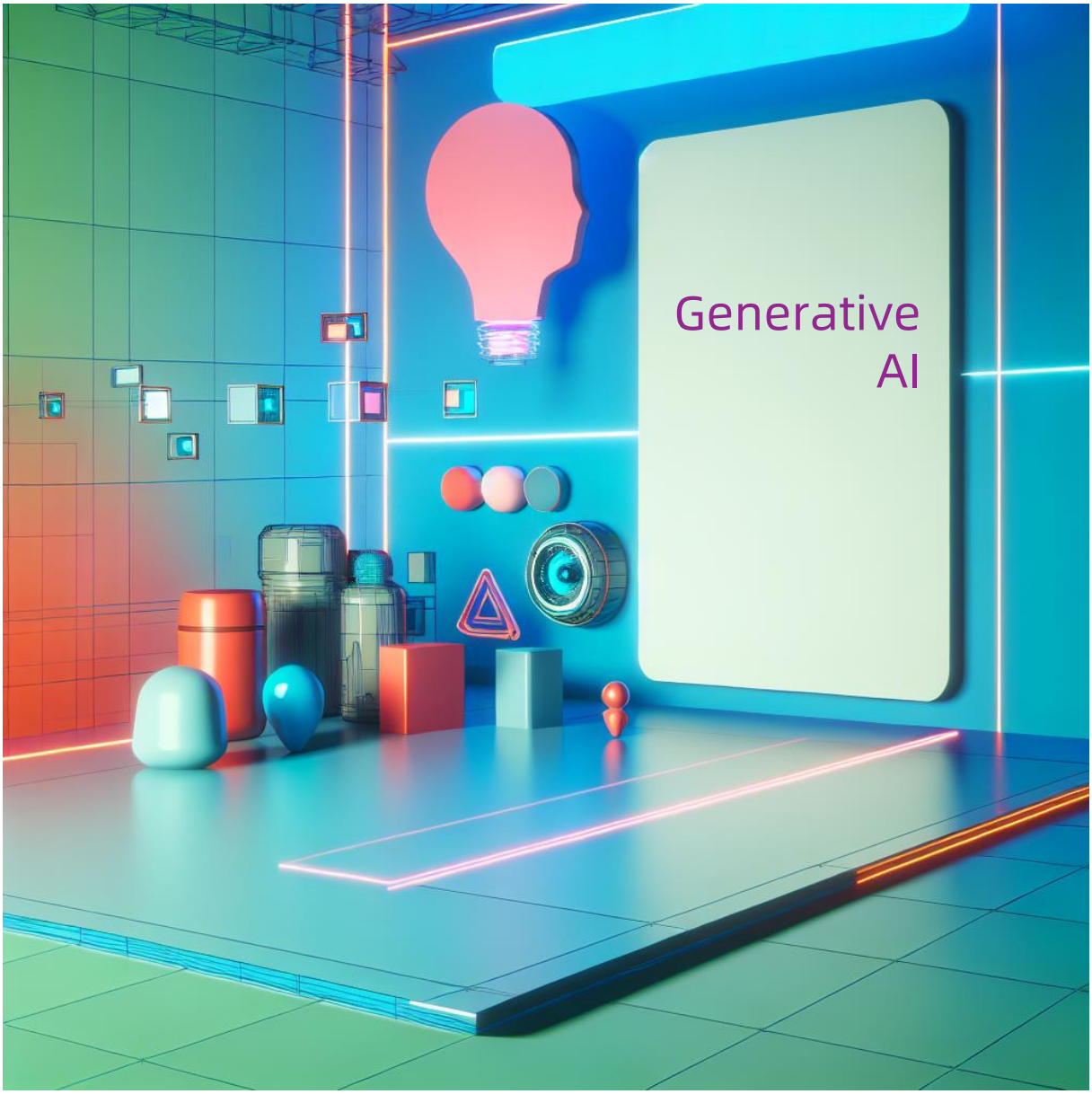
MIT 科技评论
Technology Review

生成式人工智能部署：迎接面向企业和大众的人工智能新时代

——中国AI大模型先进应用案例

Dec. 2023



Preface

2023年以来，中国已有近200家企业入局发布AI大模型。一边以百度、阿里巴巴、腾讯、华为等为代表的科技大厂投身其中，另一边科大讯飞、商汤科技等AI企业也纷纷加入。互联网大佬们更不畏惧“从0到1”，搜狗创始人王小川、美团联合创始人王慧文、创新工场董事长李开复等人高调入场生成式人工智能方向。与以往不同的是，这次的军备竞赛不仅仅在科技巨头和企业中展开，中国高校、新研机构也纷纷推出自己的大语言模型，清华大学、复旦大学、浙江大学、北京智源人工智能研究院、上海人工智能实验室等一众一流高校和机构都加入了这一激烈的竞争。以「大语言模型」作为典型代表的生成式人工智能，一时间成了全球科技公司和科研机构的“角逐场”。

毫无疑问，过去一年生成式人工智能已经成为人们讨论的中心话题，行业有「最」关心的话题、评论这是「最」卷的行业、企业开展「最」激烈的竞争……同时我们也观察到，正在「变」务实的大众和企业、正在「变」快的立法速度……时至今日，大众和企业更加关心基础模型、行业模型的实际落地应用，以及它将如何以及在多大程度上影响企业运营，企业管理者正在思考最合适自身的布局策略、选型策略、企业内部应用部署的速度、大模型的产品能力等切实问题。

我们预计2024年开始，生成式人工智能将引发真正的企业级人工智能开端。2023下半年，全球企业确实在投资、试验生成式人工智能技术，中国企业也在关注生成式人工智能部署策略，但经过调研发现生成式人工智能在企业的部署率仍处于低位，“共识”与“谨慎”并存。

本次研究《麻省理工科技评论》中国团队尝试探讨一些在部署过程中企业关心的问题，以期为企业部署使用户生成式人工智能提供一些帮助，如当前部署阶段、生成式人工智能对企业的影响、企业部署生成式人工智能的商业价值、部署时的挑战、部署策略等，以及正式发布「中国AI大模型先进应用案例」，这些企业有的已经在内部部署大模型、有些企业正在为行业（to B、to C）提供通用/定制化大模型能力。

过去一年，生成式人工智能已经成为人们讨论的中心话题
我们观察到了哪些现象、感受到了哪些变化？

「最」快速的增长与热潮

2022年11月，OpenAI发布了GPT 3.5大语言模型 (LLMs)，开始点燃大众和资本对生成式人工智能的关注和热情。ChatGPT病毒式传播，5天内获得100万用户，2个月内达到1亿。ChatGPT的推出将生成式人工智能——生成新内容（文本、代码、图像、音频等）带到了新高潮。

「最」激烈的军备竞赛

微软、谷歌、Meta、亚马逊、英伟达、阿里、百度等全球范围的巨头科技公司，纷纷下场训练大语言模型、投资生成式人工智能公司、在产品生态中部署生成式人工智能技术功能。大语言模型变成了“强者”游戏，行业当中「百模大战」的局面形成，OpenAI、Hugging Face、Anthropic、Runway等当红明星公司的一举一动也成为了2023年话题中心。

「最」集中的一些关注焦点

时下科研界、产业界讨论最集中的一些话题包括，开源模型 vs 闭源模型、通用模型 vs 垂直模型、落地场景及方向、算力、耗能、通用人工智能的到来、人工智能意识难题、政策与监管手段等。

「变」快的立法速度

各国政府正忙于起草法规、成立工作组，甚至创建新的政府部门来管理人工智能。此前，整个数字经济领域在数据隐私、网络安全和竞争政策方面的立法潮已持续多年。但是，生成式人工智能的出现似乎正在促使这些年的辩论付诸行动，2022年全球有37 项与人工智能相关的法案通过变成法律。2023年以及即将到来的2024年，中国、欧盟、美国和印度在内的主要国家都在努力制定全面的AI政策。例如，2023年10月30日，美国签署了首份关于AI的行政命令；2023年12月8日，欧盟成员国及欧洲议会议员就全球首个监管包括ChatGPT在内的AI的全面法规达成初步协议。

「变」务实的大众和企业

三个重心的转移：（1）2023上半年，在ChatGPT面世之时，大众更加关注国产大模型和国外模型在技术层面的差距，在算法、问答能力方面的差距，如今大众更关心这样的技术对生活、工作的具体改变方式；（2）大模型企业长期在卷模型的“大”、“更大”，如今开始思考更多的商业模型、产品化能力，例如AI Agent、AI原生应用、生成式人工智能超级APP等成为落地时讨论更多的内容；（3）企业高管更关心模型的实际落地，以及它将如何以及在多大程度上影响企业运营，他们思考最合适的布局策略、选型策略、企业内部应用部署的速度、大模型的产品能力等问题。

我们是如何走到今天的？
生成式人工智能发展时间线，以及过去一年重要事件

1 **2001年**
自然语言处理技术，入选当年《麻省理工科技评论》全球十大突破性技术

2 **2013年**
深度学习技术，入选当年《麻省理工科技评论》全球十大突破性技术

3 **2014年**
生成对抗网络（GANs），Ian Goodfellow推出

4 **2016年**
WaveNet 和音频生成 ，DeepMind推出

5 **2017年**
新的神经网络 “Transformer”，谷歌研究人员推出

6 **2018年**
谷歌AI发布BERT，机器理解语言上下文的能力实现了飞跃

7 **2018年**
对抗性神经网络（Dueling Neural Networks），入选当年《麻省理工科技评论》全球十大突破性技术

8 **2019年**
OpenAI发布GPT-2，文本生成能力受到关注

9 **2020年**
OpenAI 发布 GPT-3，加速大众对大语言模型的兴趣

10 **2020年**
“Deepfakes” 被大众所熟知

11 **2021年**
OpenAI 发布了文本生成图像模型DALL-E

12 **2021年**
GPT-3入选当年《麻省理工科技评论》全球十大突破性技术

13 **2022年**
Google、Midjourney、Stability AI、OpenAI 对文本生成图像模型的关注激增

14 **2022年11月**
OpenAI 推出基于 GPT-3.5 的聊天机器人ChatGPT，掀起生成式人工智能热潮

15 **2023年**
文本生成图像AI，入选当年《麻省理工科技评论》全球十大突破性技术

2023年7月，
- Anthropic推出最新版AI聊天机器人Claude 2
- Hugging Face融资2亿美金，估值40亿美金
- Meta加码开源大模型生态，发布Llama 2，免费开放给商业和研究使用

2023年11月，
- OpenAI首届开发者日，正式公布自定义GPT
- 谷歌DeepMind实验室推出的天气预测大模型，已在*Science*发表，只需要不到1分钟，就能直接预测出未来10天的天气
- OpenAI “闹剧” 经历大约5天之后，Sam Altman确认回归重新担任CEO，并组建了一个新的（但不完全新）董事会。原董事会成员Adam还在其中，Ilya出局

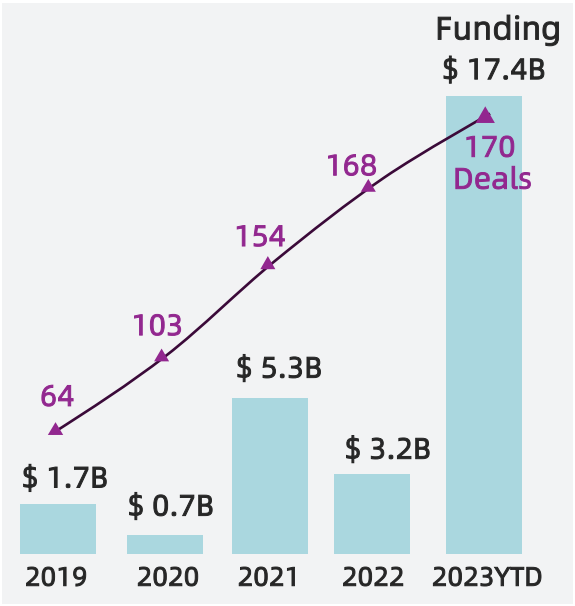
2023年12月
- 谷歌发布双子座Gemini大模型
- AMD正式发布Instinct MI300X GPU，直接对标英伟达H100，专供生成式AI领域
- ChatGPT与Ilya一同入选*Nature*年度十大 “人物”

图 | 生成式人工智能发展时间线以及过去一年领域内重要事件；来源：《麻省理工科技评论》、CB Insights

全球生成式人工智能领域发展图景

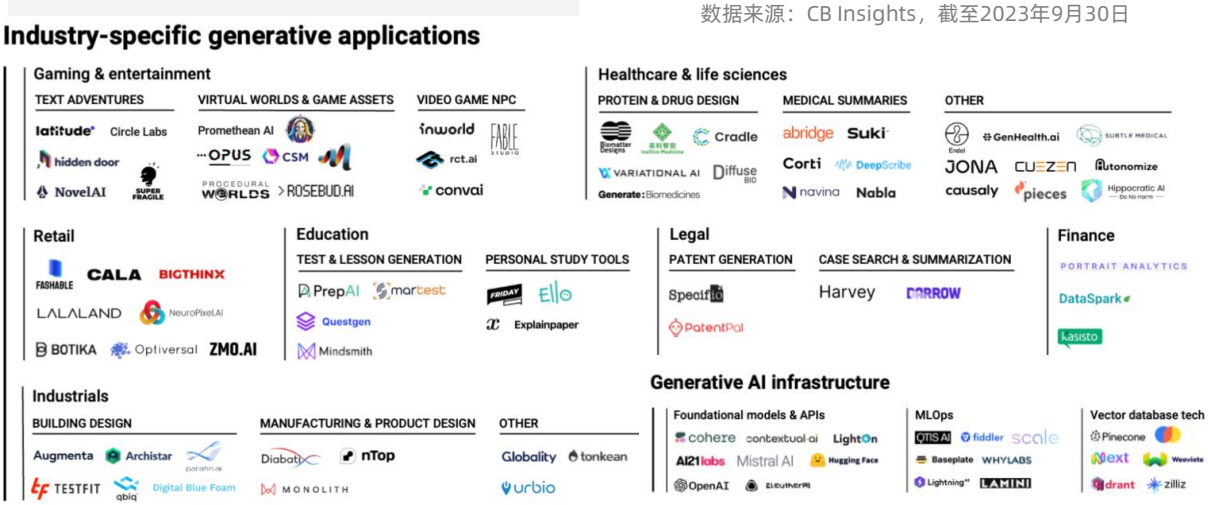
基础模型企业领跑独角兽名单，垂直方向应用逐步分化

通用性大模型的适用性较广，且一旦先发优势建立，会带动整体数据、算力以及商业“飞轮”。不断加固护城河，将大模型更好地整合入自身技术栈与生态，变成有场景侧重的底座通用大模型，而后开放服务赋能B端或C端用户，这或许将成为通用大模型的落地方式；行业大模型则更符合垂类场景的需求。当前越来越多垂直行业龙头企业加入大模型的赛道，这类企业即使在整体的数据、算法、算力设施不如科技巨头，但其拥有大量的专有数据集和高价值、特定领域的工作流程（数据护城河）。垂直企业正在针对这些工作流程开发垂直场景下的解决方案，以及进行行业数据训练，使其能迅速响应市场需求，推出具有差异化和竞争力的行业/垂直大模型产品与服务。



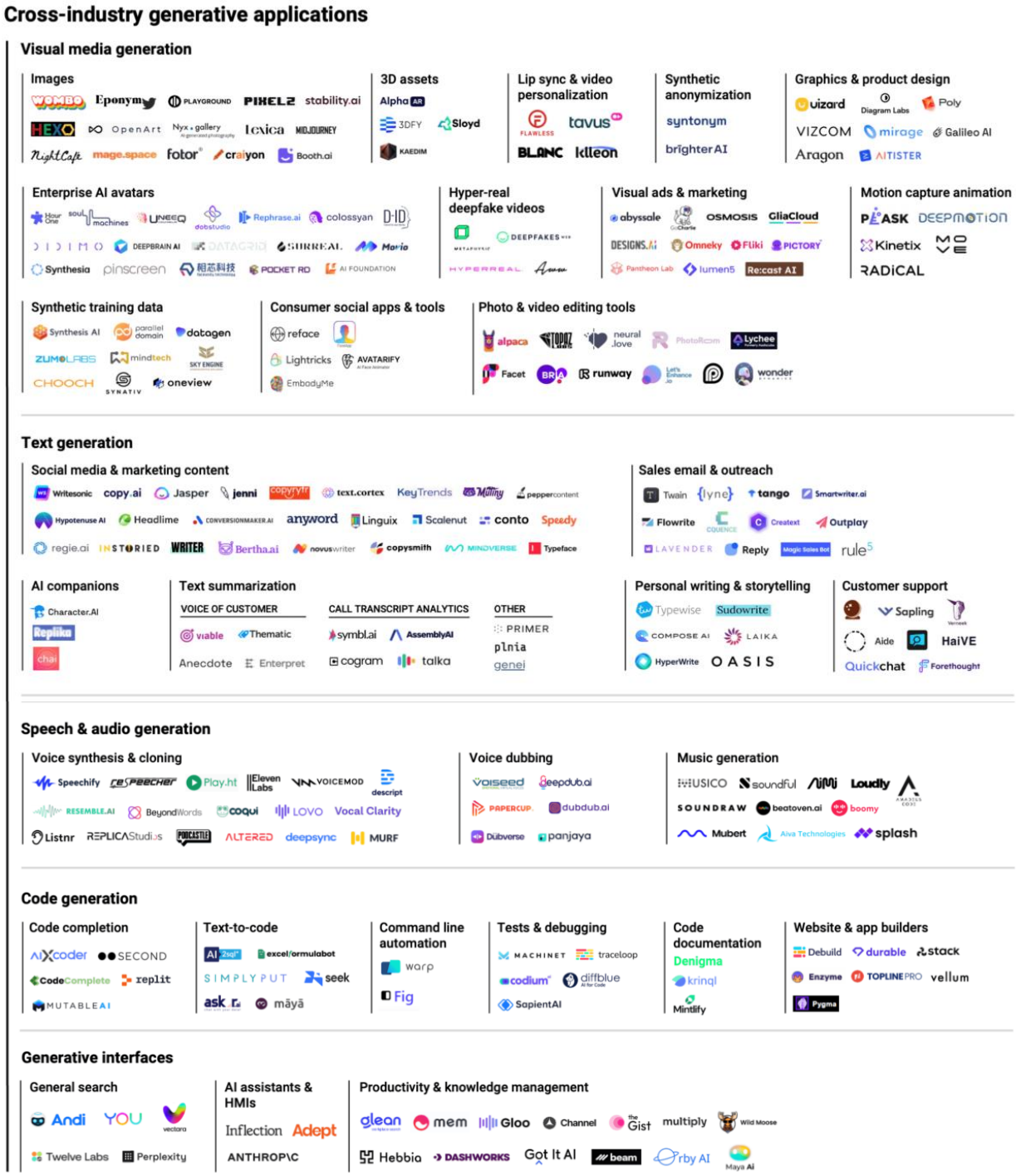
	Company	Valuation
🇺🇸	ANTHROP\C	$4.4B
🇨🇦	cohere	$2.2B
🇺🇸	runway	$1.5B
🇮🇱	AI21labs	$1.4B
🇺🇸	replit	$1.2B
🇺🇸	ADEPT	$1.0B
🇺🇸	Character.AI	$1.0B
🇬🇧	synthesia	$1.0B
🇺🇸	Typeface	$1.0B
🇺🇸	imbue	$1.0B
🇨🇳	ZHIPU·AI	$1.0B

图左丨全球生成式人工智能公司投融资市场；
图右丨2023年全球生成式人工智能领域新晋独角兽；
数据来源：CB Insights，截至2023年9月30日



图丨全球生成式人工智能企业图谱；数据来源：CB Insights，截至2023年9月30日

全球生成式人工智能领域发展图景
跨领域应用方向，全球涌现出一批技术型企业

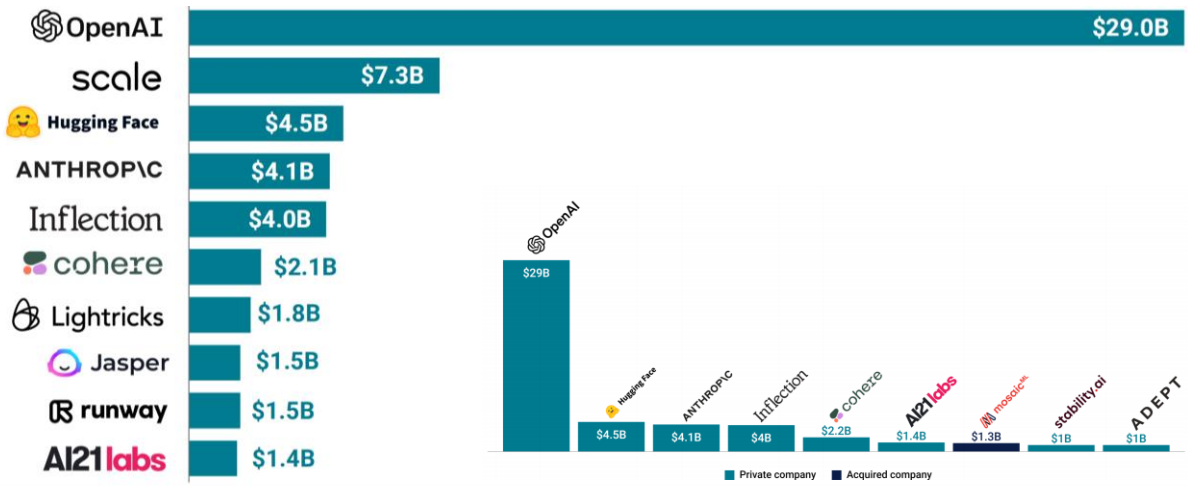


图丨全球生成式人工智能企业图谱；数据来源：CB Insights，截至2023年9月30日

全球生成式人工智能领域发展图景

海外主流大模型一览，目前还没有基础模型的最终赢家

估值最高的生成式人工智能独角兽，主要在基础设施层面/基础模型竞争，投身于大语言模型（LLMs）。OpenAI、Hugging Face、Anthropic、Inflection、Cohere、AI21 Labs，不仅是全球估值最高的独角兽，也都发布了基础大语言模型。但是目前大语言模型还是以Meta LIama为主的开源生态。



图丨全球生成式人工智能领域估值最高的独角兽企业；
数据来源：CB Insights，截至2023年9月30日

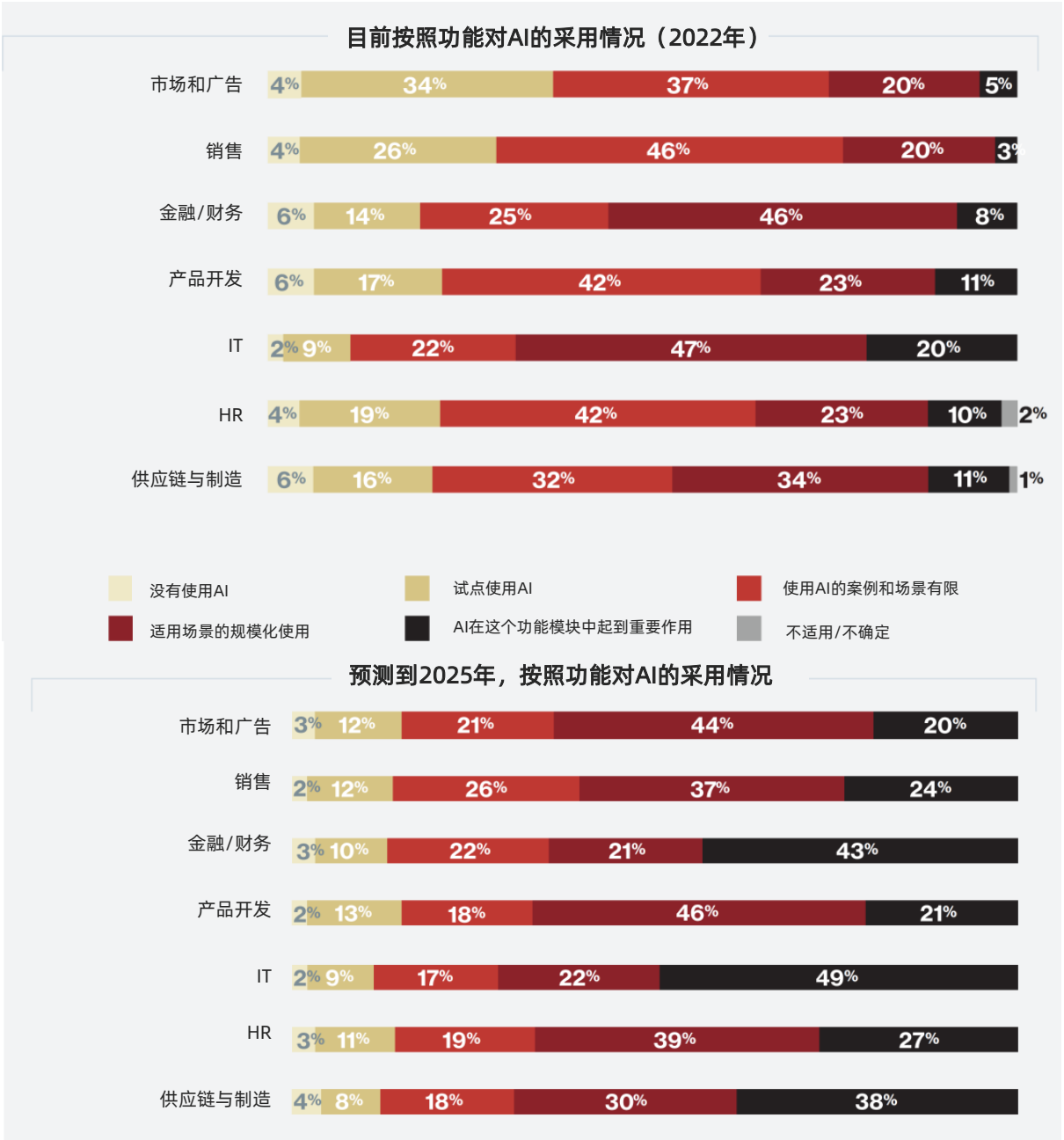
图丨全球生成式人工智能领域估值最高的LLMs独角兽企业；
数据来源：CB Insights，截至2023年9月30日

公司	基础模型
AI21 Labs	Jurassic-2
Amazon	Titan
Anthropic	Claude 2
Bloomberg	BloombergGPT
Cohere	Command
Google	Gemini
Inflection	Inflection-2
Intel	Aurora genAI
Meta	Llama, Llama 2, CodeLLaMA
Microsoft	Phi-2
Mistral AI	Mistral
MosaicML	MPT
Nvidia	ChipNeMo
OpenAI	GPT-4、DALL·E3
Stability AI	StableLM
UC Berkeley, Microsoft Research	Gorilla
xAI	Grok

表丨海外代表大模型盘点（不完全统计）；
数据来源：公开资料，统计截至2023年12月15日，如模型等信息有更新或变化，请读者自行关注，不代表任何投资建议

在ChatGPT为代表的生成式人工智能产品推出后，企业对人工智能的态度发生了什么转变？

在《麻省理工科技评论》2022年对全球600位企业高管的调研中可以看出，2022年（有影响力的产品/生成式人工智能能力出现之前），很少有组织采用人工智能作为任何业务功能的关键部分；同样到2025年，也很少有组织计划使人工智能成为跨关键职能的核心能力，企业高管对AI未来的判断也没有步入“必不可少”的阶段。



图丨2022年企业高管对AI采用的态度；
数据来源：《麻省理工科技评论》2022年对全球600位企业高管的调研统计，不代表任何投资建议

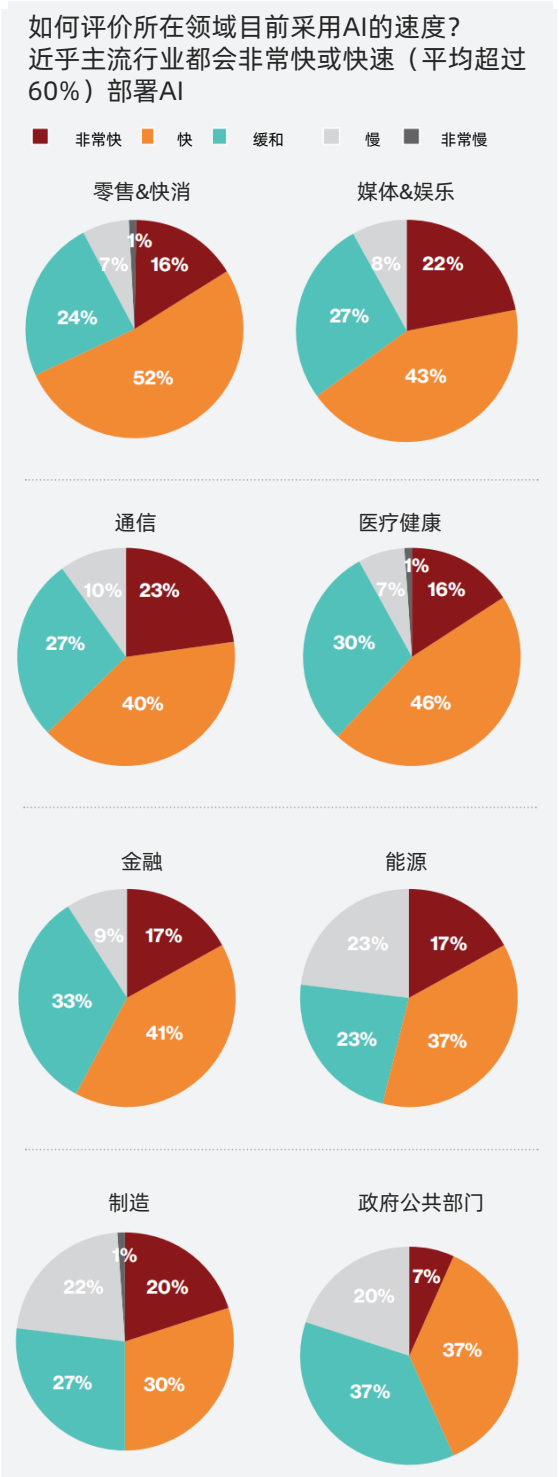
在ChatGPT为代表的生成式人工智能产品推出后，企业对人工智能的态度发生了什么转变？

2022年底和2023年初，面向消费者/大众C端的生成式人工智能工具的出现从根本上改变了公众对人工智能力量和潜力的讨论。甚至某种程度上加速了企业对于AI的部署速度。

尽管自2019年推出GPT-2以来，生成式人工智能一直在行业专家的谈论中掀起波澜，但直到现在，企业才意识到其革命性机遇。这一颠覆性时刻的影响力及其引发的连锁反应将在未来几十年内产生回响。

人们清楚地这件事情会在未来两年内，对我们的工作和生活带来极大的影响，但是却对如何影响、发生路径没有明确定论，这种确定但又不明晰的认知是当前企业界的共识，大家承认了它的重要性，也更加关注后续的风险和经济价值。

虽然增加新价值的影响还未全部实现，但生成式人工智能确实正在影响一些事情，例如组织的数据和技术基础设置，以及高层和CIO等技术高管为实现人工智能现代化而做出的投资决策。



图丨2023年企业高管对AI采用的态度；
数据来源：《麻省理工科技评论》洞察团队2023年对全球1000位企业高管的调研统计，不代表任何投资建议

生成式人工智能将引发真正的企业级人工智能开端

2023年企业对于人工智能的态度出现大转变，并且从当前市场的采用情况来看，企业在快速部署生成式人工智能能力、产品，背后的逻辑在于企业开始快速理解生成式人工智能带来的商业价值：

（1）对于个人来说，生成式人工智能让公众意识到了人工智能技术的具象价值，因为它背后有了可用工具体现，ChatGPT的出现或许拉开了开发工具平台的下一次变革大幕，从大型机、台式机、移动设备（它使服务消费无处不在，但它并没有转化为无处不在的服务创建），再到自然语言人工智能工具/生成式人工智能（带来广泛的生产力提升）。

（2）对于企业来说，以ChatGPT为代表的个人侧交互新工具的出现，给组织带来了本质变化——人工智能终于从试点项目和“卓越孤岛”转变为能够集成到组织工作流程中的通用能力。以ChatGPT为起点的一批新工具真正开始实现AI for everyone，企业明确具象地感知到这样新生产力工具对于个人、组织、大规模群体的经济意义。

（3）对于业务来说，大量非结构化和隐藏的数据现在变得清晰可见，能够释放商业价值。以前只有在结构化数据已准备就绪、且数据量充足的场景中，人工智能才能够发挥很好的效果；收集、注释和合成异构数据集的复杂性，使得更广泛的人工智能使用变得不可行。相比之下，生成式人工智能的新能力——揭示并利用（企业内外部）曾经隐藏的数据，将为企业取得新的效率进步和商业成就提供动力。

（4）生成式人工智能工具已经可以完成复杂多样的工作负载，更广泛的劳动力将从耗时的工作中得到解放，转向专注于洞察力、战略和商业价值等有更高价值的领域，让员工做更多增值业务。生成式人工智能可以开始从事创意性工作，这曾被认为是人类独有的能力和事业。

（5）未来，对于生成式人工智能的应用需求，更多将由企业成员自发地“需求拉动”，而不是技术人员、或CIO将这个技术引入、带到团队中去，这将提升企业数字化升级转型动力、速度和效率。语言类任务占社会主要行业比重较大（银行、保险、软件、通讯传媒、零售、健康、公共服务等），而在这些任务中，生成式人工智能有很大能力为它们带来能力增强、或实现自动化。

（6）企业对于劳动力的理解升级，有了ChatGPT的示范，企业管理者们现在很大程度上将人工智能视为人类员工的副驾驶，而不是竞争对手。

企业确实在投资、试验生成式人工智能技术，当前需要关注生成式人工智能部署策略，“共识”与“谨慎”并存

经过广泛的调研，当前市场的反馈是多数企业高管了解生成式人工智能够提高生产力和改进业务流程，这一点成为共识。但企业基本都还处于早期采用阶段，因为领导者正在评估采用这项新技术的安全且有价值的部署方法。高管们的确认识到了它的潜力，也在积极尝试，但他们正在谨慎地部署。

从试点、产品原型到企业内部大范围部署，企业当前的速度有多快？有哪些战略考虑？《麻省理工科技评论》洞察团队在2023年7月和8月对全球1000名企业高管进行调研，尝试了解其组织实施生成式人工智能技术的方法。这1000名受访者是CEO、副总裁或高级经理等高层管理角色，受访者分布在11个行业，包括消费品和零售、金融、制造业、医疗健康。调研范围包括美洲、欧洲 、亚太地区以及中东和非洲。受访者的公司规模不等，全球年收入（以美元计）从不足5亿美元（31%）到超过100亿美元（15%）。

结果显示，几乎所有被调研的公司都认为生成式人工智能是技术领域的重大变化，1000位被调研人中，仅有4%的人表示这不会影响他们的公司；但同时也只有9%的公司在整个组织中完全部署了单个用例。整体来说，部署阶段处于初步渗透阶段，还比较早期，生成式人工智能部署率仍处于低位。尽管受访者几乎一致认识到生成式人工智能拥有彻底改变企业的潜力，而当我们仔细讨论他们在组织内部署技术的速度时，也发现他们的热情因巨大的挑战和担忧（特别是多种风险）以及技术最终效果的不确定性而受到削弱。

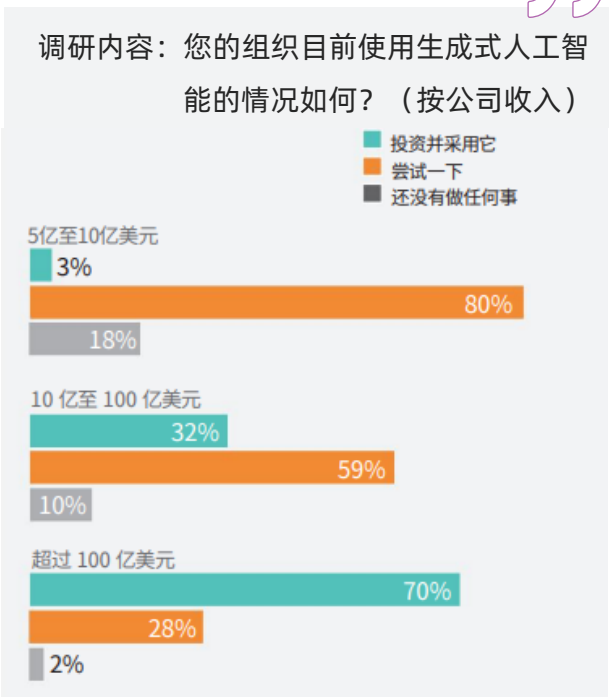


图 | 2023年企业对生成式人工智能采用的情况；数据来源：《麻省理工科技评论》洞察团队2023年对全球1000位企业高管的调研统计，不代表任何投资建议

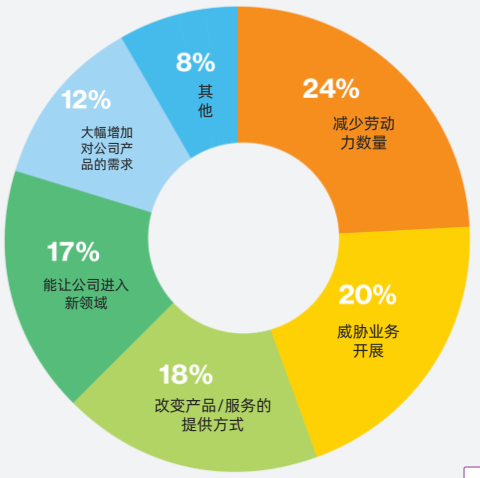
生成式人工智能对企业的影响

生成式人工智能的创新能力能够为企业和个人用户提供有价值的参考建议、开拓新的应用视角，起到提升企业的决策能力和个人工作效率的作用，甚至能够重塑各行各业的未来蓝图。当前最主要的应用方式是，许多企业正在尝试使用相对简单的大语言模型来搜索内部数据并提供基本的对话体验。

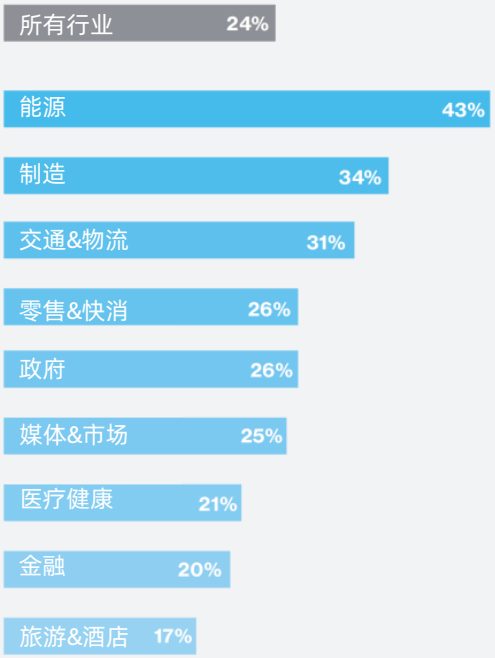
关于生成式人工智能对企业的影响，积极的看法占大多数，但也有担忧的声音，担心会对企业的业务开展造成威胁。其中来自制造、零售、媒体和娱乐以及电信行业的高管预计生成式人工智能，能够在自动化和效率方面带来可观的价值。金融服务和能源提供商尤其看好此类模型所带来的风险管理优势。

近四分之一的受访者预计生成式人工智能对其业务的主要影响是减少他们的劳动力，这个数字在能源（43%）、制造业（34%）、交通和物流（31%）等工业制造业比例最高，在旅游&酒店（17%）、IT和电信行业的比例最低 (7%)。

调研内容：使用生成式人工智能，对您的组织有哪些影响？（只能选择一项）



选择“减少劳动力”这一选项的比例按照行业分布（部分展示）



图丨2023年企业对生成式人工智能产生影响的看法；数据来源：《麻省理工科技评论》洞察团队2023年对全球1000位企业高管的调研统计，不代表任何投资建议

企业部署生成式人工智能的商业价值

调查内容：在未来两年里，哪些使用场景最能给您的组织带来价值？

	所有受访者	金融	政府&公共部门	医疗健康
个性化或者定制体验	1	1	3 tie	1
供应链优化	2			2
质量控制	3			3
实时数据分析和洞察		2	1	
自动化和提升效率			2	
产品和服务创新				
预测性维护			3 tie	
风险管理		3		

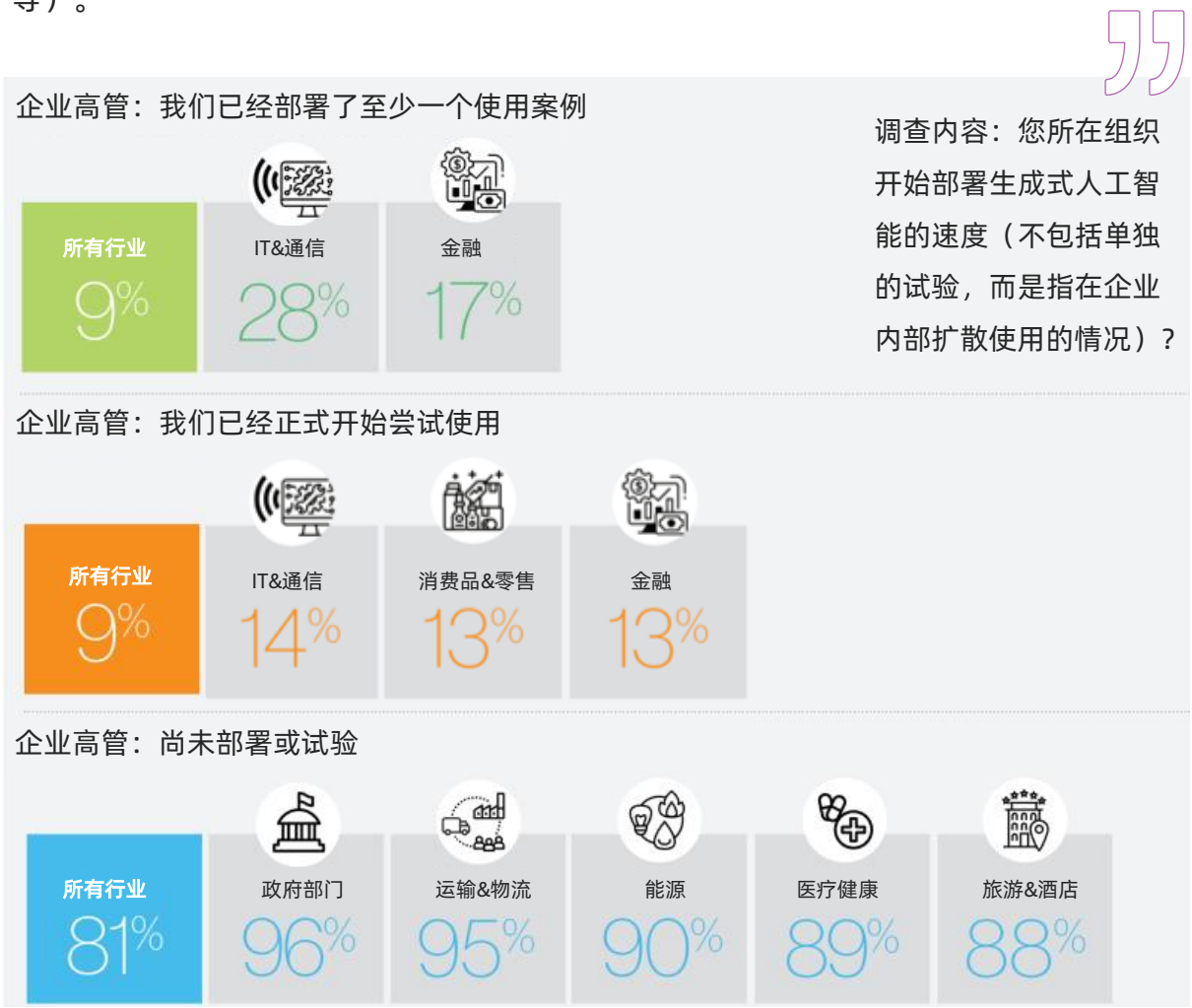
	零售&快消	制造	传媒&娱乐	能源	通信
个性化或者定制体验	2		1	3 tie	
供应链优化	1	1		1	
质量控制		2		2	1
实时数据分析和洞察				3 tie	2
自动化和提升效率	3	3	2		3
产品和服务创新			3		
预测性维护					
风险管理				3 tie	

图丨生成式人工智能最能给组织带来价值的应用场景；
数据来源：《麻省理工科技评论》洞察团队2023年对全球1000位企业高管的调研统计，不代表任何投资建议

生成式人工智能在企业的部署率仍处于低位

埃森哲公开发布的分析表示，各行业40%的工作时间可以通过生成式人工智能实现自动化或增强，其中银行、保险、资本市场和软件行业，显示出最大的潜力。

在本次调研中，IT和通信 (28%) 以及金融服务 (17%) 企业最有可能部署生成式人工智能，在政府部门，这个数字低至2% 。目前只有9%的受访者表示在他们的组织中，完成部署了生成式人工智能。部署相较广泛的行业呈现出两大特点：一是数字化程度较高且数字基础设施较完善（如通信、传媒等），二是数据积累量较大（如金融、电商等）。



图丨按行业划分的企业部署生成式人工智能速度；
数据来源：《麻省理工科技评论》洞察团队2023年对全球1000位企业高管的调研统计，不代表任何投资建议

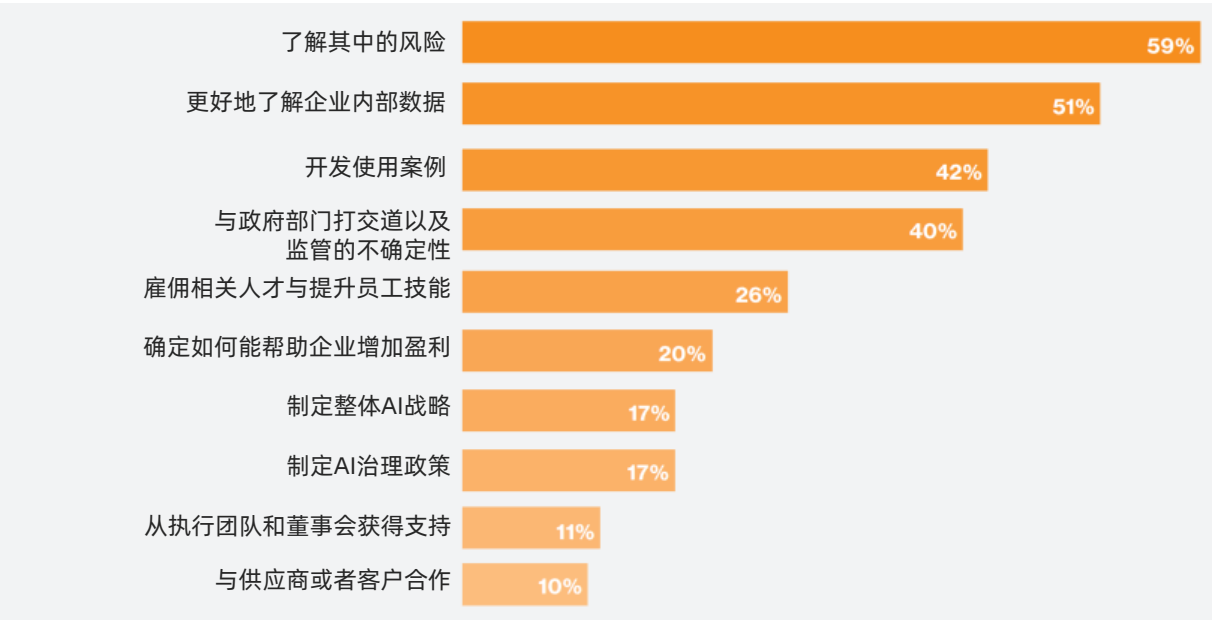
生成式人工智能部署实施时候的挑战

对话体验是极其狭隘的认识，将这个功能变成企业级解决方案还有很多工作要做

AI等技术在经济上为企业提供了“平等”的使用机会，不同规模的公司都有能力使用生成式人工智能。公司规模与生成式人工智能的部署可能性没有直接必然关系。

实际上中小型企业更有可能已经试验、或者部署生成式人工智能。中型企业在进行试验时，面临的运营和品牌风险比小企业更大，但与大型公司相比，安全推进部署的能力和基础设施更受限。对于小型企业来说，它们最不担心人工智能对他们的业务构成威胁，并且他们也不会将风险或不确定的监管列为主要的挑战。但是，规模较小的公司更有可能在制定和部署总体人工智能战略方面遇到困难，并且对确定生成式人工智能如何帮助其盈利缺乏信心。

部署的最大障碍是理解生成式人工智能风险，59%的受访者将其选为前三挑战之一。企业对这项技术可以从根本上改变工作方式，以及可以创造的服务和产品的潜力持乐观态度是正确的。但他们还需要现实地应对和思考这对组织运作方式所带来的挑战，以及对IT、组织、文化和责任的影响。



图丨组织实施生成式人工智能时候面临的挑战；
数据来源：《麻省理工科技评论》洞察团队2023年对全球1000位企业高管的调研统计，不代表任何投资建议

企业是否需要“大”模型？
建立数千亿参数的模型对大多企业来说仍然是遥不可及的

据报道，OpenAI使用10000个GPU来训练ChatGPT，训练GPT-3成本超过400万美元；Meta的Llama模型训练成本超过240万美元。较大的模型训练和运营成本更高，每次交互的成本大致与模型的大小相关。现阶段，建立大型基础模型（文本生成、文生图、文生音频/视频等）基本只能由资源最丰富的大型科技公司来完成。中小型公司不能简单地生产这些大语言模型的“自己版本”，参数规模和投入成本超过了当前近乎所有中小企业的能力范围，然而，较小的模型提供了可行的替代方案。从“我需要模型中的5000亿个参数” 转变为“我可能需要我实际拥有的数据的 7、10、30、500 亿个参数”是可行的。

复杂性的降低，来自于将企业的注意力从了解所有人类知识的大通用模型中，缩小到只为企业服务的高质量模型中，这也是企业中个人真正需要的模型。值得注意的是，“较小” 并不意味着“较弱”。现有的一些大模型已经针对需要较少数据的领域进行了微调，如BERT等模型那样，用于生物医学内容(BioBERT)、法律内容 (Legal-BERT)和法语文本（CamemBERT）。在某些情况下，企业可能会选择牺牲广泛的知识来换取其业务领域的特殊性。

2023年3月，Databricks发布了Dolly，表现出类似ChatGPT的对话能力（即遵循用户指令的能力），这是基于Meta的Llama，再根据Databricks的输入进行微调。Dolly只有60亿个参数——不到GPT-3使用的1750亿个参数的3.5%。该模型使用的是开源的代码和数据，但也允许商业使用。

> *“All the large models that you can get from third-party providers are trained on data from the web. But within your organization, you have a lot of internal concepts and data that these models won’t know about. And the interesting thing is the model doesn’t need a huge amount of additional data or training time to learn something about a new domain”*
>
> *Matei Zaharia, Cofounder and Chief Technology Officer, Databricks, and Associate Professor of Computer Science, University of California, Berkeley*

部署策略：公司不会单打独斗，与初创企业和大型科技公司的合作对于顺利部署至关重要，其中开源模型的地位关键

作为生成式人工智能模型的开发者和人工智能软件的提供者，大科技公司拥有生态系统优势，而初创企业则在几个专业领域享有优势。Google、Meta、Microsoft等公司为基础模型/大语言模型、开源生态做出了重大贡献，这些科技巨头以及Salesforce、Adobe等软件巨头能够快速将生成式人工智能嵌入到其现有的产品生态系统中。

目前我们能够观察到的现状是，企业大多倾向于混合的方法来开发这些生成式人工智能功能，在大多数行业的一些合适的场合，企业倾向于使用供应商的大型语言模型（LLMs），而在知识产权、隐私、数据安全性和准确性要求更严格时的行业和场景下，企业必须得构建他们自己的模型（例如金融）。因此，我们认为当前的部署策略属于在秉持实用主义的观点下，从为业务带来的经济价值和增量出发，企业倾向于选择多模型叠加态，后期再考虑功能模块的逐步分化。这意味着一家企业可以完全依靠供应商（无论是通用模型、还是垂直行业/功能模型）；也可以选择采用供应商、并且同步自建；还可以选择同时部署几个不同功能、针对不同部门、针对不同业务流程的“小模型”；甚至可以在已部署通用模型的基础之上，再有针对性的叠加垂直功能“小模型”，这些状态在目前都是存在的。

Adobe作为软件和创意产业的代表，快速将生成式AI功能融入产品体系

Adobe提供了一套称名为Adobe Firefly的生成式人工智能设计模型，通过100多种语言的简单文本提示，用户可以生成图像、添加或删除对象、转换文本等。2023年10月，Adobe推出了新功能“文生图2.0” Firefly Image 2，与市面上主流的几大AI绘画工具（Midjourney、Stable Diffusion XL、DALL·E 3）一较高下。

Firefly应用场景

文字转图像	生成填充	文字效果
根据详细的文本描述生成图像	使用画笔去除物体或描绘新物体	将样式或纹理应用于单词/短语
生成重新着色	**3D转变成图像**	**Project Stardust**
生成矢量图稿的颜色变化	通过 3D 元素的交互定位生成图像	移动任意物体进入到画面

来源：Adobe

部署策略：公司不会单打独斗，与初创企业和大型科技公司的合作对于顺利部署至关重要，其中开源模型的地位关键

对于基础模型来说，由于其训练成本极高，在海量算力的成本压力下，闭源模式成为谷歌和OpenAI保证商业投入的有效方式。它们通常在规模上更为庞大，这也使得其在性能方面表现出卓越的优势。这些模型被设计为可直接应用，从而使其易于使用和部署。但闭源基础模型的主要使用成本是客户需要为API支付昂贵的费用，且适应性较低，开发人员的选择性较少，限制了其灵活性和定制化能力。

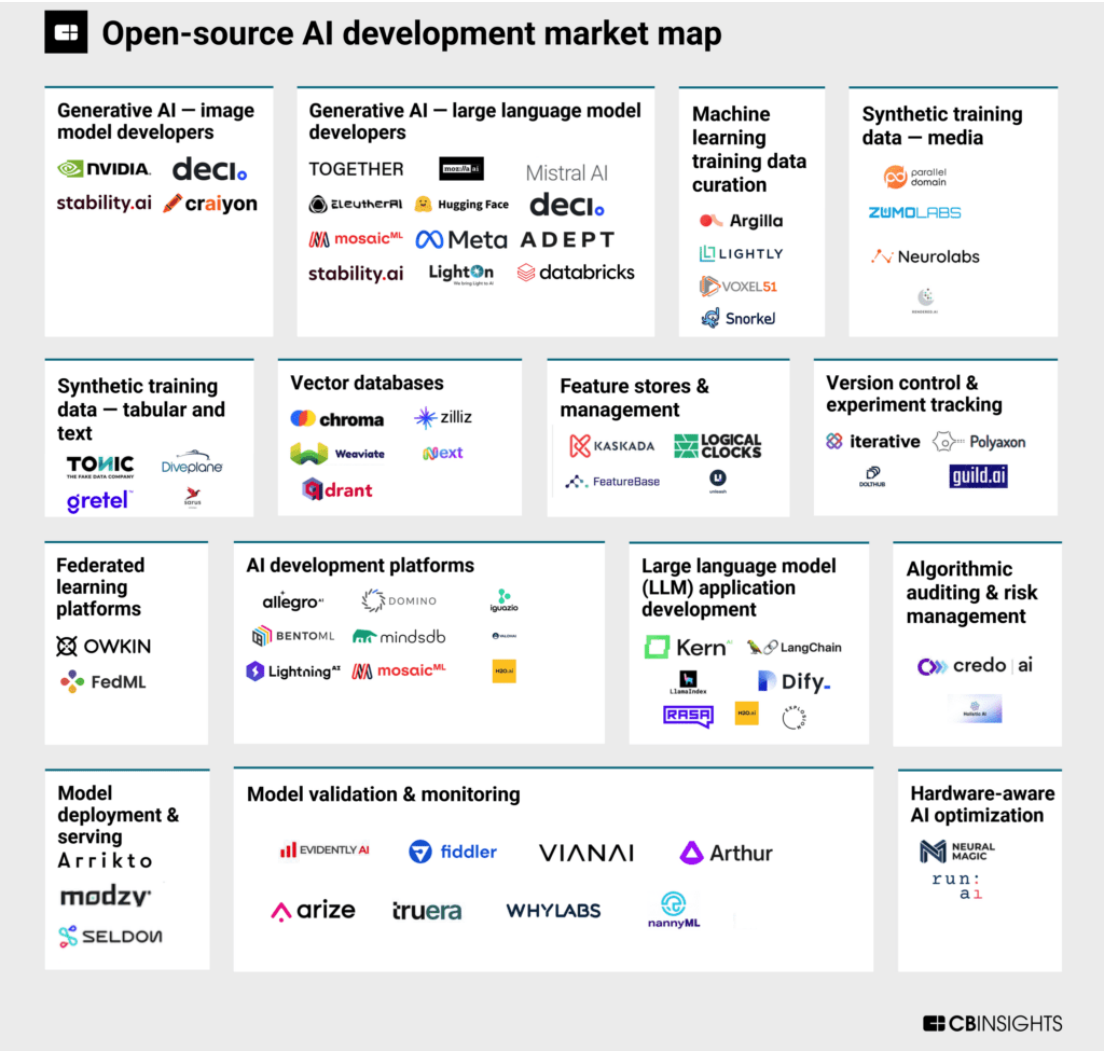
从目前已开源的模型中分析可得，开源大模型小型化趋势明显，在十亿-百亿参数级别居多。开源大模型通常具有较少的参数，在设计、训练和部署上，需要的资源和成本都相对较低。在迭代速度上，当闭源模型还在以月为单位迭代时，开源模型已经以周速度迭代。此外，基于已有的开源预训练模型进行微调也是开源大模型的优势之一。在预训练模型基础上进行微调和优化，以适应不同的任务和应用场景，这种方法不仅可以大大缩短模型的训练时间和成本，而且还可以提高模型的性能和效率。

开源模型与闭源模型之间的差距正快速缩小。闭源模式在性能、规模取胜，开源模型作为基础“母体”，迭代速度更快、更可定制化、更安全可靠。

谈及大模型时代未来的主流方式，以史为鉴，不论PC时代还是互联网时代，两种模式发展均行之有效。在确定性的个人计算机时代，微软以基于工程范式的封闭开发模型，验证了闭源的可行性。在不确定的互联网时代，开源吸引了全球开发者的参与其中，使得服务器操作系统、云操作系统等蓬勃发展，开源在不确定时代更具备创新力。在AI模型层，计算机视觉、语言翻译领域之所以发展到今天如此成熟，开源在其中起到了至关重要的作用。以计算机视觉为例，CNN（卷积神经网络）一问世就是开源的，后续的一系列模型也基本开源，加速了多轮创新和技术迭代。未来大模型闭源与开源并存，已是行业共识。但从生态发展的角度来看，开源或许可以让行业易于使用，从而实现与行业落地的紧密结合，进而发生很强的优化。与此同时，开源大模型也为生态系统中的开发者提供更多机遇。

部署策略：公司不会单打独斗，与初创企业和大型科技公司的合作对于顺利部署至关重要，其中开源模型的地位关键

开源支持者强调大型开发者社区带来的加速创新，以及该方法提高人工智能模型、数据和代码漏洞透明度的能力。反对者则担心，随着强大的生成技术占据中心位置，它会被滥用，从而助长网络攻击、人工智能生成的仇恨言论等。Meta在今年早些时候推出了开源大型语言模型Llama-2（许多模型都是建立在Llama之上），而谷歌和OpenAI到目前为止采取了封闭的方法。除了文本生成模型（LLMs）、文生图模型（例如Stable Diffusion）等基础模型开源之外，近两年还涌现出许多供应商，为AI开发过程的不同部分提供开源工具，从合成训练数据平台、矢量数据库、联邦学习平台到AI部署软件和模型监控平台。

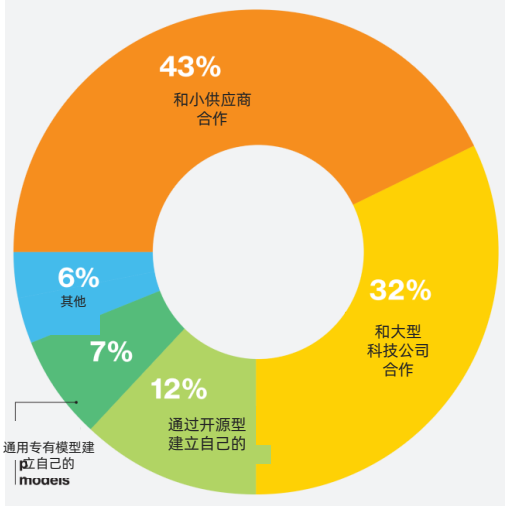


图丨全球生成式人工智能领域开源市场企业图谱；数据来源：CB Insights，截至2023年9月30日；不代表任何投资建议

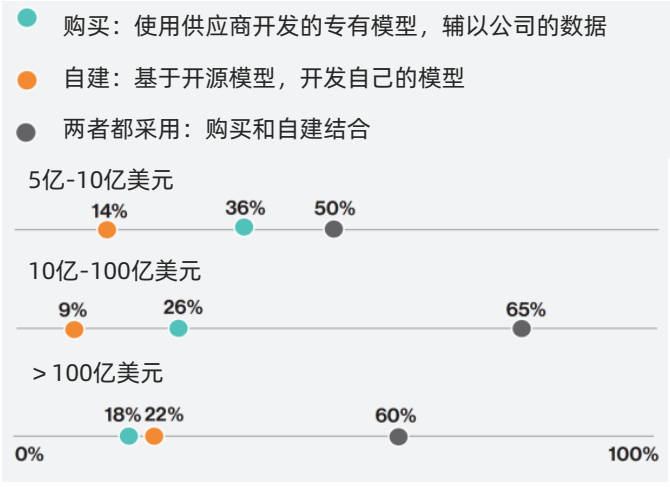
部署策略：公司不会单打独斗，与初创企业和大型科技公司的合作对于顺利部署至关重要，其中开源模型的地位关键

大型科技公司正在迅速采取行动，将生成式人工智能嵌入到其产品生态系统中，但许多企业高管正在寻找专业的初创公司，尤其是在数据敏感或高度监管的行业。一些CIO正在采取措施禁止公司使用外部生成式人工智能平台，在员工使用ChatGPT处理商业敏感信息后，三星禁止了内部使用ChatGPT。其他包括JPMorgan Chase、Amazon和Verizon也颁布限制条令或禁令。本次调研中，高管（75%）表示他们计划与大型科技公司或者专业的初创公司合作来部署生成式人工智能。对比两者，他们更有可能与小型提供商合作（43%）。受访者中有不少数量的技术和战略方向高管，因此19%的受访者表示他们的组织可能会选择单打独斗做自己的模型。这些受访者预计他们将构建自己的生成式人工智能技术和工具，要么结合公开可用的模型（12%），要么依赖专有或定制的模型（7%）。

很少有公司会在生成式人工智能时代独行，大型科技公司和初创公司都没有绝对领先优势。开源模型下一步不一定是“更大”，企业自身部署的模型下一步也不是“更大”，从大型科技公司到初创公司，蓬勃发展的支持生态系统可以帮助组织提升开发模型调整、产品化、安全和治理的能力。最近在业内，人们对于大模型的应用方向逐渐形成了一种思路：利用业内领先的通用大模型作为基础模型（Foundation Model），配合自有数据进行训练和调优，进而构建出面向不同业务场景的应用。



图丨企业部署生成式AI实例的主要技术策略



图丨企业部署生成式AI部署策略按企业收入分布
来源：《麻省理工科技评论》洞察团队2023年对全球1000位企业高管的调研统计，不代表任何投资建议

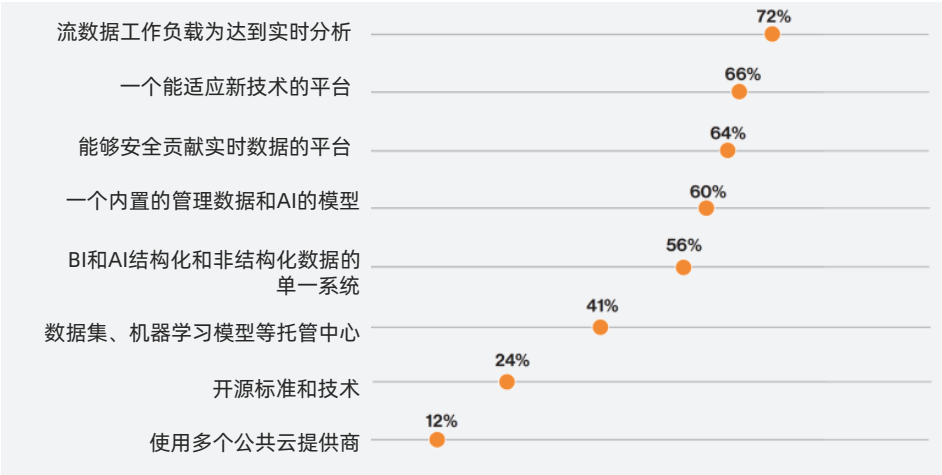
未来部署建议：更多思考对数据基础设施的要求

为了更好地生成式人工智能的部署，企业需要灵活、可扩展且高效的数据基础设施

埃森哲的公开研究表明，大多数公司在部署人工智能能力上还有很长的路要走。其2022年对全球850名高管进行的调查显示，人们普遍意识到负责人工智能的重要性，但只有 6% 的组织认为他们拥有强大的负责任的人工智能基础。《麻省理工科技评论》2023年的调研同样显示，如果为了更好地生成式人工智能的部署，企业需要灵活、可扩展且高效的数据基础设施。调研显示湖仓一体（Lakehouse）已经成为生成式人工智能时代的首选数据架构。近四分之三的受访组织采用了湖仓一体架构，其中99%的湖仓结构用户表示，这种架构正在帮助他们实现数据和人工智能目标，74%的人认为这种帮助“意义重大”。受访者表示，他们需要自己的数据架构来支持实时分析的流式数据工作负载（72%的人认为这种能力“非常重要”）、新兴技术的轻松集成（66%）以及跨平台共享实时数据（64%）。

不仅关心监管政策的变化，企业在部署时也更关心治理问题。对企业来说，数据和人工智能系统的整合是一个优先事项。数据和人工智能系统的激增在调查中最大的组织（那些年收入超过100亿美元的组织）中尤其广泛。其中，81%的组织运行10个或更多的系统，28%的组织使用超过20个系统。受访的高管们表示要减少他们的多个系统间的隔阂，将来自整个企业的数据连接到统一平台上，使人工智能部署能够扩大规模。

调查内容：未来两年里，这些现代化所需的基础设施，哪些对于实现您组织的总体技术目标是非常重要的？



来源：《麻省理工科技评论》洞察团队2023年对全球1000位企业高管的调研统计，不代表任何投资建议

生成式人工智能在金融方向的应用和部署

在历史长河中，金融业一直站在数字化技术创新的前沿。金融行业数据丰富、应用场景多，适合生成式人工智能落地推广，作为数据密集型行业，其也会率先体会到大模型带来的价值。金融领域降本增效需求强烈，行业整体进入数字化转型深水区。生产工具的改变，将大幅提升企业运营效率并创造更具竞争力的营商环境，如将生成式人工智能应用于获客营销、风控、产品设计等领域将有助于提高效率、降低运营成本。银行、保险、券商、资管等企业/机构正深刻认识到生成式人工智能的颠覆性潜力，并积极将其融入运营体系，涵盖了客户与业务增长、销售市场及营销、风险评估与管理、合规与监管、产品设计等多个应用场景。随着大模型技术与业务的深度融合，头部金融科技企业有望实现产品和商业模式的革新，相关行业已涌现大量应用案例，例如BloombergGPT、Morgan Stanley、Lemonade等。

BloombergGPT：自研大模型，开创垂直+通用混合训练范式

BloombergGPT是一个由Bloomberg公司开发的人工智能（AI）语言模型，专门针对金融领域的自然语言处理（NLP）任务进行训练。该模型拥有500亿个参数，可以快速分析金融数据，帮助进行风险评估、情感分析、问答等功能。BloombergGPT训练语料包括 3450亿的公共数据集和3630亿的金融数据集。

BloombergGPT应用场景

生成查询语言	提供新闻标题建议	金融问答
得益于训练集积累了大量历史查询记录，BloombergGPT可根据用户需求自动生成查询语言，降低Bloomberg金融数据库的使用门槛	基于丰富的新闻文章训练集，BloombergGPT可以赋能新闻应用程序，协助记者完成新闻标题撰写等日常工作，极大提高用户工作效率，减少内容编辑等琐碎工作，将更多时间聚焦于核心内容	受益于金融垂直领域知识的训练优化，BloombergGPT可以更加准确地理解并回答金融世界的问题，因此，BloombergGPT可以便利金融业的知识获取，帮助从业人员快速获得相对准确的结果

来源：Bloomberg

生成式人工智能在金融方向的应用和部署

NBA（Next Best Action）：摩根斯坦利银行大模型

NBA是一种基于数据分析和机器学习的个性化营销策略，通过实时决策引擎和个性化推荐系统，根据客户的实时需求和行为，提供最适合的营销行动，以最大程度地提高客户满意度和销售业绩。

NBA应用场景

提供投资建议	实时操作预警	辅助解决客户日常事务
快速分析客户数据，预测客户行为，并生成符合客户偏好的投资建议，财富顾问可以从多个建议中进行选择，并运用自己的判断为客户做出最合适的选择	确保客户及时获悉重要事件，例如追加保证金的通知、投资组合变化、重大市场波动等。通过将个性化文本与提醒相结合，财富顾问可以提供量身定制的见解和指导，从而加强客户关系	根据客户的独特情况，提供有关医疗机构、教育机构的指导，以及提供量身定制的理财方法，从而提升客户信任感

来源：Morgan Stanley

AI.MAYA：Lemonade保险大模型

AI.MAYA是基于GPT-3打造的销售机器人，它为客户提供个性化的保险推 荐和咨询服务。同时，MAYA还能通过向客户提出有限且高质量的问题，并根据回答进行算法调整，大幅度减少客户管理时间。

AI.MAYA应用场景

风险评估	自主理赔	获客营销
分析客户信息，如房屋面积、住房地址、历史索赔等，并生成风险分析报告，以决定是否承保和保费	通过智能语音识别、自然语言理解等技术，基于客户提供索赔信息自动处理理赔，提高理赔效率	分析客户的行为和需求，为其提供个性化的保险建议和服务，并帮助客户完成保险购买流程

来源：Lemonade

金融机构应用大模型的定位不是替代人，而是扮演“副驾驶”的功能，发挥功能途径为先内部使用再过渡到外部

全球范围来看，金融大模型的参与者分为三类：金融机构、金融资讯公司以及科技公司，其中金融资讯公司，如Bloomberg GPT，由于具备海量金融数据积累的先天优势，因此在此次生成式人工智能变革中具备明显优势。从中国范围来看，金融大模型核心玩家可以分为传统金融机构、互联网金融厂商、通用大模型及垂类大模型开发商以及开源大模型。其中，传统金融机构金融大模型以内部赋能为主，还未形成较好的对外商用案例；百度、华为等通用大模型开发商也结合过往行业积累研发金融垂类行业大模型。当前企业内部部署和使用大模型的路径是先内后外，当前模型功能仍无法直接对外，更多以企业内部人员使用为主。

机构/公司	模型名称	所在地
度小满	轩辕	北京
达观数据	曹植	北京
恒生电子	LightGPT	浙江杭州
虎博科技	TigerBot	上海
澜舟科技	孟子	北京
马上消费	天镜	重庆
蚂蚁集团	贞仪、CodeFuse	浙江杭州
天云数据	Elpis	北京
文因互联	文因	安徽合肥
星环科技	无涯、求索	上海
阳光保险集团	正言	广东深圳
有连云	麒麟	上海
中国科学院成都计算机应用研究所	聚宝盆	四川成都
中国农业银行	小数（ChatABC）	北京

图丨中国金融方向大模型（不完全）统计（来源：公开资料）；仅按机构/公司名称首字母排序
统计截至2023年12月15日，如模型等信息有更新或变化，请读者自行关注，本次统计不代表任何投资建议

以金融行业为例，
中国金融机构布局AI大模型常见三大主流方式

中国金融机构布局AI大模型常见三种模式：自主研发、结合模型深度微调以及按需接入解决方案。**行业Know-how、数据安全、持续迭代、综合成本是金融机构布局大模型需要综合考虑的四大因素。**采用完全自主研发的方式来布局对金融企业本身的人才、制度、数据以及技术基础都有着较高的要求，同时研发活动伴随周期长、成本高、不确定性强的特点，因此基于通用大模型自建或基于金融行业大模型微调成为相对稳妥的路径。

中国金融机构布局AI大模型三种模式对比

	自主研发	结合模型微调	按需接入解决方案
适用企业	• 数据量庞大 • 科技基础坚实 • AI创新能力体系完善	• 数据量庞大 • 科技基础坚实 • AI创新能力体系完善	• 数据量有限 • 科技技术基础薄弱 • 人才相对匮乏
优势	• 全栈AI技术自主可控 • 自由度高	• 确定性较高 • 投入周期相对较小 • 自由度高	• 确定性高 • 综合成本低

来源：公开材料

农业银行：自主创新的金融AI大模型应用ChatABC

ChatABC依托农业银行人工智能服务体系的算力、算法、数据、人才四位一体的基础能力，重点着眼于大模型在金融领域的知识理解能力、内容生成能力以及安全问答能力，对于大模型精调、提示工程、知识增强、检索增强、人类反馈的强化学习（RLHF）等大模型相关新技术进行了深入探索和综合应用，结合农业银行研发支持知识库、内部问答数据以及人工标注数据等金融知识进行融合训练调优，实现了金融知识理解和问答应用，同时实现了全栈AI技术的自主可控。

1.0版本ChatABC大模型拥有百亿级参数，可初步具备自由闲聊、行内知识问答、内容摘要等多类型任务的服务能力，已在行内多个渠道以多轮问答助手、工单自动化回复助手等形式面向内部员工开放试用，并可通过MaaS（Model as a Service）方式面向其他场景提供一站式决策辅助服务，未来将逐步形成大模型服务生态。

以金融行业为例，中国金融机构布局AI大模型常见三大主流方式

工商银行：合作共建，基于昇腾AI金融行业大模型实现企业级金融大模型研制投产

工商银行将大模型应用于客户服务、智慧办公、职能研发等业务领域。

- **远程银行客服：**工商银行围绕远程银行中心数千人的客服团队，贯穿坐席事前、事中、事后全流程，聚焦对客服务中枢的运营团队、群体基数较大的人工坐席、工作量较为繁重的质检人员，重新定义该群体的作业和生产模式，基于大模型能力赋能全流程业务场景。

生成式AI在工商银行客户服务领域的应用

事前	事中	事后
智能客服知识运营阶段	服务客户阶段	质量检查阶段
利用大模型自动完成数据标注与知识维护，助力提升传统智能客服分流质效	利用大模型打造前情摘要功能、知识随行功能、工单智能填写功能，极大提升坐席运营效率，压降通话时间	生成传统质检AI模型数据，即模拟坐席及客户问答，提升传统质检模型准确率

来源：中国工商银行首席技术官吕仲涛在“2023中国智能金融论坛”上的主题分享

- **智慧办公：**利用大模型的文本生成、问答能力，围绕邮件、文档、会议、员工日常事务等方面，优化行内办公工具的交互体验，助力全行40万员工的办公效率提升。比如会议纪要生成，根据会议对话内容，大模型快速生成会议纪要初稿，降低人工记录会议纪要的成本，助力全行办公领域智能化。
- **智能研发：**利用大模型代码生成、代码识别与检测、代码转自然语言等领域的全方位能力，构建基于大模型的智能研发体系。截至目前人工智能编码助手共收集到超2100万个字符编码数据，录入代码超80万行，编码助手生成代码量占总代码量的比值从20%提升至40%，有效提升一线开发人员编码效率和质量。

以金融行业为例，
中国金融机构布局AI大模型常见三大主流方式

百度智能云开放文心一言API接口，度小满开源金融大模型轩辕

百度新一代大语言模型“文心一言”后，百信银行、新网银行、邮储银行、兴业银行以及中信银行等先后宣布接入百度“文心一言”。

文心一言落地案例

百信银行	将智能对话技术成果应用在数字金融、AI数字人、数字营业厅等领域
新网银行	将智能对话技术成果应用在数字普惠金融领域
邮储银行	通过“邮储大脑”接入并应用“文心一言”的能力，将打造更为丰富、个性化的智能金融产品，提供更智能更有温度的普惠金融服务
兴业银行	推进前沿人机对话AI技术在金融场景的应用，持续提升服务智能化水平
中信银行	接入“文心一言”，将百度领先的智能对话技术成果应用在财富管理领域，提供有温度的财富管理服务

来源：公开材料

轩辕大模型是在1760亿参数的Bloom大模型基础上训练而来，在金融名词理解、金融市场评论、金融数据分析和金融新闻理解等任务上，效果相较于通用大模型大幅提升，表现出明显的金融领域优势。

为了提升轩辕大模型对金融领域问题的理解能力，度小满将自身业务中积累的金融领域的千亿tokens的中文预训练数据集用来训练模型。该数据集涵盖了金融研报、股票、基金、银行、保险等各个方向的专业知识。度小满表示，经过清洗和标注的高质量数据集，不仅在通用性方面与ChatGPT达到持平成为可能，且显著提升了模型在金融垂直领域的性能。

生成式人工智能在医疗健康方向的应用

生成式人工智能擅长学习模型里所训练数据的模式和结构，“消化”掉这些海量数据之后，从而自主生成全新的、质量水平能满足使用的数据、图像和文本（即创造新知识、新内容）。这种能力使得生成式人工智能对于理解人类生物学、帮助科学家理解疾病、阐明治疗、疾病和患者之间的因果关系非常有帮助。

医疗是大模型应用最广泛、最有潜力的领域之一，但医疗行业的严肃性、复杂性、数据敏感性、优质数据稀缺性等特质，让生成式人工智能在医疗场景的产业落地仍充满重重挑战。目前在医疗行业还存在基础要素的不足，核心是由于缺乏高质量的医疗数据，将影响到医疗大模型在知识领域的能力。此外，患者隐私的保护也限制了当前医院、保险公司、制药公司对这些数据的使用方式。

机构/公司	模型名称	所在地
澳门理工大学	XrayGLM、IvyGPT	澳门
大经中医	岐黄问道	江苏南京
福建医科大学孟超肝胆医院	孟超	福建福州
深圳市大数据研究院、 香港中文大学（深圳）	华佗、凤凰	广东深圳
光启慧语	光语	上海
哈尔滨工业大学	本草、活字	黑龙江哈尔滨
华南理工大学	扁鹊、灵心SoulChat	广东广州
赛灵力科技（清华珠三角研究院）	达尔文	广东广州
上海科技大学	DoctorGLM	上海
微脉	CareGPT	浙江杭州
医联	MedGPT	四川成都
云知声	山海	北京

图丨中国医学、医疗方向大模型（不完全）统计（来源：公开资料）；仅按机构/公司名称首字母排序；
统计截至2023年12月15日，如模型等信息有更新或变化，请读者自行关注，本次统计不代表任何投资建议

生成式人工智能在工业制造方向的应用

工业制造领域和金融、医疗健康行业都产生大量的数据、以及同样面临效率提升的困境。生成式人工智能技术在工业制造场景下，产生的主要价值在于，帮助企业在研发到产品设计的一系列业务功能中节省时间和成本。主要的应用途径包括，建筑设计（建筑开发商根据位置、气候、预算等自动生成建筑计划）、零部件的设计和开发、基础设施设计、材料发现、自动驾驶的数据生成、采购环节的效率提升（使用文本生成工具根据项目规划说明和过去的项目编写采购计划）、实操培训等。例如，现代汽车已使用Autodesk的生成式设计软件来研究如何高效地制造新型车辆的零部件、IBM正在使用生成式人工智能来识别和测试计算机芯片材料。

机构/公司	模型名称	所在地
创新奇智	奇智孔明AInnoGC	山东青岛
格创东智	章鱼智脑	广东广州
华为	盘古	广东深圳
慧安股份	蜂巢知元	北京
江苏欧软	WISE	江苏苏州
理想汽车	MindGPT	北京
思必驰	DFM-2	江苏苏州
台智云（华硕子公司）	福尔摩斯	台湾
西北工业大学+华为	秦岭·翱翔	陕西西安
新华三H3C	百业灵犀	浙江杭州
中工互联	智工	北京
中科创达	魔方Rubik	北京

图丨中国工业方向大模型（不完全）统计（来源：公开资料）；仅按机构/公司名称首字母排序；
统计截至2023年12月20日，如模型等信息有更新或变化，请读者自行关注，本次统计不代表任何投资建议

阿里云	阿里云「通义」
百川智能 BAICHUAN AI	百川智能「Baichuan」
Bai du 百度	百度「文心」
创新奇智 AInnovation	创新奇智「奇智孔明AInnoGC」
4Paradigm 第四范式	第四范式「式说」
度小满	度小满「轩辕」
HUAWEI	华为「盘古」
科大讯飞 iFLYTEK	科大讯飞「星火认知」
面壁智能 ModelBest	面壁智能「CPM」
MINIMAX	MiniMax「abab」
商汤 sensetime	商汤科技「日日新SenseNova」
vivo	vivo「BlueLM」
医联	医联「MedGPT」
Moonshot AI	月之暗面「moonshot」
智谱·AI	智谱AI「GLM」

备注：
仅按机构/公司名称首字母排序；案例统计征集时间截至2023年12月20日；
以上企业以及大模型应用案例仅来自于2023年7月至12月的公开案例征集与公开统计；如模型等信息有更新或变化，请读者自行关注；此处不代表所有应用案例、且不代表任何投资建议。

扫码查看更多

百川智能

开放API全面进军To B领域

百川智能成立于2023年4月10日，由前搜狗公司CEO王小川创立。成立仅四个月，百川智能便相继发布了三款通用大语言模型。2023年6月15日，百川智能推出了首款70亿参数量的中英文语言模型Baichuan-7B，并一举拿下多个世界权威Benchmark榜单同量级测试榜首。7月，百川智能发布Baichuan-13B，凭借百亿参数量展现出可以媲美千亿模型的能力，大大降低企业部署和调试的使用成本，让中国开源大模型商业化进入真正可用阶段。8月8日，发布参数量530亿的大语言模型Baichuan-53B，在知识性上表现优异，擅长知识问答、文本创作等领域。

9月，百川开源Baichuan2、并接连发布Baichuan2-53B闭源大模型。作为首批通过备案的大模型企业，百川智能还开放了Baichuan2-53B API接口，正式进军To B领域，开启商业化进程。企业和开发者可以通过API将Baichuan2-53B集成至他们的应用程序和服务中。Baichuan2-53B融合了最前沿的大模型技术，可以适配不同企业的各种业务需求，无论是智能客服、智能写作还是智能推荐。Baichuan API接口便捷易用，客户只需要简单的配置和集成即可接入，同时其对OpenAI的接口高度兼容，客户可以快速迁移，极大降低了模型的部署和转换成本。

对于企业用户最关注的安全合规问题，作为首批通过《生成式人工智能服务管理暂行办法》备案的大模型企业，百川智能为Baichuan2-53B打造了覆盖大模型预训练、精调、推理全周期的安全增强，能够为客户和合作伙伴提供全流程的安全保障。借助Baichuan2-53B丰富且强大的模型能力，企业用户不仅可以升级自身已有业务，提高效率、减少成本，还能够探索更多应用场景，拓展创新的边界。此次开放API后，百川智能将把行业领先的大模型能力赋能给各行各业的合作伙伴，助力万千企业智能化发展。12月19日，百川智能宣布开放基于搜索增强的Baichuan2-Turbo系列API，包含Baichuan2-Turbo-192K 及Baichuan2-Turbo。在支持192K超长上下文窗口的基础上，还增加了搜索增强知识库的能力。即日起，API用户可上传文本资料来创建自身专属知识库，从而根据自身业务需求打造更完整、高效的智能解决方案。百川智能在引领国内大模型开源生态之后，再次引领行业开启了企业定制化的新生态。

创新奇智

要做更懂制造业的工业大模型

创新奇智成立于2018年2月，以“人工智能赋能商业价值”为使命，是中国快速发展的企业级AI解决方案供应商和“AI+制造”解决方案供应商。致力于用前沿的人工智能技术为企业提供AI产品及解决方案，包括AI平台、算法、软件及AI赋能设备，提高客户运营效率和商业价值，实现数字化转型。

2023年9月，创新奇智发布拥有150+亿参数量的工业大模型（AInno-15B）、大模型服务引擎以及三款基于大模型的生成式AI应用产品 –“奇智明达ChatRobot”生成式工业机器人任务编排应用、“奇智明数ChatBI”生成式企业私域数据分析应用、“奇智明睿ChatDoc”生成式企业私域知识问答应用，在工厂物流、智造BI、智造实训等多个领域落地，赋能行业解决方案。创新奇智结合多年企业智能化转型服务经验和积累的工业大数据，在对开源免费大模型进行知识蒸馏的基础上，设计了适合工业场景的大模型神经网络结构，然后通过预训练、指令微调、人类反馈强化学习，获得工业大模型AInno-15B，它拥有150+亿参数量，具有行业化、轻量化、多模态的特点，是一款更懂制造业的行业大模型。

其中，“AInnoGC+智造实训”方案入选国家工信安全中心大模型应用创新典型案例。在整个中国制造业产业升级的背景下，得益于智能化、自动化工厂的效率在大幅提升，但是智能化自动化不是凭空生成，需要背后很多机械工程师、电气工程师、自动化工程师，这些工程师培训是一个很重要的领域。在互联网、人工智能时代，知识更新迭代的速度超越以往任何时候。当前国家正在大力深化推进产教融合，培训课程如何与时俱进，跟上产业一线快速的知识迭代步伐，对智造实训行业是一个很大的挑战。

“AInnoGC+智造实训”开展前期，创新奇智推出智造实训中心MTStudio，结合创新奇智项目落地过程中形成的方法论以及真实案例，面向AI机器视觉和AI安全生产两大领域，通过课程教学、案例讲解、工具实训等，推出一套完整的实训中心解决方案，全面助力制造业挖掘培养AI人才。

创新奇智

要做更懂制造业的工业大模型

图丨智造实训中的两大工业实训中心（来源：创新奇智）

在过往培训过程中，学员由于理解能力参差不齐，导致学习进度不一致，出现老师难以因材施教的问题；课后老师和学生答疑，需要一对一完成，培训效率比较低。创新奇智将“奇智孔明工业大模型”注入到智造实训中心的日常培训中去，构建AI2.0智造实训平台，实现在学生、教师、教学平台三端的AI生成辅助教学。

在教师端，工业大模型可以基于垂类场景模型，生产响应的数据集，供学生练习训练模型；还可以自动给学生的作业打分，并根据每个人的薄弱环节进行强化训练，实现真正的因材施教。在学生端，工业大模型可以作为一个虚拟助教，给学生答疑解惑，并且根据学生提交的作业，进行纠错和讲解。在教学平台端，工业大模型可以根据我们的智造实训数据，自动生成课件，还可以让学生在VR程序上虚拟调试机械设备，更好的衔接实操课程。

以前实训中心的老师需要十几天的时间准备教学大纲，现在借助“AInnoGC教学助手”几天就完成所有的备课工作，学生也因为“AInnoGC学习助手”更扎实的掌握了技能。

第四范式

用AIGS改造企业软件

第四范式成立于2014年9月，提供以平台为中心的人工智能解决方案，并运用核心技术开发端到端的企业级人工智能产品，使企业能够开发自己的决策类人工智能应用，致力于解决企业智能化转型中面临的效率、成本、价值问题，提升企业的决策水平。

在2023年2月底，第四范式推出了一个专为业务场景设计的企业级生成式人工智能产品SageGPT，布局大模型AI产品。2023年4月26日，第四范式首次向公众展示其大模型产品式说3.0，并首次提出AIGS战略（AI-Generated Software）：以生成式AI重构企业软件。第四范式创始人兼首席执行官戴文渊博士表示，第四范式希望可以用生成式AI来重构企业软件，也把自身定位为基于多模态大模型的新型开发平台，提升企业软件的体验和开发效率，从而实现AIGS。

图丨式说大模型背后核心能力（来源：第四范式）

Copilot的字面意思是“副机长”，也可以理解成一个二号位。大模型能力对于企业软件的变革首先是交互方式上的。通过式说Copilot，员工可以通过语音、文本、图像、表格、视频等多模态方式，向式说发起询问或下达指令，式说在精准理解其意图后，联网企业多模态的信息、企业软件及其他专用AI能力，分析出答案，并以所需要的形式来输出答案。对于企业软件来说，Copilot概念的进入使得产品设计这件事有机会变得轻量化，不会成为软件功能迭代的阻碍。CoT（Chain-of-Thought，思维链）能力是大模型所涌现出的逻辑推理能力，在接受过大量同类型数据的训练后，大模型会开始形成对于此类问题的推理能力，再遇到一个类似的新问题，能够像人类一样将其拆解。Copilot与CoT一起，即是第四范式对于用AIGS颠覆企业软件困境的两个支点。从Copilot到COT，第四范式在大模型上的发展方向逐渐清晰——瞄准的市场不是ChatGPT，而是用AIGS改造企业软件。

第四范式

用AIGS改造企业软件

营销/销售

- 营销人员的智能助手，提升服务能力与效率
- 从对话中分析用户画像与需求，个性化服务
- 进行自动化客户服务，减少人工成本
- 优化用户沟通体验
- 参与营销内容创作

产品创新

- 对产品进行模拟、优化和预测
- 快速高效地找到产品的最佳设计
- 改变现有人机交互模式
- 帮助人更好地进行创造

业务助手

- 企业或特定业务(如零售)的知识助手
- 企业内数据和业务策略的沉淀与积累
- 精简并优化复杂的业务流程
- 高效执行给定活动

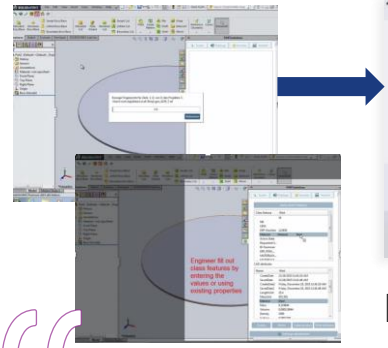
图丨式说大模型主要应用场景（来源：第四范式）

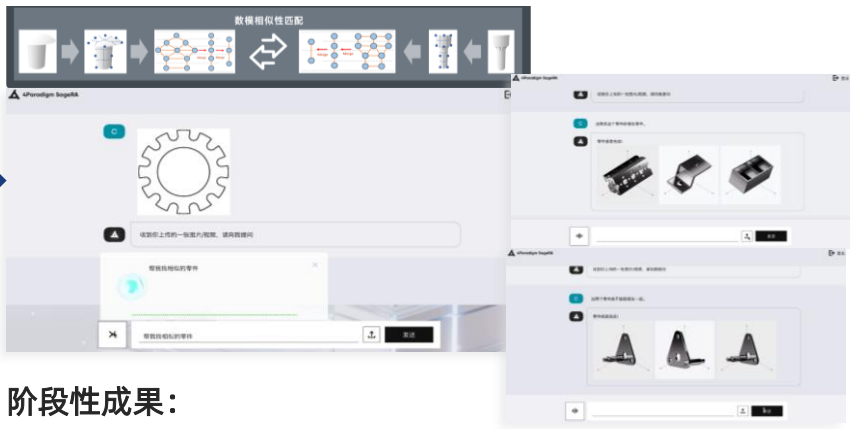
客户面临问题

- 传统模型管理软件过于复杂，查找相似软件需要多步操作
- 数模搜索不够精准，常导致在设计环节新增零件物料，带来设计、生产、制造全流程周期的高额成本

解决方案

- 软件界面升级：只需对话框下达指令，即可轻松实现零件搜索、自动化装配等 功能，无需层层点击菜单
- 软件内核升级：基于图学习的几何相似性搜索能力，做到以三维搜三维，高效精准





中国商飞：基于生成式AI的数模搜索及管理系统

阶段性成果：

•减少在数模搜索上的时间，提升零件重用率，

•节省整个零件生命周期成本上亿元

•进一步形成自主可控的三维辅助设计工具，拓展行业外的应用

图丨第四范式大模型部署应用案例（来源：第四范式）

度小满

开源国内首个千亿级金融大模型”轩辕“

2018年4月，百度宣布旗下金融服务事业群组正式完成拆分融资协议签署，拆分后百度金融启用全新品牌“度小满”，实现独立运营。2018年5月21日，度小满正式成立，延承百度技术基因，在智能金融时代，充分发挥AI优势和技术实力，携手金融机构合作伙伴，用科技为更多人提供值得信赖的金融服务。

2023年5月，度小满开源了国内首个千亿级金融大模型“轩辕”。在金融场景中的任务评测中，“轩辕”全面超越了市场上的主流开源大模型。“轩辕”用度小满实际业务场景积累的海量金融数据训练而来，通过独创的hybrid-tuning的创新训练方式，实现在大大增强金融能力的同时，不损失通用能力。自开源以来，已经有上百家金融机构申请试用“轩辕”大模型。

轩辕金融大模型的基础模型能力建设包括，（1）金融理解：增量预训练和指令微调阶段，加入大量金融数据，提升金融理解能力；（2）知识增强：外挂实时更新的业务知识库，实现低成本干预，同时降低幻觉影响；（3）应用增强：面向金融应用场景，定向增强摘要、逻辑、计算等金融场景核心能力；（4）对话能力：使用百万级经人工构建和校验的高质量指令数据进行指令微调和对齐。

2023年9月，度小满开源「轩辕-70B」金融大模型。其在通用能力表现上优异，在C-Eval和CMMLU两大权威榜单上，轩辕70B均名列所有开源模型第一；在金融能力方面，轩辕70B已经通过注册会计师、银行/证券/保险/基金/期货从业资格、理财规划师、经济师等金融领域权威考试，且考试得分领先于其他通用模型；在场景能力上，轩辕70B在度小满自有金融业务场景测试中表现领先，特别金融知识问答、NL2SQL等场景表现优异。2023年11月，度小满继续开源70B相关版本，即轩辕-70B-chat及8-bit和4-bit量化模型。

目前，大模型技术已经应用在度小满各个业务场景，从营销、客服、风控、办公再到研发，已经初见成效。在代码助手方面，用大模型辅助生成的代码，采纳率能够达到42%，帮助公司整体研发效率提升了20%；在客服领域，大模型推动服务效率提升了25%。在智能办公领域，大模型目前的意图识别准确率已达到97%。

科大讯飞

用通用人工智能解放生产力、释放想象力

2023年5月6日，讯飞星火认知大模型V1.0发布，到10月为止历经两次升级。其在6月9日、8月15日分别对开放式问答、多轮对话、数学、代码以及多模态能力都做了重大提升。10月24日，科大讯飞发布讯飞星火认知大模型V3.0，从文本生成、语言理解、知识问答、逻辑推理、数学能力、代码能力以及多模态能力方面都有了持续的提升。讯飞星火现在有1200多万的用户；已开放603项AI能力，在星火大模型的加持下，开发者团队的总数已经达到了550多万家，其中跟大模型直接相关的开发者是17.8万，其中企业级用户超过10万家、个人用户7万多；企业级用户涵盖了各个行业和领域，科大讯飞在个人用户中专门统计一个数据，有1.5万的助手开发者已经开发了2.9万的应用，他们不需要任何的软件编程能力，只要有想法有创意，就可以用大模型把他的创意变成产品、变成服务，从而加入到整个通用人工智能时代的创新和创业之中。2023年10月24日，科大讯飞还发布12个行业大模型，包括金融、汽车、运营商、工业、住建、物业、法律、科技文献、传媒、政务、文旅、水利，进一步强调在大模型时代要跟各个行业更深度地对接。

图丨讯飞星火认知大模型用户数量以及部分企业客户示例（来源：科大讯飞1024开发者节）

智谱AI

打造新一代认知智能大模型，专注于做大模型的中国创新

智谱AI（北京智谱华章科技有限公司）由清华大学计算机系的技术成果转化而来，致力于打造新一代认知智能通用模型，提出了Model as a Service（MaaS）的市场理念。其于2020年底开始研发GLM预训练架构，并训练了百亿参数模型GLM-10B，2021年利用MoE架构成功训练出万亿稀疏模型，于2022年合作研发了双语千亿级超大规模预训练模型GLM-130B，并基于此千亿基座模型打造大模型平台及产品矩阵。

基于GLM-130B，智谱AI主导构建了高精度通用知识图谱，把两者有机融合为数据与知识双轮驱动的认知引擎，并基于此千亿基座模型打造ChatGLM（*chatglm.cn*）。在斯坦福大学大模型中心的评测中，GLM-130B在准确性和公平性指标上与GPT-3 175B（davinci）接近或持平，鲁棒性、校准误差和无偏性优于GPT-3 175B。其团队研发的对话模型ChatGLM在GLM-130B基础上通过有监督微调等技术实现人类意图对齐。开源的ChatGLM-6B支持在单张消费级显卡上进行推理使用，全球下载量超过200万。此外，智谱AI也推出了认知大模型平台Bigmodel.ai，形成AIGC产品矩阵，包括高效率代码模型CodeGeeX、高精度文图生成模型CogView等，提供智能API服务。通过认知大模型链接物理世界的亿级用户、赋能元宇宙数字人、成为具身机器人的基座，赋予机器像人一样“思考”的能力。

智谱AI注重对国内科研及企事业单位的合作与支撑，中国移动、美团、360、联想、金山WPS等企业已基于ChatGLM模型从事领域大模型应用的研发。智谱AI还与首都之窗等机构基于ChatGLM共同进行政务大模型和应用的探索与服务。截至目前，ChatGLM模型先后为中科院多个院所、之江实验室、上海人工智能实验室、北京智源及多家知名高校、企业提供科研支持。自2022年初开始，GLM系列模型已支持在昇腾、神威超算、海光DCU架构上进行大规模预训练和推理，当前已支持10余种国产硬件生态，包括昇腾、神威超算、海光DCU、海飞科、沐曦曦云、算能科技、天数智芯、寒武纪、摩尔线程、百度昆仑芯、灵汐科技、长城超云等。

References

1. Generative AI deployment: Strategies for smooth scaling，MIT Technology Review Insights，2023

2. Laying the foundation for data- and AI-led growth，MIT Technology Review Insights，2023

3. The great acceleration: CIO perspectives on generative AI，MIT Technology Review Insights，2023

4. Humans at the heart of generative AI， MIT Technology Review Insights，2023

5. Hugging Face
https://huggingface.co/

6. The open-source AI development market map
The open-source AI development market map - CB Insights Research

7. 创新奇智“AInnoGC+智造实训”方案入选国家工信安全中心大模型应用创新典型案例
https://mp.weixin.qq.com/s/N6LoHKYVqokwacA5jGGyog

8.第四范式决定把大模型扔到一块无人地
https://www.4paradigm.com/content/details_739_36995.html

9. 科大讯飞1024星火V3.0发布刘庆峰演讲实录

Find Out More

https://www.mittrchina.com/

Contact Us

如您希望与我们交流或有任何问题，以及获取更多中国AI大模型企业统计信息，

请与我们联系：*research@deeptechchina.com*

Office

北京市朝阳区亮马河大厦2栋17层

浙江省杭州市余杭区仓前街道梦想小镇创业大街8幢B座

上海市徐汇区淮海中路1325号瑞丽大厦7层

广东省深圳市南山区云科技大厦7楼

MIT 科技评论
Technology
Review

版权声明

免责申明

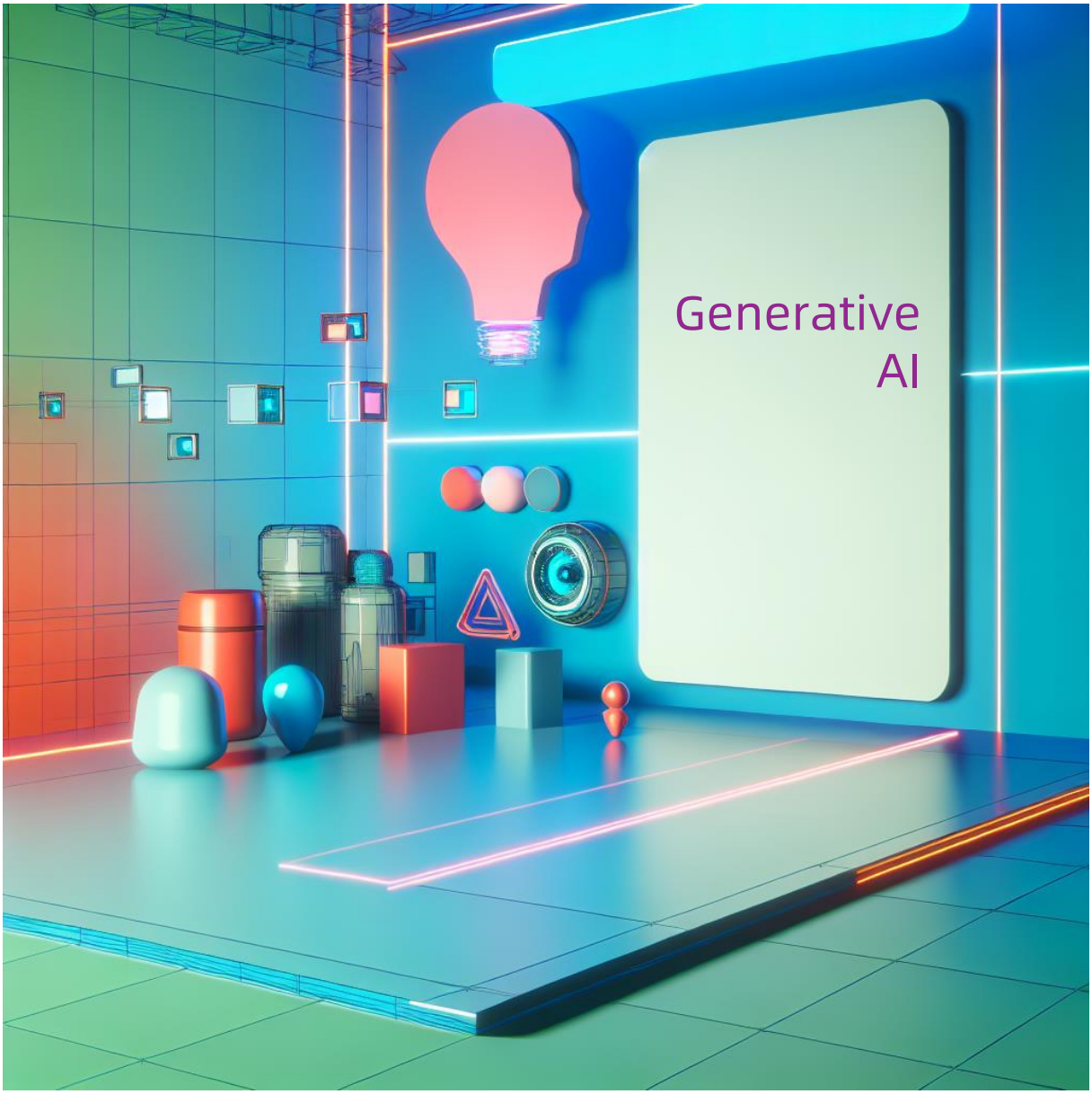
关于《麻省理工科技评论》中国

《麻省理工科技评论》（MIT Technology Review）依托麻省理工学院的学术和产业资源，于 1899 年在美国麻省理工学院创刊，是世界上历史悠久的科技商业智库。自成立之初，《麻省理工科技评论》就一直关注那些正在颠覆现有格局并创造新的市场机会影响人类社会的技术，以及那些正在从实验室走向市场即将商业化的技术。在此基础上，也高度关注将这些技术落地，并用这些技术影响我们生活的人和聪明企业。《麻省理工科技评论》于 2016 年落地中国，由 DeepTech 独家运营，开展媒体、研究及会议业务，围绕技术话题辐射和影响新兴科技圈层，重点关注新兴科技的商业化和社会价值，聚焦中国市场，为中国科技从业者带来与全球百万科技领域研究者、从业者及商业领袖进行前沿科技国际化交流的机会。

关于 DEEPTECH

DeepTech成立于2016年，是一家专注新兴科技的资源赋能与服务机构，以科学、技术、人才为核心，聚焦全球新兴科技要素的自由链接，为产业、政府、高校、科研院所、资本等科技生态的关键角色提供服务，通过科技数据与咨询、出版与影响力、科创资本实验室三大业务板块，推动科学与技术的创新进程。

MIT 科技评论
Technology
Review



DEEPTECH