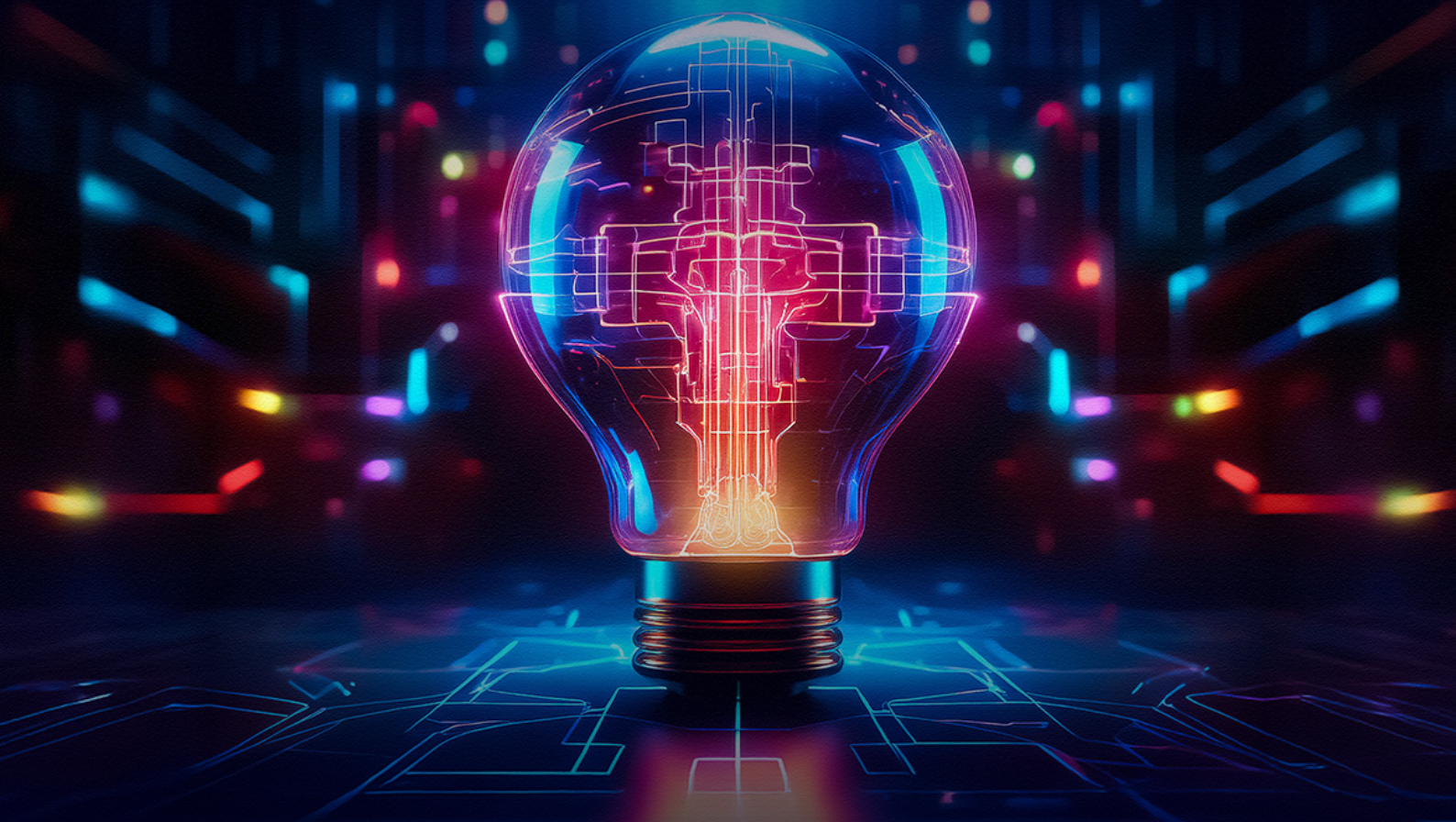
GENERATIVE AI WEEK

2025年11月11日 – 12日
圣莫尼卡奥斯汀吉恩敦，德克萨斯州

企业生成式人工智能的全球状态

行业报告2026



访问网站 注册

内容

前言

当我们在2023年发布行业首份报告时，企业正经历着一股实验浪潮，试图在工作流程中识别生成式AI的变革性应用场景。2024年的报告显示试点研究和概念验证呈爆炸式增长，企业正寻求定义治理政策、基础设施要求和价值创造。

今年的报告显示了景观是如何快速演变的，随着企业从试点转向全面生产，有效地部署和扩展生成式人工智能计划，以提供实际的业务价值。

在本报告中，我们旨在突出塑造生成式人工智能采纳的关键力量：

- **核心行业的生成式AI** 哪些特定领域的用例正在发展，哪些有效
- **人工智能行业趋势：** 技术将走向何方以及推动下一波创新的动力是什么
- **AI在企业中：** 什么才是顶尖的运营实践——从架构到治理
- **生成式人工智能投资：** 资本流动在哪里以及它如何重塑竞争格局
- **生成式人工智能基础设施：** 领导者如何构建可用于人工智能部署的可扩展、灵活且经济的平台

在我们召开生成式人工智能周之际，本报告旨在使我们的对话基于真实数据、真实战略和真实成果。这不仅仅是我们今天的快照——它是为寻求在整个端到端工作流程中实施生成式人工智能的企业领导者提供下一步指南。

山姆·莱曼
活动导演
生成式人工智能周

表格

图形

企业市场与技术格局

根据麦肯锡一项针对美国的研究，截至2025年，高达71%的组织在至少一个业务功能中使用GenAI，较2024年初的65%有所上升。因此，预计企业GenAI支出从2025年的40亿美元增长到2030年的192亿美元，年复合增长率为36.8%，这一增长并不令人意外。

每年在资本投资上花费超过70亿美元，仅用于促进与GenAI相关的新能源发电。此外，他们还需要建设支持性基础设施，例如用于长距离输电的输电线路和用于将电力输送到家庭的配电线，使整体投资额要高得多。

虽然2024年是企业将GenAI作为战略必需的年份，随着企业扩大规模并从试点项目中学到经验，2025年开始见证通过大规模部署GenAI来提供切实的投资回报（ROI）的努力。然而，高级决策者预计不会立即要求切实的价值和财务结果，并且他们正在以中期至长期的时间表进行运作。

毕竟，尽管过去两年生成式人工智能（GenAI）发展迅猛，但它仍然处于发展和应用的前期阶段，这一点从目前60%的企业生成式人工智能投资来自创新预算这一事实就可以看出。然而，由于40%的支出来自更稳定的预算，其中58%是从现有分配中转移而来，企业正日益展现出对人工智能转型的承诺。生成式人工智能需要很长时间才能产生实际价值的一个原因是，公司需要将其有限的资源分配到各种竞争激烈的转型优先事项中，同时还要应对复杂且不断变化的监管环境。

虽然2024年是企业将生成式人工智能（GenAI）提升为战略要务的一年，随着企业扩大规模并从试点项目中获得经验，2025年开始见证通过大规模部署GenAI实现可衡量的投资回报（ROI）的努力。然而，高层决策者不会立即要求可衡量的价值与财务成果，并且他们正在根据中到长期的时间表进行运作。

另一个需要考虑的是，并非所有企业的 GenAI 投资都会取得成果。实际上，根据 Gartner 的估计，到 2025 年底，至少有 30% 的 GenAI 项目会因为数据质量差、风险控制不足、成本和电力需求激增或商业价值不明确而在概念验证后被放弃。事实上，根据高盛副总裁卡莉·达文波特的说法，美国将不得不

尽管基础模型投资仍然主导着企业的生成式人工智能支出，但应用层现在正在增长得更快。生成式人工智能应用支出的前三个领域如下所述。

1 代码协作者

人工智能与编程的交叉已成为科技领域在风险投资方面最热门的领域之一。人工智能编程工具可以自动化各种常规开发任务，如代码生成、测试和调试，鉴于全球对软件的巨大需求以及熟练开发人员的短缺，这已被证明特别有用。GitHub Copilot快速增长至3亿美元的营收运行率验证了这一轨迹，而Codeium和Cursor等新兴工具也在迅速发展。除了通用的编程助手，企业还在投资特定任务的协作工具，如Harness的AI DevOps工程师和用于管道生成和测试自动化的QA助手，以及可以执行更全面的软件开发的人工智能代理，如All Hands。

2 支持聊天机器人

根据 Menlo Ventures 的研究，支持聊天机器人吸引了 2024 年 31% 的企业采用。一个很好的例子是全球银行荷兰国际集团（ING），它仅通过聊天机器人就成功处理了其在荷兰每周 85,000 次客户互动中的约 45%。Aisera、Decagon 和 Sierra 是直接与终端客户交互的代理人的例子，而 Observe AI 则在通话期间为联络中心代理提供实时指导。

3 企业搜索与检索和 数据提取与转换

企业正显著投资这些解决方案，以解锁并利用组织内部数据孤岛中常常隐藏的知识。好的例子包括 Glean 和 Sana 等解决方案，它们连接电子邮件、通讯工具和文档存储，以实现跨不同系统的统一语义搜索，并提供人工智能驱动的知识管理。

表1：按行业使用企业生成式人工智能的业务功能（%）

行业 商业功能	科技	专业服务	高级行业	媒体和电信	消费者商品和零售	金融服务	医疗保健医药，医药	能和材料	总体
市场和销售	55	49	48	45	46	40	29	33	42
产品和/或者服务开发	39	41	39	26	21	25	22	17	28
IT	31	16	26	22	20	24	30	26	23
服务操作	30	23	24	37	13	26	14	13	22
知识管理	26	34	17	26	12	16	24	13	21
软件工程学	36	9	17	30	8	20	13	8	18
人类资源	16	17	13	22	8	11	7	16	13
风险、法律、合规	12	9	6	6	11	21	5	9	11
策略与公司金融	14	14	21	10	7	7	6	5	11
供应链/库存管理	10	4	15	3	14	4	2	6	7
制造	5	3	13	3	8	0	5	7	5
使用生成式AI在至少1函数	88	80	79	79	68	65	63	59	71

源文件：需清洁本：

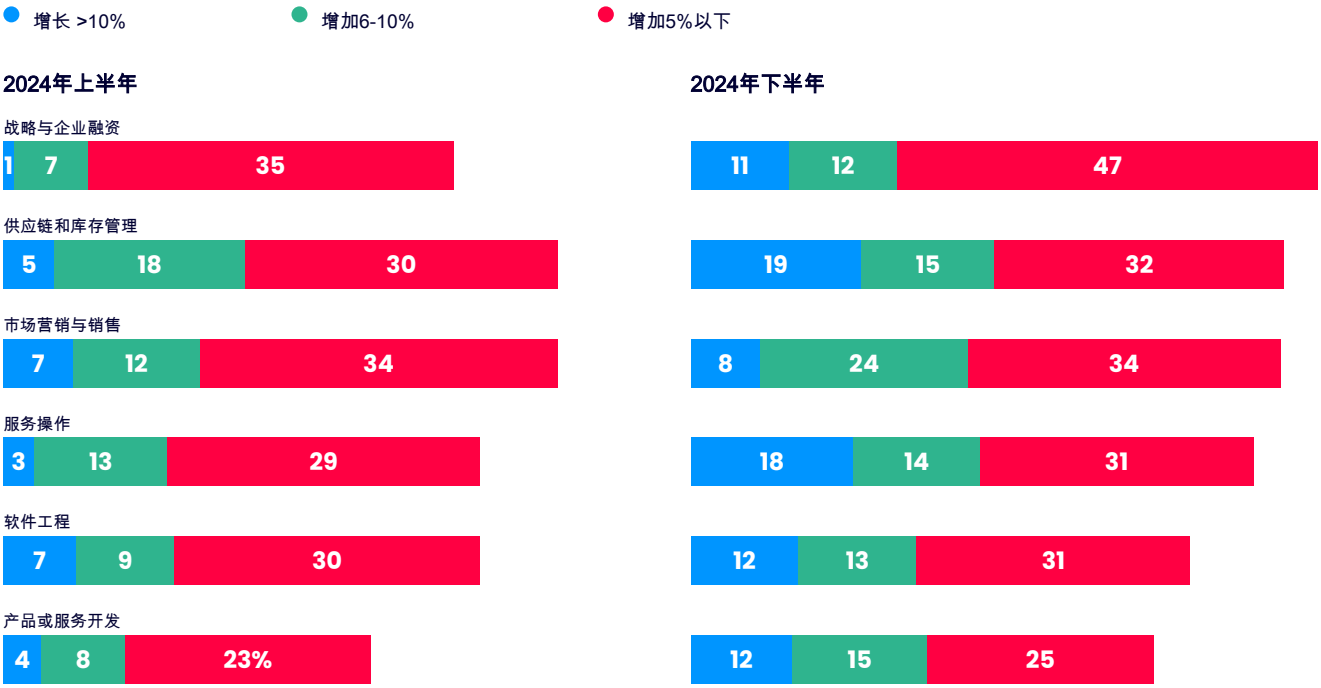
市场规模

全球企业级AI市场预计将从2025年的40亿美元增长到2030年的192亿美元，年复合增长率为36.8%。该技术在企业中日益普及的主要原因之一是像ChatGPT、谷歌的Gemini和微软的Copilot等先进和突破性的企业级AI工具的公开可用性，这些工具使专业人员对更多面向行业的潜在用例感到舒适。

尽管各行各业都存在一致的应用，但其中一些领域，如信息技术（IT）、网络安全、运营、营销和客户服务，比其他领域更为成熟。此外，那些在其规模最大的倡议中报告了更高投资回报率的企业，在他们的生成式人工智能（GenAI）旅程中普遍更进一步。

软件领域预计将在2025年占据最大67%份额，其余部分由服务构成。人工智能代理的出现及其预期迅猛发展是软件领域短期内到中期的主要驱动力，随着这项技术作为一项突破性创新而获得关注，并具有释放全功能生成式人工智能潜力的潜力。然而，需要注意的是，代理式人工智能不能被视为万能药，目前生成式人工智能所面临的所有更广泛挑战仍然适用。

全球企业级AI市场预计将从2025年的40亿美元增长到2030年的192亿美元，年复合增长率为36.8%。该技术增长的主要原因之一是像ChatGPT、谷歌的Gemini和微软的Copilot这样先进和突破性GenAI工具的公开可用性，这些工具让专业人士对面向更多行业使用场景的潜在应用案例感到舒适。



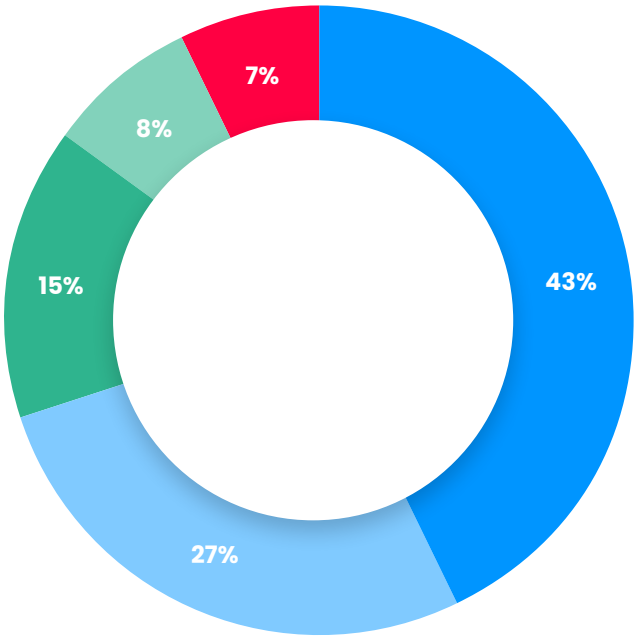
注意： 在2月22日至3月6日（2024年第一季度）和7月16日至31日（2024年第二季度）期间进行全球调查。针对那些表示其组织在特定职能中定期使用GenAI的人提出了一个问题。

麦肯锡公司

图2：GenAI实施状态

- 直播/飞行员
- 迈向全面实施
- 完全实现，
- 评估阶段

非直播（测试阶段）

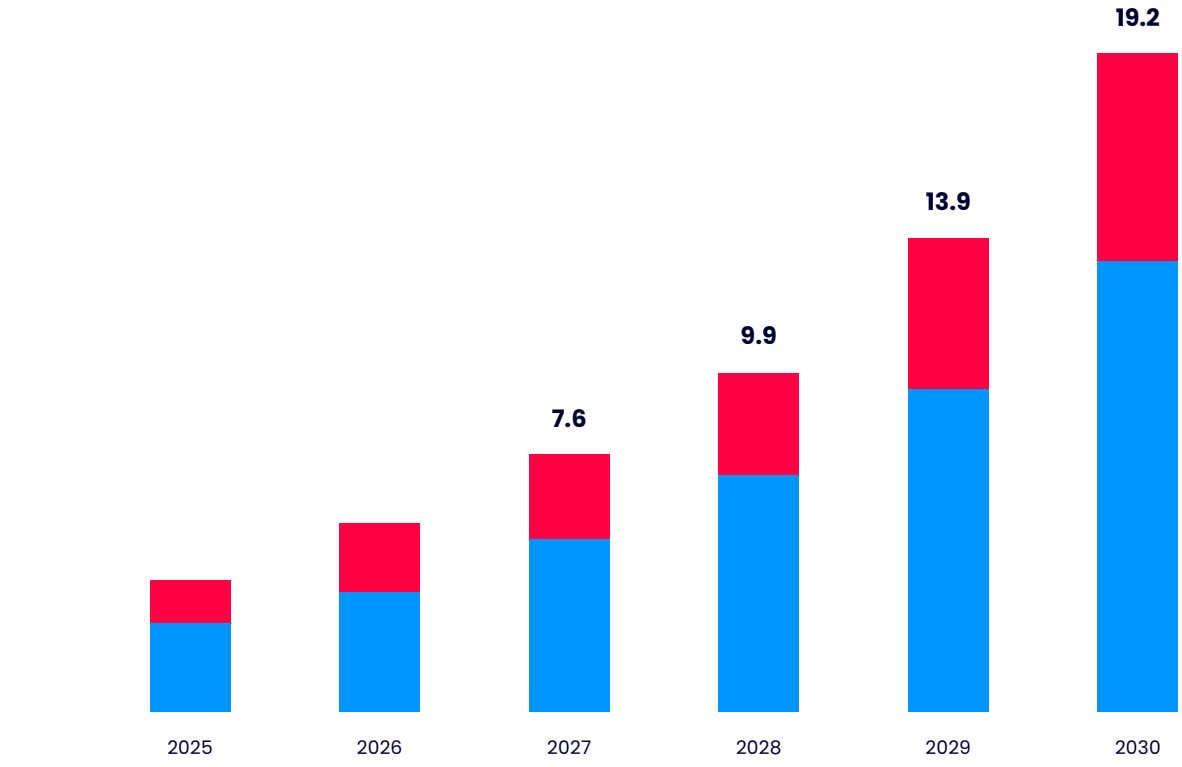


注意：ISG 生成式人工智能用例研究，于2024年8月进行在全球范围内调查了2000家公司

源文件： **翻译文本4:** 市场透镜人工智能研究

图3：全球企业GenAI市场按细分领域划分，2025-2030年，单位：百亿美元

- 软件
- 服务

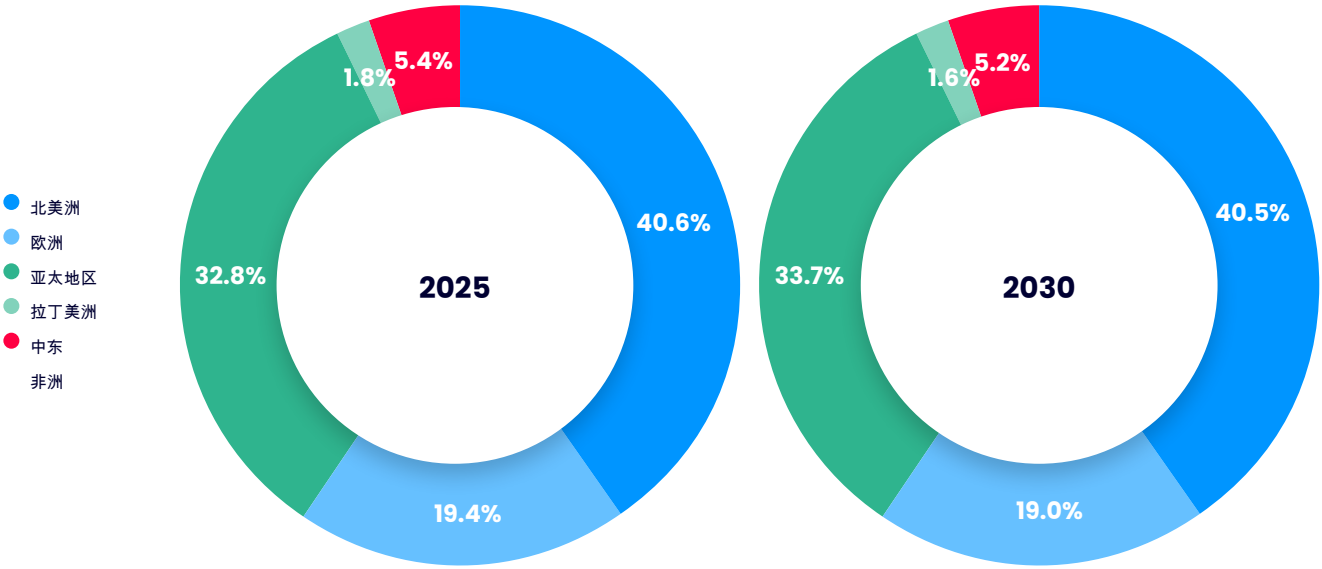


来源：AgileIntel

北美预计将主导全球市场，到2025年市场份额约占41%。与此同时，亚太地区预计将在2025年至2030年期间成为增长最快的地区，贡献显著

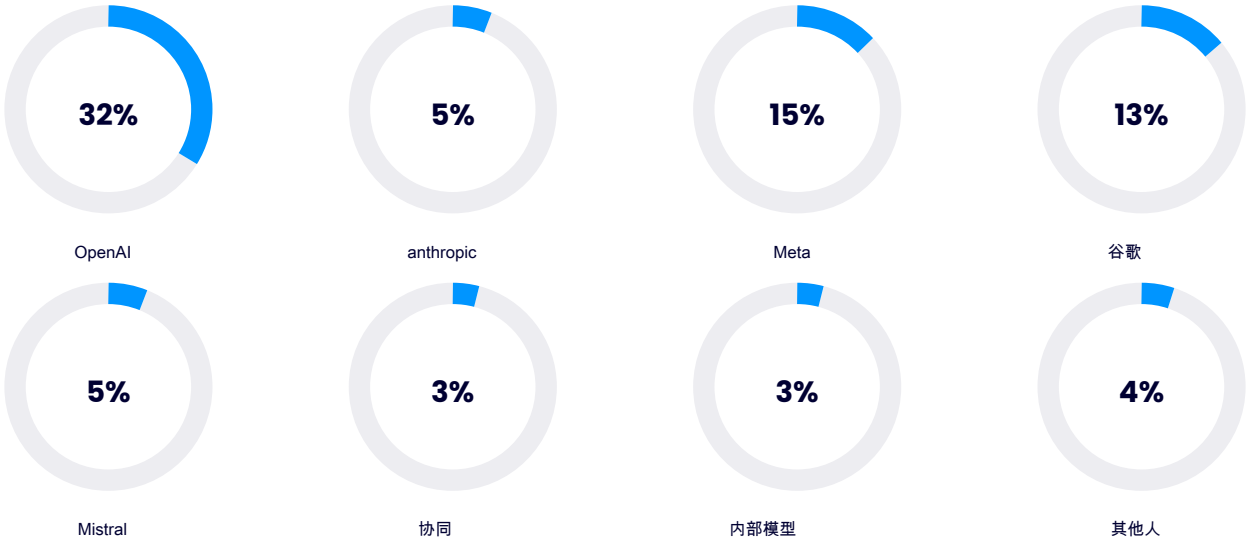
来自中国、日本、韩国和印度，由 substantial government initiatives 推动。OpenAI 以 32% 的份额占据市场份额主导地位，其次是 Anthropic (25%)、Meta (15%)、Google (13%) 和 Mistral AI (5%)。

图4：2025-2030年全球企业GenAI市场份额按地区分布，%



来源：AgileIntel

图 5：企业通用人工智能：2024 年 LLM 的市场份额百分比



注意：Meta的Llama 3和Mistral是开源的大型语言模型

敏捷智联

核心产业的生成式人工智能

金融科技领域的生成式人工智能

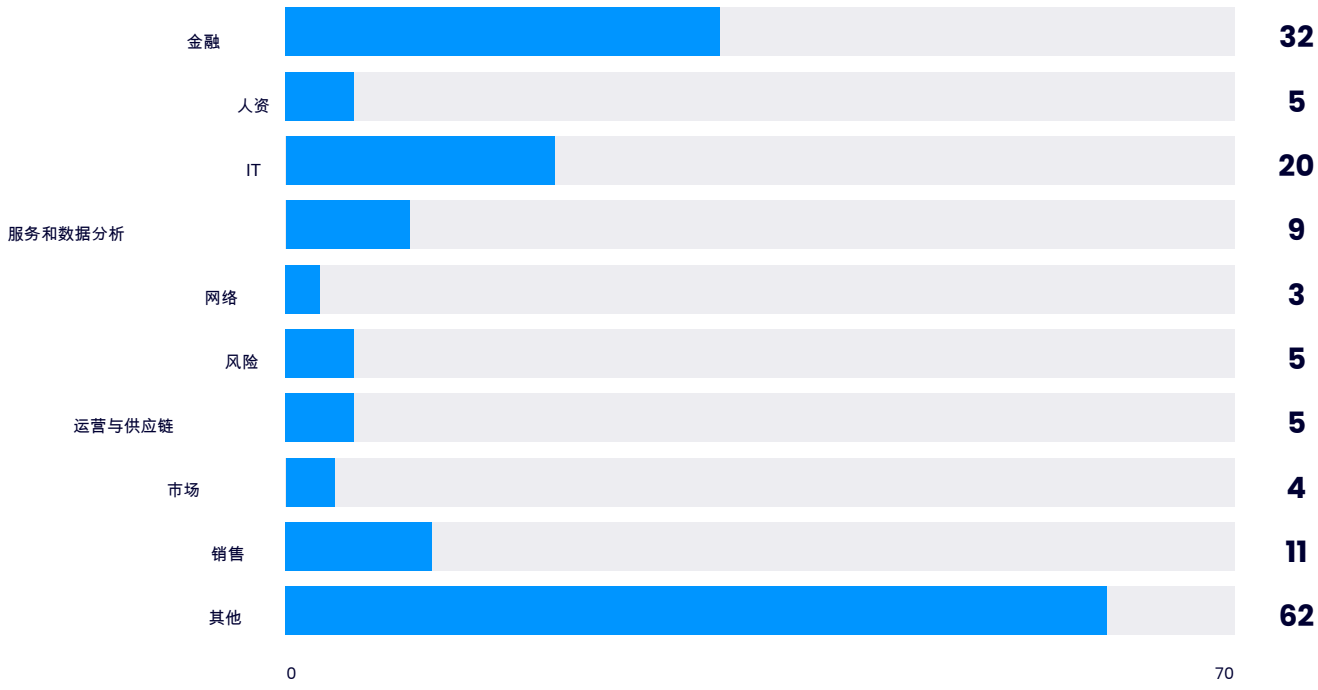
在全球经济波动加剧的大环境下，金融服务行业持续运营，特点是利率突然上调和贸易紧张局势加剧。欧洲和印度银行正从利率上升中获益，而北美银行则因结果更加两极分化而面临 mixed bag。另一方面，正当日本银行开始显现复苏迹象时，美国关税担忧导致该国银行指数在2025年4月4日当周暴跌超过20%。

在这种不确定性中，只有适应的银行才能繁荣，而其他银行则有被进一步落后的风险。全球金融服务行业采用的一项关键适应策略是整合GenAI，它已成为一个核心。

银行转型的推动者。该技术有潜力不仅推动运营转型和重塑商业模式，还能节省成本、产生更高收入，并满足风险和合规要求。

此外，随着行业日益数字化，通用人工智能（GenAI）为自动化复杂流程、提供定制化客户体验以及加强安全措施提供了机遇，从而使其能与更敏捷的数字优先竞争对手竞争。这在当前动荡的宏观经济环境下尤为重要，该环境给全球金融机构带来了巨大的压力，要求其向利益相关者提供充足的回报。根据麦肯锡的估计，GenAI每年可为全球银行业贡献2000亿至3400亿美元。

图6：按功能划分的美国生成式人工智能机会（单位：十亿美元）：银行



来源：毕马威，2025年2月

即使几家金融服务公司已经成功实施了GenAI并开始实现效率提升，也只有少数公司报告称其GenAI投资带来了收入增长。因此，他们现在面临来自股东的巨大压力，要求立即展示其投资的回报率，而毕马威在2025年的一项研究中估计这一比例约为70%。

补充其他机器学习（ML）模型和应用。因此，它们将GenAI集成不是作为独立的模型，而是作为包括机器人流程自动化（RPA）和自主代理AI解决方案的模型和技术网络的一部分。在这里，一个的见解和输出用于指导另一个的功能和方向。

这种方法已经开始以24/7虚拟顾问的形式交付成果，提供定制化金融指导，自动化常规交易，并基于实时数据和预测洞察主动管理客户需求。此外，通过分析海量数据，欺诈检测、合规监控和风险评估等后台流程正得到简化，速度和精度得到提升。

实现ROI

尽管多家金融服务公司已成功将GenAI应用于其运营并开始实现效率提升，但只有少数公司报告称其GenAI投资带来了收入增长。因此，他们现在面临股东对其投资立即展示投资回报率的巨大压力。然而，尽管存在这些压力以及AI技术快速演变带来的不确定性，全球金融机构在短期至中期内仍准备增加其GenAI预算。事实上，根据BCG 2025年的研究，三分之一银行计划在2025年投入超过2500万美元来提升其GenAI能力。但是，GenAI在银行业内的部署方式正在发生重大转变，银行和其他组织正从广泛的试验转向优先考虑目标应用的战略企业方法，特别是在机构与客户之间的界面。现在，GenAI赋能的工具支持超越预定义脚本的自主聊天代理、实时贷款审批以及提交文档的自动化处理。

创意产业历来严重依赖人类的直觉、情感和原创性，以此保护自己免受人工智能及相关技术的颠覆。然而，生成式人工智能带来了许多机遇，该行业现已成熟，即将受到颠覆性影响。

在创意产业中，GenAI最有前景的应用之一是使用对话界面来创建新内容或翻译现有内容。例如，该技术可用于根据文章和博客文章生成视频或播客，或生成脚本或故事板的变体，使创作者能够更快地探索选项。

有趣的是，企业将GenAI在金融服务行业的潜在价值不仅视为下游应用，更将其视为一个工具

这主要归因于这项技术不仅能够自动化诸如调整图像大小、移除背景和生成设计变体等重复性任务，还能为创意人士提供一个用于实验的新调色板。这包括生成几乎无法与人类制作的内容相区分的个性化内容、图片和视频，从而提高运营效率，并使公司能够快速适应不断变化的趋势。

事实上，根据BCG 2024年6月的一篇文章，现在GenAI能够以接近零的边际成本创建高质量内容，使企业能够实现大规模个性化。世界经济论坛（WEF）的另一项研究表明，GenAI工具可以帮助创意专业人士每周节省高达11个小时的时间，用于头脑风暴、原型设计和内容完善等任务。这些优势使更多的人，包括那些不具备深厚技术或艺术技能的人，能够加入创作者的行列。

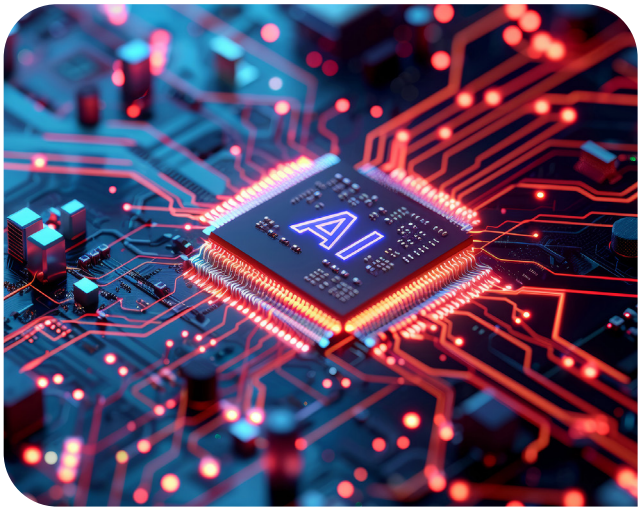
在创意产业中，GenAI 最有前景的应用之一是使用对话界面创作新内容或将现有内容翻译出来。例如，这项技术可用于将文章和博客文章转换为视频或播客，或生成脚本或故事板的变体，使创作者能够更快地探索选项。像 OpenAI 的 Sora 这样的文生视频 GenAI 模型，在广告业引发了地壳级别的变革，品牌和机构正以惊人的速度进行创新，利用 AI 生成的视频内容进行广告宣传。

大语言模型（LLMs）、生成对抗网络（GANs）、深度强化学习（DRL）和多模态生成式人工智能（GenAI）是支撑大部分这一颠覆的四大主要生成式人工智能技术。大语言模型可以比人类快得多地生成人类质量的内容，例如诗歌或剧本，并且还可以翻译语言。生成对抗网络更进一步，将两个神经网络对立起来，其中一个创建新内容，另一个评估其真实性。它们也可以生成高级图像，范围从照片级真实风景到抽象构图。DRL采用基于奖励的、试错系统，通过该系统，人工智能代理可以创建与特定美学偏好或用户行为模式相符的内容。多模态人工智能通过学习模式和文本描述与相应图像、视频或音频录音之间的关联而工作。

GenAI对创意产业的影响已经可见。一个很好的例子是Adobe在其创意云套件中集成了相关功能，例如生成填充和文生图等工具，这些工具正在改变设计师的工作方式。据Adobe前首席产品官斯科特·贝尔斯基称，该公司希望在中短期内为所有应用程序提供一个语言用户界面。另一个例子是图形设计软件公司Canva的魔法工作室，它通过使复杂的AI工具易于非设计师使用，实现了设计的民主化。

2025年，预计将获得关注的用例包括：

1. 像美国 Runway AI 的文本转视频工具和 Cinelytic 的分析及预测电影智能平台这样的应用程序旨在接入制作工作流程，使制片公司和电影制作人能够简化制作任务并做出明智的商业决策。



2 基于 OpenAI 的 GPT 大型语言模型构建的 Pencil AI 等工具，可以快速创建高质量、低成本广告，并利用预测分析来测试性能。ChatGPT 也提供分析功能，允许行业参与者创建受众原型来测试新电视节目。

3 从后期制作的角度来看，提供配音和字幕解决方案的AI应用预计将得到更广泛的使用。像Speechify、ElevenLabs和Panjaya.ai这样的平台简化并加速了配音音频和创建字幕的过程。这使得分销公司能够在以往本地化成本高昂的领域产生增量收入。

4 基于 GenAI 的音乐生成工具，例如 MuseNet、Magenta Studio 和 Musicfy 等

通过学习复杂的音乐模式，预测序列中的下一个词语或音符，以及混合指定的乐器来辅助创作音乐。它们还可以将一种声音转换为另一种声音，例如从口哨声转换为小提琴声，或从长笛转换为萨克斯风。这项功能对于可能不精通他们想要融入的所有乐器的艺术家来说是有益的，可以节省时间和成本。由于在大数据集上的无监督学习和使用Transformer，这个领域发展迅速。

5 以扩散模型（DMs）为基础的图像生成工具，如Stable Diffusion、Midjourney、DALL·E和Ideogram，正迅速获得关注。这些开源工具采用多模态扩散Transformer（MM-DiT）架构开发，对文本和图像均有益处。

表2：创意产业价值链上的GenAI应用用例

商业收养	前期制作	生产	后期制作	商业策略	商业操作
低	• 概念开发用于营销活动 • 市场分析 • 市场测试	• 媒体内容出版 (文本 & 图片) • 音频内容生成	• 人工智能配音内容本地化 • 内容审核	• 个性化内容发现 • 动态和个性化广告	• 客户服务聊天机器人 • 内容审核
中等	• 集成人工智能特效工作流程 (故事板, 运动捕捉) • 电影预测分析 • 剧本分析	• 新闻文章生成 • 音乐创作 • 基于人工智能的虚拟现实体验 • 人工智能渲染	• 音频克隆 • 创建逼真的音效电影、电视或游戏 • 视频剪辑过程自动化	• 对话摘要工具	• 网络安全与保护 • 流畅播放优化
高	• 游戏原型设计 • 剧本写作	• 人工智能新闻广播者 • 自动补全辅助代码游戏内编程 • 人工智能游戏非玩家角色	• 上色和评分 • 视觉效果 (VFX)工作流	• 对话摘要工具	• 预算管理

根据麦肯锡2024年4月的一项研究，涉及众多财富500强零售高管，高达82%的受访者表示，尽管他们仍处于试点和测试阶段，但该技术具有巨大潜力，主要在于增强其内部价值链。预计到2025年，大多数试点项目和概念验证将扩大规模并开始产生投资回报，尤其是在更快、实时可操作洞察方面，可在几分钟或几天内完成，而此前则需要几周或几个月。该技术，尤其是对话式人工智能，正在使数据分析民主化，使非技术人员能够在无需专业技能的情况下得出有意义的洞察。这不仅加速了内部决策，还使信息在零售行业的应用更加灵活和创新。

生成式人工智能也被预计将通过自动化员工排班、预测性维护、客户咨询和boarding新员工等日常任务，影响零售价值链的其他领域，并见证最大程度的颠覆。

根据麦肯锡的估计，生成式人工智能（GenAI）有望为零售商解锁4000亿至6000亿美元的经济价值，并解决数十亿美元的低效问题。预计它还将将预测误差降低高达50%，帮助零售商跟上消费者趋势。因此，据德勤最近的一项研究表明，全球45%的零售营销领导者计划在未来12至24个月内投资生成式人工智能，这并不令人意外。研究咨询公司IHL集团的一项研究则发现，生成式人工智能有望在2023年至2029年期间将零售销售额提升51%，将毛利润提升20%，同时将销售和管理（S&A）成本降低29%。

行业在GenAI部署方面面临的主要挑战是，大多数公司严重依赖现有的通用工具。PYMNTS Intelligence于2024年底的一份报告涉及在美国零售行业中，超过500名C级高管发现61%的人仅使用现有的基准模型，这限制了他们实现更具变革性的ROI的能力。相比之下，信息和制造业等行业发展专有解决方案的情况更为领先，分别有70%和69%的企业这样做。

关键用例机会

零售媒体: 为零售商提供一个高利润率的机会，这些零售商越来越多地将他们的数据出售给品牌，然后品牌可以利用这些数据来更接近购买点接触消费者。生成式人工智能的进步有望通过自动化广告活动创建和优化，以及帮助品牌提高其广告支出回报率（RoAS），来增强零售媒体。由于其能够在几秒钟内处理数百万个数据点，它还可能由于其能力而改善自助服务和程序化广告购买基础设施，帮助媒体购买者选择最佳广告格式，包括广告播出的时间和地点。根据Coresight Research 2025年1月的一项研究，美国零售媒体市场预计到2025年底将达到678亿美元，最终将在2028年增长到1044亿美元，复合年增长率为16%。

品牌在新的产品开发方面，以各种方式提升其创意流程。虽然像 Midjourney 这样的多模态模型已经提供了一段时间的图像生成功能，但新的 GenAI 应用允许创意专业人员无需学习如何设计提示和与之交互即可部署这些模型。此外，像 Digital Wave Technology 的 Maestro 这样的应用程序允许品牌生成与品牌故事更一致、避免幻觉和毒性的创意新产品想法。下方是 NIKE 的一张图片，揭示了 GenAI 在新产品构思中的艺术可能性。此外，GenAI 模型可以通过挖掘社交媒体帖子中的主要或新兴客户趋势或分析产品评论来促进新产品开发，这些信息随后可以输入到图像生成应用程序中用于新产品想法。预计 2025 年将出现能够管理和控制多个 GenAI 模型的应用程序，这将使图像生成技术平民化，使其能够被更广泛的非技术用户使用。

图7：由GenAI设计的空气概念鞋

注意： 网球选手郑钦文的空气概念
源文件：
翻译文本： 耐克

语音电商： 2025年预计将见证基于GenAI的语音购物或V-Commerce的扩张，允许用户完成购物、接收定制推荐，并使用语音指令管理订单。一个很好的例子是苹果智能，它在Siri中集成了先进的自然语言功能，提供高度定制化的购物推荐，甚至预测未来的购买行为。另一个颠覆性的例子是SoundHound AI，它集成在汽车中，允许驾驶员和乘客通过语音指令直接从汽车的娱乐信息系统中订购外卖自取。

使用 GenAI 的零售商示例

亚马逊： 已经开发了一款名为Rufus的AI虚拟助手，该助手在公司的产品目录和客户评论等资源上进行训练。该应用利用了亚马逊云服务（AWS）的Trainium和Inferentia芯片，以及一个定制构建的大型语言模型，使其能够在个性化设置中回答与产品相关的问题并进行产品比较。

卡玛克斯： 一家美国汽车零售商，是业内最早开始使用GenAI的公司之一，并自那以后发展了这项技术的应用，以创建包含规格、功能、优势和客户评价的详细汽车对比。其内部工具Rhode简化了员工对公司知识库的访问，而Skye则在车辆交易过程中增强了客户体验。

北面： 已在其线上购物平台部署了由IBM Watson驱动的GenAI模型，提供对话式购物助手。该AI助手会询问客户关于户外装备的偏好、计划活动和预期用途等问题，然后根据回答给出产品推荐。

eBay: 公司的GenAI驱动的购物助手ShopBot，通过文本、语音，甚至通过分享照片来指示他们正在寻找什么，帮助客户浏览超过十亿的列表。该助手还可以发起进一步的对话，以增强对客户需求的理解，从而允许它提供定制化的建议。

Shopify: 已推出一款名为"魔力"的生成式人工智能工具，该工具利用自动文本生成功能创建自动化内容，例如产品描述、电子邮件主题行和在线商店的页眉。它还允许商家修改照片背景以匹配其品牌，而无需精通像Photoshop这样复杂的软件。

零售价值链	在生成式人工智能之前	生成式人工智能之后
采购	• 人工处理供应商谈判 (包括端到端合同创建), 经常 导致细节被忽视 • 基于有限的供应商评估很繁琐 数据，导致次优选择	• 通用人工智能聊天机器人处理供应商谈判的初步回合 • 基于通用人工智能的供应商条款简报和摘要 协助采购助理完成交易。
分布	• 处理与 第三方物流提供商 • 对分销中断的延迟响应 由于供应链运营的复杂性	• 初次沟通和致第三方的电子邮件消息 由 Gen 人工智能聊天机器人处理的物流 • 返品管理流程，以及一个回复 分发中断，由Ge AI支持
店内运营	• 信息搜索，例如价格、店内 位置，以及库存水平由人工处理 关联，导致客户服务延迟	• 人们使用人工智能助手进行即时语音 信息获取
电子商务	• 数百小时用于生成 电子商务内容 • 基于手动规则网站个性化 消耗员工资源	• 自动生成电子商务内容（例如，产品） 几分钟内完成资料，描述） • 电商客户体验自发个性化 通过自动化前端开发技术
市场	• 由于有限，适用所有情况的营销方法 从结构化数据中获得的客户洞察 • 通过创建营销材料 漫长的迭代过程	• 从不同非结构化中提取的无限洞察 来源（例如，产品评论） • 完全个性化的营销材料生成 为每位客户提升效率
后台	• 耗时冗长的行政流程 例如人力资源和工资核算，容易出错和 低效	•下一代"白领"精益—转移 支持功能的行政流程到GenAI- 支持聊天机器人和界面，例如开发 助手，人力资源/财务助手。

麦肯锡公司

根据麦肯锡的估计，GenAI有望为零售商解锁4000亿至6000亿美元的经济价值，并解决数十亿美元的低效问题。它还预计将使预测误差减少高达50%，帮助零售商跟上消费者趋势。

制造中的生成式人工智能

在过去的几年里，生成式人工智能已从一项未来概念转变为一种切实的变革力量，以前所未有的方式塑造着制造业格局。这项技术现在使制造商能够通过支持编程和机器维护（包括预测性维护）、自主工厂管理、智能质量控制、智能供应商合同管理以及产品研发等职能，自动化和提升工厂活动。一个很好的例子是德国制造商博世，它正在使用生成式人工智能创建一个包含合成产品缺陷图像的综合数据集，以训练其人工智能系统进行最佳质量控制。

根据德勤2025年发布的一份名为《制造业的未来》的研究报告，该研究涉及全球600家制造商，其中87%的受访者表示他们已经启动了生成式人工智能（GenAI）试点项目，而24%的受访者表示他们在至少一个设施中采用了生成式人工智能用例。此外，50%的受访者表示，生成式人工智能解决方案在未来24个月内被评为其组织优先考虑的解决方案之一，其优先级高于其他备受关注的技术，如数字孪生、元宇宙和元宇宙。

另一项由科技公司NTT DATA于2025年进行的研究，涉及来自34个国家的500多名制造业领袖和决策者，令人震惊的是，其中95%的人表示GenAI已经直接提高了效率和企业绩效。有趣的是，94%的人预计将物联网数据集成到GenAI模型中，将显著提高AI生成输出的准确性和相关性。制造商们还使用GenAI通过在其内部工业物联网（IIoT）设备上训练LLM（大型语言模型）的小数据集来个性化运营，而不是传统的庞大数据集。这实现了遗留机器和不使用开源AI工具及GenAI系统的设备之间的无缝信息交换。此外，这些较小的

语言模型可以微调以更靠近边缘（最终用户）运行，在那里延迟和安全对工业物联网解决方案很重要。

基于人工智能的机器人在制造业中使用，通过利用这项技术中固有的自然语言提示。这使得机器操作员能够使用自然语言与机器进行交流，而这些操作员未必接受过机器人或软件代码的培训。

使用 GenAI 在制造业中的主要优势：

更快的产品发布： GenAI工具能够通过自动化和优化产品开发的各个阶段（包括创新、设计、原型制作和测试），帮助制造商更快地将产品推向市场。一旦GenAI模型在产品的材料清单、原材料使用、工艺参数、内部研究数据及其他数据（例如产品专利或先前产品试验数据）上完成训练，它就能识别最适合新产品成分，预测产品的功效，并推荐配方配方。一个很好的例子是阿斯利康，该公司正在使用GenAI自动化并加快药物开发过程。该技术已经帮助该公司将研发周期缩短了50%，并将实验中活性药物成分的使用量减少了75%。另一家领先制药公司正在使用GenAI分析生产线瓶颈并优化其片剂包装过程。这提高了生产效率20%，同时最大限度地减少了材料浪费。

数字孪生： 制造商正在使用生成式人工智能算法来创建其产品、生产线或整个工厂的精确数字模型。实时数据来自传感器和其他来源，用于改进设计、测试新流程以及在不中断生产流程的情况下创建新产品。一个很好的例子是印度特种化学品制造商Jubilant Ingrevia，该公司通过部署数字孪生来模拟、预测和实时管理运营，将过程可变性降低了63%。

使制造商能够在设计过程的早期阶段以高保真度可视化概念，并获得来自客户的精确反馈，从而让他们能够创造以前无法想象的产品。麦肯锡估计，仅就产品研发而言，生成式人工智能每年可以解锁600亿美元的生产力。此外，通过合成数据增强，生成式人工智能可以实现准确的模拟，使产品开发与严格的要求和客户偏好保持一致，从而节省时间和资源。

新产品开发: 通用人工智能工具分析当前市场趋势、消费者偏好和产品过往表现的大量信息，为制造商提供更清晰的新颖和先进产品设计图景，甚至发现新的商业模式。就新颖设计而言，通用人工智能

预测性维护： 以前，制造商通过根据固定周期或时间段执行计划维护来防止设备故障。随着人工智能（AI）和机器学习（ML）的出现，他们开始使用来自各种传感器的数据来识别模式，预测故障，然后主动进行维护。通用人工智能（GenAI）通过自动创建提供详细说明的文本或图像（包括所需备件清单）来进一步改进此过程。该系统使维护人员能够将更多时间用于实际任务，而不是准备说明，从而提高生产力并降低成本。由于其全面性，它还使经验不足的技术人员能够更有效地进行设备维修或维护。

规模化高效定制产品，满足个人客户的独特偏好而不影响效率。通过使用这项技术，制造商可以随时调整设计和流程以满足客户实时需求。AI驱动的洞察力使得大规模集成独特产品功能成为可能，而不会显著增加成本。随着技术的不断发展，个性化产品的潜力将得到扩展，根据特定客户偏好优化设计、性能和功能。已进入GenAI制造流程先进阶段融合的行业包括消费电子、汽车和时尚。

表4：贯穿制造业价值链的生成式人工智能应用

计划 - 产品开发
创建产品概念和工程图纸以减少研发和原型制作时间。
通过测试来定义它们作为替代原材料的适配性和功能，从而发现新材料。
使用定性消费者/市场数据预测产品市场契合度。
规划 - 生产规划与采购
根据现有材料、设备和资源制定生产计划。
跨多个来源发现新的供应商资料。
预筛选、总结并提取合同中的相关条款，评估风险。
自动处理 ERP 异常消息以实现最佳库存水平
生产 - 性能、维护和安全健康
制作员工培训视频和维护故障排除角色扮演。
编写标准操作规程和政策，并将文件翻译成其他语言。
识别危险工作条件，并通知关键利益相关者所需措施。
自动进行根本原因分析，无需手动数据分析即可确定非一致性的原因
预测精确的机器故障模式，并自动制定干预计划。
根据LoT、RFID和订单跟踪数据实时调整生产订单。
从 AI 聊天机器人那里接收性能更新、优先级和建议。
供应链 - 仓储和物流
自动化路线设计，使用路由算法来降低成本和提前期
通过聊天机器人界面提供货物运输和配送时间的更新。
生成并核实运输所需的文件。
为驾驶员提供一个交互式虚拟助手，以增强典型服务（例如，路线导航）
基于传感器和摄像头数据改进场地管理流程。
优化仓库设计以简化拣货路线。
自动重新订购物料以最小化缺货和库存水平。

内容生成 视觉生成 互动

人工智能已从未来概念转变为一种切实的变革力量，以前所未有的方式塑造着制造业格局。根据德勤2025年发布的一份题为《制造业的未来》的研究报告，该报告涵盖了全球600家制造商，高达87%的受访者表示他们已经启动了人工智能试点项目，而24%的受访者则表示他们已经在至少一个设施中采用了人工智能用例。

表 5：制造中 GenAI 模型的分类

生成式人工智能模型	在制造业中的应用
生成对抗网络 (GANs)	创建数字孪生，基于实时物理资产或过程的虚拟副本 用于产品设计及优化制造工艺的传感器数据。 **优点：**高质量逼真图像和数据增强，处理序列数据 并行 **缺点：** 难于训练，输出有限且重复，难以找到合适的平衡 在生成器和判别器之间。
变分自编码器 (VAEs)	通过在机器上训练的机器学习算法预测设备故障 数据。 **优点：**生成与训练数据相似的相似数据，克服传统图像的局限性 处理方法 **缺点：** 比GAN灵活度低，无法处理序列数据，难以控制质量
基于 Transformer 的模型	生产场景模拟、需求预测、缺陷检测和材料 断裂力学。 **优点：**并行处理序列数据，处理多种数据类型，功能强大 多样化的多模态任务 **缺点：** 需要大量高质量的训练数据，缓慢且计算量大 强化过程

来源：ScienceDirect

医疗保健中的通用人工智能

生成式人工智能正在迅速改变医疗保健行业。在麦肯锡对美国的付费者、医疗系统和医疗保健服务与技术（HST）集团的2024年第四季度调查中，高达85%的受访者已经将在整个企业中实施该技术。德勤在2024年底进行的一项研究也显示类似结果，高达75%的医疗行业的公司已经正在试验生成式人工智能。

技术在远程医疗和数字疗法领域多年增长停滞的背景下被广泛接受，这正推动着该行业的快速演变。

随着生成式人工智能的成熟，它正在催生新型解决方案，特别是在针对心力衰竭、糖尿病和心理健康等慢性病领域的空白方面。预计到2025年，生成式人工智能将对医疗保健的各个方面进行颠覆，从个性化护理到自动化工作流程，在不同层面上都将受到影响。与2024年主导的单模态模型相比，行业预计将看到能够同时分析并生成文本、图像、基因组数据和甚至实时患者生命体征的多模态生成式人工智能模型的更大规模采用。

据《医疗互联网研究杂志》（JMIR）2025年的一项研究，使用GenAI提供照护的患者参加了比其他治疗多42%的理疗课程，并实现了25%更高的康复率。这些发现展示了GenAI改善临床结果和整体护理标准的能力。

如果2023年关于生成式人工智能的实验，2024年关于点解决方案，预计2025年将通过端到端转型来实现价值交付。行业将不再看到孤立的生成式人工智能工具来完成像医生记录笔记或安排日程等特定任务，而是预计会见证集成系统自动化从患者接待到治疗方案整个工作流程的普及。这些智能代理将跨部门协调，从每次交互中学习以提高效率和结果。例如，在制药行业，将用生成式人工智能进行转型的关键流程包括临床试验、监管提交、医疗法律监管审查和全渠道互动。

总体而言，随着全球医疗保健行业面临劳动力短缺、临床医生倦怠、盈利能力下降和健康状况恶化等挑战，GenAI提供了一种变革性的企业方法来应对这些问题。该技术通过普及知识、增加互操作性、加快药物发现，并实现护理体验的高度个性化。在短期到长期内可能见证重大颠覆的各个领域包括患者和成员体验、日常行政任务以及临床医生和临床生产力。

尽管该技术在医疗保健行业的规模化整合备受热捧，但波士顿咨询公司2025年初的一份报告预测，到2025年，超过33%的正在进行中的生成式人工智能项目将无法创造价值。这些失败最终可能会为更可持续和更有影响力的转型铺平道路，从而推动将生成式人工智能更紧密地整合到现有的医疗保健工作流程中。生成式人工智能应用在短期到长期：

如果2023年是关于生成式人工智能实验探索，2024年是关于点解决方案，那么预计2025年将是通过端到端转型实现价值交付。行业将不再看到用于履行特定任务（如医生笔记记录或日程安排）的孤立生成式人工智能工具，而是将见证集成系统自动化整个工作流程的普及。

短期： 该技术的即时应用主要集中在医疗环境中自然语言处理（NLP）的应用，使诸如环境记录等功能得以减轻手动临床文档的负担。其他应用场景包括自动消费者消息、临床消息自动回复和文档自动生成。

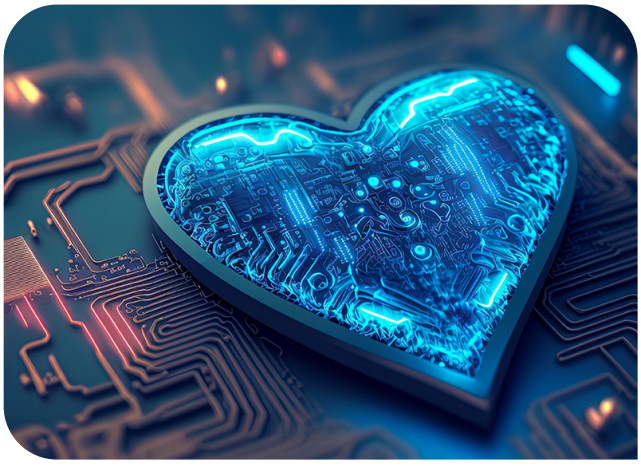
中期： 从中期来看，预计该技术将促进数据科学在医院各职能中的整合，以从病历、研究研究和患者生成数据等数据源中提取相关信息，从而制定更个性化和有效的治疗方案。

长期： 从长远来看，预计通用人工智能（GenAI）将在诊断和预后方面取代医生。事实上，在某些情况下，人工智能（AI）和机器学习（ML）在某些癌症诊断中的准确率已经达到98.4%，为未来快速颠覆奠定了基础。

GenAI在医疗保健行业的核心应用场景

药物和治疗发现： 2024年，人工智能驱动的药物研发取得了许多进展。然而，预计到2025年，通用人工智能将通过实时设计新型药物化合物带来快速颠覆。制药和生物技术公司越来越多地使用定制化语言模型来增强他们对疾病生物学的理解，并加速识别有前景化合物的过程。商业和开放的通用人工智能模型现在已经能够分析庞大的生物医学数据集，以提出新型分子结构，预测药物相互作用，并设计针对特定靶点或疾病的定制化合物。其中许多化合物通过传统方法难以发现。这减轻了巨大的成本和时间限制。当与因果建模方法结合使用时，模型使公司能够识别临床数据中以前未被发现的线索或代表性不足的线索，从而揭示先前被忽视的治疗机会。根据近期波士顿咨询公司的报告，到2025年，这一趋势将进一步缩短发现周期，并在临床环境中揭示更多有潜力的候选药物进行测试。

药物研发： 除了发现之外，GenAI 可以通过所有领域（如临床前测试、临床研究设计和监管提交）来增强药物开发过程。在临床前测试方面，GenAI 模型可以通过分析化学结构和候选疗法相关的潜在风险来估计药物化合物的毒性。它们还可以预测药物候选物的药代动力学特性和 ADME（吸收、分布、代谢和排泄）特性，这些特性可以预测药物对其靶点和相关安全水平的影响。在临床研究设计方面，GenAI 通过识别最相关的患者群体、终点和给药方案来提高成功的可能性。最后，该技术可以通过自动化合规性检查和主动对照指南进行检查来加快监管提交过程。此外，GenAI 工具有望通过处理工程师来优化制造疗法的流程，包括单克隆抗体和细胞疗法。一个很好的例子是 Exscientia，一家利用 Google Cloud GenAI 能力进行药物设计和开发的公司，它通过设计-制造-测试-学习（DMTL）循环实现更快的药物发现。



质量控制： 现在，通用人工智能正在通过标准化制造工艺和改进相关偏差的检测与缓解，在药品和医疗器械产品的质量控制中发挥更大的作用。这种质量控制方法将使制造商能够调整工艺、减少浪费、提高产量，并提升产品质量。例如，一个使用历史数据训练的问题解决通用人工智能模型能够使组织识别微小变化对产品结果的影响，从而在不进行广泛且通常的手动试错测试的情况下重新构想工艺。

聊天机器人： 根据2024年8月健康政策研究公司KFF的一项研究，超过16%的受访成年人表示他们至少每月使用一次AI聊天机器人来获取健康信息或建议，而对于30岁以下的成年人，这一比例上升至25%。随着基于GenAI的聊天机器人的发展和改进，这些消费者行为模式可能会迫使现有的在线健康信息

各门户若不提供其定制的AI工具，将面临失去网络流量的风险。这将使医疗服务提供者能够通过使用这些经过训练的聊天机器人来吸引患者并将他们引导至最合适的护理来源，从而开始实现显著的运营效率和竞争优势，同时减轻他们在提供24/7分诊能力时的人力负担。

个性化护理： 近期在自主智能领域取得的进展正通过分析包含患者特定数据（如基因型、病历和实时健康数据）的大数据集来驱动个性化治疗方案。这有助于医疗专业人员进行针对性治疗（如化疗、放疗或手术）的推荐，具体取决于每位患者的独特特征。根据2025年3月在ScienceDirect期刊发表的一项研究，基于GenAI的个性化治疗使癌症患者的生存率提高了20%，并将无进展生存期延长了15%。

根据2024年8月健康政策研究公司KFF的一项研究，超过16%的受访成年人表示他们每月至少使用一次AI聊天机器人来获取健康信息或建议，而对于30岁以下的成年人，这一比例上升至25%。

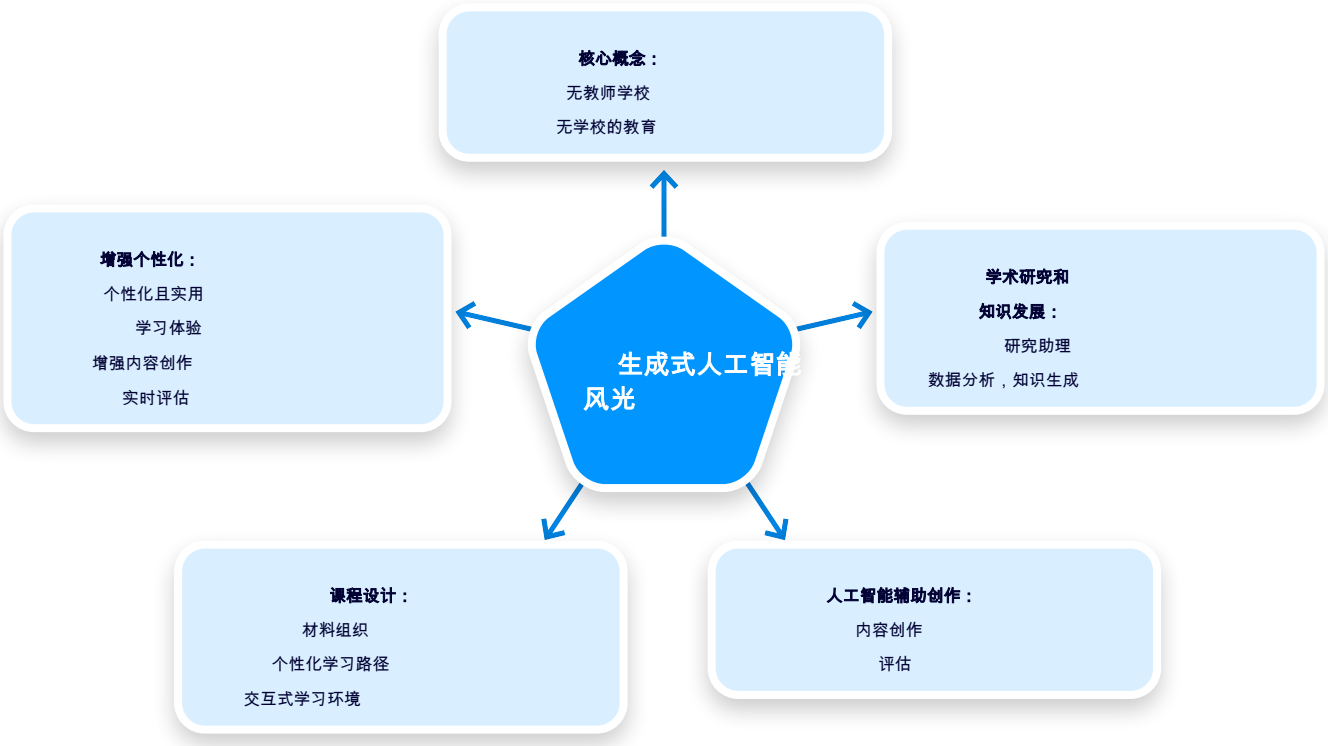
教育中的通用人工智能

人工智能正在通过颠覆传统教学方法、改善学生支持系统以及重组整体生态系统来改变教育行业。美国教育科技公司 Cengage Group 于 2024 年晚些时候发布的一份报告发现，美国高等教育中有高达 49% 的教师已经开始使用人工智能，这一比例从 2024 年的 44% 上升，而 2023 年仅为 24%。

该技术的核心能力，包括创建和传播信息，使其成为颠覆教育领域的理想选择。在过去一两年里，大语言模型展示了它们在回答各种学科问题、有逻辑地写作，甚至创建图像方面的能力。此外，ChatGPT和类似模型已证明它们在法律、医学、历史乃至运营管理等领域破解难题方面的专长。

教育科技公司和学生已经开始使用ChatGPT、TutorAI以及Poe应用等GenAI工具，这些工具通过协助头脑风暴和产生新想法来激发创造力。此外，GenAI模型已经开始协助老师创建作业和任务，向学生简单解释复杂概念，设计课程，并为每个学生创建游戏化学习体验和个性化学习计划。一个很好的例子是Speechify，它提供文本转语音或语音转文本生成功能，这对有阅读障碍或ADHD等学习障碍的学生特别有用。另一个例子是Kahoot!，它使用GenAI设计与课程目标一致的游戏，使学习既有趣又有效。

图8：教育中的生成式人工智能潜力



主要使用场景：

个性化自适应学习体验： 基于 GenAI 的智能学习平台分析各种类型的学生数据，例如历史表现、技能和教师反馈，以提供个性化自适应学习体验。通过分析大型数据集，教育者可以识别知识差距并提供建议和指导。GenAI 工具可以创建针对每个学生学习需求的练习、测验和练习题。此外，通过使用 GenAI 工具，教师可以提供实时协助、进度监控，并调整教学策略以优化学习。

虚拟实验： 通用人工智能与虚拟现实技术正被用来制作模拟环境和虚拟环境，使学生能够实时进行实验、观察结果并测试预测。

自动化评估和评分： 类似于 ChatGPT 和智能作文评估器的生成式AI工具可以可靠地审查和评分书面作业并提供反馈，从而确保速度、一致性和客观性。多项研究已经证明这些工具可以减少评分时间并提供准确且一致的评分结果。

课程创建与设计：教育者是

使用 GenAI 创建课程和教学材料，例如教学大纲、测验、练习和概念摘要。这不仅可以通过内容自动化生成来节省时间，还可以提高资源多样性。GenAI 还能够根据不同课程的需求快速创建电子学习胶囊、微视频和交互式多媒体元素。此外，提供语言学习课程的平台可以使用 GenAI 来纠正语法并创建相关的练习和问题。

通用人工智能预计将成为全球交通运输物流行业的主要增长动力之一，预计在2023年至2032年间以44%的显著复合年增长率增长，价值将接近190亿美元。随着该行业面临贸易流动变化、利润压力加剧、对可持续实践的需求上升以及托运人和监管机构要求的提高，通用人工智能提供了巨大的变革潜力。

根据2024年2月IDC进行的一项全球研究，超过50%的交通运输公司已经在使用GenAI进行知识管理、市场营销（更好的承运商/潜在客户转化、动态定价/报价增加）以及产品/服务创建，这些用例占所有用例的70%以上。另一项由德勤于2024年7月对200多名高管进行的研究发现，几乎所有人（99%）都预计这项技术将改变他们的行业，但超过三分之二（71%）的人预计这种转变将需要超过三年时间。在交通运输领域，资产管理、路线优化和仓库运营是用例中采用率和影响最高的情况。有趣的是，调查发现超过一半的公司在每一个用例中都运行着GenAI计划，并且大约80%的采用者报告称每个用例都具有极高的经济价值。

卡车和货运代理公司面临的一个长期挑战是规划高效的运输路线。GenAI模型通过分析关税、贸易协定、交通模式、公共交通和其他变量的相关数据来生成最佳路线并最小化成本，从而提供了一个解决问题的机会。

主要运输公司已经开始投资于与合同咨询、运输执行、战略和客户体验相关的用例。随着该技术仍处于非常初级的阶段，它承诺将在中期到长期内颠覆运输和物流价值链中的每一个环节。

从交付周期、需求、库存水平和其他来源出发，以提高产品可见性并防止缺货和库存积压。此外，由 GenAI 驱动的系统可以根据产品的受欢迎程度和某些物品的订单预测动态组织仓库布局，从而减少行程时间并提高效率。

主要使用场景：

路径优化： 对于卡车运输和货运代理公司而言，高效运输路线的规划一直是一项长期挑战。通过分析涉及运费、贸易协定、交通模式、公共交通及其他变量的数据，GenAI模型为解决这一问题提供了机遇，能够生成最优路线并最小化成本。GenAI在行业中的主要优势在于，通过分析交通数据、行人过街和紧急车辆位置，实现运输网络在实时中的动态优化。像DHL这样的国际航运公司正将其GenAI模型整合到其流程中，并分析有关货运量、船舶容量和港口容量的数据，以确定成本效益和环境友好的交付方式。

自动驾驶汽车：通用人工智能可以创建

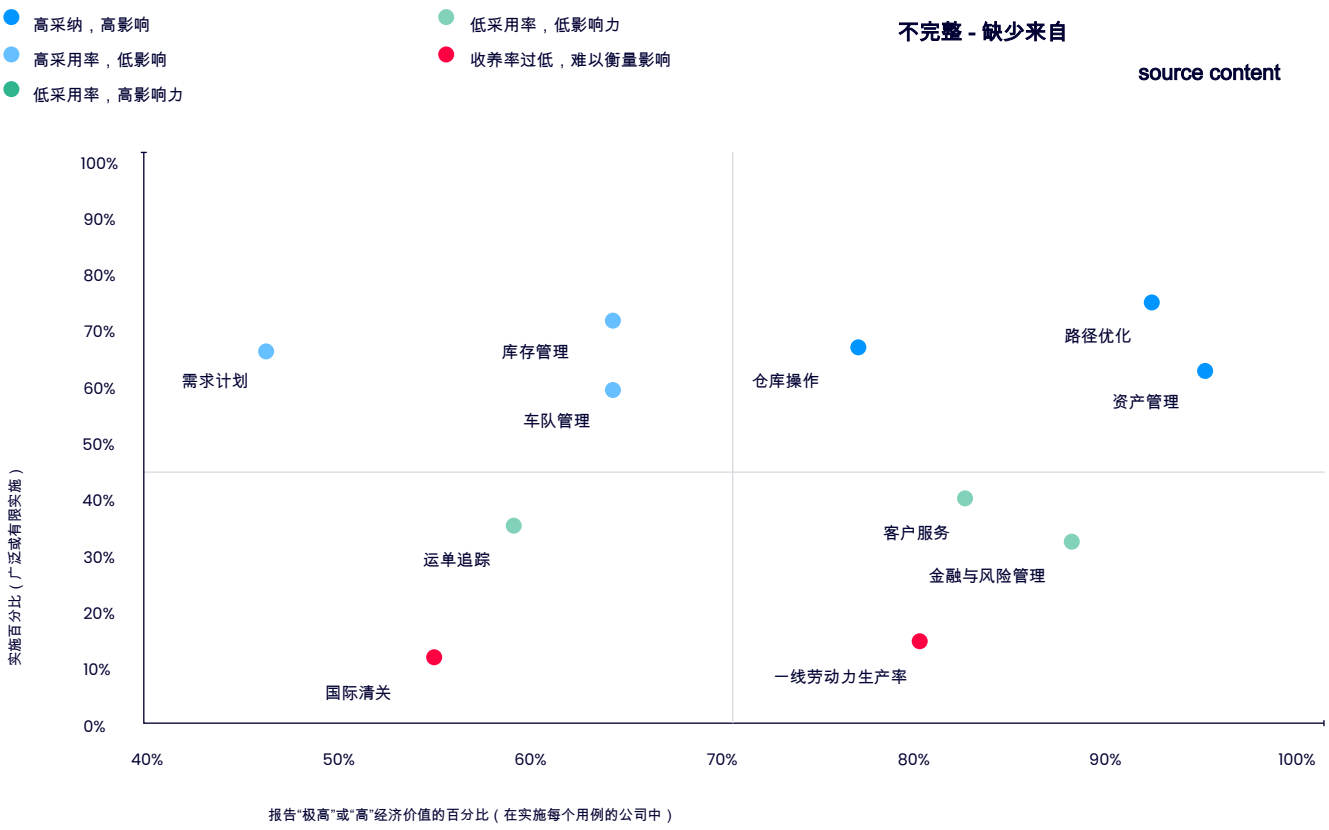
用于训练自动驾驶汽车和高级驾驶辅助系统（ADAS）以应对不可预测情况的各种逼真的虚拟驾驶场景。此外，该技术可以通过创建不同天气模式和道路条件的模拟来提高自动驾驶汽车决策能力。

通用人工智能（GenAI）还可以在问题出现之前预测基础设施和车辆维护需求，使运输公司能够采取预防措施，避免故障和停机。供应链管理者越来越多地使用这项技术来分析与季节性、促销、客户情绪和经济形势等要素相关的历史数据。这使他们能够创建高效的订购模式，精确预测未来趋势，并识别风险。

动态库存管理：具有高效

货物仓储是成功运输企业的基础，动态库存管理具有至关重要的意义。如果处理的货物数量很大，这一点尤其正确。因此，库存控制管理人员越来越多地使用GenAI来分析收集到的数据

图9：交通领域的GenAI采用与影响



注意： 全球范围内对200位高管进行的关于交通领域生成式人工智能的调研，2024年7月。
源文件：
翻译文本： 德勤

人工智能行业趋势

人工智能基础设施与架构

随着人工智能及相关技术的持续发展，企业正在进行重大投资，以开发稳健、可扩展且高效的人工智能基础设施。根据S&P全球市场情报机构2025年的研究，2024年GenAI相关投资超过560亿美元，几乎是2023年290亿美元的两倍。投资者感兴趣的领域之一是基础设施层，包括半导体、图形处理单元（GPU）云、光子织物、高密度计算解决方案、边缘计算、软件工具和可持续的GenAI基础设施。2024年GenAI基础设施投资几乎增长了近四倍，达到近260亿美元，而2023年为68.6亿美元。GenAI基础设施的前五大趋势包括：

随着传统的单体架构变得越来越昂贵和僵化，企业正转向解耦的、软件定义的基础设施，其中计算、存储和网络资源根据工作负载需求动态分配。这包括可组合的GPU工作空间，特别是在多租户环境中，由于其能够解耦计算、存储和网络资源，正快速取代传统数据中心，使组织能够重新分配GPU资源

根据当前的工作负载。对于利益相关者而言，投资可组合式GPU工作站的战略优势包括成本效益、运营敏捷性、提升ROI以及IT的面向未来。

光子网络用于人工智能加速：

通用人工智能模型的规模和复杂度的增长需要超高速、低延迟的网络。集群规模必须快速地从服务器中仅几个人工智能处理器扩展到单个机架中的数十个处理器，以及跨多个机架的数千个处理器，同时依赖高带宽、低延迟的网络连接来处理大量数据传输。光子织物正在为人工智能集群设定新标准，显著减少数据传输时间并消除网络拥塞。这些平台允许人工智能计算从处理器封装内部无缝地连接到跨多个机架的服务器。

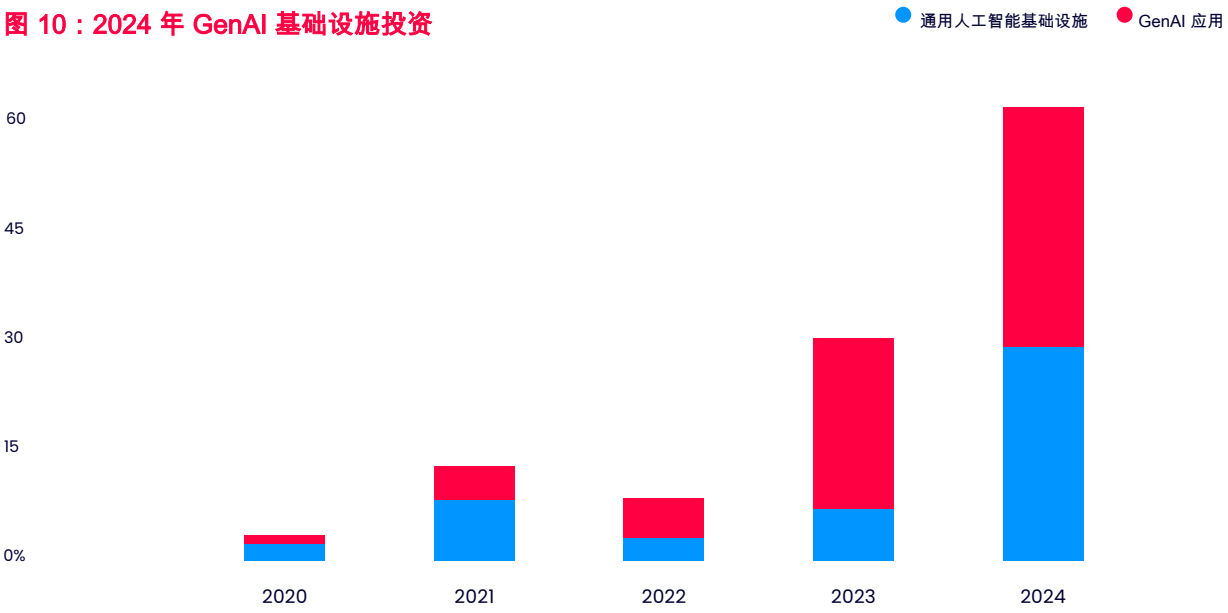
高密度计算解决方案：按照

根据德勤的最新估计，人工智能和数据中心处理效率的持续改进可能导致到2030年能源消耗达到约1,000太瓦时的水平。这些水平的AI工作负载需要大规模的硬件基础设施，使得高密度计算解决方案对于在优化电力、散热和物理空间的同时实现最大输出至关重要。这些解决方案非常适合企业通用人工智能、高性能计算（HPC）和数据中心运营。

边缘计算： 转向实时AI处理正在推动对边缘计算解决方案的需求。GenAI模型通常需要大量的计算资源和内存，具有庞大的模型参数和深度神经网络（DNN）。边缘计算通过将计算资源更靠近数据源来应对传统云中心架构的局限性，从而减少延迟和带宽消耗。

可持续基础设施： 生成式人工智能需要巨大的计算能力，使其成为一种能源密集型技术。图形处理单元（GPU）的生产需要稀土金属，其开采会导致温室气体（GHG）排放。据估计，到2030年，生成式人工智能可能产生120万至500万吨的电子垃圾，这比2023年产生的电子垃圾多约1000倍。科技公司正在采取各种举措，使生成式人工智能更加可持续的。这包括节能芯片、更小的型号、AI/Gen AI工作负载的合理配置，以及对低碳能源来源的投资。一个很好的例子是英伟达的新Blackwell芯片，其LLM工作负载性能比上一代提升了30倍，能耗降低了25倍。另一个例子是谷歌的TensorFlow和Hugging Face，它们采用了量化技术来减小模型尺寸，从而降低功耗和资源需求。

图 10：2024 年 GenAI 基础设施投资



来源：S&P Global，截至2025年1月10日

智能体式人工智能

虽然传统的大语言模型在文本、图像、音频、视频和数字的大型集合上进行训练，并对特定的人类提示做出响应，但基于先进的生成式人工智能模型的智能体（AI智能体），能够独立行动，无需持续的人类干预即可进行推理和学习。AI智能体技术之所以越来越受欢迎，仅仅是因为计算机在识别图像和理解语言方面的能力越来越强，这主要是由于基于 transformer 的技术演进。就像人类一样，这些智能体通过先进的推理和规划技能协同工作来解决复杂的多步问题，而大型语言模型 (LLM) 作为它们的“大脑”进行决策。使它们更具吸引力的地方在于，它们不仅能从数据库和网络中获取信息，还能从用户行为中学习并不断改进。OpenAI 的 GPT 模型家族、Anthropic 的 Claude 以及微软的 Copilot 等发布正推动着当前 Agentic AI 的热潮。

表6：智能体AI vs 生成式AI vs 传统AI

特性	代理式人工智能	生成式人工智能	传统人工智能
主要功能	目标导向行动 & 决策	内容生成（文本、代码、图片等）	专注于自动化重复性任务
自主性	高—以最小化运行人工监督	变量–可能需要用户提示或指导	低 – 依赖于特定算法和集合规则
学习	强化学习 – 改进通过经验	数据驱动学习——学习从现有数据	依赖预定义规则和人为干预

源：AISERA

根据IBM Watsonx.ai产品管理总监玛丽亚姆·阿舒里，2025年被预计是公司开始探索和部署智能体AI解决方案的年份。IBM与商业智能公司Morning Consult联手进行的一项2025年初的美国专注研究，涉及1000名开发企业级AI应用程序的开发者，发现高达99%的人正在探索或开发AI智能体。德勤在2024年底进行的一项研究表明，使用GenAI的25%公司将可能在2025年启动智能体AI试点或概念验证，到2027年这一比例将增长到50%。此外，在某些行业，针对某些用例的一些应用程序，可能会在2025年实际被采纳到现有工作流程中，尤其是在下半年的时候。另一项由埃森哲进行的全球研究发现，50%的受访者将在2025年实施AI智能体，预计到2028年这一数字将增长到82%。

在全球范围内进行投资，当前的智能体模型容易出错并陷入循环。在多智能体系统中，“幻觉”经常从一个智能体传播到另一个智能体，从而导致错误行动和结果的循环。一个很好的例子是Cognition Software于2024年3月推出的AI智能体Devin，它能够根据人类程序员的自然语言提示，无需协助地执行编程工作。在最近的基准测试中，Devin能够解决来自真实世界代码仓库的近14%的GitHub问题，尽管其性能是基于LLM的聊天机器人的两倍，但仍远未实现完全自主。

然而，人工智能技术解决方案架构师维约马·加吉尔告诫不要盲目乐观，他说该技术的普及需要不仅仅是更好的算法。它需要语境推理和边缘案例测试的显著进步，而这些领域的缺乏能力是广泛采用的 主要障碍之一。此外，尽管该技术正引起高度关注

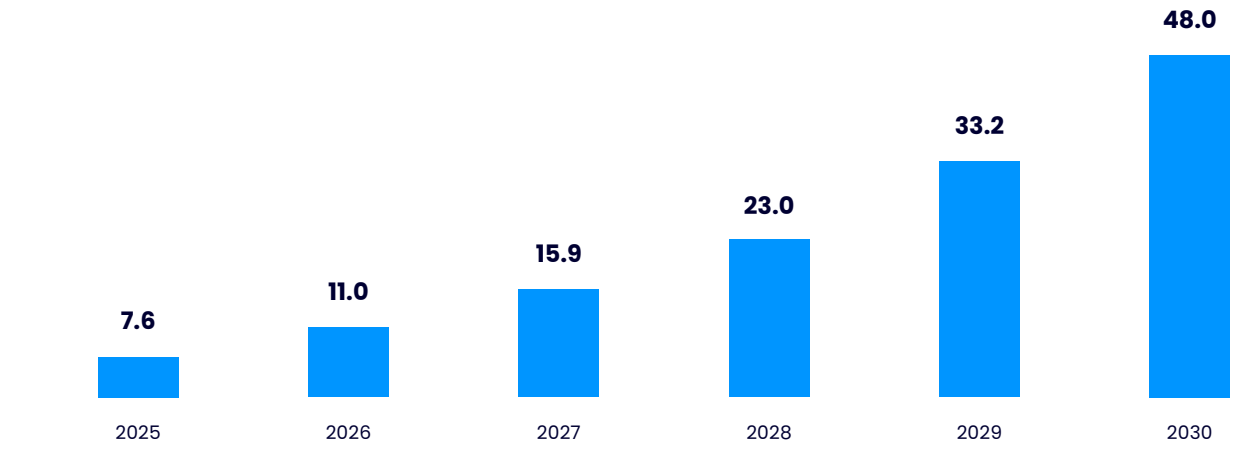
尽管目前存在局限性，自主AI的愿景是引人入胜的。该技术正在快速发展中，一些最新的自主AI模型采用了思维链功能，虽然与更传统的大型模型相比速度较慢且更审慎，但可以对复杂问题进行高阶推理。此外，多模态数据分析有可能通过增加可分析和创建的数据类型，使自主AI更加灵活。多模态AI还表明，当自主AI与其他类型的AI技术（如计算机视觉（图像识别）、语音转写和翻译）相结合时，其能力将更加强大。

全球Agentic AI市场规模预计将从2025年的76亿美元增长到2030年的480亿美元，复合年增长率为44.5%。

用例

客户服务： 像 Sierra、Ema 和 Decagon 这样的美国初创公司正在开发具有自主能力的 AI 聊天机器人，这些机器人可以根据其对客户意图和情感的理解独立行动。它们通过多个专业代理运作，每个代理负责对话的不同方面，例如意图识别、知识检索和情感理解。例如，一个 AI 代理可以预期到配送延迟，主动通知客户，并提供折扣以提高满意度。它还可以通过富有同情心和个性化的对话支持来改变与客户的互动。自主 AI 聊天机器人可以是多种类型：反应型、记忆增强型、工具使用型、半自主型、多代理网络和自我改进型。

图 11：2025-2030 年全球自主人工智能市场规模（以十亿美元计）



尽管存在局限，自主型人工智能的愿景具有强大的吸引力。该技术正在飞速发展，一些最新的自主型人工智能模型采用了思维链功能，虽然与更传统的超大规模模型相比速度较慢且更深思熟虑，但可以对复杂问题进行高阶推理。

采购： 尽管当前的采购工具侧重于数据分析与引导自动化，但像Zip这样的代理式AI系统已经能够自主运行，通过审查公司政策与要求，引导员工完成复杂的采购决策。

销售支持： 代理型CRM系统如Rox不仅存储客户数据，还通过预测其需求并主动与其互动，帮助公司更好地了解客户。总部位于美国的11x已开发了两个代理型AI系统，Alice和Mike。前者作为数字销售发展代表运行，可自主识别关键决策者并安排会议；而Mike则通过个性化、低延迟的语音通话，在28种语言中自动化进行呼入和呼出电话。

科学和材料发现： 即使机器学习和非代理AI已在药物发现和新材料创建等领域长期使用，代理AI也即将颠覆该领域。代理不仅可以分析特定材料的特性，还可以根据用户寻求的特征提出新材料或组合。此外，它还能根据成本或时间等优先级识别最佳供应商，甚至订购必要的材料。一个有前景的应用是ADME（吸收、分布、代谢、排泄）分析，它预测药物在体内的行为。一个主要的障碍是药物候选物因ADME特性差或毒性突显而在后期失败。代理AI可以通过分析分子结构和历史数据来提前预测这些特性，筛选掉不利候选物并优先处理有前景的候选物。

娱乐： 完全自主的AI智能体已因其在为非玩家角色（NPC）提供类人行为和游戏玩法方面的能力而被应用于游戏行业。例如，研究人员通过构建一个沙盒环境，创建了一个由AI居住的小型虚拟城镇

类似于《模拟人生》，有25个名为“斯坦福AI村”的代理。在这个村庄里，用户可以观察并与代理互动，他们分享新闻、建立关系以及安排小组活动。

应用与网络安全： 根据天箱安全研究实验室的报告，在2024年6月之前的一年里发现了超过30,000个新的漏洞。随着网络威胁数量和复杂性的增加，智能代理型人工智能在加强安全态势中正扮演着关键角色。这主要是因为这项技术比传统安全系统（如防火墙和杀毒软件）表现更出色，提供了一种新的自动化防御级别。它不仅分析应用程序代码、网络流量、用户行为和系统日志等因素以检测异常，还根据风险级别对漏洞进行优先级排序，并自动应用补丁或推荐修复方案。

完全自主的AI代理因其在非玩家角色（NPC）上提供类人行为和游戏玩法的能力，已在游戏产业中得到应用。例如，研究人员通过构建一个类似于《模拟人生》的沙盒环境，创建了一个由AI居住的小型虚拟城镇，该环境中有25个名为“斯坦福AI村”的代理。

图12：多模态GenAI代理的进化

进化可以分为三个关键阶段：

2000年代

机器学习（ML）的集成

从数据中学习： 机器学习（ML）的集成使智能体能够从大型数据集中学习，从而提高了它们做出决策和执行任务的能力。这对于基于规则的系统来说是一个重要的进步，因为智能体现在能够适应新信息并随着时间推移而改进。
自然语言处理（NLP）赋能用户交互： 自然语言处理领域的进步使得智能体能够更有效地理解和生成人类语言，使交互更加自然和直观。

2010年代

多模态介绍

结合文本、图像和音频： 多模态智能体出现，能够处理和整合来自各种来源的信息。例如，一个智能体可以分析文本描述，识别图像中的物体，并理解语音指令。这种多模态特性使智能体更具多功能性，并能够处理复杂任务。

增强用户交互： 多模态智能体可以以更动态的方式与用户交互，例如在响应用户的文本查询时提供视觉辅助，或从语音和视觉输入的组合中理解上下文。

2000年代
当前

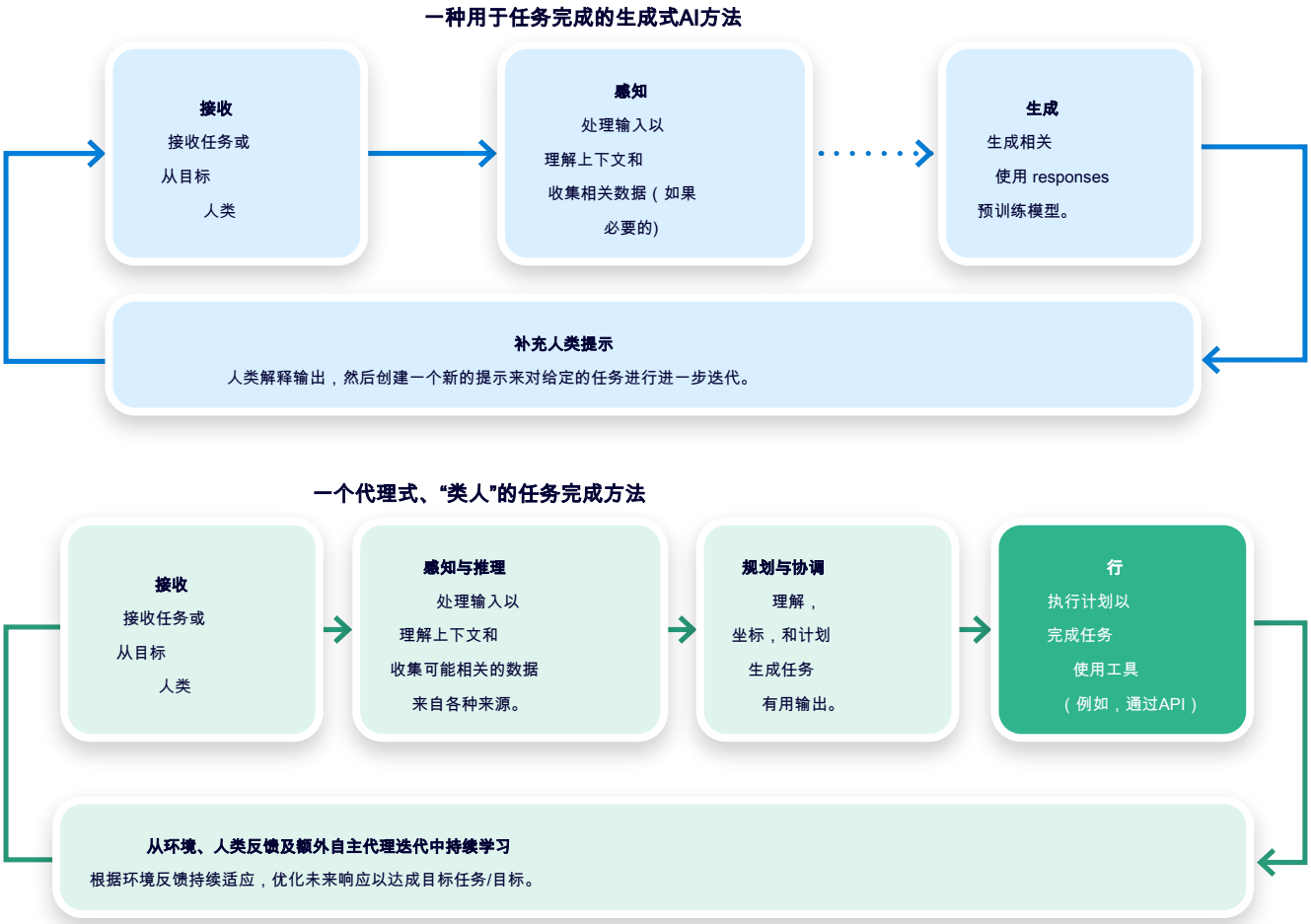
先进的自主性和实时交互

高级自主性： 智能体可以独立运行，合理化并设定自己的目标，开发实现这些目标的路径，并在无需持续人工干预的情况下做出独立决策，利用来自多个来源或合成数据集的数据。在一个多智能体编排系统中，第一组智能体专注于模仿人类行为（例如 ChatGPT-4o），也就是说快速思考以提出解决方案方法，而第二组智能体专注于慢速推理（例如 ChatGPT-1o）以提出经过审核的解决方案5。结合快速思考和慢速推理，智能体可以实时处理信息并做出最佳决策——这对于自动驾驶、实时客户服务以及各种关键任务业务流程至关重要。这种自主性使智能体AI在动态和复杂的现实世界环境中尤其强大。

在道德和负责任的AI控制环境下的人机交互： 随着能力的提升，人们也开始关注确保代理系统在考虑偏见、透明度和问责制等因素的情况下，以合乎道德和负责任的方式运行。

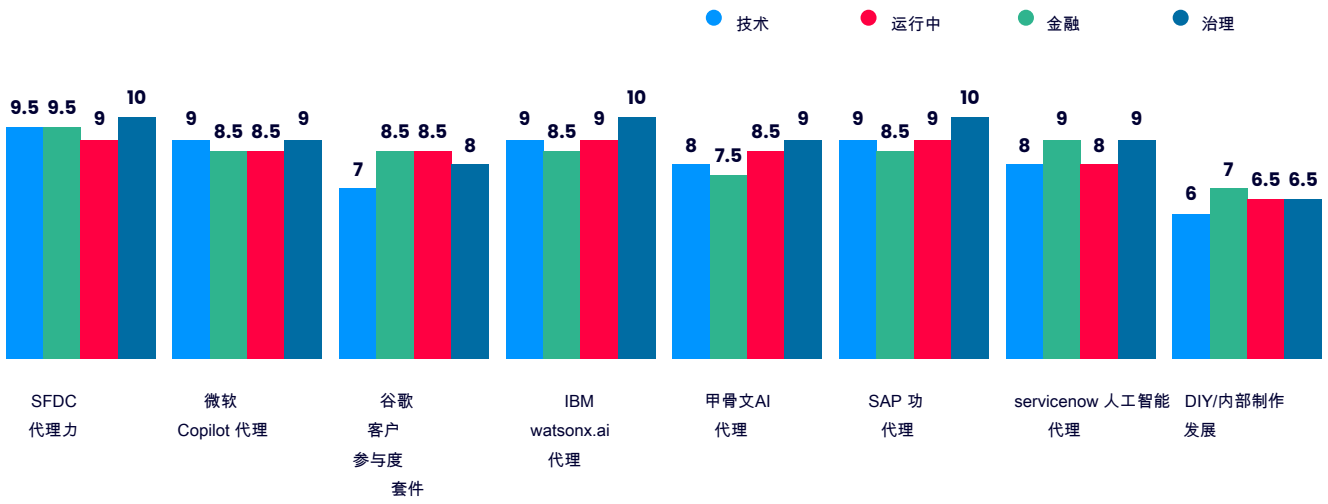
来源：AgileIntel

图13：GenAI与代理式AI在任务完成上的方法



剑桥替代金融中心

图14：领先代理式AI解决方案的对比评分



来源：The Futurum Group

人工智能治理——风险、合规、负责任的AI

负责任的AI

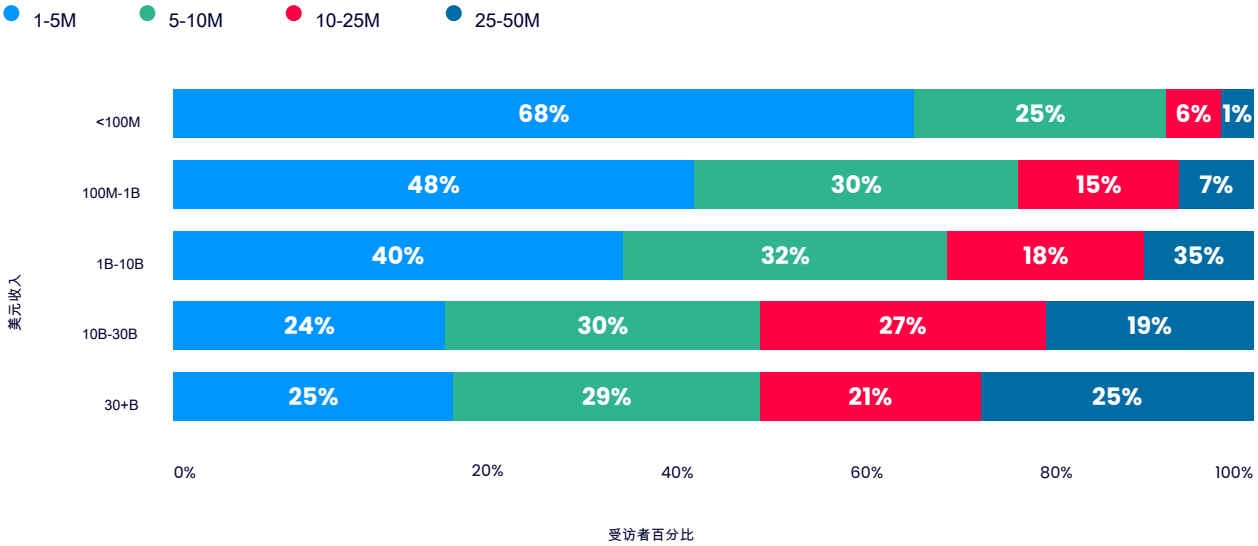
根据一份2024年的麦肯锡报告，研究前一年的全球GenAI使用量翻了一番，截止到2024年8月，ChatGPT每周有2亿活跃用户，是2023年的两倍。另据汤森路透在2025年进行的一项研究显示，95%的受访者认为GenAI将在未来五年内成为其组织工作流程的核心。相关地，GenAI的采用速度比个人电脑和互联网更快。

然而，正是GenAI的潜力诱惑，导致组织们一头扎进了采用中没有充分缓解风险。鉴于这项技术的整体新兴性，这已经被证明尤其具有挑战性。此外，与传统人工智能相比，GenAI的更高复杂性从技术角度来看构成了巨大挑战。毕竟，人工智能模型已经从机器学习中的少数几个参数发展到深度学习中的数万个参数，现在又发展到大型语言模型中的数百万、数十亿，甚至有时是数万亿参数。

因此，公司和组织正越来越负责任地设计通用人工智能应用程序，解决潜在风险并透明地分享经验教训，以帮助建立最佳实践。根据麦肯锡2025年的报告，能够从该技术的使用中捕获显著价值的公司一直更关注解决已知风险并识别和预防新风险。

根据2024年麦肯锡报告，研究前一年全球GenAI使用量翻了一番，截至2024年8月，ChatGPT拥有2亿周活跃用户，是2023年的两倍。相关地，GenAI的采用速度比个人电脑和互联网更快。

图15：公司收入对负责任人工智能的投资，2024



注意： 这项调查在来自30多个国家的商业领袖之间进行，N=759
源文件：
翻译文本： 斯坦福大学人工智能指数报告2025

负责任的AI（RAI）是一个全面且整体的框架，指导公司和其他组织以能够从AI系统中受益、降低风险并保持与企业价值观一致的方式实施AI。为了实现GenAI在跨行业的规模化集成，公司必须通过管理其数据、保护公司知识产权，在整个应用生命周期中实施RAI的原则。

(ip),保护用户隐私，并遵守法律法规。一种方法是自动化和扩展人工智能治理、安全和风险管理计划的部分内容，以更有效地检测和监控配置的护栏和控制机制。另一种方法是采用风险分层方法，根据风险和对客户、合作伙伴和员工的影响，将不同的监控标准应用于人工智能系统。

日期	利益相关者	范围	描述
五月 2024	OECD	全球	经合组织更新了其人工智能原则，并完善了其框架，以反映最新的发展。人工智能治理。这些原则强调了构建考虑包容性的AI系统增长、透明度以及可解释性，以及尊重法治、人权民主价值观。
五月 2024	委员会 欧洲	欧洲	欧洲理事会通过了一项具有法律约束力的AI条约（欧洲理事会框架）关于人工智能、人权、民主与法治公约)。该条约是为了确保人工智能系统生命周期内的活动与人类权利保持一致而制定的民主，法治。
君 2024	欧洲 并集	欧洲	欧盟通过了人工智能法案（欧盟人工智能法案），这是主要国家中第一个人工智能全面监管框架全球经济。该法案按风险对人工智能进行分类，相应地对其进行监管，并确保高风险系统的提供者——或者开发者——承担了大部分义务。
七 2024	非洲联盟	非洲	非洲联盟发布了其大陆人工智能战略（AU人工智能战略），概述了一个统一的愿景跨 continent 的 AI 发展、伦理与治理。该战略强调伦理，非洲内部负责任、公平的AI发展。
SEP 2024	联合 国家	全球	联合国更新了其《为人类治理人工智能报告》（联合国人工智能咨询委员会），概述建立全球人工智能治理机制的举措。报告建议制定应对人工智能相关风险的蓝图,呼吁国家和国际标准化组织科技公司、民间社会和政策制定者合作制定人工智能标准。
八 2024	G7	全球	G7数字竞争公报（G7人工智能合作）重申了对公平的承诺在open AI市场，强调协调监管方法的需要。之前的讨论专注于竞争和人工智能快速发展带来的监管挑战。
八 2024	东盟和 美国	亚洲 和 美国	在东盟-美国第12次峰会后，东盟与美国领导人发表了一份声明促进安全、可靠和值得信赖的人工智能。他们致力于合作开发制定国际人工智能治理框架和标准，以推进这些目标。
十一 2024	国际 网络 of 人工智能安全 机构	全球	第一个人工智能安全 институтов 国际网络已成立，汇集了九个各国和欧盟正式化全球人工智能安全合作。该网络联合技术致力于推动人工智能安全、帮助政府和理解社会的组织先进人工智能系统的风险，并提出解决方案。
二月份 2025	阿拉伯国家联盟	阿拉伯 国家	阿拉伯对话圈关于“阿拉伯世界的人工智能：创新应用和”“伦理挑战”在阿拉伯国家联盟总部发布，重点关注人工智能创新在强调道德考量的同时。

来源：斯坦福AI指数报告2025

企业中的生成式人工智能：
案例研究

西岩的生成式人工智能集成带来了更高的生产力和更低的成本。

保罗·麦克朗，全球可持续、基于纤维的包装解决方案公司WestRock的内部审计副总裁，于2022年首次了解到GenAI，但认为不应将其用于增强公司的审计职能。然而，该公司IT部门于2023年底开发了一个安全的GenAI平台，供所有内部部门进行实验。

该技术的首批应用之一是在审计流程的前端起草目标。当这被证明是成功的时，Paul决定通过输入数据并点击按钮运行无缝模型来自动化整个审计流程。然而，团队发现将多个任务链接在一起而不是单独执行将更有效。另一种有效的策略是在内部流程中整合高度标准化，从编写审计目标和执行方法的标准化提示开始。这使得WestRock能够自动化创建样本风险和控制矩阵、起草审计计划，甚至建议公司考虑的技术工具和脚本。

通过使用 GenAI 所捕获的一些早期价值，毫不奇怪地体现在了更高的生产力和更低的成本上。然而，保罗警告说，这些优势仍然处于非常初级的阶段，为了实现最佳价值，公司将不得不对其流程、时间表、里程碑和资源部署模式进行全面的再设计。它还将不得不从让程序员根据需求开发脚本的先前策略，转向一种更迭代的实时开发脚本并按需调整的过程。这需要团队协作

当多人挑战GenAI模型的结果，但基于技术速度，在压缩的时间范围内进行。

根据保罗的说法，WestRock与GenAI技术的未来涉及与Agentic AI集成，为平台添加一种学习机制，该机制基于历史教训来改进并扩大其未来运营范围。另一个立即目标是利用公司的数据分析经验来改进GenAI的实施。这包括利用平台进行持续监控和全面人口评估，而不仅仅是抽样。

西岩公司运用生成式人工智能技术的未来，涉及与代理式人工智能集成，为平台添加学习机制，该机制基于历史教训来改进和扩大未来运营范围。另一个即时目标是利用公司的数据分析经验来改进生成式人工智能的实施。

在中期，公司预计将开发一个动态系统，该系统根据实时收集和分析的行业和外部环境数据生成风险评估问题。这可能导致后续报告、行动跟踪和趋势分析通过交互式聊天机器人实现自动化。

麦当劳广受赞誉的对话式人工智能解决方案已于2024年6月在美国撤出

2019年，麦当劳收购了基于语音的对话式人工智能技术公司Apprente，标志着该公司与GenAI探索的开始。Apprente专注于开发复杂的语音识别和自然语言处理（NLP）系统，旨在处理复杂的、多语言的和上下文敏感的交互。这些解决方案预计将自动化麦当劳的驾车购买系统并简化订单流程。2021年10月，该公司与IBM建立了战略合作伙伴关系，以利用其人工智能和云计算专业知识，在更多地点扩展人工智能驱动的驾车购买系统的部署。

然而，尽管集成了先进技术，这款人工智能系统仍经常因背景噪音、口音差异以及复杂的订单而误解客户指令，导致会导致严重的误解。事实上，许多人工智能失败的视频被记录下来并在社交媒体上广泛传播，为麦当劳带来了很多负面宣传。2024年6月，这家快餐连锁店从美国100多个地点撤回了自动化系统。失败的主要原因如下：

实际测试： 这一失败的主要原因之一是缺乏真实世界的测试来确保系统能够处理实际客户交互的可变性。这包括模拟不同的口音、背景噪音和复杂的订单场景。此外，该系统没有在经过充分且多样化的数据集上进行训练，这些数据集会定期更新，以使其能够适应新的语言模式和客户行为。

以用户为中心的设计和反馈循环：

公司未能将用户反馈纳入开发周期以持续改进和优化系统。这对像麦当劳这样的公司尤其重要，因为理解用户需求和期望对于设计人工智能系统至关重要。人工智能系统应根据真实世界性能数据和使用者反馈持续更新和改进。建立反馈回路可以促进持续改进，适应不断变化的条件和用户行为。这种迭代过程有助于保持系统在一段时间内的相关性和有效性。

导致此次失败的主要原因之一是缺乏现实世界测试，以确保系统能够处理实际客户互动的变异性。这包括模拟不同的口音、背景噪音以及复杂的订单场景。

生成式人工智能技术

最新的 llm 模型，如 gpt-4（1.8 万亿参数）、claude 3（2 万亿参数）和 meta 的 llama 3（4050 亿参数），现在正被训练在数十亿甚至数万亿的参数上，从而在自然语言理解、代码生成和推理方面取得了显著进步。事实上，这些模型中的一些现在在阅读、图像识别、语音识别和语言理解等功能上已经达到了或接近人类水平的准确性。

2024年5月发布了最新版本的Gemini，即Gemini 1.5 Pro。

Llama: 由 Meta 开发，Llama 于 2023 年首次发布，随后于 2024 年 7 月分别作为 4050 亿和 70 亿参数模型发布。最新版本是 Llama 3.2，于 2024 年 9 月发布，最初具有 110 亿和 90 亿较小的参数数量。Llama 采用 Transformer 架构，并在包括 Common Crawl、GitHub、Wikipedia 和 Project Gutenberg 的许多公共数据源上进行训练。

当前一些顶尖的LLM包括：

克劳德： 由 Anthropic 创建的 Claude 专注于宪法人工智能，有三个主要分支 - Opus、Haiku 和 Sonnet。其最新版本是能够比以往版本更准确地解读细微差别、幽默感以及复杂指令的 Claude 3.5 Sonnet。该大型语言模型还具备广泛的编程能力，非常适合应用程序开发。2024 年 10 月，Claude 增加了一个计算机使用 AI 工具，使其能够像人类一样使用计算机。

聊天机器人gpt继续作为市场领导者，但由于谷歌和微软对其人工智能助手进行了改进，它的增长已经放缓。在初创公司中，通用人工智能聊天机器人正在逐渐但持续地获取用户，而专注于商业的claude人工智能目前在该领域增长领先。

深搜-1 是一个使用强化学习来提供数学问题解决和逻辑推理能力的开源推理大型语言模型。DeepSeek-R1可以通过自我验证、思维链推理和反思来执行关键问题解决。

ernie: 2023年8月由中国科技公司百度发布的人工智能Ernie，据称具有100万亿个参数，并已在全球吸引了4500万名用户。

双子座： 谷歌家族LLM产品，Gemini模型是多模态的，以网络聊天机器人为形式、Google Vertex AI服务以及通过API提供。它们有三种变体Ultra、Pro和Nano。Ultra是最大和最强大的，Pro是中端模型，Nano是最小的模型，专为设备上的任务效率而设计。

表8：重要模型和数据集发布

特性	Copilot（微软）	ChatGPT（OpenAI）	杰尼米 (谷歌)	Llama（Meta）
开发者	微软	OpenAI	谷歌深度思维	Meta AI
最新型号	Microsoft 365 助手 (2025)	GPT-4.5（2025）	Gemini 2.5 (2025)	llama 4 (2025)
主要焦点	人工智能的集成 微软应用程序	通用人工智能, 对话 编码	多模态人工智能，谷歌 生态系统	–
训练数据	基于 OpenAI 的 GPT-4 构建 专有	广阔，多模态，多样	Web规模，多模态	–
关键特性	深度集成与 Microsoft 365 和 GitHub	网络浏览，DALL-E， 文档分析，和 语音交互	多模态（文本，图像， 音频，视频），Google 服务集成	开源LLM 为研究优化 设备端部署
代码生成	精通 Python JavaScript，C++，Java	精通（Python，JS，SQL）	强（Python、JS、SQL）	平均
多模态支持	支持文本和图像 生成	强力 (图片，文本)	非常强（文本，图片， 音频，视频)	纯文本；非母语 多模态支持 图片、音频或视频。
内存特性	是	是 (适用于Plus用户)	是	是
API可用性	不	是	是	是
免费版	是 (Microsoft Edge)	是 (GPT-3.5)	是 (Gemini 1.0)	是 (Llama 2和Llama 3.2)
优势	代码特定协助	通用，可靠通用 AI，强大的对话能力 能力，集成 插件	最好的多模态AI，谷歌 生态系统集成 严谨的逻辑推理	隐私 & 移动 部署
缺点	过度依赖风险	无实时浏览 在免费版本中，可以 产生幻觉 以及高级功能 被付费墙挡住了。	需要谷歌 集成，一些 精度问题	有限的复杂性 处理
最适合	软件开发	通用人工智能 聊天机器人、写作和 研究	多模态任务，搜索 和生产力	离线，低资源 环境
成本结构	随附Microsoft 365 订阅	订阅制(Plus/) 团队)	使用谷歌账号免费	基于订阅

来源：瑞士德语大学，网络搜索

表9：领先的GenAI模型和规格

模型	创作者	上下文窗口	人工智能分析 智能指数	端到端 响应时间
命令-R	协同	128k	15	6.97
Jamba 1.6 mini	AI21 labs	256k	18	2.89
DBRX	Databricks	33k	20	NA
codestral（2024年5月）	Mistral AI	33k	20	5.01
LFM 40B	液体	32000	22	3.23
Qwen3 0.6B	阿里巴巴	32000	23	NA
一大型	阿里巴巴	32000	28	7.78
新星微	AWS	130k	28	1.79
Tulu3 405B	Ai2	128k	40	NA
Φ-4	MS Azure	16k	40	12.98
Φ-4	MiniMax	4m	40	16.51
声纳专业	困惑度	200k	43	7.98
Reka Flash 3	瑞卡	128k	47	45.06
Claude 3.7十四行诗	Anthropic	200k	48	7.44
GPT-4o	OpenAI	128k	50	3.83
Llama 4 骁猛	Meta	1m	51	4.21
grok 3	X.AI	1m	51	10.36
DeepSeek V3	深Seek	128k	53	22.47
Gemini 2.5 Pro	谷歌	1m	68	39.73
OpenChat 3.5	Openchat	8k	NA	10.87
北极	雪花	4k	NA	NA
太阳能迷你	上舞台	4k	NA	38.52

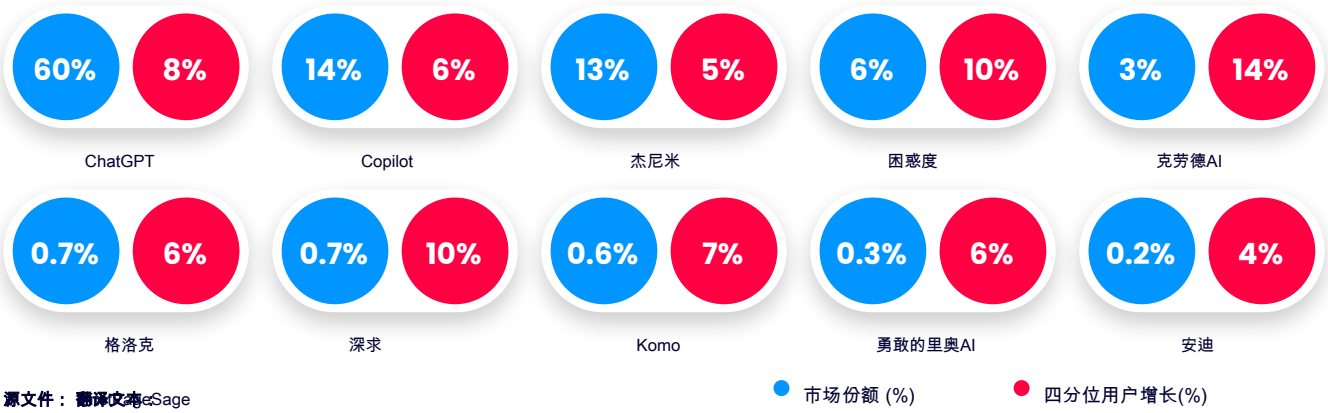
注意： 上下文窗口 - 最大输入和输出token数量组合数。人工智能分析指数 - 一种综合基准，用于评估和比较语言模型的智能。
源文件：
翻译文本： 人工智能分析

表10： 精选前沿实验室的GenAI平台说明能力 2022-23

		2025年1月
anthropic	**Claude** • 非多模态（仅文本） • 有限的上下文理解（难以复杂的对话） • 无工具使用	**Claude 3.5** • 多模态（文本、音频和图像） • 增强了长时段内的上下文理解与连贯性交互 • 部分用户的实验性计算机使用能力
谷歌 杰尼米	**谷歌巴德** • 非多模态（仅文本） • 公正推理 • 有限的上下文理解（难以复杂的对话） • 有限的实时数据集成 • 低个性化（适应性有限）	**Gemini 2.0 闪存** • 多模态（文本、音频和图像） • 高级推理（能够进行多步问题解决） 精妙分析） • 提升了上下文理解能力（在长диалоги） • 实时数据集成（来自谷歌搜索） • 高级个性化（用户上下文）
Meta	**Llama 1** • 非多模态（仅文本） • 公正推理 • 有限的上下文理解（难以复杂的对话） • 没有API访问权限	**Llama 3.3** • 基于文本（早期版本是多模态，LLaMa 3.2） • 高级推理（能够进行多步问题解决） 精妙分析） • 提升了上下文理解能力（在长диалоги） • API访问（用于模型和代理开发的工具）
Meta	**微软 Phi-1** • 非多模态（仅文本） • 公平推理（即，限制在编码任务中） • 聚焦训练（更小、以编码为重点的数据）set）	**Phi-4** • 多模态（文本、音频和图像） • 高级推理（能够进行多步问题解决） 精妙分析） • 全面培训（多样数据）
OpenAI	**GPT-3.5** • 非多模态（仅文本） • 公平的推理能力（例如，在SAT中得分很高，但在律师考试中排名后10%） • 有限的上下文理解（难以复杂对话中的连贯性 • 标准API访问（用于文本生成）	**OpenAI 0** • 多模态（文本、音频和图像） • 高级推理（例如，通过律师资格考试前10%） • 增强上下文理解与连贯性在长时间交互中 • 高级 API 访问（支持多模态输入）

--

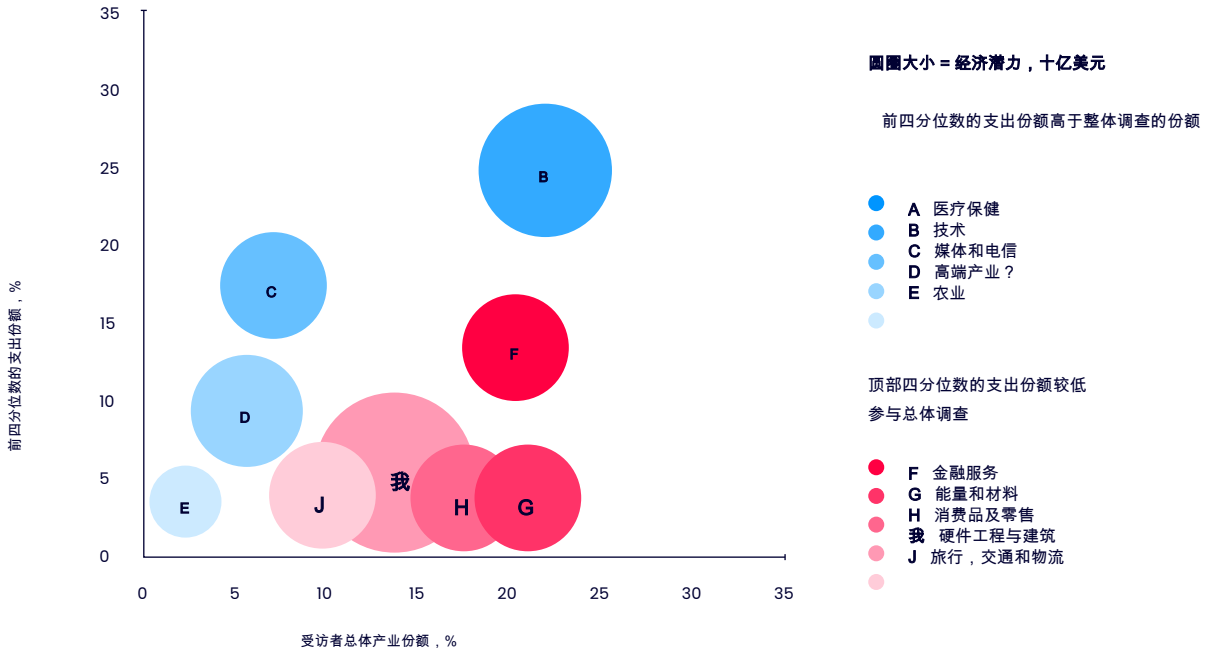
图16: 美国领先通用人工智能聊天机器人市场份额和用户增长，2025年4月



源文件： 翻译文本 Sage

生成式人工智能与投资

图17：生成式AI支出与行业经济潜力

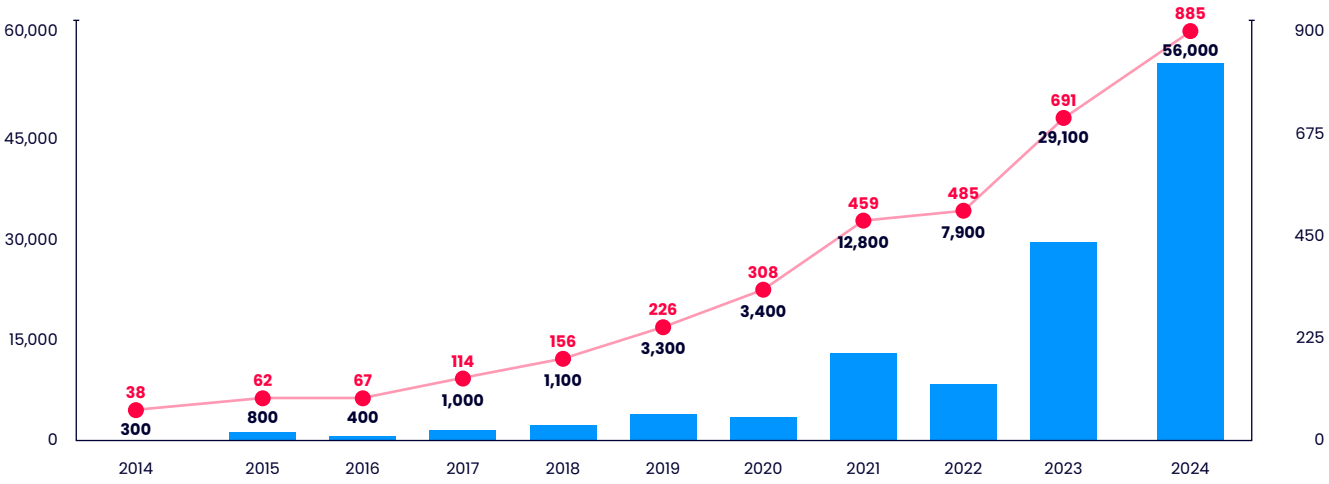


注意： 麦肯锡美国高管调查，2024年10月-11月，n=118
源文件：
翻译文本： 麦肯锡

表11：GenAI中最活跃的10位投资者

公司	VC 投资	风险投资支持的出口	中位数交易大小（美元百万）	选择投资组合公司
红杉	84	7	16	OpenAI，xAI，Glean，安全超级智能
Gainels	76	6	10	Adbridge，Cerebras Systems，People.ai，Figure
开拓基金	74	2	3.6	Moonvalley,Agent Labs,模型ML
安德森·霍洛维茨	73	4	30	mistral，cursor，openai
科斯拉风险投资公司	57	3	15	OpenAI，Curai Health，Replika
Soma Capital	55	5	4.2	工匠，灌注，月光谷
校友企业	54	3	11.7	Cohere, Lambda, Groq
通用催化剂	41	7	26.3	Cohere, Lambda, Groq
光速	40	2	32.4	Anthropic, Granola, xAI
光速	40	-	0.2	Zealth-ai, Omma, Banqora

图18：2014-2024年生成式人工智能领域的风险投资，单位：百万美元



来源：TechCrunch，截至2025年1月3日

表12：2025年第一季度生成式AI领域顶尖私募股权交易

公司	圆形金额	圆形日期	圆形估值	选择投资者	国家
OpenAI	40.0亿美元	风险投资 2025-03-31	30亿美元	软银、Altimeter Capital、Coatue、微软、Thrive 资本	美国
anthropic	3.5亿美元	系列E 2025-03-03	61.5亿美元	Lightspeed Venture Partners，Bessemer Venture 合作伙伴，通用催化剂，门洛风险投资公司 Salesforce Ventures	美国
安全超级智能	2.0亿美元	B轮融资 2025-03-09	30.0亿美元	绿洲资本，Andreessen Horowitz，红杉资本	美国
Groq	1.5亿美元	未公布 2025-02-10	N/A	沙特阿拉伯王国	美国
anthropic	1.0亿美元	公司少数股东 2025-01-22	N/A	谷歌	美国
同构实验室	600.0百万美元	系列A 2025-03-31	N/A	thrive capital，google ventures，alphabet	英国
萨罗尼克	600.0百万美元	系列C 2025-02-18	40亿美元	伊兰·吉尔，安德森·霍洛维茨，通用 catalyst 8VC，咖啡因资本	美国
Lambda	480.0百万美元	系列D 2025-02-19	2.5亿美元	安德拉资本，SGW，1517基金，新月湾 顾问，超微计算机	美国
Apptronik	403.0百万美元	系列A 2025-02-12	N/A	B资本，资本工厂，韩国投资 合作伙伴，ARK Invest，Atinum Investment	美国
CoreWeave	350.0百万美元	公司少数股东 2025-03-10	N/A	OpenAI	美国

生成式AI基础设施建设

深度求的R1模型近期问世，与前沿模型相比，其能以更低成本提供先进性能，为生成式人工智能（GenAI）格局带来了重大转变。突然间，需要巨额投资的大规模数据中心不再是生成式人工智能进步的制约因素。事实上，行业专家们认为，R1开启了领先模型训练和部署对资源要求显著降低的时代，有可能终结生成式人工智能基础设施的万亿美元军备竞赛。

然而，专家和行业参与者将DeepSeek的效率提升视为推动更激进GenAI部署的催化剂，计算能力和相关基础设施正迅速成为本世纪最重要的资源之一。这是因为数百万台服务器持续运行以处理支撑人工智能及相关技术的底层模型和ML应用。这就是为什么据称OpenAI首席执行官Sam Altman在2030年之前讨论了为GenAI投资建立一笔700亿美元基金的原因。

根据麦肯锡2025年的研究，预计到2030年，用于支持人工智能相关数据中心容量需求的资本投资将介于约3万亿美元至8万亿美元之间。其中，预计约15%的投资将投向建设者，用于土地、材料和场地开发；25%将投向发电和输电、冷却以及电气设备的能源供应者；而最大份额的60%将投向技术开发商和设计者，他们为数据中心生产芯片和计算硬件。

这项投资是由通用人工智能的广泛采用、跨各行业的AI赋能应用集成以及企业竞争所推动的

为了构建有竞争力的基础设施。此外，政府现在越来越多地大力投资人工智能基础设施，以改善其安全和经济态势，以及技术独立。

近期的DeepSeek R1模型的问世，其相较于前沿模型能够以更低成本提供先进的性能，为GenAI格局带来了重大转变。 Suddenly, 需要大量投资的超大规模数据中心不再是GenAI发展的限制因素。

投资挑战：

技术不确定性：中断或

模型架构的进步，包括计算利用效率的提升，可能导致预期硬件和能源需求的减少。

供应链约束： 劳动力短缺、供应链瓶颈和监管障碍可能会延误电网连接、芯片供应和数据中心的扩张，从而减缓整体人工智能的采纳和创新。

地缘政治紧张局势： 特朗普政府最近的关税和技术出口管制给计算能力需求带来了不确定性，可能影响基础设施投资和人工智能增长。

治理与投资回报率： 人工智能治理问题，包括偏见、安全和监管，可能会增加额外的复杂性，从而减缓发展。此外，随着人工智能推理预计到2030年成为主要的工作负载，它构成了一个重大的不可预测的成本组成部分。因此，公司可能面临难以展示相关人工智能投资清晰的投资回报率的问题。

市场供需： 全球半导体制造业由少数几家公司控制，抑制了竞争。因此，市场的产能建设能力仍然不足以满足当前需求，同时，人工智能模型训练方法和工作量的变化也使得预测未来特定芯片的需求变得困难。

竞争优势： 为了在日益拥挤的市场中获得竞争优势，公司正在创建定制模型、微调现有模型，或使用检索增强生成（RAG）嵌入来使GenAI系统获取最新的准确的企业信息。这些举措需要大量投入用于训练和部署这些系统。

网格弱点： 为数据中心供电可能因现有电网的薄弱以及处理器密度上升带来的热管理挑战而面临中断。

通过生成式人工智能创造价值

根据2024年9月的一篇BCG文章，生成式人工智能转型相比基准可以带来1至2个百分点的收入增长和8%至12%的成本降低。一项更近期的2025年BCG研究发现，生成式人工智能可以解决30%至50%的行业无关IT成本，从而在技术职能方面触发高达10%的潜在节约。一项2024年的麦肯锡研究提供了定量视角，估计生成式人工智能将在客户服务、研发、制造、供应链和采购等职能方面带来1.4万亿至2.6万亿美元的节约机会。

谷歌云又一项研究涉及2500多名美国公司高管，这些公司年收入超过1000万美元。该研究发现，实施生成式人工智能（GenAI）的公司中，86%的总收入增加了6%以上。此外，77%的公司见证了其潜在客户和客户获取的改善，45%的公司员工生产力至少翻了一番，56%的公司报告了网络安全状况的改善，71%的公司表示能够更快地解决问题。

此外，越来越强大的小型LLMs的出现已经将执行GPT-3.5水平系统的推理成本在2022年11月至2024年10月之间降低了280倍。在硬件层面，成本每年降低了30%，而能源效率每年提高了40%。开放权重的模型正在缩小与封闭模型的差距，在一年内将某些基准测试上的性能差异从8%减少到仅1.7%。所有这些趋势正在迅速降低先进人工智能的门槛。

GenAI工具可以通过自动化重复性任务、准确预测下一代支出、创建各种运营场景的详细模拟来为企业节省成本

监控实时支出、分类战略性支出并提升供应商管理。例如，基于GenAI的聊天机器人可以处理大量客户咨询，降低运营成本。GenAI模型可以分析历史支出数据并为未来支出提供准确预测，使企业能够更有效地管理预算、避免不必要的成本并有效配置资源。

生成式人工智能模型还可以创建详细的模拟，生成逼真的场景，帮助企业在将决策实施到现实环境中之前，在虚拟环境中测试各种决策的影响。最后，该技术可以监控实时支出并标记偏差，以防止成本超支。

随着越来越强大的小型LLM的出现，一个运行在GPT-3.5水平上的系统在2022年11月至2024年10月期间的推理成本降低了280倍。在硬件层面，成本每年降低了30%，而能源效率每年提高了40%。

供应商格局

表 13：2024 年起重要 AI 模型和数据集发布

日期	名称	分类	创作者 (s)
2024年9月11日	笔记本LM播客工具	文本到播客	谷歌实验室
2024年9月12日	o1-preview	语言，数学，生物学	OpenAI
2024年9月17日	NVLM (D, H, X)	愿景，语言	英伟达
2024年9月19日	Qwen2.5	LLM	阿里巴巴
2024年10月16日	部长	LLM	Mistral
2024年10月22日	人机使用	自主能力	anthropic
2024年10月28日	苹果智能	iPhone功能	苹果
2024年12月3日	Nova Pro	多模态	亚马逊
2024年12月11日	Gemini 2	LLM	谷歌深度学习研究院
2024年12月12日	Sora	文字转视频	OpenAI
2024年12月13日	全球MMLU	数据集	协同
2024年12月20日	o3 (beta)	多模态	OpenAI
2024年12月27日	深寻-V3	LLM	深寻
2025年2月3日	深度研究	多模态	OpenAI
2025年2月5日	Gemini 2.0 闪存	LLM	谷歌深度学习研究院
2025年2月6日	猫	LLM	Mistral
2025年2月18日	grok-3	聊天机器人	xAI
2025年2月24日	Claude 3.7十四行诗	LLM	anthropic
2025年2月27日	GPT 4.5	LLM	OpenAI
2025年3月4日	爱亚视觉	多模态	协同
2025年3月25日	Gemini 2.5 Pro	LLM	谷歌深度学习研究院
2025年4月16日	o3, o4-mini/高	多模态	OpenAI
2025年4月17日	Gemini 2.5 快闪版	LLM	谷歌深度学习研究院
2025年4月22日	Gemini Veo 2	大语言模型，文本到视频	谷歌深度学习研究院
2025年4月30日	Nova Pro	基础模型	亚马逊
2025年5月1日	梅卢姆	LLM	JetBrain
2025年5月30日	我看到了3	大语言模型，文本到视频	谷歌深度学习研究院

表14：主要供应商：生成式人工智能

供应商	国家	专长
OpenAI	美国	开发先进语言模型，包括GPT，用于生成式人工智能任务。
微软	美国	提供人工智能工具和云服务，专注于企业人工智能解决方案。
AWS	美国	提供基于云的人工智能服务，包括机器学习和NLP模型。
谷歌	美国	研发人工智能和机器学习技术，包括像BERT这样的语言模型。
anthropic	美国	专注于开发安全且可控的AI模型，尤其强调对齐。
AI21 labs	以色列	为企业构建先进的语言模型和生成式AI解决方案
协同	加拿大	开发专注于自然语言处理的私有、可扩展的AI解决方案
阿里云	中国	提供AI服务和云计算基础设施，包括NLP模型
百度	中国	专注于自然语言处理和自主系统的AI解决方案开发
Alpha Alpha	德国	专精于先进人工智能语言模型的研究与开发
Meta	Llama	一个可根据用户需求进行定制和部署的开源AI模型。
Hugging Face	美国	专注于提供开源人工智能模型和自然语言处理工具。
合成	英国	提供一套强大的工具，用于快速、专业的视频创作。
指南	美国	帮助团队快速轻松地创建和共享基于视频的文档。
深寻	中国	专注于开发大型语言模型（LLMs）。
困惑AI	美国	结合传统网络搜索与大型语言模型，提供对话式答案 完成并附有来源引用。

附录
领先 GenAI 供应商简介

OpenAI

成立年份： 2015
总部： 美国
员工数量： 5,328 (2025年4月)
CEO: 山姆·奥尔特曼
收入： 10亿美元（2025年）

OpenAI，成立于2015年12月，总部位于加利福尼亚州旧金山，是一家领先的私营研究机构，专注于开发人工智能产品。该公司由CEO山姆·奥特曼领导，布莱特·泰勒担任董事长，格雷格·布罗克曼担任总裁。

自成立以来，OpenAI凭借颠覆性产品取得了重大进展。这包括GPT系列（如GPT-3和GPT-4等先进语言模型）、DALL·E（一个从文本提示中创建图像的工具）、OpenAI Codex（为代码编写工具提供支持）以及ChatGPT（一种参与类人类对话的对话式人工智能）。凭借这些创新，OpenAI正在改变科技、娱乐和教育等行业，同时继续推动人工智能发展，朝着使社会各领域受益的未来迈进。

OpenAI成功进行了10轮融资，累计筹集资金219亿美元，最新一轮融资是在2024年10月3日进行的债务融资。该公司获得了39家投资者的支持，其中花旗集团和摩根大通近期的贡献尤为突出。

微软

成立年份： 1975
总部： 美国
员工数量： 228,000（2024年6月）
CEO: 萨提亚·纳德拉
收入： 24.51亿美元（2024年6月）
收入智能云： 105.4亿美元（2024年6月）

1975年4月4日由比尔·盖茨和保罗·艾伦创立，微软是一家全球科技领导者，总部位于华盛顿州雷德蒙德。作为在纳斯达克上市的公司，股票代码为MSFT，微软是信息技术行业的关键参与者。公司多样化的产品和服务的范围包括软件开发、消费电子产品、云计算、社交网络和视频游戏。知名品牌和服务包括Windows、Microsoft 365、Azure、Xbox、LinkedIn和GitHub。

在首席执行官萨提亚·纳德拉的领导下，微软在2024年实现了显著增长，收入达2451亿美元，净利润达881亿美元。该公司在全球范围内运营，拥有领英、GitHub和 Skype 技术等子公司，以及超过228000名员工。

截至2025年初，微软已收购256家公司。 đáng chú ý là việc mua lại Activision Blizzard trị giá687亿美元以加强其游戏部门，以及在2023年以1.9亿美元收购Fungible以扩展其云计算和人工智能能力。其他例子包括收购Xandr和Ally.io，进一步增强了其在广告技术和劳动力解决方案领域的投资组合。

AWS

成立年份： 2002，(云计算—2006)
总部： 美国
员工数量： 155.6万（2024年12月）
CEO: 马特·加曼
收入： 60.38亿美元（2024年12月）
收入AWS： 107.6亿美元（2024年12月）

亚马逊网络服务（AWS），成立于2002年，是亚马逊的主要子公司之一，已成为云计算和网页服务行业的领军企业。2023年，AWS创造了

90.8亿美元收入和24.6亿美元营业利润。该部门以其提供的综合云解决方案而闻名，包括计算能力、存储和机器学习服务。

亚马逊有数家子公司，帮助扩展其服务和能力。例如，安纳普尔纳实验室开发云计算定制芯片，提升了亚马逊的硬件能力。亚马逊元素提供视频处理工具，帮助媒体公司流式传输高质量内容。NICE软件提供数据分析和决策解决方案。Wickr是一个安全的通讯平台，加强了亚马逊对安全和隐私的专注。

截至2025年1月，公司总共进行了145项投资，其中97项为领投。公司主要投资于人工智能、云计算和技术初创企业。此外，公司还收购了9家公司，最新的收购是2021年6月25日收购的Wickr。

-

成立年份： 1998，(谷歌云平台 – 2008)
总部： 美利坚合众国
员工数量： 183,323 (2024年12月) Google Cloud; 54,000 (2024年1月)
CEO： Sundar Pichai，(谷歌云——托马斯·库里克)
收入： 35亿美元（2024年12月）
收入 Google Cloud： 43.2亿美元（2024年12月）

1998年9月4日，由拉里·佩奇和谢尔盖·布林创立的谷歌是字母表公司的子公司，总部位于加利福尼亚州山景城。该公司的产品和服务涵盖多个类别。其基于网络的工具包括谷歌搜索、谷歌地图和谷歌驱动等搜索引擎，以及像 Gmail 和谷歌文档这样的生产力工具。该公司还通过谷歌广告和 AdSense 等平台提供广告服务，同时提供像 Google Meet 和 Google Voice 这样的通讯工具。

在硬件方面，谷歌生产像Pixel智能手机、谷歌Nest智能家居产品这样的设备

以及 Fitbit 智能穿戴设备。此外，谷歌在云计算和人工智能方面进行了大量投资，拥有像 Google Cloud 和 TensorFlow 这样的产品。

截至2025年初，谷歌已进行了266次收购，最近的一次是在2024年6月5日收购了Cameyo。该公司经常在人工智能、云计算和安全等领域收购组织。 đáng chú ý的收购包括2022年的Mandiant（价值54亿美元）、Raxium（价值10亿美元）和Alter（价值1亿美元）。此外，该公司还进行了306次投资，最近一次是在2024年12月24日对Hazeltree进行投资，Hazeltree是一家财富管理解决方案提供商。

anthropic

成立年份： 2021
总部： 美国
员工数量： 1,097 (2025)
CEO: 达里奥·阿莫迪
收入： 14亿美元（2025）

Anthropic是一家于2021年由前OpenAI员工Dario和Daniela Amodei在美国创立的人工智能初创公司。该公司致力于提升人工智能的安全性和可靠性，特别是在大型语言模型（LLMs）方面。Anthropic的旗舰产品是Claude，这是一系列旨在优先考虑安全、透明和与人类价值观一致的AI模型。

多年来，公司获得了大量投资，包括来自亚马逊的40亿美元和来自谷歌的20亿美元，突显了其在人工智能领域的日益增长的影响力。截至2025年1月，公司在12轮融资中共筹集了137亿美元。其他知名投资者包括Ventioneers、曼哈顿风险投资公司（MVP）、Stackpoint风险投资公司、联盟全球合作伙伴和TeleSoft合作伙伴。

AI121实验室

成立年份： 2017
总部： 以色列
员工数量： 268（2025年5月）
CEO: 奥里·戈申

收入：3500万美元（2025年5月）

由阿姆农·沙舒亚、约阿夫·沙赫姆和奥里·戈申于2017年创立，AI21 Labs致力于创建先进的AI系统和模型，帮助企业将在真实世界的应用中使用生成式AI。多年来，该公司推出了各种产品，包括用于增强写作的生成式AI工具Wordtune Spices，以及用于构建各种应用和服务的开发者平台AI21 Studio。

截至2025年1月，AI21 Labs共进行了8轮融资，累计筹集资金326.5百万美元。其中最大的一笔投资来自C轮融资，2023年8月筹集了155百万美元，2023年11月筹集了53百万美元。该公司吸引了包括康卡斯特风险投资、英特尔资本、三星NEXT和Pitango VC在内的知名投资者的资金，仅列举部分。

-

成立年份： 2019
总部： 加拿大
员工数量： 796（2025年5月）
CEO: 爱登·戈麦斯
收入： 3500万美元（2025年5月）

协同是一家优先考虑数据安全的人工智能公司，致力于创建可扩展和私有的AI解决方案，旨在解决实际商业问题。它在多个行业开发AI解决方案，包括金融服务、制造业、能源与公用事业以及医疗保健。在金融领域，协同通过自动化任务、改善风险管理和提供实时洞察来提高效率。在医疗保健领域，它通过连接数据源、加速研究和优化工作流程来改善患者护理，从而实现更好的患者成果。

2025年1月，加拿大皇家银行与该公司合作，开发面向金融行业的生成式人工智能产品，具体针对风险管理和安全。

该公司共获得了 11 亿美元融资，分 7 轮进行，最新一轮是在 2024 年 12 月 6 日获得的一项拨款。该公司是

得到34家投资者的支持，包括加拿大政府和英伟达。此外，Cohere已经进行了两项投资，最新的一项是在2024年7月8日对Questflow进行的150万美元投资。

阿里云

成立年份： 1999，(阿里云 – 2009)
总部： 中国
员工数量： 124,320 (2025年3月)
阿里云： 4,656 (2025)
CEO : 刘迪武
收入：1373亿美元（2025年3月）
收入云智能集团： 163亿美国美元 (2025年3月)

成立于2009年，阿里云是全球领先的云计算服务提供商，同时也是阿里巴巴集团的子公司。公司提供多样化的产品和解决方案，以满足各种业务需求。其产品包括高性能虚拟服务器弹性计算服务（ECS）、对象存储服务（OSS）和可扩展计算能力的弹性GPU服务。其他值得注意的产品包括用于安全的Web应用防火墙（WAF）、用于无缝连接的云企业网络（CEN）以及用于团队协作的钉钉企业版。

在最近推出的解决方案中包括用于网络安全的安全访问服务边缘（sase）、智能媒体服务（ims）以及用于人工智能模型开发的阿里云模型工作室。此外，阿里云还提供专门服务，如用于选择数据库的ApsaraDB以及用于通信的短信服务（sms）。

截至2025年1月，阿里云已进行9次投资，最近一次是在2024年12月12日对中国云操作系统提供商Sealos的投资。此外，该公司还收购了3家公司，主要集中在网络安全和技术相关领域。迄今为止，通过两轮融资，它总共筹集了12亿美元资金，主要来自阿里巴巴集团。

2000

成立年份：

总部： 中国

员工数量： 35,900（2024年12月）

CEO: 李彦宏

收入：182亿美元（2024年12月）

收入云服务： 30亿美元（2024年12月）

2000年1月18日成立，百度是一家领先的中文跨国公司，提供跨各行业的广泛产品和服务，专注于互联网服务、人工智能和云计算。其移动生态系统包括百度App、好看（短视频平台）和全民（快进视频应用），而百度百科全书和百度知道等知识型平台提供专家驱动内容和用户生成问答。

该公司也为企业提供人工智能驱动服务，如智能小程序、百佳号和托管页面。此外，百度由 Apollo 平台领导的智能驾驶部门是中国自动驾驶技术市场的领导者。其他服务包括百度健康医疗服务和智能助手平台小度OS。

截至2025年1月，百度已经通过三轮融资总共筹集了2620万美元。该公司由九家投资者支持，其中最近的两家是创投TDF和ePlanet Capital。此外，该公司已进行128次投资、3次多元化投资以及32次退出，其中 đáng chú ý的收购包括2023年2月收购的医疗数据提供商GBI。

Alpha Alpha

成立年份： 2019
总部： 德国
员工数量： 298 (2025)
CEO： 乔纳斯·安德鲁利斯
收入： 147万美元（2025）

阿莱夫阿尔法有限公司是一家由乔纳斯·安德鲁利斯和塞缪尔·韦纳巴赫创立的德国AI先锋初创公司。该公司最近推出了

其下一代控制模型，旨在提供更拟人化的交互并使用大型语言模型解决复杂任务。这些模型配备了增强的自然语言处理功能，非常适合聊天机器人和数字助手等应用。此外，它们还具备可解释人工智能技术，能够追踪和验证人工智能生成的内容。这一突破确保了透明度，减少了幻觉现象，并支持符合即将到来的欧盟法规。通过这些创新，Aleph Alpha结合了高性能、信任和效率，在生成式人工智能领域树立了新的标杆。

截至2025年1月，公司已通过6轮融资总共筹集了5.336亿美元。最大的一轮是2023年11月的B轮融资，筹集了5亿美元，主要投资者包括博世风险投资、创新园区人工智能和施瓦茨集团。之前的轮次包括2021年的A轮融资，筹集了2540万美元，以及同年筹集了583万美元的天使轮。最新的一轮融资是2024年11月的二级市场轮。公司有15位投资者，施瓦茨集团和博尔达主要投资是最近加入的。

Meta

成立年份： 2004
总部： 美利坚合众国
员工数量： 74067 (2024年12月)
CEO: 马克·扎克伯格
收入： 164.5亿美元（2024年12月）

Meta将自己定位为人工智能领域的主要全球投资者，每年投入约400亿美元用于人工智能和虚拟现实研究。这项重大投资突显了他们致力于拓展数字交互边界的承诺。

这项投资的一个关键产品是Meta AI聊天机器人，于2023年末推出，并集成了WhatsApp、Instagram和Facebook Messenger。

这个聊天机器人提供上下文理解、多语言交流、图像生成和实时信息处理，为用户提供对话协助和创意支持。

此外，Meta正积极开发生成式AI工具，如AI图像编辑、用于定制AI角色的AI工作室，以及实验性文本到视频生成。该公司还与Google Cloud合作，提供其Llama模型。

Hugging Face

成立年份： 2016
总部： 美国
员工数量： 534 (2025)
CEO: 克莱门特·德朗格
收入： 468万美元（2025年）

位于纽约市的爱荷华州，提供机器学习开源工具，主要专注于自然语言处理（NLP）。凭借其流行的Transformers库而闻名，该平台使用户能够构建、训练和分享机器学习模型、数据集和项目。

截至2025年1月，该公司已完成7轮融资，总共筹集了3952万美元，最新一轮D轮融资于2024年1月16日完成。该公司获得了包括Bossa Invest和PremjiInvest在内的38家投资机构的支持。此外，该公司进行了4次收购，最近一次是2024年8月8日收购的XetHub。该公司经常收购与机器学习、自然语言处理和AI工具相关的组织。 đáng chú ý的收购包括2024年6月收购的Argilla，2021年12月收购的Gradio以及2017年9月收购的Sam。

成立年份： 2017
总部： 英国
员工数量： 511 (2025年6月)
CEO: 维克托·里帕贝利
收入： 3500万美元（2025年6月）

由Lourdes Agapito、Matthias Niessner、Steffen Tjerrild和Victor Riparbelli创立，Synthesia是英国领先的AI驱动视频创作技术企业。该公司的平台提供一套强大的工具，用于快速、专业的视频创作。例如，其AI视频编辑器和屏幕录制器使内容创作和编辑直接在浏览器中变得容易，同时品牌套件和集中的媒体库有助于保持一致性。此外，用户可以选择超过230个AI虚拟形象（包括个人和自拍形象），并使用包括声音克隆在内的选项，以140多种语言生成旁白。一键翻译、AI配音和字幕的本地化非常简单。内置功能，如审阅工作流、实时协作和版本控制，支持高效团队生产和反馈。

截至2025年，Synthesia已通过七轮融资筹集了3.366亿美元，主要投资来自顶级公司，包括Accel、Kleiner Perkins、NEA、First Mark、Seedcamp和Adobe Ventures。其最新的2025年1月D轮系列融资获得了1.8亿美元，由New Enterprise Associates领投。

指南

成立年份： 2020
总部： 美国
员工数量： 52
CEO: Yoav Einav
收入： 550万美元（2025）

Guidde 是一家成立于 2020 年的加州初创公司，帮助团队快速简便地创建和分享基于视频的文档。借助 Guidde Create 和 Guidde Broadcast 等人工智能驱动工具，用户可以一键捕捉工作流程，生成分步骤指南，并在几分钟内分享专业教程。该平台支持超过 100 种语言，包含智能编辑工具，并通过自动模糊处理确保内容安全。该平台被客户成功、产品和预售团队广泛使用，以改进入职流程、减少支持工单并提高生产力。

该公司通过四轮融资共筹集了2660万美元，以支持其成长和创新。公司的融资历程始于2021年的种子轮融资，其中包括Entrée Capital的支持。2023年，该公司由Norwest Venture Partners领投，成功获得了1160万美元的A轮融资，随后在2025年初，该公司通过A轮融资筹集了1500万美元，投资者包括Qualcomm Ventures和其他投资者。

困惑之AI是一家成立于2022年的美国人工智能搜索引擎公司，由Aravind Srinivas、Denis Yarats、Johnny Ho和Andy Konwinski共同创立。总部位于旧金山，该公司结合传统网络搜索与大型语言模型，提供包含来源引用的会话式答案。用户可以提出后续问题，使体验更像对话而非标准搜索。

深寻

成立年份： 2023
总部： 中国
员工数量： 200 (2025年1月)
CEO: 梁文锋
收入： 2亿美元（2024年）

该平台于2022年12月7日启动，可通过网页、Google Chrome扩展程序以及iOS和Android的手机应用程序使用。它使用微软必应提供搜索结果，并在微软Azure上运行。免费版由OpenAI的GPT-3.5提供支持，而Pro订阅则提供对更高级模型（包括GPT-4）的访问权限。

深Seek 是一家总部位于浙江省杭州市的中国人工智能公司，专注于研发大型语言模型（LLMs）。该公司由梁文峰于 2023 年 7 月创立。2025 年 1 月，该公司发布了自己的 AI 聊天机器人和 DeepSeek-R1 模型，从而获得了广泛关注。

该公司总共完成了五轮融资，累计募集资金6.65亿美元，主要投资者包括英伟达、IVP、软银愿景基金、NEA和贝塞默尔风险投资公司。在2024年12月最近的一轮融资中，该公司筹集了5亿美元，估值达到90亿美元，成为人工智能搜索领域增长最快的初创公司之一。

其产品线包括先进的LLM，如DeepSeek R1、V2和V3，以及专用工具，例如用于编程的DeepSeek Coder、用于解决数学问题的DeepSeek Math和用于视觉与语言任务的DeepSeek VL。这些模型可通过DeepSeek App、DeepSeek Chat和DeepSeek Platform访问，用户可通过API集成该技术。该公司还提供服务性能数据，以帮助用户监控可靠性和使用情况。

2025年5月，该公司宣布与PayPal合作，以实现聊天支付功能，允许美国用户直接通过其AI聊天购买旅游和活动门票，这标志着向AI驱动电子商务迈出的一步。

该公司经历了飞速增长，在发布后的14天内就达到了1亿用户，使其成为历史上增长最快的平台之一。

困惑AI

成立年份： 2022
总部： 美国
员工数量： 1,292 (2025)
CEO: 阿равинд·辛里瓦斯
收入： 1亿美元（2024年）

GENERATIVE AI WEEK

2025年11月11日 – 12日 | 德克萨斯州奥斯汀市乔治城喜来登酒店

北美企业级生成式AI规模化领先活动

生成式人工智能正在改变行业，但如何从实验阶段转向企业级影响？生成式人工智能周2025是北美商业和技术领导者寻求可衡量的商业价值，将生成式人工智能解决方案投入运营、实施和扩展的首届峰会。这个独特的两天活动汇集了来自领先企业的顶尖人工智能和数据高管，分享他们成功将人工智能嵌入运营的见解。无论您是在探索人工智能的采用还是优化现有的战略，#GenAIWeek 2025为您提供所需的路线图、专业知识和实际案例研究，以推动规模的转型。

下载日程

参考文献

遵循的风格：

作者姓氏，名字。“页面标题。”网站名称。月日，年。URL。

1. allen，leanne，hõck，benedikt和clamp，adrian。 通过人工智能驱动的转型创造价值的蓝图。KPMG。 https:// assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2025/02/ intelligent-banking-report.pdf.

2. 人工智能分析AI。 “LLM排行榜 - GPT-4o、Llama 3、Mistral、Gemini以及30多种模型的对比。” https:// artificialanalysis.ai/leaderboards/models.

3. 贝林恩，伊文。 “市场份额最高的生成式AI聊天机器人——2025年5月。”Firstpagesage。2025年5月9日。 https://firstpagesage.com/ reports/top-generative-ai-chatbots/.

4. 巴贾派，拉胡尔，蒂瓦里，阿普兰和萨雷尔，巴里斯。 “边缘人工智能的未来。”德勤。 https://www2.deloitte.com/content/dam/Deloitte/ us/Documents/technology-media-telecommunications/ deloitte-the-future-of-edge-ai.pdf.

5. 巴图拉基斯，米诺斯和文图里尼，弗朗切斯科。 “生成式人工智能对创意产业的影响，以及我们必须建立起来的伦理和治理。”世界经济论坛。2025年1月21日。 https://www. weforum.org/stories/2025/01/the-impact-of-genai-on-the- creative-industries/.

6. BCG。 “数字和人工智能将如何重塑2025年的医疗保健。”2025年1月。 https://web-assets.bcg.com/8c/f8/ ae51ffb44ca59cb8abd751940441/bcg-how-digital-and-ai- solutions-will-reshape-health-care-in-2025.pdf.

7. 贝尔奇奇，伊万和斯特雷克，科尔。 “2025年的人工智能代理：期望与现实的对比。”IBM。2025年3月4日。 https://www.ibm.com/think/insights/ ai-agents-2025-expectations-vs-reality.

8. 鲍比尔，让-弗朗索瓦，查特吉，阿比克，和埃贝林，鲁斯。
“首席信息官在人工智能价值创造中的作用。”BCG。2025年2月19日。
https://www.bcg.com/publications/2025/cios-role-in-ai- transformation-and-productivity.

9.商业 wire。 “首项现实世界多中心研究显示：基于生成式AI的心理健康治疗优于常规治疗。”2025年3月10日。 https:/www.businesswire.com/news/ home/20250310848349/en/First-Real-World-Multisite- Study-Shows-GenAI-Powered-Mental-Health-Treatment- Outperforms-Standard-of-Care.

10. 普华永道。 “德勤加速企业采用智能AI，为行业赋能NVIDIA。”2025年3月19日。 https://www. capgemini.com/news/press-releases/capgemini-accelerates- enterprise-adoption-of-agentic-ai-for-industries-with-nvidia/.

11. CB Insights. “人工智能的状态。” https://www.cbinsights.com/reports/ CB-Insights_Artificial-Intelligence-Report-2024.pdf.

12. Cengage Group. “高等教育中的通用人工智能 - 随着快速转型，积极情绪逐渐积累。”2025年4月7日。 https://www.cengagegroup. com/news/perspectives/2025/genai-in-higher-education— positive-sentiment-builds-with-rapid-transformation/.

13. 柴廷，恩维尔。 “自主人工智能与个性化医疗的未来。” Ciklum。2025年5月19日。 https://www.ciklum.com/ resources/blog/future-of-personalized-healthcare.

14.钱德拉塞卡兰，阿Run。 “关于 GenAI 未来发展的 3 个大胆且可执行的预测。” Gartner。2024 年 4 月 12 日。 https://www.gartner. com/en/articles/3-bold-and-actionable-predictions-for-the- future-of-genai.

15. 智华。 “交通和物流行业的生成式人工智能。” https://www.cognizant.com/en_us/industries/ documents/generative-ai-in-transport-logistics-industry.pdf.

16. Coshow，Tom，和高，阿诺德。 “2025年顶级战略技术趋势：智能代理式AI。”Gartner。2024年10月21日。 https://www. gartner.com/doc/reprints.

17. 数据跨度。 “制造业中的生成式人工智能：7个实际应用案例。”2024年12月27日。 https://www.dataspan.ai/blog/7-use-cases- of-genai-in-manufacturing.

18. 迪尔梅格尼，杰姆。 零售中的生成式人工智能：2025年的用例、实例与优势。AI Multiple Research。2025年5月5日。 https:// research.aimultiple.com/generative-ai-in-retail/.

19. 迪尔梅加尼，杰姆。 “2025年生成式AI在教育领域的10大应用场景。”AI多重研究。2025年5月6日。 https://research. aimultiple.com/generative-ai-in-education/.

20. 迪尔梅加尼, 塞姆。 “自主AI：2025年的8种应用场景与真实案例”。AI多重研究。2025年4月28日。 https://research. aimultiple.com/agentic-ai/.

21. 为您完成。 “AI模型比较：2025年哪个AI称霸？” https://doneforyou.com/ai-model-comparison-which-ai-reigns-supreme-in-2025/.

22. 德鲁特。 “AI基础设施的未来：2025年需要关注的趋势。”2025年2月19日。 https://drut.io/drut-blog/f/the-future-of-ai- infrastructure-trends-to-watch-in-2025.

23.达德利，布莱恩，和德尔马斯特罗，托马斯。 “下一场前沿：代理式人工智能的崛起。”亚当斯街合作伙伴。2025年3月12日。 https:// www.adamsstreetpartners.com/insights/the-next-frontier-the- rise-of-agentic-ai/.

24. 多克，维塔利。 “生成式人工智能：企业降低成本和提升销售额的实用方法。”立即获取Dynamiq。2025年2月6日。 https://www.getdynamiq.ai/post/generative-ai-practical-ways-for-enterprises-to-cut-costs-and-boost-sales。

25. 内德曼，马克和布里恩，凯瑟琳。 “人工智能在创意产业中的应用：强化而非取代人类创造力——电视与电影。”阿利克斯合伙公司。2025年1月10日。 https://www.alixpartners.com/insights/102jsme/ai-in-creative- industries-enhancing-rather-than-replacing-human- creativity-in/.

26. 欧空局自动化。 “生成式人工智能赋能智能制造。”2025年3月27日。 https://www.esa-automation.com/en/ generative-ai-powers-smart-manufacturing/.

27. 费尔南德斯，胡安昆。 “领先的生成式AI公司。”IOT Analytics。2025年3月4日。 https://iot-analytics.com/leading- generative-ai-companies/.

28. 加西亚，西里尔，沙普蒂奥，文森特和安德里隆，弗洛朗。
“开发可持续的生成式人工智能。”普华永道。 https://www. capgemini.com/dk-en/wp-content/uploads/sites/7/2025/02/ Final-Web-Version-Report-Sustainable-Gen-AI-2-1.pdf.

29. Gartner. “Gartner 专家回答您企业的顶级生成式AI问题。” https://www.gartner.com/en/ topics/generative-ai.

30. 高斯，蒂姆。 “超越自动化：生成式人工智能如何重塑制造业。”德勤。2025年4月22日。 https://www2.deloitte. com/us/en/blog/business-operations-room-blog/2025/ generative-ai-in-manufacturing.html.

31. 高盛。 “生成式人工智能：支出过多，收益过少？”2024年6月25日。 https://www.goldmansachs.com/images/ migrated/insights/pages/gs-research/gen-ai--too-much- spend,-too-little-benefit-/TOM_AI%202.0_ForRedaction.pdf.

32. 高夫，乔纳森·D. “十大自主人工智能示例及应用案例。” Converge Technology Solutions。2025年5月6日。 https:// convergetp.com/2025/05/06/top-10-agentic-ai-examples-and- use-cases/.

33. 雷奇科克，拉里，加尔萨，毛里西奥，和 克劳利，埃伊琳。
“生成式人工智能改变交通：新兴技术前沿的启示。”德勤。2024年11月21日。 https:// www2.deloitte.com/us/en/insights/focus/transportation/ai-in- transportation.html.

34. 赫尔曼，法比安。 “生成式人工智能将西门子的预测性维护解决方案提升到新的水平。”西门子。2024年2月5日。 https://press.siemens.com/global/en/ pressrelease/generative-artificial-intelligence-takes-siemens- predictive-maintenance-solution-next.

35. Intellias. 零售中的生成式人工智能：用例、示例和实施。2025年3月31日。 https://intellias.com/ generative-ai-in-retail/.

36. ISG. “企业对生成式人工智能的支出预计将在 2025 年增长 50%，随着重点从效率转向专业能力。”2024 年 9 月 23 日。 https://ir.isg-one.com/news-market-information/press- releases/news-details/2024/Enterprise-Spending-on-GenAI- Expected-to-Rise-50-in-2025-as-Focus-Shifts-From-Efficiency- to-Expertise/default.aspx.

37. 科尔，贾格里特。 使用代理式人工智能构建聊天机器人。Xenonstack。2025年4月3日。 https://www.xenonstack.com/blog/chatbot-agentic-ai.

38. 肯纳，肖恩·迈克尔。 “2025年前25佳大型语言模型。”Techtarget。2025年1月31日。 https://www.techtarget.com/ whatis/feature/12-of-the-best-large-language-models.

39. 科罗廖夫，玛丽亚。 随着人工智能的扩展，基础设施挑战浮现。首席信息官。2024年10月23日。 https://www.cio.com/article/3577669/as- ai-scales-infrastructure-challenges-emerge.html.

40. 库杜马拉，阿迪提亚，以色列，亚当和萊拉，赛伊。 “从人工智能与生成式人工智能中实现生命科学领域的变革价值。”德勤。 https://www2.deloitte.com/content/dam/Deloitte/us/ Documents/us-realizing-transformative-value-from-AI-GenAI- in-life-sciences-032124.pdf.

41. 拉达吉，纳文。 “2025年企业需要拥有自己生成式人工智能平台的前五大原因。” Quantiphi。2025年2月4日。 https:// quantiphi.com/top-5-reasons-why-enterprises-need-their-own-

42. 劳顿，乔治。 “2025年8大类生成式人工智能工具。”TechTarget。2025年1月7日。 https://www.techtarget.com/ searchenterpriseai/tip/Top-generative-AI-tool-categories.

43. 李惠仪。 赋能未来制造：面向2025年及以后的AI与运营技术。IDC。2025年2月10日。 https://blogs. idc.com/2025/02/10/empowering-future-manufacturing-ai-and-operational-technologies-for-2025-and-beyond/.

44.洛多切,马特,和莫兰,梅根. “Gartner预测，全球生成式AI支出将在2025年达到6440亿美元。” Gartner。2025年3月31日。 https://www.gartner.com/en/newsroom/press- releases/2025-03-31-gartner-forecasts-worldwide-genai- spending-to-reach-644-billion-in-2025.

45.路易森特，伊达。 “医疗保健中的生成式人工智能：用例、益处与挑战。”约翰·斯诺实验室。2025年5月22日。 https://www. johnsnowlabs.com/generative-ai-healthcare/.

46. 曼尼亚尔，斯维塔。 “人工智能将如何变革2025年生命科学。”Pharma Phorum。2025年1月7日。 https://pharmaphorum.com/ digital/人工智能如何改变2025年生命科学。

47. 马克汉姆，艾莉丝贝尔。 “如何西岩公司利用生成式人工智能提升内部审计。”德勤。2024年3月23日。 https://deloitte.wsj. com/riskandcompliance/how-westrock-harnessed-genai-to- enhance-internal-audit-f0926363.

48. 马丁，卡洛斯·帕多，兰布，杰西卡和达哈布，阿明。

“医疗保健中的生成式人工智能：当前趋势和未来展望。”麦肯锡。2025年3月26日。 https://www.mckinsey.com/industries/ healthcare/our-insights/generative-ai-in-healthcare-current- trends-and-future-outlook.

49. 米格里奥，安德烈亚·德·，焦文，卡洛，和豪泽，

斯蒂芬妮。 “以创新为基石：ING如何利用生成式AI以人为本。”麦肯锡。 https://www.mckinsey.com/ industries/financial-services/how-we-help-clients/banking- on-innovation-how-ing-uses-generative-ai-to-put-people- first.

摩根森，保罗。 “零售2025：6大趋势重塑购物未来。”WNS。 https://www.wns.com/perspectives/ articles/retail-2025-6-trends-re-defining-the-future-of- shopping.

51.诺夫辛格，杰西，帕特尔，马克和萨奇德瓦，潘卡杰。

计算成本：一场7000亿美元的数据中心扩展竞赛。麦肯锡。2025年4月28日。 https://www.mckinsey.com/industries/ technology-media-and-telecommunications/our-insights/ the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data- centers.

52. ntt data. “制造业的‘完整革命’：NTT DATA研究报告揭示GenAI的变革潜力及其对核心功能的影响。”2025年5月1日。 https://www.nttdata.com/global/ en/news/press-release/2025/may/050100.

53.佩特鲁克，马克西姆。 “如何比较来自 OpenAI、Google 等的 AI 模型。”We Soft You。2024年7月1日。 https://wesoftyou. com/ai/how-to-compare-ai-models/.

54. 普拉特，玛丽·K. “10 个真实世界的自主人工智能示例和用例。”TechTarget。2025年3月7日。 https://www.techtarget.com/ searchenterpriseai/feature/Real-world-agentic-AI-examples- and-use-cases.

55. 普雷特，玛丽·K. “生成式人工智能的未来：2025年需要关注的10大趋势。” TechTarget。2025年2月4日。 https://www.techtarget. com/searchenterpriseai/feature/The-future-of-generative-AI- Trends-to-follow.

。和西罗尼，保罗，كر拉马莫尔蒂，尚.56 “2025年银行业和金融市场全球展望。”IBM。 https://www.ibm.com/ downloads/documents/us-en/115dcc7faf363f21.

57. 里默、斯蒂内、科波拉、马特奥和罗格，尤尔根。 “对银行而言，人工智能的审判已至。”BCG。2025年5月。 https://web-assets.bcg. com/3e/6f/9dfa63434eb7a00e1cf1cdcb3754/for-banks-the-ai- reckoning-is-here-may-2025.pdf.

58. 罗宾斯，雅各布。 “ Meet the 10 most active investors in generative AI. ” Pitchbook. June 12, 2024. https://pitchbook.com/news/articles/ top-generative-ai-vc-investors-list.

59. 赛，莫古鲁。 “ChatGPT与Gemini AI Pro与Llama与Copilot与DeepSeek R1。”Medium。2025年2月7日。 https://medium.com/@ saimoguloju2/chatgpt-vs-gemini-ai-pro-vs-llama-vs-copilot- vs-deepseek-r1-9ce268b3492d.

60. SGU. “主流人工智能模型的比较：DeepSeek AI、ChatGPT、Gemini 和 Perplexity AI.” 2025年2月7日。 https:// sgu.ac.id/a-comparison-of-leading-ai-models-deepseek-ai- chatgpt-gemini-and-perplexity-ai/.

61. 夏尔马，苏拉杰。 “交通领域生成式人工智能的九种提升方式。”Nextgen Invent https://nextgeninvent. com/blogs/generative-ai-in-transportation-enhancing-the- sector/.

62. Shubham. “自主智能系统导论：Agent型人工智能”。学习Open CV。2025年2月11日。 https://learnopencv. com/agentic-ai/.

。亚历山大，和 雅伊，拉瑞娜，كر唐纳利，亚历克斯，苏哈列夫斯 .63
“人工智能的现状：组织如何重塑以获取价值。”麦肯锡。2025年3月12日。 https://www.mckinsey.com/ capabilities/quantumblack/our-insights/the-state-of-ai.

64. “2024年生成式人工智能融资额创新高，得益于 **Struta，Luri。**
基础设施利益。” SP Global。2025年1月22日。 https:// www.spglobal.com/market-intelligence/en/news-insights/ articles/2025/1/genai-funding-hits-record-in-2024-boosted- by-infrastructure-interest-87132257.

65. Talkai信息。 “最佳语言模型的比较分析：ChatGPT、Gemini、Claude 和 Llama。” https://talkai. info/blog/comparative_analysis_of_chatgpt_gemini_claude_llama/.

66. 图利，蒂姆，雷德菲尔，乔夫，和肖，德雷克。 “2024：企业生成式人工智能现状。”门洛帕克。2024年11月20日。 https://menlovc.com/2024-the-state-of-generative-ai-in-the- enterprise/.

67。泰隆。 在多租户环境中设计可组合的GPU工作空间：敏捷AI基础设施的蓝图。2025年3月31日。 https://blog.tyronesystems.com/designing- composable-gpu-workspaces-in-multi-tenant-environments- a-blueprint-for-agile-ai-infrastructure/.

68. Vals AI. “GPQA基准。”2025年3月26日。 https://www.vals. ai/benchmarks/gpqa-03-26-2025.

69. 维尔桑特。 “人工智能在创意产业：我们所知的创造力的终结？”2025年4月22日。 https://www.virtasant.com/ai-today/ai- in-creative-industries-end-of-creativity-as-we-know-it.

70.沃伦，扎克，阿伯特，麦克，和利奇。 “2025 年专业服务生成式人工智能报告。”汤森路透。 https:// www.thomsonreuters.com/content/dam/ewp-m/documents/thomsonreuters/en/pdf/reports/2025-generative-ai-in- professional-services-report-tr5433489-rgb.pdf.

71. 泽默, 维克特里. 德勤全球2025年预测报告：生成式人工智能：为技术、媒体和电信业塑造变革性未来。德勤。2024年11月19日。 https://www.deloitte.com/global/en/about/press-room/ deloitte-globals-2025-predictions-report.html.