., Wang, J., & Yu, P. S. (2017). *HashNet: Deep Learning to Hash by Continuation* (arXiv:1702.00758). arXiv. <https://doi.org/10.48550/arXiv.1702.00758>

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). *Evaluating Large Language Models Trained on Code* (arXiv:2107.03374). arXiv. <https://doi.org/10.48550/arXiv.2107.03374>

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference* (arXiv:2403.04132). arXiv. <https://doi.org/10.48550/arXiv.2403.04132>

Chollet, F., Knoop, M., Kamradt, G., & Landers, B. (2025). *ARC Prize 2024: Technical Report* (arXiv:2412.04604). arXiv. <https://doi.org/10.48550/arXiv.2412.04604>

Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). “Lip Reading Sentences in the Wild.” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3444–53. <https://doi.org/10.1109/CVPR.2017.367>

附录

第二章：技术性能

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). *Training Verifiers to Solve Math Word Problems* (arXiv:2110.14168). arXiv. <https://doi.org/10.48550/arXiv.2110.14168>

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., ... Florence, P. (2023). *PaLM-E: An Embodied Multimodal Language Model* (arXiv:2303.03378). arXiv. <https://doi.org/10.48550/arXiv.2303.03378>

Fang, H., Grotz, M., Pumacay, W., Wang, Y. R., Fox, D., Krishna, R., & Duan, J. (2025). *SAM2Act: Integrating Visual Foundation Model With a Memory Architecture for Robotic Manipulation* (arXiv:2501.18564). arXiv. <https://doi.org/10.48550/arXiv.2501.18564>

Fattorini, L., Maslej, N., Perrault, R., Parli, V., Etchemendy, J., Shoham, Y., & Ligett, K. (2024). *The Global AI Vibrancy Tool* (arXiv:2412.04486). arXiv. <https://doi.org/10.48550/arXiv.2412.04486>

Glazer, E., Erdil, E., Besiroglu, T., Chicharro, D., Chen, E., Gunning, A., Olsson, C. F., Denain, J.-S., Ho, A., Santos, E. de O., Järvinen, O., Barnett, M., Sandler, R., Vrzala, M., Sevilla, J., Ren, Q., Pratt, E., Levine, L., Barkley, G., ... Wildon, M. (2024). *Frontier-Math: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI* (arXiv:2411.04872). arXiv. <https://doi.org/10.48550/arXiv.2411.04872>

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring Massive Multitask Language Understanding* (arXiv:2009.03300). arXiv. <https://doi.org/10.48550/arXiv.2009.03300>

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). *Measuring Mathematical Problem Solving With the MATH Dataset* (arXiv:2103.03874). arXiv. <https://doi.org/10.48550/arXiv.2103.03874>

Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Zhang, Y., & Ginsburg, B. (2024). *RULER: What's the Real Context Size of Your Long-Context Language Models?* (arXiv:2404.06654). arXiv. <https://doi.org/10.48550/arXiv.2404.06654>

Huang, Q., Vora, J., Liang, P., & Leskovec, J. (2024). *MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation* (arXiv:2310.03302). arXiv. <https://doi.org/10.48550/arXiv.2310.03302>

Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., & Vidgen, B. (2023). *FinanceBench: A New Benchmark for Financial Question Answering* (arXiv:2311.11944). arXiv. <https://doi.org/10.48550/arXiv.2311.11944>

James, S., Ma, Z., Arrojo, D. R., & Davison, A. J. (2019). *RLBench: The Robot Learning Benchmark & Learning Environment* (arXiv:1909.12271; Version 1). arXiv. <https://doi.org/10.48550/arXiv.1909.12271>

Jia, X., Yang, Z., Li, Q., Zhang, Z., & Yan, J. (2024). *Bench2Drive: Towards Multi-ability Benchmarking of Closed-Loop End-to-End Autonomous Driving* (arXiv:2406.03877). arXiv. <https://doi.org/10.48550/arXiv.2406.03877>

附录

第二章：技术性能

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2024). *SWE-bench: Can Language Models Resolve Real-World GitHub Issues?* (arXiv:2310.06770). arXiv. <https://doi.org/10.48550/arXiv.2310.06770>

Jones, C. R., & Bergen, B. K. (2024). *People Cannot Distinguish GPT-4 From a Human in a Turing Test* (arXiv:2405.08007). arXiv. <https://doi.org/10.48550/arXiv.2405.08007>

Karnchanachari, N., Geromichalos, D., Tan, K. S., Li, N., Eriksen, C., Yaghoubi, S., Meh-dipour, N., Bernasconi, G., Fong, W. K., Guo, Y., & Caesar, H. (2024). *Towards Learning-Based Planning: The nuPlan Benchmark for Real-World Autonomous Driving* (arXiv:2403.04133). arXiv. <https://doi.org/10.48550/arXiv.2403.04133>

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., & Farhadi, A. (2024). *Matryoshka Representation Learning* (arXiv:2205.13147). arXiv. <https://doi.org/10.48550/arXiv.2205.13147>

Leal, I., Choromanski, K., Jain, D., Dubey, A., Varley, J., Ryoo, M., Lu, Y., Liu, F., Sindhwani, V., Vuong, Q., Sarlos, T., Oslund, K., Hausman, K., & Rao, K. (2023). *SARA-RT: Scaling Up Robotics Transformers With Self-Adaptive Robust Attention* (arXiv:2312.01990). arXiv. <https://doi.org/10.48550/arXiv.2312.01990>

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., & Qiao, Y. (2024). *MVBench: A Comprehensive Multi-modal Video Understanding Benchmark* (arXiv:2311.17005). arXiv. <https://doi.org/10.48550/arXiv.2311.17005>

Li, T., Chiang, W.-L., Frick, E., Dunlap, L., Wu, T., Zhu, B., Gonzalez, J. E., & Stoica, I. (2024). *From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline* (arXiv:2406.11939). arXiv. <https://doi.org/10.48550/arXiv.2406.11939>

Li, X., Mata, C., Park, J., Kahatapitiya, K., Jang, Y. S., Shang, J., Ranasinghe, K., Burgert, R., Cai, M., Lee, Y. J., & Ryoo, M. S. (2025). *LLaRA: Supercharging Robot Learning Data for Vision-Language Policy* (arXiv:2406.20095). arXiv. <https://doi.org/10.48550/arXiv.2406.20095>

Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., ... Tang, J. (2023). *AgentBench: Evaluating LLMs as Agents* (arXiv:2308.03688). arXiv. <https://doi.org/10.48550/arXiv.2308.03688>

Liu, X., Zhang, T., Gu, Y., long, I. L., Xu, Y., Song, X., Zhang, S., Lai, H., Liu, X., Zhao, H., Sun, J., Yang, X., Yang, Y., Qi, Z., Yao, S., Sun, X., Cheng, S., Zheng, Q., Yu, H., ... Tang, J. (2024). *VisualAgentBench: Towards Large Multimodal Models as Visual Foundation Agents* (arXiv:2408.06327). arXiv. <https://doi.org/10.48550/arXiv.2408.06327>

Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., & Scialom, T. (2023). *GAIA: A Benchmark for General AI Assistants* (arXiv:2311.12983). arXiv. <https://doi.org/10.48550/arXiv.2311.12983>

Mitchell, M. (2024). “The Turing Test and Our Shifting Conceptions of Intelligence.” *Science*, 385(6710), eadq9356. <https://www.science.org/doi/10.1126/science.adq9356>

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). *MTEB: Massive Text Embedding Benchmark* (arXiv:2210.07316). arXiv. <https://doi.org/10.48550/arXiv.2210.07316>

附录

第二章：技术性能

Ni, J., Xue, F., Yue, X., Deng, Y., Shah, M., Jain, K., Neubig, G., & You, Y. (2024). *MixEval: Deriving Wisdom of the Crowd From LLM Benchmark Mixtures* (arXiv:2406.06565). arXiv. <https://doi.org/10.48550/arXiv.2406.06565>

O'Neill, A., Rehman, A., Gupta, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., Tung, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Gupta, A., ... Lin, Z. (2024). *Open X-Embodiment: Robotic Learning Datasets and RT-X Models* (arXiv:2310.08864). arXiv. <https://doi.org/10.48550/arXiv.2310.08864>

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Ren, R., Hausenloy, J., Zhang, O., Mazeika, M., ... Hendrycks, D. (2025). *Humanity's Last Exam* (arXiv:2501.14249). arXiv. <https://doi.org/10.48550/arXiv.2501.14249>

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., & Bowman, S. R. (2023). *GPQA: A Graduate-Level Google-Proof Q&A Benchmark* (arXiv:2311.12022). arXiv. <https://doi.org/10.48550/arXiv.2311.12022>

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., & Kochenderfer, M. J. (2024). *BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices* (arXiv:2411.12990). arXiv. <https://doi.org/10.48550/arXiv.2411.12990>

Turing, A. M. (2009). Computing Machinery and Intelligence. In Epstein, R., Roberts, G., & Beber, G., eds., *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer* (23–65). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5_3

Valmeeckam, K., Stechly, K., & Kambhampati, S. (2024). *LLMs Still Can't Plan; Can LRM? A Preliminary Evaluation of OpenAI's o1 on PlanBench* (arXiv:2409.13373). arXiv. <https://doi.org/10.48550/arXiv.2409.13373>

Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley, T., Chan, L., Chen, M., Clymer, J., Dhyani, J., Ericheva, E., Garcia, K., Goodrich, B., Jurkovic, N., Kinniment, M., Lajko, A., Nix, S., Sato, L., Saunders, W., ... Barnes, E. (2024). *RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents Against Human Experts* (arXiv:2411.15114). arXiv. <https://doi.org/10.48550/arXiv.2411.15114>

Xia, Z., Li, J., Lin, Z., Wang, X., Wang, Y., & Yang, M.-H. (2024). *OpenAD: Open-World Autonomous Driving Benchmark for 3D Object Detection* (arXiv:2411.17761). arXiv. <https://doi.org/10.48550/arXiv.2411.17761>

Xu, C., Guan, S., Greene, D., & Kechadi, M.-T. (2024). *Benchmark Data Contamination of Large Language Models: A Survey* (arXiv:2406.04244). arXiv. <https://doi.org/10.48550/arXiv.2406.04244>

Yang, X., Sun, K., Xin, H., Sun, Y., Bhalla, N., Chen, X., Choudhary, S., Gui, R. D., Jiang, Z. W., Jiang, Z., Kong, L., Moran, B., Wang, J., Xu, Y. E., Yan, A., Yang, C., Yuan, E., Zha, H., Tang, N., ... Dong, X. L. (2024). *CRAG—Comprehensive RAG Benchmark* (arXiv:2406.04744). arXiv. <https://doi.org/10.48550/arXiv.2406.04744>

Yen, H., Gao, T., Hou, M., Ding, K., Fleischer, D., Izsak, P., Wasserblat, M., & Chen, D. (2025). *HELMET: How to Evaluate Long-Context Language Models Effectively and Thoroughly* (arXiv:2410.02694). arXiv. <https://doi.org/10.48550/arXiv.2410.02694>

附录

第二章：技术性能

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., ... Chen, W. (2024). *MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI* (arXiv:2311.16502). arXiv. <https://doi.org/10.48550/arXiv.2311.16502>

Zellers, R., Bisk, Y., Far-hadi, A., & Choi, Y. (2019). *From Recognition to Cognition: Visual Commonsense Reasoning* (arXiv:1811.10830). arXiv. <https://doi.org/10.48550/arXiv.1811.10830>

Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., Zhao, T., Raja, P., Zhuang, C., Slack, D., Lyu, Q., Hendryx, S., Kaplan, R., Lunati, M., & Yue, S. (2024). *A Careful Examination of Large Language Model Performance on Grade School Arithmetic* (arXiv:2405.00332). arXiv. <https://doi.org/10.48550/arXiv.2405.00332>

附录

第三章：负责任的人工智能

第三章：负责任的人工智能

致谢

人工智能指数谨此致谢Andrew Shi在负责任的人工智能相关会议报告分析方面所做的工作。人工智能指数承认“全球负责任的人工智能状况”分析是与埃森哲合作进行的。人工智能指数特别强调了埃森哲Chief Responsible AI Officer，Arnab Chakraborty，以及埃森哲研究团队（包括Patrick Connolly、Jakub Wiatrak、Dikshita Venkatesh和Shekhar Tewari）在数据收集和分析方面做出的贡献。人工智能指数谨此致谢麦肯锡团队（特别是Medha Bankhwal、Emily Capstick、Katherine Ottenbreit、Brittany Presten、Roger Roberts和Cayla Volandes）在负责任的人工智能生态系统调查中的合作。

会议材料分析

关于负责任的人工智能相关会议论文的分析，人工智能指数研究了以下会议中负责任的人工智能相关的数量：AAAI, AIES, FAccT, ICML, ICLR, 和 NeurIPS。具体来说，团队从会议网站或会议投稿库中搜索包含相关关键字的论文，这些关键字表明这些论文可能属于特定的负责任的人工智能类别。然后，由一个人工团队对论文进行人工验证，以确认其分类。一篇论文有可能属于多个负责任的人工智能类别。

搜索的关键词包括

公平与偏见：algorithmic fairness, bias detection, bias mitigation, discrimination, equity in AI, ethical algorithm design, fair data practices, fair ML, fairness and bias, group fairness, individual fairness, justice, nondiscrimination, representational fairness, unfair, unfairness.

隐私和数据管理：anonymity, confidentiality, data breach, data ethics, data governance, data integrity, data privacy, data protection, data transparency, differential privacy, inferenceprivacy, machine unlearning, privacy by design, privacy-preserving, secure data storage, trustworthy data curation.

安全：adversarial attack, adversarial learning, AI incident, attacks, audits, cybersecurity, ethical hacking, forensic analysis, fraud detection, red teaming, safety, security, security ethics, threat detection, vulnerability assessment.

透明度和可解释性：algorithmic transparency, audit, auditing, causal reasoning, causality, explainability, explainable AI, explainable models, human-understandable decisions, interpretability, interpretable models, model explainability, outcome explanation, transparency, xAI.

埃森哲全球负责任的人工智能状况调查

斯坦福大学的研究人员与埃森哲公司合作进行第二次“全球负责任的人工智能状况”调查。调查收集了来自 20 个国家和 19 个行业的 1,500 家企业的回复，每家企业的总收入至少为 5 亿美元。调查于 2025 年 1 月至 2 月进行。负责任的人工智能全球状况调查的目的是了解采用 RAI 原则和实践所面临的挑战，并对组织和运营 RAI 活动的 10 个方面进行长期比较。

调查共涉及 10 个 RAI 维度：可靠性、隐私和数据管理、公平和非歧视、透明度和可解释性、人际互动、社会和环境福祉、问责制、领导力/原则/文化、合法性和合规性以及组织管理。有关该方法的详细信息，请访问[这里](#)。

附录

第三章：负责任的人工智能

麦肯锡负责任的人工智能调查

麦肯锡公司最近对 38 个国家的 750 多位领导者进行了调查，深入了解了企业 RAI 的现状。这些领导者代表了从技术到医疗保健等各行各业，包括法律、数据 / 人工智能、工程、风险和财务领域的专业人士。麦肯锡 RAI 成熟度模型是一个负责任的人工智能框架，包含 RAI 的四个维度——战略、风险管理、数据和技术以及运营模式——以及 21 个子维度。RAI 成熟度分为四个等级，从开发基础 RAI 实践到全面、积极的计划。

附录

第三章：负责任的人工智能

引用作品

Alanazi, S., & Asif, S. (2024). “Exploring Deepfake Technology: Creation, Consequences and Countermeasures.” *Human–Intelligent Systems Integration*, 6(1), 49–60. <https://doi.org/10.1007/s42454-024-00054-8>

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024). *Measuring Implicit Bias in Explicitly Unbiased Large Language Models* (arXiv:2402.04105). arXiv. <https://doi.org/10.48550/arXiv.2402.04105>

Birhane, A., Dehdashtian, S., Prabhu, V. U., & Boddeti, V. (2024). “The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models.” *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1229–44. <https://doi.org/10.1145/3630106.3658968>

Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Xiong, B., Maslej, N., & Liang, P. (2025). The 2024 Foundation Model Transparency Index (arXiv:2407.12929). arXiv. <https://doi.org/10.48550/arXiv.2407.12929>

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Grifn, C., ... Manyika, J. (2024). *The Ethics of Advanced AI Assistants* (arXiv:2404.16244). arXiv. <https://doi.org/10.48550/arXiv.2404.16244>

Germani, F., Spitale, G., & Biller-Andorno, N. (2024). *The Dual Nature of AI in Information Dissemination: Ethical Considerations*. *Jmir Ai*, 3, e53505. <https://doi.org/10.2196/53505>

Gu, X., Zheng, X., Pang, T., Du, C., Liu, Q., Wang, Y., Jiang, J., & Lin, M. (2024). *Agent Smith: A Single Image Can Jailbreak One Million Multimodal LLM Agents Exponentially Fast* (arXiv:2402.08567). arXiv. <https://doi.org/10.48550/arXiv.2402.08567>

Lafer, J., & Rehman, A. (2023). “Deepfakes and Harm to Women.” *Journal of Digital Life and Learning*, 3(1), Article 1. <https://doi.org/10.51357/jdll.v3i1.218>

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2023). *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models* (arXiv:2305.11747). arXiv. <https://doi.org/10.48550/arXiv.2305.11747>

Liebowitz, J., ed. (2024). *Regulating Hate Speech Created by Generative AI*. Auerbach Publications. <https://doi.org/10.1201/9781032654829>

Lin, S., Hilton, J., & Evans, O. (2022). *TruthfulQA: Measuring How Models Mimic Human Falsehoods* (arXiv:2109.07958). arXiv. <https://doi.org/10.48550/arXiv.2109.07958>

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighof, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2023). *The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing and Attribution in AI* (arXiv:2310.16787). arXiv. <https://doi.org/10.48550/arXiv.2310.16787>

Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., Klyman, K., Klamm, C., Schoelkopf, H., Singh, N., Cherep, M., Anis, A., Dinh, A., Chitongo, C., Yin, D., ... Pentland, S. (2024). *Consent in Crisis: The Rapid Decline of the AI Data Commons* (arXiv:2407.14933). arXiv. <https://doi.org/10.48550/arXiv.2407.14933>

附录

第三章：负责任的人工智能

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., & Hendrycks, D. (2024a). *HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal* (arXiv:2402.04249). arXiv. <https://doi.org/10.48550/arXiv.2402.04249>

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2022). *BBQ: A Hand-Built Bias Benchmark for Question Answering* (arXiv:2110.08193). arXiv. <https://doi.org/10.48550/arXiv.2110.08193>

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., & Henderson, P. (2024). *Safety Alignment Should Be Made More Than Just a Few Tokens Deep* (arXiv:2406.05946). arXiv. <https://doi.org/10.48550/arXiv.2406.05946>

Reuel, A., Connolly, P., Meimandi, K. J., Tewari, S., Wiatrak, J., Venkatesh, D., & Kochenderfer, M. (2024). *Responsible AI in the Global Context: Maturity Model and Survey* (arXiv:2410.09985). arXiv. <https://doi.org/10.48550/arXiv.2410.09985>

Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2024). *XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models* (arXiv:2308.01263). arXiv. <https://doi.org/10.48550/arXiv.2308.01263>

Ruan, Y., Dong, H., Wang, A., Pitis, S., Zhou, Y., Ba, J., Dubois, Y., Maddison, C. J., & Hashimoto, T. (2024). *Identifying the Risks of LM Agents with an LM-Emulated Sandbox* (arXiv:2309.15817). arXiv. <https://doi.org/10.48550/arXiv.2309.15817>

Sheshadri, A., Ewart, A., Guo, P., Lynch, A., Wu, C., Hebbar, V., Sleight, H., Stickland, A. C., Perez, E., Hadfeld-Menell, D., & Casper, S. (2024). *Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs* (arXiv:2407.15549). arXiv. <https://doi.org/10.48550/arXiv.2407.15549>

Simchon, A., Edwards, M., & Lewandowsky, S. (2024). *The Persuasive Effects of Political Microtargeting in the Age of Generative Artificial Intelligence*. PNAS Nexus, 3(2), pgae035. <https://doi.org/10.1093/pnasnexus/pgae035>

Spivak, R. (2018). “Deepfakes”: The Newest Way to Commit One of the Oldest Crimes. *Georgetown Law Technology Review*, 3, 339. <https://georgetownlawtechreview.org/wp-content/uploads/2019/05/3.1-Spivak-pp-339-400.pdf>

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1), 2056305120903408. <https://doi.org/10.1177/2056305120903408>

Vidgen, B., Scherrer, N., Kirk, H. R., Qian, R., Kannappan, A., Hale, S. A., & Röttger, P. (2024). *SimpleSafetyTests: A Test Suite for Identifying Critical Safety Risks in Large Language Models* (arXiv:2311.08370). arXiv. <https://doi.org/10.48550/arXiv.2311.08370>

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., & Fedus, W. (2024). *Measuring Short-Form Factuality in Large Language Models* (arXiv:2411.04368). arXiv. <https://doi.org/10.48550/arXiv.2411.04368>

Zeng, Y., Yang, Y., Zhou, A., Tan, J. Z., Tu, Y., Mai, Y., Klyman, K., Pan, M., Jia, R., Song, D., Liang, P., & Li, B. (2024). *AIR-Bench 2024: A Safety Benchmark Based on Risk Categories From Regulations and Policies* (arXiv:2407.17436). arXiv. <https://doi.org/10.48550/arXiv.2407.17436>

附录

第四章：经济

第四章：经济

国际机器人联合会(IFR)

机器人安装量部分的数据来自 World Robotics 2024 报告。

Lightcast

Vishy Kamalapuram 和 Elena Magrini 编写

Lightcast 提供就业市场分析，使雇主、工人和教育工作者能够做出数据驱动的决策。公司的人工智能技术分析了数以亿计的招聘信息和现实生活中的职业转换，以提供对劳动力市场模式的洞察力。这种实时战略情报提供了重要的洞察力，如哪些工作最有需求、雇主需要的具体技能以及能为工人提供最大潜力的职业方向。欲了解更多信息，请访问 <https://lightcast.io>。

职位发布数据

为了支持这些分析，Lightcast 挖掘了其自 2010 年以来收集的数百万个招聘信息数据集。Lightcast 收集了 51,000 多个在线招聘网站的招聘信息，对劳动力市场需求进行全面、实时的描绘。它汇总招聘信息，删除重复信息，并从招聘信息文本中提取数据。其中包括职位名称、雇主、行业和地区信息，以及所需的经验、教育和技能。

招聘启事有助于了解劳动力市场的趋势，因为通过招聘启事可以详细、实时地了解雇主所需的技能。为了评估职位发布数据的代表性，Lightcast 进行了一系列分析，将美国职位发布的分布情况与政府官方数据和其他第三方的分布情况进行比较。美国职位发布的主要政府

数据来源是劳工统计局开展的职位空缺和劳动力流动调查 (JOLTS) 项目。根据 JOLTS 和 Lightcast 之间的比较，Lightcast 数据捕捉到的劳动力市场需求占劳动力的 99% 以上。未在网上发布的职位通常出现在小企业（如餐馆橱窗上的 "Help Wanted" 招牌）和工会招聘大厅。

衡量对人工智能的需求

为了衡量雇主对人工智能技能的需求，Lightcast 使用其包含 33,000 多种技能的技能分类法。¹ 以下是 Lightcast 提供的人工智能技能列表，以及相关的技能集群。在本报告中，以下所有技能均被视为人工智能技能。如果招聘信息中提到这些技能中的任何一项，则被视为人工智能职位。

人工智能伦理、治理和监管：ethical AI, data sovereignty, AI security, artificial intelligence risk.

人工智能：agentic systems, AI/ML inference, AIOps (artificial intelligence for IT operations), AI personalization, AI testing, applications of artificial intelligence, artificial general intelligence, artificial intelligence, artificial intelligence development, Artificial Intelligence Markup Language (AIML), artificial intelligence systems, automated data cleaning, Azure Cognitive Services, Baidu, cognitive automation, cognitive computing, computational intelligence, Cortana, Data Version Control (DVC), Edge Intelligence, embedded AI, expert systems, explainable AI (XAI), intelligent control, intelligent systems, interactive kiosk, IPSoft Amelia, knowledge distillation, knowledge engineering, knowledge-based configuration, knowledge-based systems, knowledge representation, multi-agent systems, neuro-symbolic AI,

1. <https://lightcast.io/open-skills>

附录

第四章：经济

Open Neural Network Exchange (ONNX), OpenAI Gym, operationalizing AI, PineCone, Qdrant, reasoning systems, swarm intelligence, synthetic data generation, Watson Conversation, Watson Studio, Weka Weaviate.

自动驾驶： advanced driver-assistance systems, autonomous cruise control systems, autonomous system, autonomous vehicles, dynamic routing, guidance navigation and control systems, light detection and ranging (LiDAR), object tracking, OpenCV, path analysis, path finding, remote sensing, scene understanding, unmanned aerial systems (UAS)..

生成式人工智能： Adobe Sensei, ChatGPT, CrewAI, DALL-E image generator, generative adversarial networks, generative AI agents, generative artificial intelligence, Google Bard, image inpainting, image super-resolution, LangGraph, large language modeling, Microsoft Copilot, multimodal learning, multimodal models, prompt engineering, retrieval-augmented generation, Stable Diffusion, text summarization, text to speech (TTS), variational autoencoders (VAEs).

机器学习： AdaBoost (adaptive boosting), adversarial machine learning, Apache MADlib, Apache Mahout, Apache SINGA, Apache Spark, association rule learning, attention mechanisms, AutoGen, automated machine learning, autonomic computing, AWS SageMaker, Azure Machine Learning, bagging techniques, Bayesian belief networks, Boltzmann Machine, boosting, Chi-Squared Automatic Interaction Detection (CHAID), Classification and Regression Tree (CART), cluster analysis, collaborative filtering, concept drift detection, confusion matrix, cyber-physical systems, Dask (Software), data classification, Dbscan, decision models, decision-tree learning, dimensionality reduction, distributed machine learning, Dlib (C++ library), embedded intelligence, ensemble methods, evolutionary programming, expectation maximization algorithm, feature engineering, feature extraction, feature learning, feature selection, federated learning, game AI, Gaussian process, genetic algorithm, Google AutoML, Google Cloud ML Engine, gradient boosting, gradient boosting machines (GBM), H2O.ai,

ai, hidden Markov model, hyperparameter optimization, incremental learning, inference engine, k-means clustering, kernel methods, Kubeflow, LIBSVM, loss functions, machine learning, machine learning algorithms, machine learning methods, machine learning model monitoring and evaluation, machine learning model training, Markov chain, matrix factorization, meta learning, Microsoft Cognitive Toolkit (CNTK), MLflow, MLOps (machine learning operations), mlpack (C++ library), ModelOps, Naive Bayes Classifier, neural architecture compression, neural architecture search (NAS), objective function, Oracle Autonomous Database, Perceptron, Predictionio, predictive modeling, programmatic media buying, Pydata, PyTorch (machine learning library), PyTorch Lightning, Random Forest Algorithm, recommender systems, reinforcement learning, Scikit-Learn (Python package), semi-supervised learning, soft computing, sorting algorithm, supervised learning, support vector machines (SVM), t-SNE (t-distributed Stochastic Neighbor Embedding), test datasets, topological data analysis (TDA), Torch (machine learning), training datasets, transfer learning, transformer (machine learning model), unsupervised learning, Vowpal Wabbit, Xgboost, Theano (software).

自然语言处理： AI copywriting, Amazon Alexa, Amazon Textract, ANTLR, Apache OpenNLP, BERT (NLP Model), chatbot, computational linguistics, conversational AI, DeepSpeech, dialog systems, fastText, fuzzy logic, handwriting recognition, Hugging Face (NLP framework), Hugging Face Transformers, intelligent agent, intelligent virtual assistant, Kaldi, language model, latent Dirichlet allocation, Lexalytics, machine translation, Microsoft LUIS, natural language generation (NLG), natural language processing (NLP), natural language programming, natural language toolkits, natural language understanding (NLU), natural language user interface, nearest neighbour algorithm, Nuance Mix, optical character recognition (OCR), screen reader, semantic analysis, semantic interpretation for speech recognition, semantic kernel, semantic parsing, semantic search, sentence transformers, sentiment analysis, Seq2Seq, Shogun, small language model, speech recognition, speech recognition software, speech synthesis, statistical language

附录

第四章：经济

statistical language acquisition, summarization methods, text mining, text retrieval systems, text to speech (TTS), tokenization, Vespa, voice assistant technology, voice interaction, voice user interface, word embedding, Word2Vec models.

神经网络：Apache MXNet, artificial neural networks, autoencoders, Cafe (framework), Cafe2, Chainer (Deep Learning Framework), convolutional neural networks (CNN), Cudnn, deep learning, deep learning methods, Deeplearning4j, deep reinforcement learning (DRL), evolutionary acquisition of neural topologies, Fast. AI, graph neural networks (GNNs), Keras (neural network library), Long Short-Term Memory (LSTM), neural ordinary differential equations, OpenVINO, PaddlePaddle, Pybrain, recurrent neural network (RNN), reinforcement learning (RL), residual networks (ResNet), sequence-to-sequence models (seq2seq), spiking neural networks, TensorFlow.

机器人：advanced robotics, bot framework, cognitive robotics, meta-reinforcement learning, motion planning, Nvidia Jetson, OpenAI Gym environments, reinforcement learning from human feedback (RLHF), robot framework, robot operating systems, robotic automation software, robotic liquid handling systems, robotic programming, robotic systems, servomotor, SLAM algorithms (Simultaneous Localization and Mapping).

视觉图像识别：3D reconstruction, activity recognition, computer vision, contextual image classification, Deck.gl, digital image processing, digital twin technology, eye tracking, face detection, facial recognition, general-purpose computing on graphics processing units, gesture recognition, image analysis, image captioning, image matching, image recognition, image segmentation, image sensor, ImageNet, instance segmentation, machine vision, MNIST, motion analysis, object recognition, OmniPage, pose estimation, RealSense, thermal imaging analysis.

领英

Rosie Hood、Akash Kaura 和 Mar Carpanelli 编写

领英 数据

这组作品代表了通过领英数据所看到的世界，这些数据来自领英全球超过 10 亿会员的匿名和汇总的个人资料信息。因此，它受到会员选择使用平台方式的影响，而会员选择使用平台的方式可能因职业、社交和地区文化以及网站的整体可用性和可访问性而有所不同。在发布 领英经济图谱的洞察时，领英提供准确的统计数据，同时确保 领英会员的隐私。因此，所有数据显示的都是相应时期汇总信息，遵循严格的数据质量阈值，不会披露任何特定个人的信息。

国家样本

领英 提供了阿根廷、澳大利亚、奥地利、比利时、巴西、加拿大、智利、哥斯达黎加、克罗地亚、塞浦路斯、捷克、丹麦、爱沙尼亚、芬兰、法国、德国、希腊、香港特别行政区、匈牙利、冰岛、印度、印度尼西亚、爱尔兰、以色列、意大利、拉脱维亚、立陶宛、卢森堡、墨西哥、荷兰、新西兰、挪威、波兰、葡萄牙、罗马尼亚、沙特阿拉伯、新加坡、斯洛文尼亚、南非、韩国、西班牙、瑞典、瑞士、土耳其、阿拉伯联合酋长国、英国、美国和乌拉圭的数据。

技能

领英会员在他们的 领英 简介中自我报告他们的技能。目前，领英 已识别出 41,000 多种不同的标准化技能。

领英 将人工智能技能分为两个相互排斥的类别：“人工智能工程”和“人工智能素养”。从广义上讲，人工智能工程技能指的是设计、开发、部署和维护人工智能系统所需的专业技术和实践能力，而人工智能素养技能指的是理解、评估和有效地与人工智能技术互动所需的知识、能力和批判性思维能力。随着技能的不断发展，我们会定期维护和更新这些分类。有关本分析中包含的技能列表，请参阅下面的 领英 人工智能技能列表。

附录

第四章：经济

行业

领英的行业分类法是共享经济活动并对特定产品或服务做出贡献的实体的集合。一个行业代表一家公司提供或销售的产品或服务。领英分析了人工智能背景下的以下行业：教育；金融服务；制造业；专业服务；以及技术、信息和媒体。

性别

领英认识到，一些领英会员的性别认同超越了“男性”和“女性”的传统性别建构。如果没有明确的自我认同，领英会根据领英资料中使用的代词或根据名字来推断本次分析中会员的性别。性别无法被推断为男性或女性的会员，将被排除在任何性别分析之外。请注意，领英过滤了其性别归因算法覆盖范围不足的国家。

人工智能工作或职业

领英会员的头衔是标准化的，分为16,000多种职业。这些职业并不针对特定行业或国家。人工智能工作需要具备人工智能技能才能高效注意力地完成工作。这类职业的例子包括（但不限于）：机器学习工程师、人工智能专家、数据科学家和计算机视觉工程师。

人工智能人才

如果领英会员在个人档案中明确添加了至少两项人工智能技能，并且 / 或者他们现在或曾经受雇于人工智能工作，那么他们就被视为人工智能人才。

2. 增长最快的人工智能技能

所有会员最常添加的人工智能技能的同比增长率。请注意，领英对最近一年的技能添加量设置了阈值，该阈值设定为最近一年按国家分类的人工智能技能添加量分布的第50百分位数。

解读：全球增长最快的人工智能工程技能是定制GPT、人工智能生产力和人工智能智能体。

3. 人工智能人才集聚

人工智能人才计数用于计算人才集中度指标。换句话说，在计算国家级人工智能人才集中度时，领英使用特定国家的人工智能人才数量除以该国的领英会员数量。请注意，集中度指标可能会受到领英在这些国家覆盖范围的影响，因此应谨慎使用。

解读：拥有人工智能工程技能的人工智能人才占美国领英会员的0.78%。

4. 相对人工智能人才聘用率同比比率

领英雇用率是按领英会员人数归一化的雇用衡量标准。它的计算方法是在工作开始的同期新增雇主的领英会员的百分比除以相应地点的领英会员总数。

人工智能聘用率采用总体聘用率方法计算，但只考虑被归类为

附录

第四章：经济

人工智能人才的成员。相对人工智能人才招聘率同比是指人工智能人才招聘率相对于同一国家整体招聘率的同比变化。领英 分享的是 12 个月的移动平均值。

解读：在美国，人工智能人才招聘相对于整体招聘的比例增长了 24.7%。

5. 技能穿透

技能基因组 (SKILLS GENOME)

对于任何类别 (职业、国家、行业等)，技能基因组都是该类别最具特征的 50 种技能的有序列表 (向量)。这些最具特征的技能是通过 TF-IDF 算法确定的，该算法会将那些对特定实体 (如 Microsoft Word) 增加信息很少的无处不在技能向下排序，而将特定实体 (如人工智能) 独有的技能向上排序。更多详情，请参阅 LinkedIn's skills genome 和 LinkedIn—World Bank Methodology 说明。

例如，表 1 详细列出了 2024 年美国技术、信息和媒体行业的技能基因组，显示了按 TF-IDF 排序的前 10 项技能。

技能名称	TF-IDF 技能等级
Amazon Web Services (AWS)	1
Software as a Service (SaaS)	2
Artificial intelligence (AI)	3
Python (programming language)	4
Go-to-market strategy	5
Customer success	6
Large language models (LLM)	7
Salesforce.com	8
SQL	9
Generative AI	10

人工智能技能渗透

这一指标的目的是采用以下方法衡量某一类别中人工智能技能的强度：

- 领英 计算了 2015 年以来 领英会员在特定实体 (职业、行业等) 中自我添加的所有技能的频率。
- 领英 使用 TF-IDF 模型对技能频率进行重新加权，以获得该实体中最具代表性的前 50 种技能。这 50 种技能构成了该实体的 " 技能基因组 "。
- 领英 会计算属于人工智能技能组的技能在所选实体的顶级技能中所占的比例。

解释：人工智能技能渗透率表明人工智能技能在各职业中的普及程度，或者说 领英会员在工作中使用人工智能技能的强度。例如，工程师这一职业的前 50 项技能是根据它们在 领英会员资料中出现的加权频率计算出来。如果工程师所掌握的技能中有四项属于人工智能技能组，那么这一测量结果表明，人工智能技能在工程师中的渗透率估计为 8% (即 4/50)。

相对人工智能技能渗透率

为了对各国的技能渗透率进行比较，需要计算技能基因组，并选择一个相关的比较基准 (如全球平均值)。然后，在控制职业的情况下，构建一个国家与比较基准的人工智能技能渗透率之间的比率。

解释：如果一个国家的人工智能技能相对渗透率为 1.5，这意味着在一组重叠的职业中，人工智能技能的使用频率是比较基准的 1.5 倍。

全球比较

为了进行跨国比较，领英 展示了人工智能技能的相对渗透率，即特定国家中每种人工智能技能在各职业中的渗透率之和，除以全球人工智能在样本国家重叠职业中的平均渗透率。

附录

第四章：经济

解释：相对渗透率为 2 意味着在同一组职业中，该国人工智能技能的平均渗透率是全球平均水平的 2 倍。

全球比较：按行业

按国家分类特定行业的相对人工智能技能渗透率，对各行业和各国的人工智能技能渗透率进行了深入的行业分解。

解释：一个国家教育领域的人工智能技能相对渗透率为 2，意味着在该领域的同一组职业中，该国人工智能技能的平均渗透率是全球平均水平的 2 倍。

全球比较：按性别分类

按性别划分的人工智能技能相对渗透率提供了一个性别内人工智能技能渗透率的跨国家比较。由于每个性别的全球平均值不同，因此该指标只能用于比较每个性别内的国家排名，而不能用于国家内的跨性别比较。

解释：一个国家的女性人工智能技能渗透率为 1.5，这意味着该国女性成员掌握人工智能技能的可能性是所有国家女性成员平均水平的 1.5 倍。

全球比较：跨性别

由于 领英 将一个国家的人工智能技能渗透率按性别与相同的全球平均水平进行比较，而不分性别，因此性别间的相对人工智能技能渗透率可以在全球各国内部和各国之间跨性别比较。

6.人工智能领域的女性代表

这是指女性在人工智能人才中所占的比例。

解释：在全球范围内，拥有人工智能工程技能的人工智能人才中，女性所占比例为 30.5%。

7.人工智能人才迁移

移民数据来自 World Bank Group— LinkedIn “数字数据促进发展 (Digital Data for Development) ” 合作项目 (见 <https://linkedindata.worldbank.org/> 和 Zhu 等 (2018))。领英的迁移率来自领英会员档案中自我认定的地点。例如，当领英会员将其所在地从巴黎更新为伦敦时，这将被算作一次迁移。迁移数据自 2019 年起可用。

领英 数据可帮助各国深入了解因迁移趋势而增减的人工智能人才。人工智能人才迁移是指在时间 “t”，A 国为相关国家，B 国为流入来源国和流出目的国，所有拥有人工智能技能 / 从事人工智能工作的成员。因此，A 国与 B 国之间的人工智能人才净迁徙计算如下：

$$\text{Net AI Talent Migration}_{a,b,t} = \frac{\text{Net AI Talent flows}_{a,b,t}}{\text{Member count}_{a,t}}$$

净流量的定义是在特定时间段内到达人数减去离开人数。不同国家的 领英 会员人数各不相同，这对解释会员从一个国家到另一个国家的绝对流动情况具有挑战性。因此，我们对每个国家的迁徙流量进行了归一化处理。例如，就 A 国而言，所有进出 A 国的绝对净流量 (不考虑原籍国和目的地国) 都根据 A 国每年年底的 领英 会员人数进行了归一化处理，然后乘以 10,000。因此，该指标显示的是从所有国家迁入或迁出 A 国的相对人才流动情况。请注意为了使过渡具有足够的样本量，我们采用了最低阈值。

解释：美国的人工智能人才净流量与其成员规模相比呈正数，每 10,000 名成员中有 1.07 名人工智能人才净流入。

8.进入人工智能工作的职业转型

领英 考虑了为人工智能职业提供养分的来源职业，分析了五年转入人工智能职业的比例。职业转换的计算方法是将成员从一个职

附录

第四章：经济

业到另一个职业的工作转换进行汇总。领英 不包括新毕业生新增的第一份职业和职业内部的转换。

解读：在美国，有 26.9% 的人工智能工程师是从软件工程师转型而来，其次是 13.3% 的数据科学家。

领英人工智能技能列表

人工智能工程

3D reconstruction, AI agents, AI productivity, AI strategy, algorithm analysis, algorithm development, Amazon Bedrock, Apache Spark ML, applied machine learning, artificial intelligence (AI), artificial neural networks, association rules, audio synthesis, autoencoders, automated clustering, automated feature engineering, automated machine learning (AutoML), automated reasoning, autoregressive models, Azure AI Studio, Cafe, chatbot development, chatbots, classification, cognitive computing, computational geometry, computational intelligence, computational linguistics, concept drift adaptation, conditional generation, conditional image generation, convolutional neural networks (CNN), custom GPTs, decision trees, deep convolutional generative adversarial networks (DCGAN), deep convolutional neural networks (DCNN), deep learning, deep neural networks (DNN), evolutionary algorithms, expert systems, facial recognition, feature extraction, feature selection, fuzzy logic, generative adversarial imitation learning, generative adversarial networks (GANs), generative AI, generative design optimization, generative few models, generative modeling, generative neural networks, generative optimization, generative pre-training, generative query networks (GQNs), generative replay memory, generative synthesis, gesture recognition, Google Cloud AutoML, graph embeddings, graph networks, hyperparameter optimization, hyperparameter tuning, image generation, image inpainting, image processing, image synthesis, image-to-image translation, information extraction, intelligent agents, k-means clustering, Keras, knowledge discovery, knowledge representation and reasoning,

LangChain, large language model operations (LLMOps), large language models (LLM), machine learning, machine learning algorithms, machine translation, Microsoft Azure Machine Learning, MLOps, model compression, model interpretation, model training, music generation, nNatural language generation, natural language processing (NLP), natural language understanding, neural network architecture design, neural networks, NLTK, object recognition, ontologies, OpenAI API, OpenCV, parsing, pattern recognition, predictive modeling, probabilistic generative models, probabilistic programming, prompt flow, PyTorch, question answering, random forest, RapidMiner, recommender systems, recurrent neural networks (RNN), reinforcement learning, responsible AI, Scikit-Learn, semantic technologies, semantic web, sentiment analysis, speech recognition, Spring AI, statistical inference, style transfer, StyleGAN, supervised learning, support vector machine (SVM), synthetic data generation, TensorFlow, text analytics, text classification, text generation, text mining, text-to-image generation, Theano, time series forecasting, transformer models, unsupervised learning, variational autoencoders (VAEs), video generation, web mining, Weka, WordNet.

人工智能素养

AI Builder, AI prompting, Anthropic Claude, ChatGPT, DALL-E, generative AI, Generative AI Studio, generative AI tools, generative art, GitHub Copilot, Google Bard, Google Gemini, GPT-3, GPT-4, LLaMA, Microsoft Copilot, Microsoft Copilot Studio, Midjourney, multimodal prompting, prompt engineering, Stable Diffusion.

致谢

领英谨此致谢 Murat Erer 和 Carl Shan 在制定这些方法和指标方面所做的贡献，以及我们在经合组织人工智能组织(OECD.AI)、Stanford Institute for Human-Centered AI 和 Centro Nacional de Inteligencia Artificial, Cenia 的合作者提供的反馈意见。

附录

第四章：经济

Quid

由 Heather English 和 Hansen Yang 撰写的 Quid 见解

Quid 利用自身的内部 大语言模型和其他智能搜索功能以及传统的布尔查询，搜索许多数据集中的重点领域、主题和关键词：社交媒体、新闻、论坛和博客、公司、专利，以及其他自定义（如调查数据）。Quid 拥有多种可视化选项和数据交付终端，包括基于语义相似性的网络图表、平台内仪表盘功能和程序化 PostgreSQL 数据库交付。Quid 应用一流的人工智能和 NLP 来揭示大型数据集中隐藏的模式，使用户能够准确、快速、高效地做出数据驱动型决策。

搜索、数据来源和范围

对来自多个数据源的 800 多万条全球上市和私营公司简介进行索引，以便在公司描述中进行搜索，同时过滤并包含从投资信息到公司地理信息（如成立年份、总部所在地等）的元数据。公司信息每周更新一次。Quid 算法从每份文档中读取大量文本数据，根据不同文档的相似语言在它们之间建立链接。这一过程在巨大的范围内重复进行，从而产生一个由不同集群组成的网络，识别出不同的主题或重点领域。趋势是根据 Quid 识别的关键词、短语、人物、公司和机构以及其他输入软件的元数据确定的。

数据公司

组织数据来自 Capital IQ 和 Crunchbase。这些公司包括全球各种类型的组织（私营、上市、运营、作为子公司运营、停业）。投资数据包括私人投资、并购、公开募股、私募股权投资/风险投资公司持有的少数股权、企业风险部门、政府以及境内外的机构。有些数据

无法获得。例如，当投资者姓名或融资金额未披露时。Quid 默认嵌入了 Capital IQ 的数据，并将来自 Crunchbase 的数据添加到 Capital IQ 未捕获的数据点中。这不仅能获得有关所有全球机构的全面而准确的数据，还能捕捉早期初创企业和融资事件的数据。

搜索参数

布尔查询用于搜索存档公司数据库及其业务描述和网站内的重点领域、主题和关键字。Quid 可根据总部所在地区、投资金额、运营状况、组织类型（私营 / 公共）和成立年份过滤搜索结果。然后，Quid 根据语义相似性对这些公司进行可视化处理。如果搜索结果中有超过 7,000 家公司，Quid 会根据语言算法选择 7,000 家最相关的公司进行可视化。布尔搜索：“人工智能”或“AI”或“机器学习”或“深度学习”。

公司

- 2014 年 1 月 1 日至 2024 年 12 月 31 日期间获得投资（私募、IPO、并购）的全球人工智能和机器学习公司。
- 在过去 10 年（2014 年 1 月 1 日至 2024 年 12 月 31 日）融资超过 150 万美元的全球人工智能和机器学习公司
- 我们还提取了生成式人工智能查询的全球数据（布尔搜索：“生成式人工智能”或“生成式人工智能”或“生成式人工智能”）查询过去 10 年（2014 年 1 月 1 日至 2024 年 12 月 31 日）中获得超过 150 万美元的公司。

目标事件定义

- 私募投资：私募是公司向特定投资者或投资者群体私下出售新发行的证券（股权或债务）。买方在私募中获得的股权通常是少数股权（50% 以下），不过有可能通过私募获得

公司的控制权，在这种情况下，私募就是多数股权投资。

- 少数股权投资：这是指 Quid 的少数股权收购，即买方收购实体、资产产品和业务部门 50% 以下的现有所有权股权。

- 并购：指买方收购实体、资产产品和业务部门 50% 以上的现有所有权。

麦肯锡

企业活动 "部分使用的数据来自麦肯锡的两项全球调查：The State of AI in Early 2024: Gen AI Adoption Spikes and Starts to Generate Value " 和 "The State of AI: How Organizations Are Rewiring to Capture Value ." 。

2024 年的首次在线调查于 2 月 22 日至 3 月 5 日进行，共收到 1363 名参与者的回复，他们代表了不同地区、行业、公司规模、职能专业和任期。在受访者中，981 人表示他们的组织至少在一项业务职能中采用了人工智能，878 人表示他们的组织至少在一项职能中定期使用人工智能技术。

2024 年的第二次在线调查于 7 月 16 日至 7 月 31 日进行，共收到来自 101 个国家的 1491 位参与者的回复，他们代表了不同的地区、行业、公司规模、职能专业和任期。42% 的受访者表示，他们为年收入超过 5 亿美元的企业工作。

为调整答复率差异，数据按每个答复国对全球 GDP 的贡献进行加权。

人工智能指数还考虑了麦肯锡前几次调查的数据。这些数据包括：

The State of AI in 2023: Generative AI's Breakout Year
 The State of AI in 2022—and a Half Decade in Review
 The State of AI in 2021
 The State of AI in 2020
 AI Proves Its Worth, But Few Scale Impact (2019)
 AI Adoption Advances, But Foundational Barriers Remain (2018)

附录

第四章：经济

引用作品

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work. *The Quarterly Journal of Economics*, qjae044. <https://doi.org/10.1093/qje/qjae044>

Cui, Z. (Kevin), Demirer, M., Jafe, S., Musolf, L., Peng, S., & Salz, T. (2025). *The Effects of Generative AI on High-Skilled Work: Evidence From Three Field Experiments With Software Developers* (SSRN Scholarly Paper 4945566). <https://doi.org/10.2139/ssrn.4945566>

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4573321>

Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Ritchie, S., Belonax, T., Troy, K. K., Amodei, D., Kaplan, J., Clark, J., & Ganguli, D. (2025). *Which Economic Tasks Are Performed with AI? Evidence From Millions of Claude Conversations* (arXiv:2503.04761). arXiv. <https://doi.org/10.48550/arXiv.2503.04761>

Hofmann, M., Boysel, S., Nagle, F., Peng, S., & Xu, K. (2024). *Generative AI and the Nature of Work* (No. 11479). CESifo Working Paper. https://www.econstor.eu/bitstream/10419/308375/1/cesifo1_wp11479.pdf

Jafe, S., Shah, N. P., Butler, J., Farach, A., Cambon, A., Hecht, B., Schwarz, M., & Teevan, J. (eEds.). (2024). *Generative AI in Real-World Workplaces: The Second Microsoft Report on AI and Productivity Research*. Microsoft. <https://www.microsoft.com/en-us/research/wp-content/uploads/2024/07/Generative-AI-in-Real-World-Workplaces.pdf>

Necula, S.-C., Fotache, D., & Rieder, E. (2024). Assessing the Impact of Artificial Intelligence Tools on Employee Productivity: Insights From a Comprehensive Survey Analysis. *Electronics*, 13(18), Article 18. <https://doi.org/10.3390/electronics13183758>

Toner-Rodgers, A. (2024). *Artificial Intelligence, Scientific Discovery, and Product Innovation* (arXiv:2412.17866). arXiv. <https://doi.org/10.48550/arXiv.2412.17866>

附录

第五章：科学与医学

第五章：科学与医学

致谢

人工智能指数谨此致谢 Armin Hamrah 在调查与人工智能相关的科学和医学重大趋势方面所做的工作。

比较基准

1.MedQA: 有关 MedQA 的数据来自 2025 年 2 月的 MedQA Papers With Code leaderboard。要了解有关 MedQA 的更多信息, 请阅读原文。

人工智能驱动的蛋白质科学论文统计

人工智能使用 Dimensions 的人工智能文献检索功能来内发表的手稿数量。搜索范围仅限于 2024 出版年和生物科学类 (987,717 论文统计)。然后对每个关键词进行检索, 这些必须同时出现在标题和摘要。这一要求限制返回的手稿数量, 因为这些手稿可能只是顺带提到了关键词, 而不是描述了有关该的研究。确定手稿数量后, 计算每个关键词在生物科学手稿总数中所占的百分比。

图像和多模态人工智能促进科学发现

人工智能指数利用 Semantic Scholar 和谷歌学术 (Google Scholar) 来衡量 2023 年至 2025 年发表的手稿数量。然后对每个关键术语 (如 “foundation models,” “microscopy,” “electron microscopy,” “fluorescence microscopy,” “light microscopy”) 进行搜索, 要求这些术语同时出现在标题和摘要中。此外, 还对搜索进行了改进, 以严格遵守基础模型的定义 -- 具体来说, 就是在大量数据集上训练出来的、可广泛应用于各种用例的模型。为此, 任何被认为是基础模型的模型, 如果在

少于 100 万个数据点上进行过训练, 或者没有在多个任务上进行过评估, 则会被舍弃。

FDA 批准的人工智能医疗设备

FDA 批准的人工智能医疗设备数据来自 FDA 网站, 该网站跟踪支持人工智能和机器学习 (AI/ML) 的医疗设备。

伦理方面的考虑

人工智能指数使用 PubMedCentral 的 API 查询 2020 年 1 月 1 日至 2024 年 12 月 31 日期间发表的英文 – 语言索引文章, 搜索关键词涉及人工智能、医学和伦理问题。为了只获取这三个主题交叉点上的文章, 人工智能指数进一步缩小了文章范围, 只收录那些摘要包含与以下内容相关的关键词的文章: (a) 人工智能; (c) 伦理问题: (a) 人工智能, (b) 医学, c) 至少一个伦理问题。在剔除预印本、被撤回的文章和不符合纳入标准的文章后, 还剩下 2,916 篇文章。人工智能伦理指数利用这批文章摘要中提及伦理问题的频率进行分析。

API 查询:

```
( "artificial intelligence" [MeSH] OR "machine learning" [MeSH]
OR "deep learning" [All Fields] OR "AI" [All Fields] OR "ML" [All
Fields] OR "predictive analytics" [All Fields]) AND (( "ethics"
[MeSH] OR "ethical implications" [All Fields] OR "fair*" [All
Fields] OR "unfair*" [All Fields] OR "bias" [All Fields] OR "ac-
countability" [All Fields] OR "transparency" [All Fields] OR
"explainability" [All Fields] OR "privacy" [All Fields] OR "trust-
worthy AI" [All Fields]) OR ( "bioethics" [MeSH] OR "ELSI" [All
Fields] OR "autonomy" [All Fields] OR "equity" [All Fields] OR
```

附录

第五章：科学与医学

“equitab*” [All Fields] OR “justice” [All Fields] OR “beneficence” [All Fields] OR “non-malefcence” [All Fields] OR “independent review” [All Fields] OR “oversight” [All Fields] OR “racis*” [All Fields] OR “prejud*” [All Fields] OR “inequit*” [All Fields] OR “community engagement” [All Fields] OR “misuse” [All Fields] OR “dual use” [All Fields]) AND (“medicine”[MeSH] OR “medical AI” [All Fields] OR “clinical decision support” [All Fields] OR “health informatics” [All Fields]) AND (“2020/01/01” [PubDate] : “2024/12/31”[PubDate]

搜索日期：2/14/2025

摘要纳入标准：

因此，只包括讨论医学的文章、人工智能，以及摘要中至少一个伦理问题(N= 2,916)。

•**人工智能关键词：**“artifcial intelligence,” “AI,” “algorithm,” “ML,” “machine learning,” “deep learning,” predictive analytics.

•**医学关键词：**“medicine,” “medical,” “health,” “healthcare.”

•**伦理关键词：**“ethic*,” “fairness,” “bias,” “accountability,” “transparency,” “explainability,” “privacy,” “trustworthy AI,” “bioethics,” “ELSI,” “autonomy,” “equit*,” “justice,” “beneficence,” “non- malefcence,” “independent review,” “oversight,” “racism,” “inequit*,” community engagement, misuse, dual use.

附录

第五章：科学与医学

引用作品

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O’Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žogulytė, Arvaniti, E., ... Jumper, J. M. (2024). Accurate Structure Prediction of Biomolecular Interactions With AlphaFold 3. *Nature*, 630(8016), 493–500. <https://doi.org/10.1038/s41586-024-07487-w>

Acharya, R., Abanin, D. A., Aghababaie-Beni, L., Aleiner, I., Andersen, T. I., Ansmann, M., Arute, F., Arya, K., Asfaw, A., Astrakhantsev, N., Atalaya, J., Babbush, R., Bacon, D., Ballard, B., Bardin, J. C., Bausch, J., Bengtsson, A., Bilmes, A., Blackwell, S., ... Google Quantum AI and Collaborators. (2025). Quantum Error Correction Below the Surface Code Threshold. *Nature*, 638(8052), 920–26. <https://doi.org/10.1038/s41586-024-08449-y>

Blankemeier, L., Cohen, J. P., Kumar, A., Veen, D. V., Gardezi, S. J. S., Paschali, M., Chen, Z., Delbrouck, J.-B., Reis, E., Truyts, C., Bluethgen, C., Jensen, M. E. K., Ostmeier, S., Varma, M., Valanarasu, J. M. J., Fang, Z., Huo, Z., Nabulsi, Z., Ardila, D., ... Chaudhari, A. S. (2024). *Merlin: A Vision Language Foundation Model for 3D Computed Tomography* (arXiv:2406.06512). arXiv. <https://doi.org/10.48550/arXiv.2406.06512>

Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Vaughan, A., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J. A., Dong, H., Gupta, J. K., Thambiratnam, K., Archibald, A. T., Wu, C.-C., Heider, E., Welling, M., Turner, R. E., & Perdikaris, P. (2024). *A Foundation Model for the Earth System* (arXiv:2405.13063). arXiv. <https://doi.org/10.48550/arXiv.2405.13063>

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods in Molecular Biology* (Clifton, N.J.), 1607, 627–41. https://doi.org/10.1007/978-1-4939-7000-1_26

Callahan, A., McElfresh, D., Banda, J. M., Bunney, G., Char, D., Chen, J., Corbin, C. K., Dash, D., Downing, N. L., Jain, S. S., Kotecha, N., Masterson, J., Mello, M. M., Morse, K., Nallan, S., Pandya, A., Revri, A., Sharma, A., Sharp, C., ... Shah, N. H. (2024). Standing on FURM Ground: A Framework for Evaluating Fair, Useful, and Reliable AI Models in Health Care Systems. *NEJM Catalyst*, 5(10), CAT.24.0131. <https://doi.org/10.1056/CAT.24.0131>

Campanella, G., Chen, S., Verma, R., Zeng, J., Stock, A., Croken, M., Veremis, B., Elmas, A., Huang, K., Kwan, R., Houldsworth, J., Schoenfeld, A. J., & Vanderbilt, C. (2024). *A Clinical Benchmark of Public Self-Supervised Pathology Foundation Models* (arXiv:2407.06508). arXiv. <https://doi.org/10.48550/arXiv.2407.06508>

Carrillo-Perez, F., Pizurica, M., Zheng, Y., Nandi, T. N., Madduri, R., Shen, J., & Gevaert, O. (2023). RNA-to-Image Multi- cancer Synthesis Using Cascaded Diffusion Models. *bioRxiv: The Preprint Server for Biology*, 2023.01.13.523899. <https://doi.org/10.1101/2023.01.13.523899>

附录

第五章：科学与医学

Chambon, P., Bluethgen, C., Delbrouck, J.-B., Sluijs, R. V. der, Połacin, M., Chaves, J. M. Z., Abraham, T. M., Purohit, S., Langlotz, C. P., & Chaudhari, A. (2022). *RoentGen: Vision-Language Foundation Model for Chest X-ray Generation* (arXiv:2211.12737). arXiv. <https://doi.org/10.48550/arXiv.2211.12737>

Chambon, P., Delbrouck, J.-B., Sounack, T., Huang, S.-C., Chen, Z., Varma, M., Truong, S. Q., Chuong, C. T., & Langlotz, C. P. (2024). *CheXpert Plus: Augmenting a Large Chest X-ray Dataset With Text Radiology Reports, Patient Demographics and Additional Image Formats* (arXiv:2405.19538). arXiv. <https://doi.org/10.48550/arXiv.2405.19538>

Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., & Mahmood, F. (2022). *Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning* (arXiv:2206.02647). arXiv. <https://doi.org/10.48550/arXiv.2206.02647>

Chen, Z., Varma, M., Xu, J., Paschali, M., Veen, D. V., Johnston, A., Youssef, A., Blankemeier, L., Bluethgen, C., Altmayer, S., Valanarasu, J. M. J., Muneer, M. S. E., Reis, E. P., Cohen, J. P., Olsen, C., Abraham, T. M., Tsai, E. B., Beaulieu, C. F., Jitsev, J., ... Langlotz, C. P. (2024). *A Vision-Language Foundation Model to Enhance Efficiency of Chest X-ray Interpretation* (arXiv:2401.12208). arXiv. <https://doi.org/10.48550/arXiv.2401.12208>

Christensen, M., Vukadinovic, M., Yuan, N., & Ouyang, D. (2024). Vision-Language Foundation Model for Echocardiogram Interpretation. *Nature Medicine*, 30(5), 1481–88. <https://doi.org/10.1038/s41591-024-02959-y>

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maftt, D., Pringle, M., Tarbox, L., & Prior, F. (2013). The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6), 1045–57. <https://doi.org/10.1007/s10278-013-9622-7>

Ding, S., Li, J., Wang, J., Ying, S., & Shi, J. (2023). *Multi-scale Efficient Graph-Transformer for Whole Slide Image Classification* (arXiv:2305.15773). arXiv. <https://doi.org/10.48550/arXiv.2305.15773>

Ding, T., Wagner, S. J., Song, A. H., Chen, R. J., Lu, M. Y., Zhang, A., Vaidya, A. J., Jaume, G., Shaban, M., Kim, A., Williamson, D. F. K., Chen, B., Almagro-Perez, C., Doucet, P., Sahai, S., Chen, C., Komura, D., Kawabe, A., Ishikawa, S., ... Mahmood, F. (2024). *Multimodal Whole Slide Foundation Model for Pathology* (arXiv:2411.19666). arXiv. <https://doi.org/10.48550/arXiv.2411.19666>

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Rodman, A., & Chen, J. H. (2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open*, 7(10), e2440969. <https://doi.org/10.1001/jamanetworkopen.2024.40969>

Goh, E., Gallo, R. J., Strong, E., Weng, Y., Kerman, H., Freed, J. A., Cool, J. A., Kanjee, Z., Lane, K. P., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., Olson, A. P. J., Hom, J., Chen, J. H., & Rodman, A. (2025). GPT-4 Assistance for Improvement of Physician Performance on Patient Care Tasks: A Randomized Controlled Trial. *Nature Medicine*, 1–6. <https://doi.org/10.1038/s41591-024-03456-y>

附录

第五章：科学与医学

Gruver, N., Sriram, A., Madotto, A., Wilson, A. G., Zitnick, C. L., & Ulissi, Z. (2024). *Fine-Tuned Language Models Generate Stable Inorganic Materials as Text* (arXiv:2402.04379). arXiv. <https://doi.org/10.48550/arXiv.2402.04379>

Guevara, M., Chen, S., Thomas, S., Chaunzwa, T. L., Franco, I., Kann, B. H., Moningi, S., Qian, J. M., Goldstein, M., Harper, S., Aerts, H. J. W. L., Catalano, P. J., Savova, G. K., Mak, R. H., & Bitterman, D. S. (2024). Large Language Models to Identify Social Determinants of Health in Electronic Health Records. *Npj Digital Medicine*, 7(1), 1–14. <https://doi.org/10.1038/s41746-023-00970-0>

Guo, Z., Zhao, W., Wang, S., & Yu, L. (2023). *HIGT: Hierarchical Interaction Graph-Transformer for Whole Slide Image Analysis* (arXiv:2309.07400). arXiv. <https://doi.org/10.48550/arXiv.2309.07400>

Haberle, T., Cleveland, C., Snow, G. L., Barber, C., Stookey, N., Thornock, C., Younger, L., Mullahkhel, B., & Ize-Ludlow, D. (2024). The Impact of Nuance DAX Ambient Listening AI Documentation: A Cohort Study. *Journal of the American Medical Informatics Association*, 31(4), 975–79. <https://doi.org/10.1093/jamia/ocae022>

Hashmi, A. U. R., Almakky, I., Qazi, M. A., Sanjeev, S., Papineni, V. R., Jagdish, J., & Yaqub, M. (2024). *XReal: Realistic Anatomy and Pathology-Aware X-ray Generation via Controllable Diffusion Model* (arXiv:2403.09240). arXiv. <https://doi.org/10.48550/arXiv.2403.09240>

Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., Verkuil, R., Tran, V. Q., Deaton, J., Wiggert, M., Badkundri, R., Shafkat, I., Gong, J., Derry, A., Molina, R. S., Thomas, N., Khan, Y. A., Mishra, C., Kim, C., ... Rives, A. (2024). *Simulating 500 Million Years of Evolution With a Language Model* (p. 2024.07.01.600583). bioRxiv. <https://doi.org/10.1101/2024.07.01.600583>

Hellert, T., Montenegro, J., & Pollastro, A. (2024). *PhysBERT: A Text Embedding Model for Physics Scientific Literature* (arXiv:2408.09574). arXiv. <https://doi.org/10.48550/arXiv.2408.09574>

Horwick, T., Mao, C., Koynov, A., Yawman, P., Thool, P., Salish, K., Giles, M., Nagapudi, K., & Zhang, S. (2024). In Silico Formulation Optimization and Particle Engineering of Pharmaceutical Products Using a Generative Artificial Intelligence Structure Synthesis Method. *Nature Communications*, 15(1), 9622. <https://doi.org/10.1038/s41467-024-54011-9>

Istasy, P., Lee, W. S., Iansavichene, A., Upshur, R., Gyawali, B., Burkell, J., Sadikovic, B., Lazo-Langner, A., & Chin-Yee, B. (2022). The Impact of Artificial Intelligence on Health Equity in Oncology: Scoping Review. *Journal of Medical Internet Research*, 24(11), e39748. <https://doi.org/10.2196/39748>

Jiang, J. X., Qi, K., Bai, G., & Schulman, K. (2023). Pre-pandemic Assessment: A Decade of Progress in Electronic Health Record Adoption Among U.S. Hospitals. *Health Affairs Scholar*, 1(5), qxad056. <https://doi.org/10.1093/haschl/qxad056>

Jin, D., Pan, E., Oufattolle, N., Weng, W.-H., Fang, H., & Szolovits, P. (2020). *What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset From Medical Exams* (arXiv:2009.13081). arXiv. <https://doi.org/10.48550/arXiv.2009.13081>

附录

第五章：科学与医学

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., & Horng, S. (2019). MIMIC-CXR, a De-identified Publicly Available Database of Chest Radiographs With Free-Text Reports. *Scientific Data*, 6(1), 317. <https://doi.org/10.1038/s41597-019-0322-0>

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klwer, M., Lottes, J., Rasp, S., Dben, P., Hatfeld, S., Battaglia, P., Sanchez-Gonzalez, A., Willson, M., Brenner, M. P., & Hoyer, S. (2024). Neural General Circulation Models for Weather and Climate. *Nature*, 632(8027), 1060–66. <https://doi.org/10.1038/s41586-024-07744-ya>

Kudiabor, H. (2024). Virtual Lab Powered by ‘AI Scientists’ Super-Charges Biomedical Research. *Nature*, 636(8043), 532–33. <https://doi.org/10.1038/d41586-024-01684-3>

Kumar, A., Kriz, A., Havaei, M., & Arbel, T. (2025). PRISM: High-Resolution & Precise Counterfactual Medical Image Generation Using Language-Guided Stable Diffusion (arXiv:2503.00196). arXiv. <https://doi.org/10.48550/arXiv.2503.00196>

Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Zhao, M., Chow, A. K., Ikemura, K., Kim, A., Pouli, D., Patel, A., Soliman, A., Chen, C., Ding, T., Wang, J. J., Gerber, G., Liang, I., Le, L. P., Parwani, A. V., Weishaupt, L. L., & Mahmood, F. (2024). A Multimodal Generative AI Copilot for Human Pathology. *Nature*, 634(8033), 466–73. <https://doi.org/10.1038/s41586-024-07618-3>

Lutsker, G., Sapir, G., Shilo, S., Merino, J., Godneva, A., Greenfeld, J. R., Samocha-Bonet, D., Dhir, R., Gude, F., Mannor, S., Meirom, E., Chechik, G., Rossman, H., & Segal, E. (2025). From Glucose Patterns to Health Outcomes: A Generalizable Foundation Model for Continuous Glucose Monitor Data Analysis (arXiv:2408.11876). arXiv. <https://doi.org/10.48550/arXiv.2408.11876>

Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment Anything in Medical Images. *Nature Communications*, 15(1), 654. <https://doi.org/10.1038/s41467-024-44824-z>

Ma, S. P., Liang, A. S., Shah, S. J., Smith, M., Jeong, Y., Devon-Sand, A., Crowell, T., Delahaie, C., Hsia, C., Lin, S., Shanafelt, T., Pfefer, M. A., Sharp, C., & Garcia, P. (2025). Ambient Artificial Intelligence Scribes: Utilization and Impact on Documentation Time. *Journal of the American Medical Informatics Association*, 32(2), 381–85. <https://doi.org/10.1093/jamia/ocae304>

Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., & Naik, N. (2023). Large Language Models Generate Functional Protein Sequences Across Diverse Families. *Nature Biotechnology*, 41(8), 1099–1106. <https://doi.org/10.1038/s41587-022-01618-2>

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., ... Kopp-Schneider, A. (2018). Why Rankings of Biomedical Image Analysis Competitions Should Be Interpreted With Care. *Nature Communications*, 9(1), 5217. <https://doi.org/10.1038/s41467-018-07619-7>

附录

第五章：科学与医学

Mei, X., Liu, Z., Robson, P. M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K. E., Yang, T., Wang, Y., Greenspan, H., Deyer, T., Fayad, Z. A., & Yang, Y. (2022). RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. *Radiology: Artificial Intelligence*, 4(5), e210315. <https://doi.org/10.1148/ryai.210315>

Narayanan, S., Braza, J. D., Griffiths, R.-R., Ponnappati, M., Bou, A., Laurent, J., Kabeli, O., Wellawatte, G., Cox, S., Rodrigues, S. G., & White, A. D. (2024). *Aviary: Training Language Agents on Challenging Scientific Tasks* (arXiv:2412.21154). arXiv. <https://doi.org/10.48550/arXiv.2412.21154>

Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., Luo, R., McKinney, S. M., Ness, R. O., Poon, H., Qin, T., Usuyama, N., White, C., & Horvitz, E. (2023). *Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine* (arXiv:2311.16452). arXiv. <https://doi.org/10.48550/arXiv.2311.16452>

Nori, H., Usuyama, N., King, N., McKinney, S. M., Fernandes, X., Zhang, S., & Horvitz, E. (2024). *From Medprompt to o1: Exploration of Run-Time Strategies for Medical Challenge Problems and Beyond* (arXiv:2411.03590). arXiv. <https://doi.org/10.48550/arXiv.2411.03590>

Pokharel, S., Pratyush, P., Heinzinger, M., Newman, R. H., & Kc, D. B. (2022). Improving Protein Succinylation Sites Prediction Using Embeddings From Protein Language Model. *Scientific Reports*, 12(1), 16933. <https://doi.org/10.1038/s41598-022-21366-2>

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., & Willson, M. (2025). Probabilistic Weather Forecasting With Machine Learning. *Nature*, 637(8044), 84–90. <https://doi.org/10.1038/s41586-024-08252-9>

Qian, Z., Callender, T., Cebere, B., Janes, S. M., Navani, N., & van der Schaar, M. (2024). Synthetic Data for Privacy-Preserving Clinical Risk Prediction. *Scientific Reports*, 14(1), 25676. <https://doi.org/10.1038/s41598-024-72894-y>

Qiu, J., Wu, J., Wei, H., Shi, P., Zhang, M., Sun, Y., Li, L., Liu, H., Liu, H., Hou, S., Zhao, Y., Shi, X., Xian, J., Qu, X., Zhu, S., Pan, L., Chen, X., Zhang, X., Jiang, S., ... Yuan, W. (2024). Development and Validation of a Multimodal Multitask Vision Foundation Model for Generalist Ophthalmic Artificial Intelligence. *NEJM AI*, 1(12), Aloa2300221. <https://doi.org/10.1056/Aloa2300221>

Quer, G., & Topol, E. J. (2024). *The Potential for Large Language Models to Transform Cardiovascular Medicine*. *The Lancet Digital Health*, 6(10), e767–71. [https://doi.org/10.1016/S2589-7500\(24\)00151-1](https://doi.org/10.1016/S2589-7500(24)00151-1)

Rashidi, H. H., Albahra, S., Rubin, B. P., & Hu, B. (2024). A Novel and Fully Automated Platform for Synthetic Tabular Data Generation and Validation. *Scientific Reports*, 14(1), 23312. <https://doi.org/10.1038/s41598-024-73608-0>

Shah, S. J., Devon-Sand, A., Ma, S. P., Jeong, Y., Crowell, T., Smith, M., Liang, A. S., Delahaie, C., Hsia, C., Shanafelt, T., Pfeifer, M. A., Sharp, C., Lin, S., & Garcia, P. (2025). Ambient Artificial Intelligence Scribes: Physician Burnout and Perspectives on Usability and Documentation Burden. *Journal of the American Medical Informatics Association*, 32(2), 375–80. <https://doi.org/10.1093/jamia/ocae295>

附录

第五章：科学与医学

Shapson-Coe, A., Januszewski, M., Berger, D. R., Pope, A., Wu, Y., Blakely, T., Schalek, R. L., Li, P. H., Wang, S., Maitin-Shepard, J., Karlupia, N., Dorkenwald, S., Sjostedt, E., Leavitt, L., Lee, D., Troidl, J., Collman, F., Bailey, L., Fitzmaurice, A., ... Lichtman, J. W. (2024). A Petavoxel Fragment of Human Cerebral Cortex Reconstructed at Nanoscale Resolution. *Science*, 384(6696), eadk4858. <https://doi.org/10.1126/science.adk4858>

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., & Bakas, S. (2020). Federated Learning in Medicine: Facilitating Multi-institutional Collaborations Without Sharing Patient Data. *Scientific Reports*, 10(1), 12598. <https://doi.org/10.1038/s41598-020-69250-1>

Shi, J., Tang, L., Gao, Z., Li, Y., Wang, C., Gong, T., Li, C., & Fu, H. (2023). MG-Trans: Multi-scale Graph Transformer With Information Bottleneck for Whole Slide Image Classification. *IEEE Transactions on Medical Imaging*, 42(12), 3871–83. <https://doi.org/10.1109/TMI.2023.3313252>

Snel, B., Lehmann, G., Bork, P., & Huynen, M. A. (2000). STRING: A Web-Server to Retrieve and Display the Repeatedly Occurring Neighbourhood of a Gene. *Nucleic Acids Research*, 28(18), 3442–44. <https://doi.org/10.1093/nar/28.18.3442>

Snowdon, J. L., Scheufele, E. L., Pritts, J., Le, P.-T., Mensah, G. A., Zhang, X., & Dankwa-Mullan, I. (2023). Evaluating Social Determinants of Health Variables in Advanced Analytic and Artificial Intelligence Models for Cardiovascular Disease Risk and Outcomes: A Targeted Review. *Ethnicity & Disease*, 33(1), 33–43. <https://doi.org/10.18865/1704>

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large Language Models Could Change the Future of Behavioral Healthcare: A Proposal for Responsible Development and Evaluation. *Npj Mental Health Research*, 3(1), 1–12. <https://doi.org/10.1038/s44184-024-00056-z>

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3), e1001779. <https://doi.org/10.1371/journal.pmed.1001779>

Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Kipnis, P., Liu, V., & Lee, K. (2024). Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. *NEJM Catalyst*, 5(3), CAT.23.0404. <https://doi.org/10.1056/CAT.23.0404>

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Židek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutf, M., ... Velankar, S. (2022). AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space With High-Accuracy Models. *Nucleic Acids Research*, 50(D1), D439–44. <https://doi.org/10.1093/nar/gkab1061>

附录

第五章：科学与医学

Veitch, D. P., Weiner, M. W., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Morris, J. C., Petersen, R. C., Saykin, A. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., & Alzheimer's Disease Neuroimaging Initiative. (2019). Understanding Disease Progression and Improving Alzheimer's Disease Clinical Trials: Recent Highlights From the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 15(1), 106–52. <https://doi.org/10.1016/j.jalz.2018.08.005>

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N., Yang, E., Mathieu, P., van Eck, A., Lee, D., Viret, J., Robert, E., Wang, Y. K., Kunz, J. D., Lee, M. C. H., ... Fuchs, T. J. (2024). A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection. *Nature Medicine*, 30(10), 2924–35. <https://doi.org/10.1038/s41591-024-03141-0>

Wang, R., Fang, X., Lu, Y., & Wang, S. (2004). The PDBbind Database: Collection of Binding Afnities for Protein–Ligand Complexes With Known Three-Dimensional Structures. *Journal of Medicinal Chemistry*, 47(12), 2977–80. <https://doi.org/10.1021/jm030580l>

Wang, X., Liu, S., Tsaris, A., Choi, J.-Y., Aji, A., Fan, M., Zhang, W., Yin, J., Ashfaq, M., Lu, D., & Balaprakash, P. (2024). ORBIT: Oak Ridge Base Foundation Model for Earth System Predictability (arXiv:2404.14712). arXiv. <https://doi.org/10.48550/arXiv.2404.14712>

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., & Han, X. (2022a). Transformer-Based Unsupervised Contrastive Learning for Histopathological Image Classification. *Medical Image Analysis*, 81, 102559. <https://doi.org/10.1016/j.media.2022.102559>

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., & Han, X. (2022b). Transformer-Based Unsupervised Contrastive Learning for Histopathological Image Classification. *Medical Image Analysis*, 81, 102559. <https://doi.org/10.1016/j.media.2022.102559>

Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., Wang, F., Peng, Y., Zhu, J., Zhang, J., Jackson, C. R., Zhang, J., Dillon, D., Lin, N. U., Sholl, L., ... Yu, K.-H. (2024). A Pathology Foundation Model for Cancer Diagnosis and Prognosis Prediction. *Nature*, 634(8035), 970–78. <https://doi.org/10.1038/s41586-024-07894-z>

Wang, Y., He, J., Du, Y., Chen, X., Li, J. C., Liu, L.-P., Xu, X., & Hassoun, S. (2025). Large Language Model Is Secretly a Protein Sequence Optimizer (arXiv:2501.09274). arXiv. <https://doi.org/10.48550/arXiv.2501.09274>

Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., Yu, K.-H., Willens, S., Olguin, F. M., Nirschl, J. J., Neal, J., Diehn, M., Yang, S., & Li, R. (2025). A Vision–Language Foundation Model for Precision Oncology. *Nature*, 638(8051), 769–78. <https://doi.org/10.1038/s41586-024-08378-w>

Xie, Y., Wu, J., Tu, H., Yang, S., Zhao, B., Zong, Y., Jin, Q., Xie, C., & Zhou, Y. (2024). A Preliminary Study of AI in Medicine: Are We Closer to an AI Doctor? (arXiv:2409.15277). arXiv. <https://doi.org/10.48550/arXiv.2409.15277>

附录

第五章：科学与医学

Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., Gonzlez, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., ... Poon, H. (2024). A Whole-Slide Foundation Model for Digital Pathology From Real-World Data. *Nature*, 630(8015), 181–88. <https://doi.org/10.1038/s41586-024-07441-w>

Yang, L., Xu, S., Sellergren, A., Kohlberger, T., Zhou, Y., Ktena, I., Kiraly, A., Ahmed, F., Hormozdiari, F., Jaroensri, T., Wang, E., Wulczyn, E., Jamil, F., Guidroz, T., Lau, C., Qiao, S., Liu, Y., Goel, A., Park, K., ... Golden, D. (2024). *Advancing Multimodal Medical Capabilities of Gemini* (arXiv:2405.03162). arXiv. <https://doi.org/10.48550/arXiv.2405.03162>

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2022). *GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records* (arXiv:2203.03540). arXiv. <https://doi.org/10.48550/arXiv.2203.03540>

Yu, B., Baker, F. N., Chen, Z., Ning, X., & Sun, H. (2024). *LlaSMol: Advancing Large Language Models for Chemistry With a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset* (arXiv:2402.09391). arXiv. <https://doi.org/10.48550/arXiv.2402.09391>

Zambaldi, V., La, D., Chu, A. E., Patani, H., Danson, A. E., Kwan, T. O. C., Frerix, T., Schneider, R. G., Saxton, D., Thillaisundaram, A., Wu, Z., Moraes, I., Lange, O., Papa, E., Stanton, G., Martin, V., Singh, S., Wong, L. H., Bates, R., ... Wang, J. (2024). *De Novo Design of High-Afnity Protein Binders with AlphaProteo* (arXiv:2409.08022). arXiv. <https://doi.org/10.48550/arXiv.2409.08022>

Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H. H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J., Moung-Wen, C., Piening, B., Bifulco, C., Wei, M., Poon, H., & Wang, S. (2025). A Foundation Model for Joint Segmentation, Detection and Recognition of Biomedical Objects Across Nine Modalities. *Nature Methods*, 22(1), 166–76. <https://doi.org/10.1038/s41592-024-02499-w>

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., Liu, T., Xu, M., Lozano, M. G., Woodward-Court, P., Kihara, Y., Altmann, A., Lee, A. Y., Topol, E. J., Denniston, A. K., Alexander, D. C., & Keane, P. A. (2023). A Foundation Model for Generalizable Disease Detection From Retinal Images. *Nature*, 622(7981), 156–63. <https://doi.org/10.1038/s41586-023-06555-x>

附录

第六章：政策

第六章：政策

致谢

人工智能指数谨此感谢 Julia Betts Lotufo 和 Alexandra Rome 在收集人工智能重大政策事件信息方面所做的努力。此外，人工智能指数还要感谢 Lapo Santarasci 领导了对人工智能公共支出和美国赠款相关人工智能支出的分析工作。

全球人工智能提及率

对于世界各地与人工智能相关的立法程序中提及人工智能的内容，人工智能指数在 75 个地理区域的国会或议会网站上，通常在名为“minutes”、“hansard”栏目下，用各自的语言搜索关键词“人工智能”。提及次数按届计算，因此在同一届立法会议上多次“人工智能”算作一次。人工智能指数小组调查了以下数据库：

安道尔、亚美尼亚、澳大利亚、阿塞拜疆、巴巴多斯、比利时、百慕大、巴西、加拿大、开曼群岛、中国¹、捷克共和国、丹麦、多米尼加共和国、厄瓜多尔、萨尔瓦多、爱沙尼亚、斐济、芬兰、法国、德国、直布罗陀、希腊、香港、冰岛、印度、爱尔兰、马恩岛、意大利、日本、肯尼亚、科索沃、拉脱维亚、莱索托、列支敦士登、卢森堡、中国澳门特别行政区、马达加斯加、马来西亚、马尔代夫、马耳他、毛里求斯、墨西哥、摩尔多瓦、荷兰、新西兰、北马里亚纳群岛、挪威、巴基斯坦、巴拿马、巴布亚新几内亚、菲律宾、波兰、葡萄牙、罗马尼亚、俄罗斯、圣马力诺、塞舌尔、塞拉利昂、新加坡、斯洛文尼亚、南非、韩国、西班牙、斯里兰卡、瑞典、瑞士、坦桑尼亚、特立尼达和多巴哥、乌克兰、英国、美国、乌拉圭、赞比亚、津巴布韦

美国委员会提及

为了研究美国委员会提及人工智能的趋势，我们进行了以下搜索

网站：[Congress.gov](https://www.congress.gov)

关键词：人工智能

过滤器：Committee Reports

全球人工智能立法记录

对于已通过成为法律的人工智能相关法案，人工智能指数在 116 个的国会或议会网站上，以各自的语言文字和法案全文检索关键词“人工智能”。需要注意的是，只有 2016 年至 2024 年间由州级立法机构通过并签署成为法律（如由总统签署或获得皇家批准）的法律才包含在内。已获批准但随后被废除的法律不在分析之列。对于在最初颁布后增加或修订了人工智能相关条款的法律，人工智能指数在相关情况下使用纳入年份而非最初通过年份。未来的人工智能指数报告希望包括对其他类型法律文件的分析，如国家或超国家、政府机构等通过的法规和标准。

人工智能指数小组调查了以下地理区域的数据库：

阿尔及利亚、安道尔、安提瓜和巴布达、阿根廷、亚美尼亚、澳大利亚、奥地利、阿塞拜疆、巴哈马、巴林、孟加拉国、巴巴多斯、白俄罗斯、比利时、伯利兹、百慕大、不丹、玻利维亚、巴西、文莱、保加利亚、喀麦隆、加拿大、智利、中国、克罗地亚、古巴、库拉索岛、塞浦路斯、捷克共和国、丹麦、爱沙尼亚、法罗群岛、斐济、芬兰、法国、德国、直布罗陀、希腊、格陵兰、格林纳达、关岛、危地马拉、圭亚那、香港、匈牙利、冰岛、印度、伊拉克、爱尔兰、马恩岛、以色列、爱尔兰、马恩岛、以色列、意大利、牙买加、日本、哈萨克斯坦、肯尼亚、基里巴斯、大韩民国、科索沃、吉尔吉斯共和国、拉脱维亚、列支敦士登、立陶宛、卢森堡、中国澳门特别行政区、马拉维、马来西亚、马耳他、毛里求斯、墨西哥、摩纳哥、黑山、摩洛哥、莫桑比克、瑙鲁、荷兰、新西兰、北马里亚纳群岛、挪威、巴拿马、菲律宾、波兰、葡萄牙、罗马尼亚、俄罗斯、萨摩亚、沙特阿拉伯、塞尔维亚、塞舌尔、塞拉利昂、新加坡、斯洛伐克共和国斯洛文尼亚、南非、西班牙、圣基茨和尼维斯、苏里南、瑞典、瑞士、塔吉克斯坦、坦桑尼亚、多哥、汤加、土耳其、图瓦卢、乌干达、乌克兰、阿拉伯联合酋长国、英国、美国、乌拉圭、越南、也门、赞比亚、津巴布韦。

¹ 全国人民代表大会每年召开一次，不提供完整的立法程序。因此，本分析中的计数是在大会会议发布的唯一公开文件，即总理所作的《政府工作报告》中搜索提及“人工智能”的内容。

附录

第六章：政策

美国州级人工智能立法

对于已通过成为法律的人工智能相关法案，人工智能指数在美国 50 个州的网站上对法案全文中的关键词 " 人工智能 " 进行了搜索。只有当该关键词出现在法案最终版本（而不仅仅是介绍版本）中时，该法案才算通过成为法律。请注意，仅包括 2015 年至 2024 年通过的法律。拟议法律的统计包括已通过的拟议法律和尚未通过或目前处于非活动状态的拟议法律。人工智能指数团队调查了以下数据库：

阿拉巴马州、阿拉斯加州、亚利桑那州、阿肯色州、加利福尼亚州、科罗拉多州、康涅狄格州、特拉华州、佛罗里达州、佐治亚州、夏威夷州、爱达荷州、伊利诺伊州、印第安纳州、爱荷华州、堪萨斯州、肯塔基州、路易斯安那州、缅因州、马里兰州、马萨诸塞州、密歇根州、明尼苏达州、密西西比州、密苏里州、蒙大拿州、内布拉斯加州、内华达州、新罕布什尔州、新泽西州、新墨西哥州、纽约州、北卡罗来纳州、北达科他州、俄亥俄州、俄克拉荷马州、俄勒冈州、宾夕法尼亚州、罗得岛州、南卡罗来纳州、南达科他州、田纳西州、得克萨斯州、犹他州、佛蒙特州、弗吉尼亚州、华盛顿州、西弗吉尼亚州、威斯康星州、怀俄明州。

为了进行更全面的审查，人工智能指数还包括了 Multistate AI state legislation tracker 上列出的与人工智能相关的州法律，即使这些法律没有具体提及 " 人工智能 " 这一关键词。

美国人工智能监管条例

本节探讨了以下机构颁布的与人工智能相关的监管条例。从 2016 年到 2024 年美国监管机构的人工智能指数，分析了监管条例总数及其发起机构。为了编制这些数据，人工智能指数在 Federal Register 上对 " 人工智能 " 进行了关键词搜索，这是一个全面的政府文件库，收录了超过 436 个机构和美国政府几乎所有部门的文件。

美国委员会提及

为了研究美国委员会提及人工智能的趋势，我们进行了以下搜索

网站：Congress.gov

关键词：人工智能

过滤器：Committee Reports

人工智能领域的公共投资

人工智能指数分析了欧洲各国和美国的政府人工智能支出，重点关注数据更容易获取的地区。需要注意的是，由于数据的可获得性和质量可能存在很大差异，因此本分析可能无法完全代表所有国家或地区。此外，虽然本分析包括各国政府合同的数据，但只涉及美国赠款级支出。造成这一差异的原因是，从欧盟和中国等其他国家和地区收集可比赠款数据存在困难。不过，美国的情况说明，政府在人工智能方面的支出有很大一部分是通过拨款实现的。随着更多数据的出现，人工智能指数的覆盖范围将在未来迭代中扩大，但现有数据中的差异和差距可能会影响研究结果的全面性和准确率。

数据来源

对于欧洲国家，人工智能指数指数从 Tenders Electronic Daily (TED) (Publications Ofce of the European Union, 2024)——欧盟官方期刊的在线增刊，专门报道欧洲公共采购——收集公开招标数据。虽然合同有多种格式，但最详细的数据来自批量 XML 下载，其中包括招标程序、发标实体、中标承包商、批量价值、说明、中标日期和通用采购词汇 (CPV) 代码等综合信息。TED 的发布受欧盟法律 thresholds 的制约：超过特定货币价值的投标，如果被认为具有跨国界利益，必须在 TED 上公布。不过，一些国家也报告低于阈值的采购，导致各国的覆盖范围不同。

附录

第六章：政策

英国的数据来源包括 TED , Find a Tender , Contracts Finder , 和 Contracts Finder Archive 。苏格兰和威尔士的数据是通过其采购网站的 API 访问的，而北爱尔兰不提供这种服务，因此必须将其排除在分析之外，并可能导致低估英国在人工智能方面的公共投资。由于 API 限制了历史数据的访问，人工智能指数利用 Open Contracting Partnership 的数据注册表，通过 Kingfisher Collect 获得苏格兰和威尔士的全面数据。

美国的数据来源于可公开访问的 USAspending 平台，这是一个方便批量下载合同授予通知和赠款数据相关信息的官方存储库。虽然该数据集比 TED 数据集涵盖的时间更长，但必须指出的是，数据质量可能存在差异。此外，美国政府问责局 (GAO, 2023))的一项研究发现，49 个机构(包括 25 个行政部门)没有向 USAspending 报告数据，这占 2022 财年超过 50 亿美元的净支出。

数据处理

由于合同说明的存储不一致，根据发布时间和采购类型的不同，XML 标记名称也不尽相同，这给处理 TED 数据带来了巨大挑战。一些文件包含汇总说明，而另一些文件则详细说明了每批授予的合同。为了获取全面的信息，我们将每项竞争要求的主要说明与现有的部分说明合并在一起。州的标准于 2024 年底通过，不包括在本数据集中。

由于来自不同国家的数据具有语言多样性，因此需要使用 deep- translator 工具和谷歌翻译引擎将所有文本翻译成英文。翻译后，使用自然语言处理 (NLP) 技术对投标文本进行处理。这些技术包括删除停顿词和特殊字符、保留关键语法类别语篇 (POS) 标记、小写转换、词法化以及用 <NUM> 标记替换数字量。

为便于比较，所有货币金额均换算成美元，并使用购买力平价 (PPP) 指数对价格水平差异进行调整。

分类

使用正则表达式进行全文检索，对人工智能相关合同和赠款进行分类。通过生成与人工智能相关的表达式，并纳入 Yamashita 等的 (2021) 词汇表中的“核心”表达式，编制了人工智能词典。此外，Word2Vec 模型用每个基线表达式的余弦相似词扩展了词典，这些词经过人工审核后被纳入最终词汇。这一过程提供了对识别人工智能内容至关重要的关键词和共现模式。

分类采用了多步骤方法。首先，通过正则表达式 (regex) 匹配确定合同和中的人工智能术语。然后，这些文件被归类为“非人工智能相关”或“人工智能相关”。为了验证人工智能相关匹配，采用了 BERTopic 模型和预训练的 DeBERTA Transformer 来评估特定人工智能相关主题的概率分数。相关性得分低于 20% 的奖项需要经过人工审核，而得分较高的奖项则被确认为与人工智能相关。为确保更高的准确率，所有高价值的标书也都经过了人工审核。

附录

第七章：教育

第七章：教育

Code.org、CSTA、ECEP 联盟

州级数据

State of Computer Science Education 2024报告附录 2 全面介绍了 Code.org、CSTA 和 ECEP 联盟收集数据的方法。Code.org 的工作人员还维护着一个 数据库，其中收录了美国 K-12阶段阶段教育的状况，并在本policy primer中提供了有关美国各州 K-12阶段阶段教育状况的更多详细信息。

AP 计算机科学数据

AP 计算机科学数据是根据学院委员会与 Code.org 签订的协议提供给 Code.org。AP 计算机科学数据来自大学委员会的 national and state summary reports 。

获得计算机科学教育

有关接受计算机科学教育的数据来自Code.org、CSTA和 ECEP 联盟的State of Computer Science Education 2024报告。

2024 年 K-12 阶段阶段计算机科学领域教师队伍状况调查

欲了解更多信息或访问数据集，请联系membership@csteachers.org。

州标准比较

CSTA 和计算机教育促进协会 (IACE) 于 2024 年 12 月发布了州标准比较 报告。该数据集包含约 10,000 个州采用的 K-12 阶段阶段标准，以 电子表格的形式提供，以及Python notebook ，可能对数据分析有用。科罗拉多州 和 弗吉尼亚州 的标准于 2024 年底通过，不包括在本数据集中。

全球 K-12 阶段阶段人工智能教育

剑桥大学计算机科学与技术系 Raspberry Pi Computing Education Research Centre 在 Brookings Institution 为其 2021 年的报告 Building Skills for Life: How to Expand and Improve Computer Science Education Around the World 所做研究的基础上，编制了本数据集 。我们对他们的数据集做了一处修改，以澄清美国的 CS 课程在一些学校 / 地区开设，而不是在所有地方都作为选修课开设。有关方法的更多信息，请参阅 报告 。

IPEDS

高等教育综合数据系统 (IPEDS) 结合了美国教育部国家教育统计中心 (NCES) 开展的年度调查。IPEDS 从参与联邦学生资助计划的每所学院、大学、技术和职业机构收集信息。

完成数据

本章节使用的数据来自 Completions survey ，该调查收集了完成中学后教育课程的学生人数据。根据 教学计划分类 (CIP) 代码，人工智能相关领域的毕业生被认定为第一主修专业为计算机与信息科学，普通 (11.01)；计算机编程 (11.02) 或 计算机科学 (11.07)。今年报告中包含的人工智能相关专业的毕业生人数与往年不同，因为人工智能指数使用了多个 CIP 代码。

经合组织

本章节使用了经合组织数据资源管理器中的数据，特别是“按教育领域划分的在校学生、毕业生和新入学学生人数”表中的数据。该数据集的方法可参见 Education at a Glance2024 Sources, Methodologies and Technical Notes 。

附录

第八章：公众观点

第八章：公众观点

益普索

为简洁起见，2025 年人工智能指数报告选择不重新发布报告中介绍的益普索调查所使用的方法。有关益普索调查方法的更多详情，请参阅调查。