



五大关键要素

解锁生成式AI全新机遇

执行概要

从基础模型到落地的生成式 AI 应用，需要经过模型训练、模型定制、模型部署、模型推理等环节。各个环节上的行业参与者均面临着不同程度的基础设施、数据集成、应用场景选择、安全与隐私以及负责任 AI 方面的挑战：

大模型提供商在训练基础模型的过程中，数据准备工作复杂而耗时，训练基础设施的管理与优化有较高的门槛，且算力成本高昂。对于生成式 AI 应用开发者而言，需要为应用场景选择最为适配的模型，也需要集成来自不同存储库、不同格式的数据，还需要保障用户数据隐私、模型安全以及生成内容的安全。对于生成式 AI 应用使用方而言，识别应用场景、利用自有数据与大模型进行交互的过程中如何保证数据安全，如何确保模型生成内容所问即所答，都是其挑战所在。

整体来看，能够帮助生成式 AI 新生态里的产业参与者加速生成式 AI 应用落地的解决方案，必须具备五大要素：

高性价比的基础设施、丰富而灵活的模型选择、使用私有数据实现差异化定制、开箱即用的生成式 AI 驱动的应用和负责任的 AI。关于部署方式，现阶段用户更倾向于从云端开始部署：一方面大模型尚处于早期，模型能力不断突破，云服务能够助力用户快速使用到最新的模型能力；另一方面，使用云服务也可以助力用户降低基础设施搭建与运维的时间成本，因此云服务商成为各类玩家构建和部署生成式 AI 应用的首选合作伙伴。



开启云上生成式 AI 之旅：

有必要尽快将生成式 AI 融入企业级战略，确定是自建模型还是使用模型定制功能来发挥企业的差异化优势。在参考外部诸如提高生产力、增强用户体验以及优化业务流程成功实践的同时，筛选确定自身应用场景。端到端构建生成式 AI 应用，则从为应用场景选择模型开始，参考当前的成功实践，选择适当的模型定制路线进行模型的适配与调整，将经过评估验证的模型集成到企业应用系统，并设立评估指标，对应用进行持续迭代。



亚马逊云科技助力释放生成式 AI 潜力：

亚马逊云科技致力于不断降低机器学习使用门槛，面向生成式 AI，公司提供丰富的算力选择与高效的加速训练与推理；在海外区域推出 Amazon Bedrock 助力轻松构建并规模化生成式 AI 应用，以及企业级生成式 AI 助手 Amazon Q；高度重视为用户提供负责任的 AI 策略支持。此外，公司也在持续建设端到端的数据基座，以支持生成式 AI 的数据集成需求。至今，超过 1 万家客户利用亚马逊云科技 Amazon Bedrock 进行生成式 AI 创新。



CONTENTS

01 /	生成式 AI 应用：热度高、落地慢	01
	1.1 近六成企业开始拥抱生成式 AI	02
	1.2 生成式 AI 落地不及预期	03
02 /	构建生成式 AI 应用的现实挑战	05
	2.1 生成式 AI 三类主要参与者	06
	2.2 数据准备耗时，训练成本高昂	08
	2.3 数据集成复杂，模型适配难，安全要求高	09
	2.4 应用场景筛选难，数据隐私安全顾虑，缺乏成功部署实践	11
03 /	云服务助力构建生成式 AI 应用	13
	3.1 落地生成式 AI 应用的五大关键要素	14
	3.2 具备成本效益和领先 AI 应用的云服务加速构建生成式 AI	17

04 /	从云端开始构建生成式 AI	19
	4.1 制定融入生成式 AI 的人工智能战略	20
	4.2 规划生成式 AI 应用场景	22
	4.3 端到端构建生成式 AI 应用	37
 05 /	 亚马逊云科技助力企业解锁生成式 AI 潜力	 41
	5.1 高性价比且丰富的基础设施加速训练与推理	43
	5.2 丰富的模型可选,可基于私有数据差异化定制	44
	5.3 生成式 AI 驱动的应用程序	47
	5.4 亚马逊云科技注重负责任的 AI	50
	5.5 端到端数据基座打通壁垒,让所有应用可治理、可管控	51

01

第一章

生成式 AI 应用： 热度高、落地慢

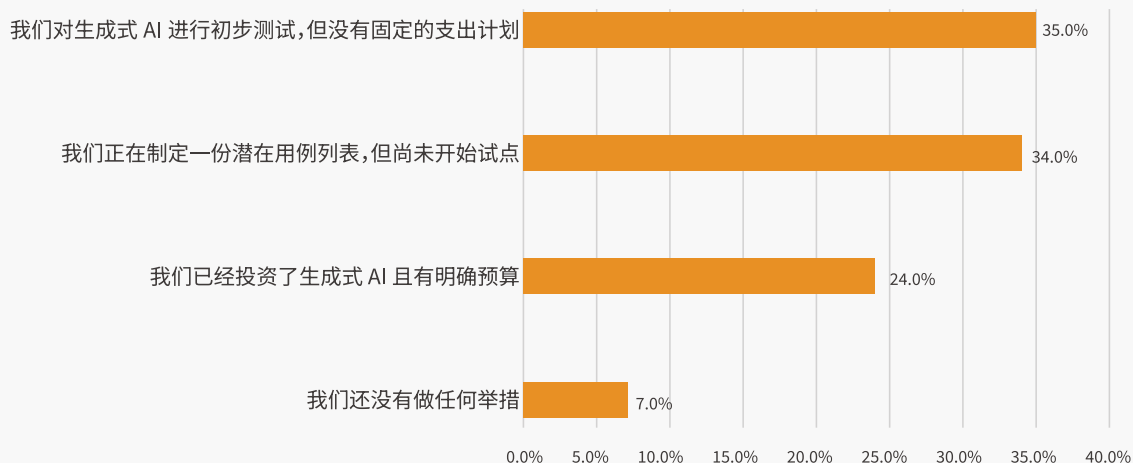
随着生成式 AI 应用快速进入市场，以及越来越多大模型的不断面世，2023年已经成为生成式 AI 的元年：不同区域、各行各业、不同领域的人们开始尝试在工作和生活中使用生成式 AI，以探索各种可能性。于企业而言，生成式 AI 也已经成为从董事会到基层员工都在探讨热议的话题，一些领先的企业甚至已经利用生成式 AI 技术实现了业绩增长。

1.1 近六成企业开始拥抱生成式 AI

在 2023 年第四季度的生成式 AI 用户调研中，IDC 发现，已经有近六成的企业开始拥抱生成式 AI，超过三成企业开始制定潜在应用场景，仅有 7% 的企业还没有任何举措。

- 24% 的企业表示已经投资生成式 AI，并且有相应的预算计划。这些预算可能来自 2023 年已有的规划，也有些是针对生成式 AI 新增的预算。
- 35% 的企业表示已经做了一些初步的测试与概念验证，但还没有明确的投资计划。
- 34% 的企业表示正在制定潜在应用场景，但还没有进入试点阶段。
- 仅有 7% 的企业表示目前还没有任何举措。

图 1 贵组织目前评估或者使用生成式 AI 的进度如何？



来源：IDC 2023, N=100

1.2 生成式 AI 落地不及预期

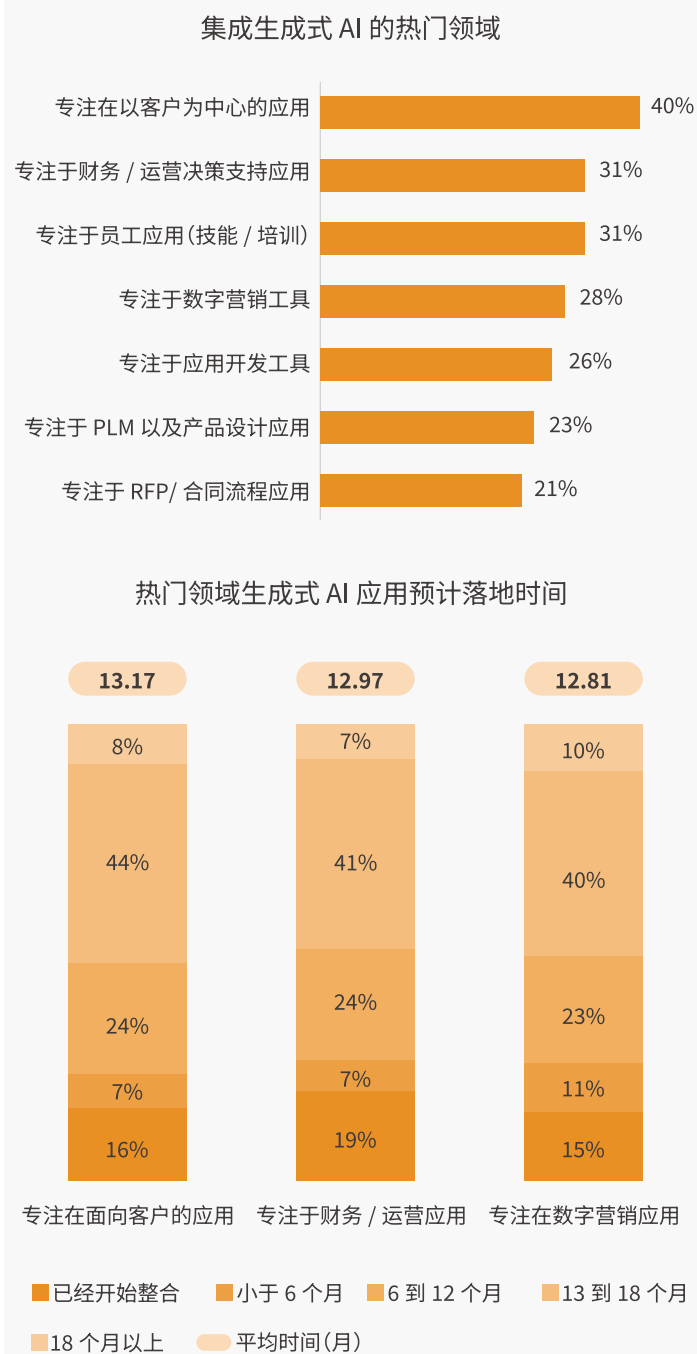
在已经开始尝试生成式 AI 的企业中,应用场景遍及企业各个业务流程,包括客户服务类应用、支持运营决策类应用、提高员工生产力的应用、辅助产品设计的应用、支持营销的应用等。其中,高管最为关注且最希望从中获取价值的前三大应用为:智能客服类应用、支持财务和运营决策的应用以及专注于提升员工生产力的应用。在实际探索中,预计能优先落地的场景主要是生成式 AI 支持数字化营销领域的应用,支持财务、运营决策类的应用,以及智能客服类应用。

在落地时间方面,五成以上的企业认为各个领域的应用落地时间都需要在 13 个月以上,这远远长于企业期望的 6 个月的时间。

在应用范围方面,即使已经上线的应用,范围也都比较局限,以单点应用为主,如基于大模型升级的智能客服;当前,单个项目的应用可能也就局限于小部分客户群体或者不足百路的坐席中。

综合来看,模型生成结果的准确性、人才技能稀缺、部署的成本等,都影响了生成式 AI 的规模化落地。深入探索企业生成式 AI 的实践,发掘落地前、中、后的全链路问题,让企业能够理性选择技术方案、大胆尝试应用场景,才能真正实现以生成式 AI 带来差异化竞争优势。

图 2 企业高管最关注的生成式 AI 应用以及预计落地时间



来源：IDC 全球生成式 AI 调研，N=1363

科技行业正处于开创性时刻，生成式 AI 使预测变得更加容易，甚至可能更加普惠、价格更加亲民。企业将把生成式 AI 应用到更多全新的领域，带来众多商业模式和竞争战略的改变。IDC 预计，生成式 AI 将在未来 10 年内为全球 GDP 增加近 10 万亿美元。然而，高关注度、高预期并没有带来相应的落地进展。厘清生成式 AI 应用落地的路径，寻找能够加速应用落地的解决方案，是行业参与者的当务之急。

02

第二章 ▲

构建生成式 AI 应用的 现实挑战

相比上一代 AI，生成式 AI 应用的开发需要重新在各种大模型之间进行选型适配，对于用于模型训练的数据的安全性要求更高，在模型定制环节的工作有更加多样性的选择，也因此形成了全新的生成式 AI 应用产业生态。而生态上的所有参与者，今天也面临着诸如数据、算力、成本、模型适配、安全隐私等挑战。

2.1 生成式 AI 三类主要参与者

从基础模型到集成到企业系统中的生成式 AI 应用，简单说就是对基础模型进行定制、调整，再进行推理的过程。其中，基础模型的训练，对于数据、算力以及技能、人才的要求都很高，因此只有少部分企业专注于模型的训练，即模型提供商。更多的企业选择依托基础模型，使用自有数据进行调优、采用提示词工程 (Prompt Engineering) 的方法，以提升模型生成内容的准确性；或者使用检索增强生成 (下文简称：RAG) 的方法以充分利用企业自有知识使模型更好地理解上下文的语境，给出更相关的答案——这两类群体为生成式 AI 应用开发者。行业应用解决方案商则是将生成式 AI 应用集成到现有的系统中，开发出 AI 赋能的行业应用；大部分行业用户也更倾向于采用融入生成式 AI 能力的行业应用，来实现智能化升级。

图 3 大模型驱动的生成式 AI 价值链



来源：IDC, 2023

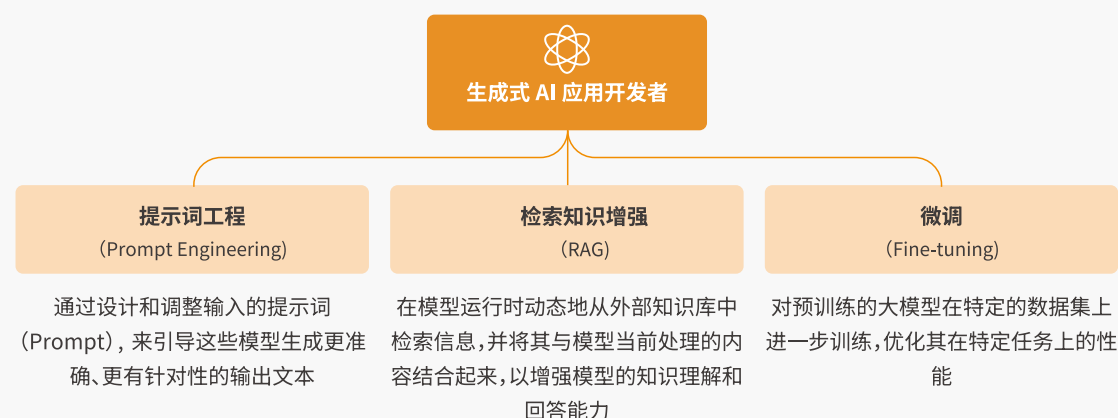
► 大模型提供商——基础模型训练

大模型提供商从零开始构建基础模型，参数量达到百万、千万甚至亿级别，典型厂商如 Anthropic、Al21 Labs、Meta、cohere、OpenAI、Stability.ai、智谱华章、百川智能等。当前阶段，通常只有科技巨头、头部创企才有能力自主进行大模型训练，这些大模型也是生成式 AI 应用创新的重要源头。从模型的供给侧来看，随着大模型提供商的逐渐发展和壮大，参数量的不断增加，这些模型应用到生产环境中的准确性也相应的越来越高，过去不可能达到的预测准确度也开始成为可能。

► 生成式 AI 应用开发者

从基础模型得到最终应用，需要将特定领域数据引入基础模型进行再次训练 / 调优，才能获得理想的效果。如前所述，当下有三种方式可以实现模型的定制：调优，提示词工程，检索增强生成 (RAG)。生成式 AI 应用开发者可以基于市场上的商业化或者开源模型，采用专有数据对模型进行调整，来构建特定场景下的生成式 AI 应用。各行业的解决方案商可以采用基础模型基于领域数据训练领域大模型，行业企业则可以使用模型定制功能开发企业专属应用，市场上也涌现出越来越多的 AI 原生企业采用大模型开发出各种创新的应用。

图 4 生成式 AI 应用开发者



来源：IDC, 2023

► 生成式 AI 应用使用方

将大模型驱动的生成式 AI 应用程序，部署到企业现有的系统中。采用构建好的生成式 AI 应用，可以将其应用在内部知识问答、自动邮件生成、代码补齐等场景中，也可以集成在诸如新药研发、预测性维护、广告推荐、智能营销等行业垂直应用中。

2.2 大模型提供商的挑战：数据准备耗时，训练成本高昂

现阶段，作为生成式 AI 应用创新的源头，模型提供商在模型训练环节面临着一系列关键挑战，比如数据准备耗时、计算存储网络需要提升性能、训练成本高昂、生成内容需要减少幻觉且需要符合监管等。

- **数据准备耗时耗力：**相比中小模型，用于大模型训练的数据有三个特征——数据量大，对数据质量的要求更高，对数据安全的要求更高。在数据准备环节，模型提供商在海量数据清洗、关键数据脱敏等环节需要消耗大量时间精力。为确保建立负责任的 AI，也需要对用于模型训练的数据进行大量的审核工作。因此，缩短数据清洗时间，提高数据准备工作效率，是当前的迫切需求。
- **基础架构管理与优化门槛高：**大模型参数量往往达到千亿、万亿级别，用于模型训练的数据会达到千百万个数据点。训练过程中，无论大小文件的存储、海量小文件的读取，还是算力集群的稳定性、数据的高速写入写出，都要求团队具备专业的基础架构相关技能。
- **用于训练的基础设施成本高昂：**大模型的单次训练通常会用到千卡甚至万卡级别，因此计算所需要的算力相比以往中小模型时代大幅提升。如何在保证计算性能、稳定性的前提下，降低整体成本，是大模型提供商的迫切需求。

- ▶ **确保模型效果不断提升：**能够确保大模型走向市场的前提在于模型生成内容的准确性，以及怎样确保模型少一些幻觉，实现“所答即所问”，这也是模型提供商重点投入的方向。此外，未来 2 ~ 3 年，如何能够确保大模型效果持续提升，也是大模型提供商的当务之急。

2.3 生成式 AI 应用开发者的挑战：

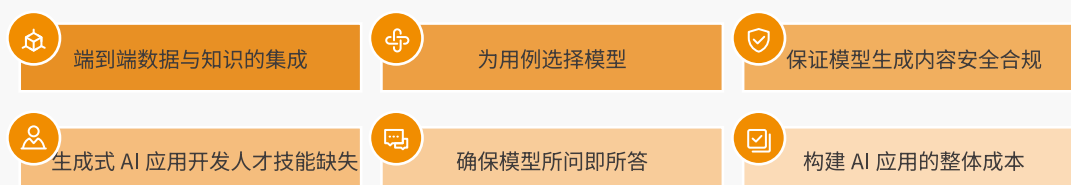
数据集成复杂，模型适配难，安全要求高

生成式 AI 应用开发者是打通最后一公里、实现技术规模化复制的重要环节，然而当前生成式 AI 的应用开发，也面临着诸多挑战。

- ▶ **模型选择与评估耗时且无参考依据：**一方面，今天的大模型应用仍然处于早期，企业中的应用场景纷繁复杂，开发人员需要花时间去尝试，以便为应用场景选择效果最优的模型；另一方面，在同一个应用中可能会用到多个模型，而不同模型的 API 也存在较大差异。如何能在模型选择与评估环节提高效率，是生成式 AI 应用开发的首要挑战。
- ▶ **模型适配与调整门槛较高：**从基础模型到能上线的生成式 AI 应用，离不开对于模型的调优，或者其他模型定制功能来发挥大模型的差异化优势。而大部分企业在这方面并没有任何储备。对于采用 RAG 的用户，在将来自企业内外部数据做成外挂知识库的过程中，需要选择专门的数据库来存储知识，这一数据库既可能是全新的向量数据库，也可能是具备向量引擎能力的传统数据库。对于微调模型的用户，模型调优环节需要对模型结构有基础的了解，调优也会带来一定的算力要求。即使只是采用提示词工程的方式来使用大模型，也需要有专业的提示词工程师。上述这三种方式，当前市场上的成功实践均比较少，企业内部也没有相关的积累，入门的门槛较高。

- ▶ **推理部署成本高昂：**未来，将有越来越多的生成式 AI 实践开始走向推理阶段，也将有越来越多的用户开始使用生成式 AI。如何提升推理性能，同时大幅降低推理成本，是整个行业的迫切需求。
- ▶ **缺乏数据 / 知识集成的经验：**在生成式 AI 时代，无论是模型定制环节还是模型推理环节，都会涉及到大量的数据交互工作。这些领域的专有数据将造就一家企业的不同。例如，对于一家数据储备完善、拥有高质量数据集的企业，若能基于自有数据对模型进行微调，或者以插件的形式结合自有数据来生成内容，则大模型生成的结果将得到显著提高。而今天，大部分企业还是存在不同程度的数据烟囱现象，数据分散、数据归集难、数据管理复杂是用户普遍会遇到的问题。建立端到端的自动化数据集成管道，减少跨数据存储库的数据集成工作，能够加快应用的上线。
- ▶ **工程化落地复杂：**在落地过程中，大模型本身只能解决 20 ~ 30% 的问题，从大模型到落地的应用，基础设施的搭建与优化，模型的部署与运维，数据工程等，这些周边的工程能力要解决剩下的 70 ~ 80%。这其中每一个环节都需要优化，很多企业恰恰缺乏这方面的经验和能力。
- ▶ **安全要求高：**生成式 AI 应用开发者，既需要保证用户用于模型定制的数据不被基础模型所吸收，自身优势不被友商所利用；又需要确保模型输出内容的合规，避免答非所问等幻觉问题。这意味着，生成式 AI 应用的开发和使用将对安全提出更高的要求。

图 5 生成式 AI 应用开发环节的挑战



来源：IDC, 2023

2.4 生成式 AI 应用使用方的挑战：

应用场景筛选难，数据隐私安全顾虑，缺乏成功部署实践

- **难以确定应用场景：**企业对生成式 AI 的期待很高，然而从哪些应用场景着手，如何筛选、判定应用场景，以及如何寻求一些可参考的成功案例，是企业着手生成式 AI 项目面临的首批挑战。
- **对于安全、隐私的担忧：**成熟地运用生成式 AI 工具可以显著提升员工工作效率，提升创造力，然而用户也会担忧将自有数据、文件上传到大模型，是否会泄露私有数据与知识。这也是很多具备海量数据的企业的首要担忧，一定程度上也影响了生成式 AI 的落地进程。
- **缺乏成功的部署实践：**训练好的模型需要在保留数据集的基础上进行性能、准确性等方面的测试。部署上线推理过程中，也需要注意可能的数据漂移。而在应用不断迭代的过程中，当前也缺乏成熟的模型效果监测机制。
- **成本高企：**生成式 AI 应用上线后，一方面用户使用频率增加，与模型的每一次交互，都会伴随推理成本的增加；另一方面，并发量、吞吐量、时延要求的不断提高，对算力的要求也会不断提高。只有不断降低使用成本，才有可能真正促进 AI 的普惠。
- **人才 / 技能不足：**今天，几乎所有参与生成式 AI 产业的企业，都面临着严重的人才短缺问题。大模型相关人才稀缺且成本高昂，导致企业相关技能的不足；具备模型调优等定制能力的人才尤为短缺。

企业落地生成式 AI 应用的现实挑战

整体来看

要加速生成式 AI 应用的构建与落地，从应用场景识别到模型选择，模型适配与定制，模型部署，各个环节都需要助力，以下五大关键挑战是当前企业拥抱生成式 AI 时面临的现实问题：

- 模型选择与适配难；
- 如何充分发挥自有数据优势；
- 面向 AI 的基础设施的优化；
- 如何确保负责任的 AI；
- 以及推进企业应用程序的 AI 规模化。

要克服这些挑战，既需要产业上下游的生态协作与优势互补，也需要依托能够应对这些挑战的技术栈，在此基础上，才能加速生成式 AI 应用的落地。

图 6 企业落地生成式 AI 应用的现实挑战



03

第三章 ▲

云服务助力 构建生成式 AI 应用

尽管现阶段使用大模型的挑战、担忧居多，展望其潜在的商业价值，用户的期待仍大过今天的担忧。截至 2023 年第四季度，已经有越来越多的企业开始从观望转向试点阶段。当然，大部分应用场景目前仍处于探索创新、试点阶段。企业希望寻求广泛的外部合作以快速获取生成式 AI 相关能力；同时，从技术的角度，也希望寻找可轻松入门的生成式 AI 开发技术栈。

3.1 落地生成式 AI 应用的五大关键要素

从基础模型到成功构建生成式 AI 应用，每一个环节的行业参与者都需要应对基础架构、端到端数据、保护数据隐私等挑战。助力生成式 AI 应用的落地，必须具备如下因素：

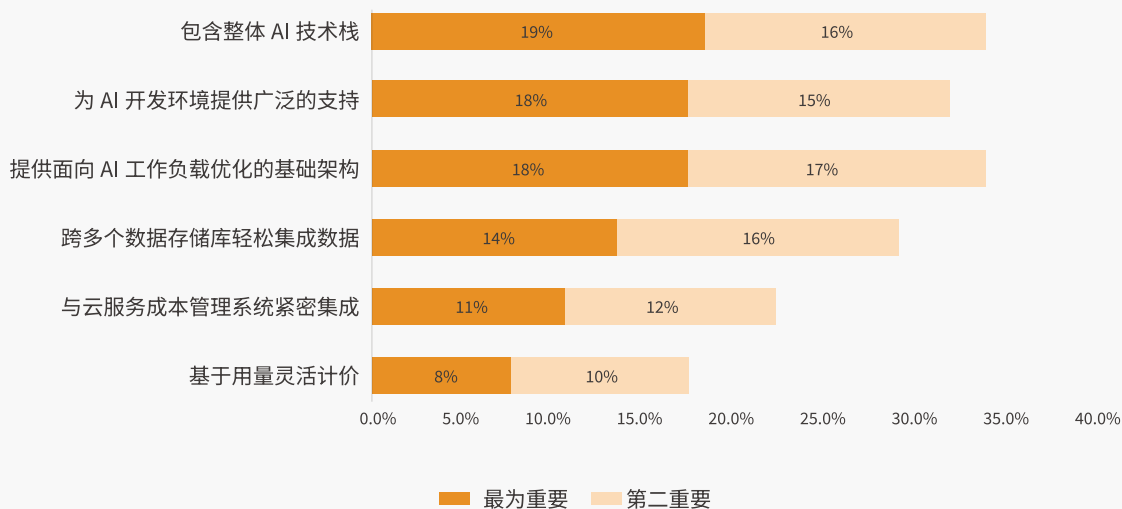
图 7 构建生成式 AI 应用的关键能力



来源：IDC, 2023

- **高性价比的基础设施：**无论是面向大模型的训练还是调优推理，低成本、高性能、高度优化的云基础设施，是试点以及真正部署生成式 AI 应用的首要条件。如下图所示，能够提供包含整体 AI 技术栈而无需用户做硬软件的适配，面向 AI 工作负载管理并优化计算、存储与网络设施，为 AI 应用开发提供从部署到上线的全面支持，是生成式 AI 解决方案必须具备的能力。此外，与云服务成本管理系统无缝集成，以帮助用户实时监测生成式 AI 基础设施的用量以及成本，提供灵活的定价模式以更好地管理生成式 AI 应用的整体成本，也是基础设施层需要具备的重要能力。

图 8 您的组织在未来 18 个月内在生成式 AI 基础架构提供商中寻找哪些重要特征？

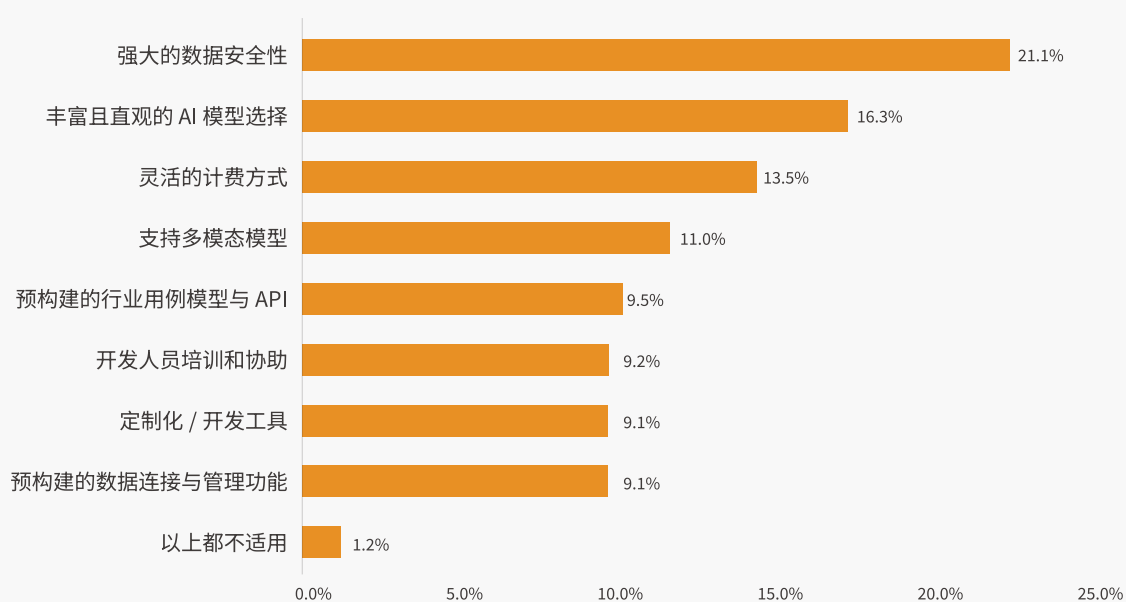


来源：IDC 中国,生成式 AI 市场调研,N=100, 2023

- **丰富的基础模型：**现阶段市场上还在涌现新的大模型，针对原有的大模型也不断有新的版本发布。然而，单一的大模型无法做到满足用户在不同业务场景的全部需求，因此，构建生成式 AI 应用需要供应商做到能够提供丰富、灵活的模型选择，并为用户提供模型评估、验证、适配的工具，以帮助用户减少试错时间，加快入门速度。
- **差异化定制生成式 AI 应用：**要构建适配企业自身业务场景和个性化需求的生成式 AI 应用，必须要具备成熟完善的模型定制功能，现阶段即提供微调、提示词工程以及知识库增强的功能。此外，开发人员都需要连接来自数据库、数仓、数据湖的数据，需要从企业现有的应用中抽取数据，最终也需将模型集成到应用系统中。因此，预训练的行业应用场景模型与 API，预置的数据连接与数据管理功能，对开发人员的培训和协助、定制化开发工具，都是企业构建和部署生成式 AI 的需求。

- **开箱即用的生成式 AI 驱动的应用：**基于生成式 AI，用户既可以构建原生的 AI 应用，也会涉及到将原有的应用系统进行逐步、持续的 AI 升级。采用生成式 AI 驱动的应用程序，更可以帮助用户提高开发效率，加速应用上线。一方面，用户需要将生成式 AI 应用由单点扩展到推进企业所有应用程序的 AI 化。另一方面，构建生成式 AI 应用程序的过程，也需要自动化的、生成式 AI 驱动的工具来加速。

图 9 根据您当前或近期使用软件平台和工具提供商开发生成式人工智能能力的计划，您的组织在未来 18 个月内对于软件供应商有哪些需求？



来源：IDC 中国，生成式 AI 市场调研，N=100,2023

- **负责任的 AI 策略：**在数据安全、云服务安全以外，开发与部署生成式 AI 应用尤其要强调建立负责任的 AI 策略，这就要求确保输入到模型的数据合规，输出的内容符合“所问即所答”，且符合监管要求。一方面，在生成式 AI 领域，数据与知识对于塑造企业差异化竞争优势至关重要。企业尤其不希望自己的优势被外部所利用，故构建与部署生成式 AI 应用过程中保证数据与知识“不出私域”十分重要。另一方面，对于生成的内容，如何减少负面 / 有害内容，确保合规，也是应用上线的前提。另外，当今社会的监管合规要求日益严格，如何保证生成式 AI 应用符合当地的监管政策，也是生成式 AI 工具需要具备的能力。

在五大要素之外，更需要注意的一点是企业也需要全面、集成、可治理的端到端数据底座：

构建生成式 AI 应用的完整流程上涉及的数据来源更广泛——不仅是训练环节要读取数据，推理环节也要读取来自内外部不同系统的知识。当下，企业的 IT 系统可能有成百上千个，海量数据存储在不同架构的存储库中。支持生成式 AI 应用开发的解决方案必须要能够端到端集成数据，以支持模型的调优与推理。

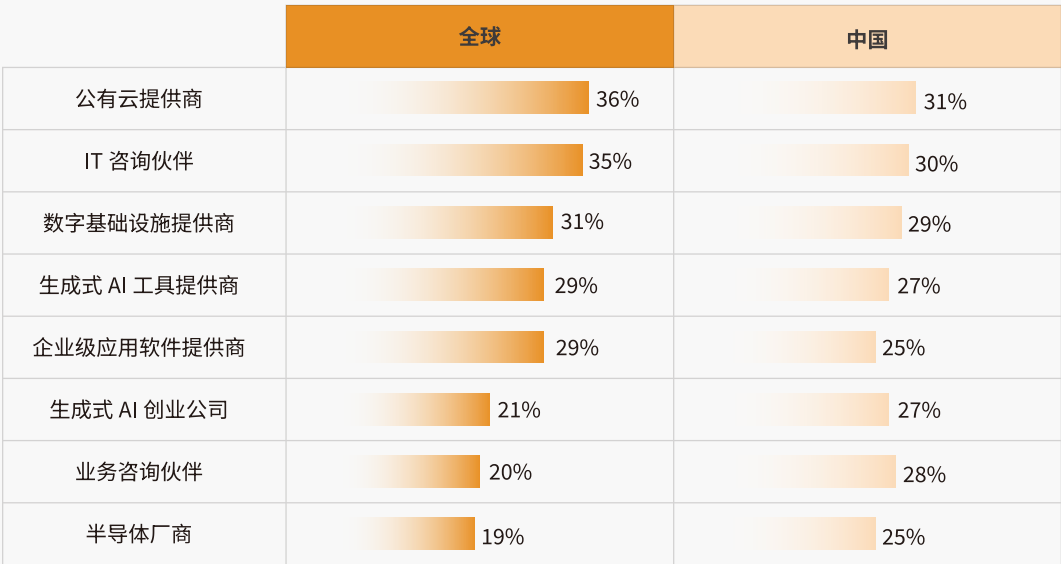
3.2 具备成本效益和领先 AI 应用的云服务加速构建生成式 AI

当前，生成式 AI 领域日益繁荣，既有全新的大模型初创企业、创新的生成式 AI 应用厂商，也有传统的解决方案商、云服务商。而在企业启动生成式 AI 应用时，根据 IDC 调研，无论在全球还是中国，云提供商都是企业用户在选择生成式 AI 战略合作伙伴时的首选。一方面，云服务既能够帮助用户降低基础设施搭建门槛、节省基础设施维护时间，又能够实现 PaaS 层产品的快速更新；另一方面，在生成式 AI 这一前沿领域，尤其会面临模型与应用快速迭代的情形。在云端构建生成式 AI 应用，可以让用户专注于使用新的能力快速构建应用，也因此成为构建生成式 AI 应用的首选合作伙伴。

对企业而言，如何搭建面向生成式 AI 的架构，从何处开始着手，甚至如何获得对生成式 AI 的基础认知，也都需要外部咨询合作伙伴提供建议。此外，生成式 AI 工具提供商、企业级应用软件提供商、初创企业等，也是建设生成式 AI 重点考虑的合作伙伴。

图 10 企业对于生成式 AI 合作伙伴的选择偏好

Q: 在未来 12 个月,您更可能选择哪类厂商作为最重要的生成式 AI 战略合作伙伴?



来源 : IDC2023, N=890, China N=10

IDC 预测,生成式 AI 将重塑传统应用程序,到 2027 年,企业将利用生成式 AI 工具和云服务提供商平台启动并执行 50% 的代码转换和开发任务。

——IDC FutureScape: 全球开发者和 DevOps 2024 年预测—亚太区(不含日本)启示

04

第四章 ▲

从云端 开始构建生成式 AI

生成式 AI 的真正价值和投资回报率取决于企业本身如何推动生成式 AI 的落地，也取决于生成式 AI 技术的应用如何为企业带来业务层面的改变。企业应该尽快评估准备情况，制定人工智能战略与落地路线图，为生成式 AI 的应用奠定必要的基础，从而在中长期内通过差异化和重点战略来建立竞争优势。

4.1 制定融入生成式 AI 的人工智能战略

4.1.1 将生成式 AI 融入企业战略

评估生成式 AI 的准备情况，需要关注智能化架构的基础、人才技能的储备和数据基础，制定融入生成式 AI 的人工智能战略和路线图，并在人才、技能、流程方面进行保障式的资源适配。

在生成式 AI 的时代，尤其要建立负责任的 AI 应用机制。大模型通常缺乏透明度（与简单的统计模型不同）来解释它们为什么会产生结果，而这种缺乏信任是在实践中使用生成式 AI 的真正风险。企业在启动生成式 AI 之前，应该首先建立负责任的 AI 策略。这要求企业理清组织数据在 AI 模型中的使用和存储方式，确保数据受到保护；要求企业对于输入模型的数据以及模型输出的内容实时进行合规审查。同时，在选择外部服务商的产品时，亦应考量其是否提供了相应的功能。

图 11 企业落地生成式 AI 的战略思考框架



来源：IDC, 2023

4.1.2 根据自身情况确定生成式 AI 构建策略

如今,市场上已经提供了丰富的开源模型选择,也有一些相对成熟的 API 类应用程序。对于有一定开发实力的企业来说,可以考虑基于大模型进行微调,将微调后的模型嵌入到生产环境。

- ▶ **直接与模型进行交互——开箱即用、成本领先：** 不对模型做任何开发或训练,将调整好的大模型 API 或者 SDK 直接嵌入应用程序, 与该应用程序直接交互。这种“开箱即用”的方式无需任何级别的编程或数据工程即可实现必要的功能, 同时还可将基础设施管理及其相关运营支出“外包”给应用程序提供商。
- ▶ **模型定制以实现差异化优势：** 通过微调、知识增强生成或者提示词工程的方式对模型进行调整, 以更好地适配自己的应用场景。该情形下, 需要企业有一定的技术能力, 对模型调整方法有一定的认知, 才能充分发挥出差异化优势。
- ▶ **自主训练模型：** 相比于基于基础模型进行调整的方式, 企业自主进行模型的预训练, 需要搭建基础设施, 准备用于模型训练的数据, 还需对用于训练的数据进行合规性审核。无论是硬件、软件栈, 还是人力投入方面, 投入程度均将明显高于前两类。

图 12 根据自身情况确定生成式 AI 构建策略

构建生成式 AI 应用的路径	人才、技能要求	整体成本	竞争优势	适用情景
 采用开箱即用的解决方案	无需任何编程、数据工程操作, 任何员工可以直接上手使用	启动成本低, 尤其从云上开始使用该项能力, 启动成本更低	无法利用到企业专有数据, 模型为企业带来的差异化竞争优势低	- 企业没有对生成式 AI 以及大模型的基础认知的情形 - 对应用场景进行早期探索的情形
 模型定制	需一定的数据工程与编程技能	涉及到模型调优, 会产生一定的算力成本。使用提示词工程、RAG 方式, 不会带来成本显著上升	可以利用企业专有数据, 能够充分发挥企业的差异化优势	企业正式开始测试生成式 AI 用例的最佳启动方式
 自主训练模型	需要数据工程以及专业的算法开发技能	模型训练到部署各个环节都会带来显著的成本上升	可以利用企业专有数据来发挥差异化竞争优势	主要适用于模型提供商

来源 : IDC, 2023

整体来看，采用开箱即用的解决方案可以让企业以较低的成本快速使用模型能力；自主训练模型需要内部有非常专业的大模型领域的人才以及技能，且整体投资成本高昂；而采用模型定制 / 调整的方法，可以在成本可控范围内充分发挥企业专有数据的优势，因此是现阶段快速启动生成式 AI 应用的较优选择。

4.2 规划生成式 AI 应用场景

4.2.1 评估可能的生成式 AI 应用场景

评估生成式 AI 的应用场景，需要充分认识到其潜在回报和可能需要的资源、时间投入，以便更好地规划和落地其在各个领域的具体应用。如下图所示，在具体的场景选择上，优先从有一定数据基础、生成式 AI 能带来一定的投资回报、以及能够获得业务部门支持的场景开始，也可以考虑从价值高、实施复杂性低的场景入手，同时避免一些可能带来高风险的应用场景。

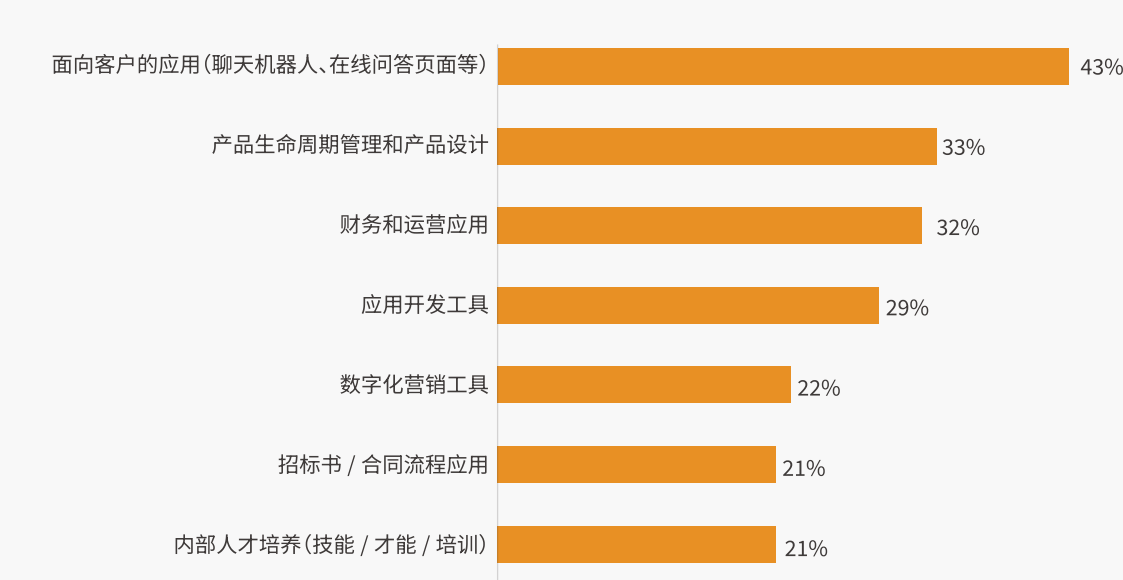
- 从价值的角度，企业应该重点考虑具备经济价值的场景，如 ROI 可衡量的场景；同时也应该从战略统一的角度来选择更具战略意义的场景。
- 从实施复杂性的角度，综合考虑企业自身的数据基础、评估是否有参考的算法模型、企业相关的流程支持以及自身具备一定技能的场景。
- 考虑从低风险的应用场景开始，首先要解决具体的现有业务问题，避免华而不实的应用场景，以简单方法让员工更有效地利用生成式 AI 提高生产力、拓宽员工创新和决策的空间。

图 13 生成式 AI 应用场景评估框架

维度		说明
 价值	经济价值	优先考虑能够快速看到投资回报的场景。 例：生成式 AI 驱动的数字员工可以代替人类员工服务更多的客户；生成式 AI 驱动的个性化推荐，助力获客等。
	战略一致	考虑与企业战略目标一致，能助力发挥出企业竞争优势的场景。
 复杂性	数据	寻找企业有数据积累的场景；同时也应考虑有迫切需求但没有数据积累的场景，流程上应该如何促进业务数字化 / 数据化。
	算法	选择有成熟算法支持的场景。
	专业技能	优选内部可轻松获取该领域技能的场景。如内部不具备该技能，则需要考虑外部合作伙伴是否具备该技能。
 风险	使用风险 / 实施风险	高风险的场景可能包括：合规风险高，实施过于复杂（涉及部门众多、流程繁琐）

来源：IDC, 2023

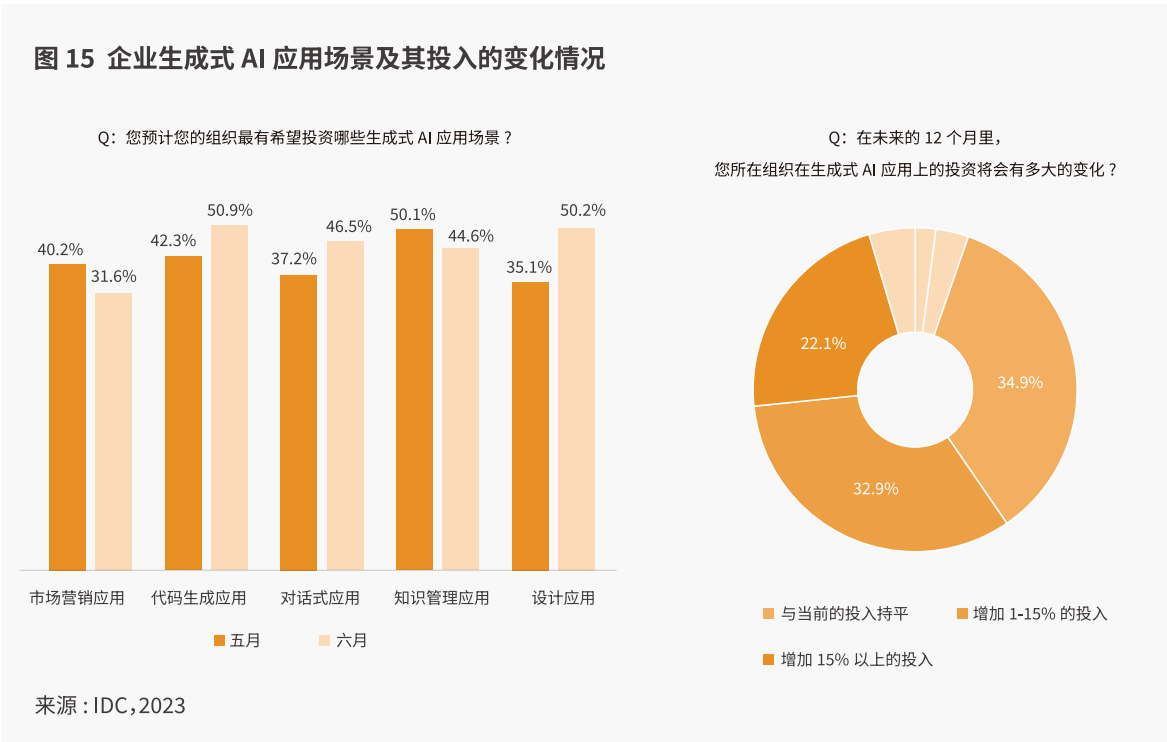
图 14 在将生成式人工智能集成到现有应用或业务流程中时，贵公司高层最关注的应用领域是什么？



来源：IDC 中国,生成式 AI 市场调研,N=100, 2023

4.2.2 参考外部实践选择应用场景

目前，不少行业已经在开展生成式 AI 应用的探索。IDC 调研显示，55% 的企业未来一年内将增加用于生成式 AI 的投资，其中 22.1% 的企业表示相比于去年将增加 15% 以上。



对话式营销成为主流：到 2026 年，普遍存在的情绪和意图人工智能将推动 50% 的 A500 企业以实时、双向对话的方式进行所有营销旅程，从而将购买转化率提高 40%。

IDC FutureScape：全球首席市场营销官 2024 年预测—亚太区（不含日本）启示

落地的典型应用场景主要围绕增强客户体验、提升员工生产力与创造力以及优化业务流程等，典型应用场景如下图所示。



➤ 增强客户体验

主要用于客户服务环节。典型的场景包括聊天机器人、联络中心呼叫分析、智能导购、数字人导购等。使用生成式 AI，可以更好地判断客户的情感、预测客户的需求，为客户提供更精准的服务。

➤ 提升员工生产力与创造力

生成式 AI 可以为企业内全体员工提供对话式搜索能力，以对话的方式快速获取企业当前运营情况的洞察；对于 IT 人员，生成式 AI 可以赋能自动生成代码、代码补全、不同开发语言的转换等业务；在具体的产品设计环节，生成式 AI 还可以辅助进行产品的设计，例如工业产品参数设计、服装图案设计等。

➤ 优化业务流程

生成式 AI 可以应用在文档处理场景中，用于自动提取、分析文档内容，以及生成内容摘要，也可以基于已有内容自动补全、生成新的内容。



案例背景

西门子股份公司(以下简称“西门子”)是一家专注于工业、基础设施、交通和医疗领域的科技公司。在知识的查询和回复这一日常场景中,由于西门子内部的数据横跨多个不同领域、涉及多个不同业务单元,长期以来,员工在进行内部资源的检索和调用时遇到诸如结构散乱、检索速度慢、交互不便等问题,知识的查询、回复生成速度慢、搜索关键词的命中率低,造成员工的使用体验不佳。



解决方案

- 亚马逊云科技提供的智能知识库暨智能会话机器人的解决方案指南;
- 解决方案指南中包括预训练的大语言模型、Amazon OpenSearch Service 的向量数据服务,以及相关系统集成等;
- 采用了检索增强生成(RAG)方式来增强模型;
- 通过应用大数据库和生成式 AI,有助于提高“智能知识库”的可用性。



业务价值

- 西门子在 3 个月内构建了基于自有模型的智能知识库暨会话机器人 " 小禹 " ；
- 首周有 4000+ 员工使用,解答了 12000+ 问题；
- " 小禹 " 避免了重复开发,提升了员工的信息获取效率,优化体验；
- 基于云的弹性资源和托管服务进一步降低了成本。

3 个月

灵活构建基于自有模型的
AI 对话机器人

4000

首周服务的用户数

12000

首周回答的问题



“在亚马逊云科技的帮助下，我们仅仅用了三个月时间就上线了生成式 AI 对话机器人‘小禹’。通过定制增强亚马逊云科技所提供的以向量数据库服务和大语言模型为基础的知识库原型，在短期内实现了快速、精准的查询和回复。”

——西门子,IT 经理,李朝明





案例背景

中科创达是一家智能操作系统、AI 技术与产品的提供商，公司的操作系统产品和技术已经全面赋能智能手机、智能汽车、智能硬件等多个智能产品和场景。中科创达紧跟前沿技术趋势，启动全新的“Smart to Intelligent”，旨在自主研发大模型以赋能各行各业的客户转型，但由于算力不足、模型部署耗时耗力、模型训练速度慢、集群管理受阻、缺乏模型训练的实战经验等诸多挑战，自主大模型的研发遇到困难。



解决方案

- 中科创达使用 Amazon SageMaker 更专注于大模型训练，无需管理计算资源；
- Amazon SageMaker 提供技术支持，确保训练过程稳定，避免训练中断。其故障重试和自动更新实例机制，提高了训练的稳定性；
- 利用 Amazon SageMaker、Amazon Lambda、Amazon API Gateway 创建无服务器解决方案，进行模型推理；
- 将 Amazon Lambda 函数与 API 关联起来，外部应用程序快速调用大模型；

- 通过 Amazon FSx for Lustre 对机器学习、分析和高性能计算进行工作负载优化,加速训练作业;
- 中科创达使用 Amazon CloudWatch 来监控 Amazon SageMaker, 优化大模型训练和任务管理,加快模型调试,缩短问题解决时间。



业务价值

- 中科创达基于亚马逊云科技基础设施成功训练全自主大模型,提升生产效率 10 倍;
- 利用各类完全托管云服务,按需扩展或缩减 IT 设施,节省 50% 的建设成本;
- 发布了 Rubik Brain(创达魔方企业大脑)等多款生成式 AI 应用;
- 中科创达的 AI 应用服务于智能汽车、智能硬件等行业,推动企业数智化转型。

首个

中科创达全自主训练的大模型

10 倍

开发团队生产效率提升

50%

IT 基础设施建设成本节省

“

随着大语言模型的出现，中科创达也进入了颠覆性创新的时代，基于亚马逊云科技高性能的算力以及 Amazon SageMaker 等产品与服务，我们研发出了更具竞争优势的生成式人工智能产品与解决方案，赋能更多行业客户的数智化转型，同时亚马逊云科技也为中科创达自身技术的创新注入了强劲动力，加速了企业“Smart to Intelligent”的战略升级。

—— 中科创达,生态合作中心总经理，郑韦

”

辱骂识别准确率提升至 95% 以上

生成式 AI 为游戏玩家体验和业务流程优化带来双重收获



案例



案例背景

上海沐瞳科技有限公司成立于 2014 年，是中国游戏公司中最早一批走向国际的企业之一，用户群体主要分布在海外，尤其是东南亚地区。作为领先的游戏出海公司，沐瞳科技一直在积极探索大模型结合旗下游戏特征、赋能游戏业务的可能性，希望借助生成式 AI，改善游戏体验，赋能业务决策。其中 MOBA 类的游戏《Mobile Legends: Bang Bang》(以下简称《MLBB》)是沐瞳科技重点使用生成式 AI 进行优化的产品。



解决方案

- 通过使用亚马逊云科技的生成式 AI 技术方案，改善辱骂识别和舆情分析中存在的判断周期长、准确率低、颗粒度低等各类问题；
- 通过 Amazon Bedrock 的完全托管生成式 AI 服务和 Claude 模型，提高文本处理任务的效率和准确性；
- Claude 模型也用于 Prompt Engineering(提示工程)优化，帮助用户有效编写和调整模型的提示语，提升模型性能和输出效果；
- 沐瞳科技将 Amazon Bedrock 与 Amazon OpenSearch Service 集成，在舆情分析和风险控制应用方面，提升数据处理和分析效率；

- 基于 Amazon Bedrock 调用 Claude2 进行文本摘要、深度分析，为《MLBB》提供新游戏角色构建的事实依据。



业务价值

- 沐瞳科技通过 Amazon Bedrock 使用 Claude 模型，提升了辱骂识别的响应速度和准确率，准确率达到 95% 以上；
- Claude 模型在舆情分析方面具有 200K Token 输入能力，相比其他主流大模型精准度更高，输出更精准；
- 通过应用 Amazon Bedrock，沐瞳科技优化了数据处理流程，提高了响应速度、准确率和舆情分析效果；
- 提升了玩家体验和业务战略决策的可靠性，保障了沐瞳注重玩家体验、因地制宜进行游戏运营战略的实现。

**7-8 秒缩短
至 1.5 秒**

辱骂识别场景相较之前的响应
速度提升

**从 80%+ 提升
到 95% 以上**

辱骂识别场景利用 Claude 次
过滤后准确率提高

分析效率提升

舆情分析场景中使用 Claude2
分析效果大幅提升

“

沐瞳拥抱变化，追求极致的用户体验。我们惊讶于大语言模型的潜力，由好奇开始，到逐渐理解，并在 Amazon Bedrock 持续构建创新且务实的 AI 应用，在提升玩家游戏体验和提升内部能效上齐头并进。希望沐瞳能在 AI 与游戏结合的领域进入领先地位，也希望亚马逊云科技能成为我们坚实的伙伴。

——沐瞳科技，智能服务负责人，张文文

”

4.2.3 参考外部实践启动生成式 AI 行动计划

考虑到当前的数据基础、可参考的成功案例以及潜在的投资回报，在未来 3 年内，大部分企业可以从通用的应用场景开始试点，同步建设生成式 AI 应用开发平台，逐渐深入到开发行业领域专有生成式 AI 应用。

► 12 个月内行动计划——着手通用、易落地的业务场景

在未来 1 年内，可以先从通用的应用场景开始试点，例如：辅助销售、营销类的生成式 AI 应用，在该场景下不需要海量专有数据。在平台建设层面，可以先从云端开始投资生成式 AI 的模型构建编排与部署平台，并注重建立端到端的数据管道。针对行业垂直的场景，可以先从 POC 开始。

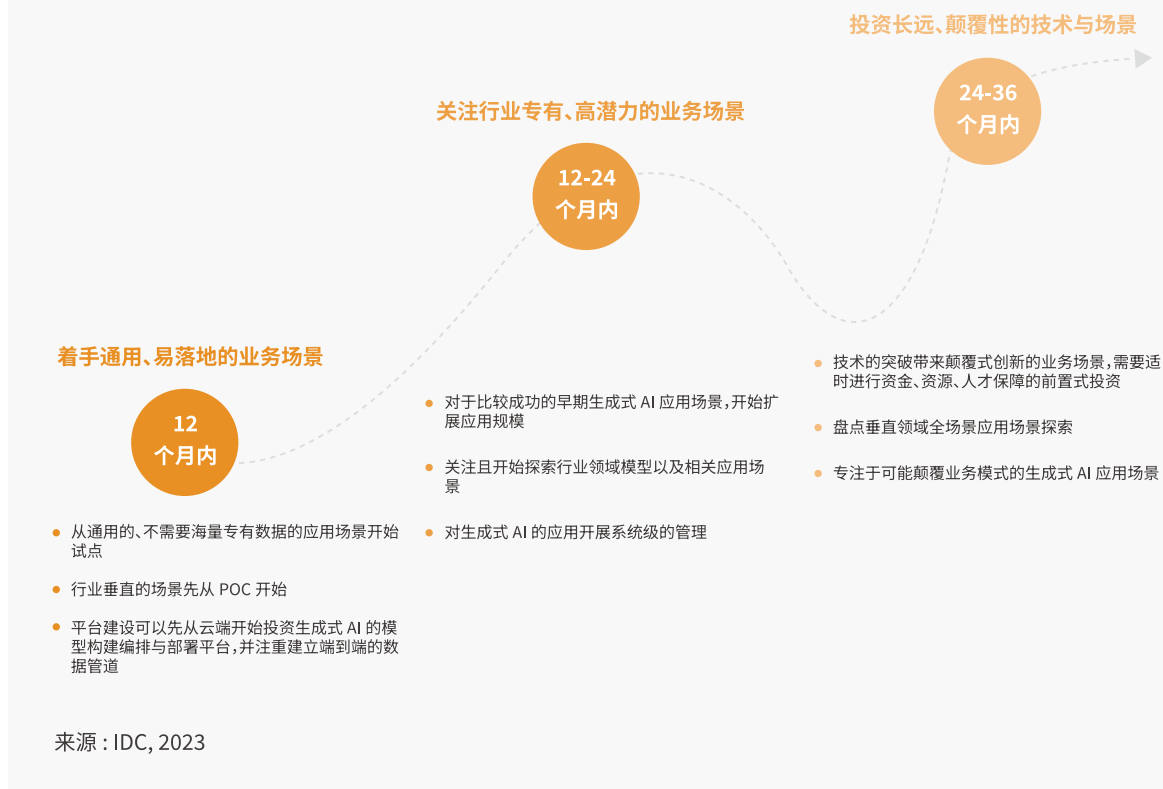
► 12 ~ 24 个月行动计划——关注行业专有、高潜力的业务场景

在未来 12 ~ 24 个月内，关注且开始探索行业领域模型以及相关应用场景。值得重点投资的场景包括知识管理、辅助产品设计、AI 数字员工等。对生成式 AI 的应用开展系统级的管理，包括基础架构、数据平台以及应用层，并探索 PaaS 级别的生成式 AI 服务。对于比较成功的早期生成式 AI 应用场景，开始扩展其应用规模，对于尚未形成清晰的落地方法与路径的、行业专有且高潜力的业务场景进行试点和投资。

► 24 ~ 36 个月行动计划——投资长远、颠覆性的技术与场景

未来 24 ~ 36 个月，由于大模型及生成式 AI 相关的技术和产品发展日新月异，企业应以更加长远的目标对当前尚未产生重大改变，但随着技术的突破可能产生颠覆式创新的业务场景进行资金、资源、人才保障的前置式投资。同时，继续加大在垂直领域应用场景的探索，对场景全盘盘点，专注于可能颠覆业务模式的生成式 AI 应用场景，例如新药研发、创新的产品设计、全新的数字服务模式等。

图 17 生成式 AI 行动计划



案例：Arçelik 规划 67 个生成式 AI 应用场景

Arçelik 是一家全球家电耐用品公司,其使命是像一家技术公司一样生产电器。因此,该公司一直希望尝试新技术。在人工智能领域,Arçelik 早前就设立了人工智能竞赛、工作团队以及相关项目。2022 年 11 月,ChatGPT 推出之后,Arçelik 立即加大了在生成式 AI 领域的举措。公司将人工智能战略集中在三大战略领域:在组织内部提高认识;建设内部能力,包括卓越中心、人工智能冠军奖项以及研讨会;开发新的项目和平台,开展具体的试点项目。公司由数据科学部门牵头,梳理企业内从研发设计、生产制造到市场营销、销售,以及覆盖包括财务、行政、人力资源等在内的职能部门的生成式 AI 应用场景,规划了 67 个应用场景,目前已开始探索其中的 14 个应用场景。

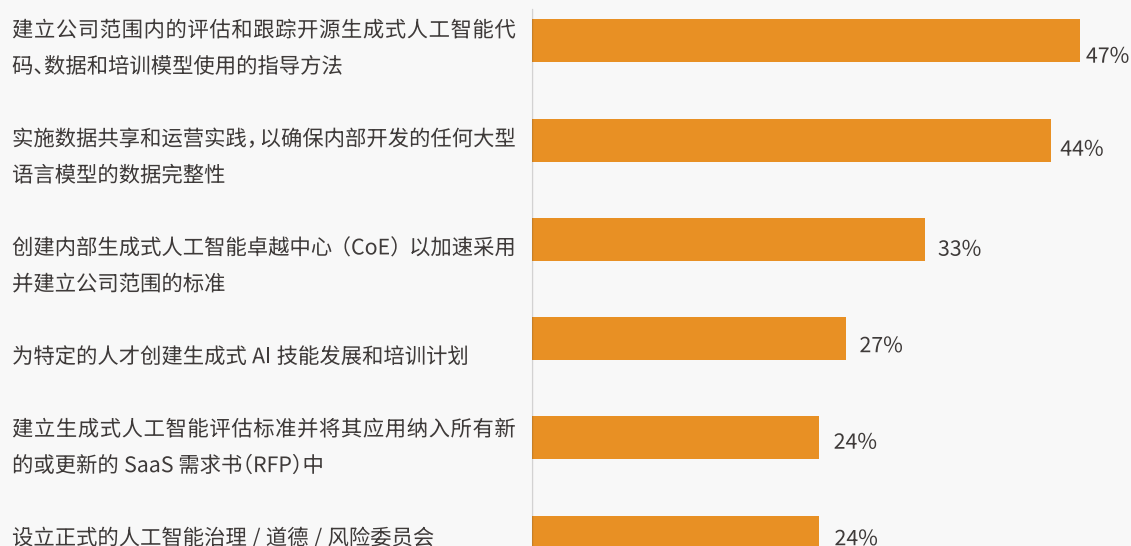
案例：某资产管理公司开发生成式 AI 财富助手

该资产管理公司在应用大模型之前，理财师团队规模约 400 人，每年财富销售规模过千亿，平均每位理财师服务客户数 50 ~ 80 人。理财师平均为每个客户每次编写持仓分析报告、资产配置建议书、营销文案耗时 30 ~ 60 分钟，能完成的复杂金融产品销售比例约 10%。

应用大模型之后，理财师团队规模保持不变，每年财富销售规模增幅 15%，平均每位理财师服务客户数 80 ~ 100 人。理财师平均为每个客户每次编写持仓分析报告、资产配置建议书、营销文案的耗时降低至 5 分钟，为理财师拓展和服务客户节省出大量的时间。与此同时，由于大模型撰写的文案内容更具针对性和多样性，复杂金融产品销售的比例也提升至 30%。

ROI：在理财师人员不增加的情况下，资产管理规模增量约 150 亿，按照收取 1% 的管理费计算，新增管理费 1.5 亿。

图 18 基于贵组织的经验，以下哪些是生成式 AI 的最佳实践？

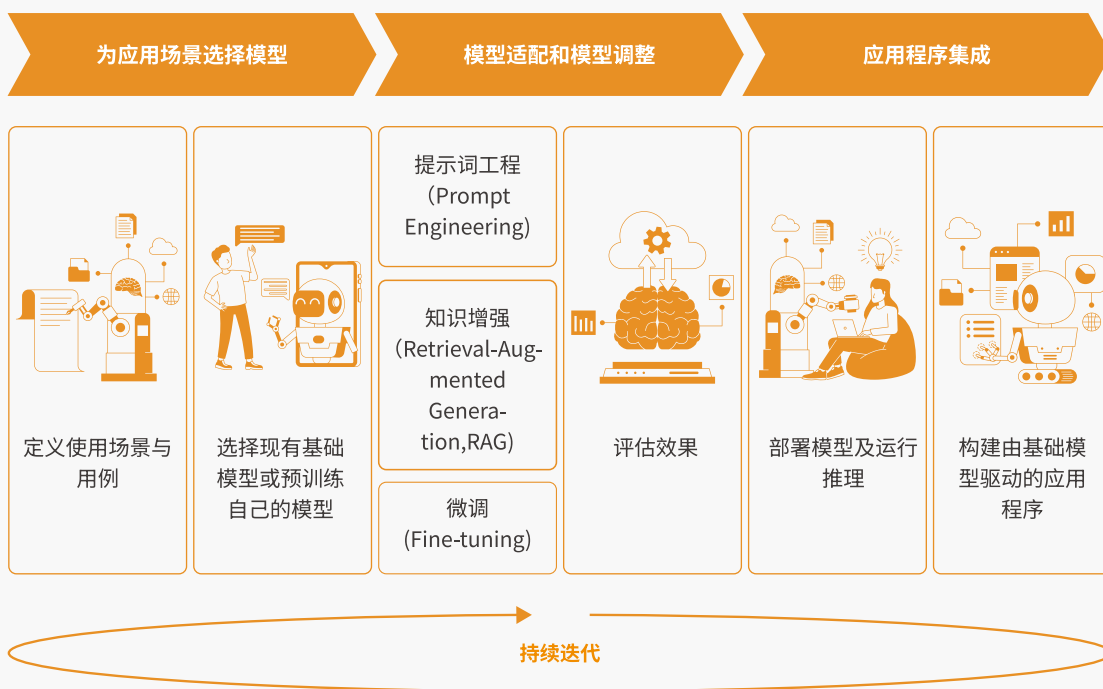


来源：IDC 中国，生成式 AI 市场调研，N=100, 2023

4.3 端到端构建生成式 AI 应用

对于有计划自主训练模型的企业，数据的收集与准备、模型的训练与调优，可以参考上一代 AI 中机器学习、深度学习模型训练与部署的流程。而对于使用基础模型进行开发的企业，本报告提出以下建议。

图 19 端到端构建生成式 AI 应用



来源：IDC, 2023

4.3.1 模型选择

根据企业制定的应用场景路线图，优选可以最先落地的场景，定义生成式 AI 的应用场景。根据场景与需求选择并评估基础模型，使用 AI 开发平台中提供的模型评估功能快速选择模型，参数量不一定越大越好，百亿参数量的模型也许更适合当前的生产环境。

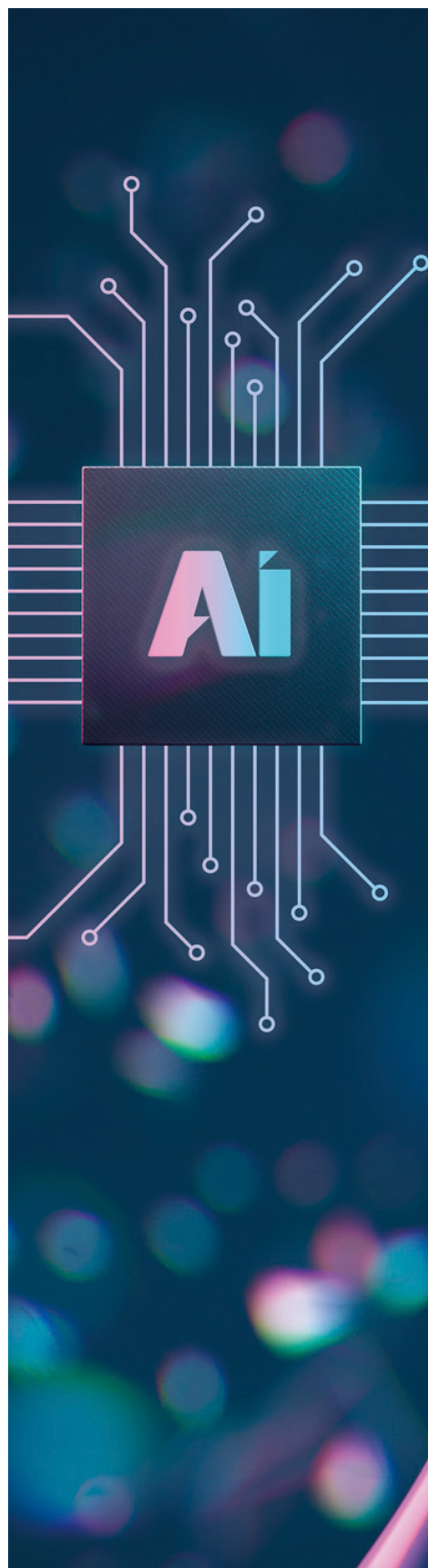
4.3.2 模型适配与调整

- ▶ **利用专有数据定制模型：**这些数据可能来自企业内部系统，也可能来自外部渠道。定制模型主要包括三种方式：提示词工程、检索增强生成(RAG)以及微调。
 - **提示词工程：**采用预训练模型的最简单方法，无需对模型或其参数进行任何定制，尤其适用于提升生产力的应用场景，如智能知识搜索和写作助手。这种方法适用于不需要特定领域或企业上下文的应用场景，同时允许某些用户级别的控制，以产生特定任务的输出。提示词工程技术的例子包括少量提示、思维链和常识提示。
 - **RAG：**获取数据源，将数据与提示词结合，增强提示词，最终返回响应。这种方法通过利用企业文件集，从通用模型中生成特定领域的输出。例如，一家律师事务所的资料库中可能有 20 年的咨询和案例文件，这些文件可以作为法律领域的专业知识加以捕获和访问。专有信息通常存储在矢量数据库中。在推理过程中，用户查询会通过嵌入模型运行，并转换成一个向量，与包含领域知识的向量数据库进行匹配；再将这种情境化查询发送到通用 LLM，以检索更有针对性的回复。这种将 LLM 和检索增强功能相结合的方法，是建立专有模型而不产生模型微调成本的好方法。
 - **微调：**使用标记数据集调整模型参数，在具体任务中进一步训练参数，使其适合业务需求。
- ▶ **评估模型：**评估模型生成内容的准确性、模型的稳定性、性能是否满足上线要求，确保模型在实际环境中能够满足预期的需求。

4.3.3 应用程序集成

为了充分利用生成式 AI 的潜力，企业需要不断更新其技术设施、对模型进行充分的验证和迭代，使其更加适应和适合企业的业务需求，包括模型验证、管理、模型治理等。

- ▶ **构建生成式 AI 应用程序：**为生成式 AI 应用准备好数据来源、数据传输管道，设置用户使用权限。如果企业开发一个全新的 AI 原生应用，可以从具备成本效益、免除基础设施搭建运维工作、具备行业领先 AI 应用实践的云服务开始搭建。在与企业现有数据库进行交互的过程中，可以选择生成式 AI 驱动的 agent 类工具来加速应用的部署。
- ▶ **模型部署：**将大模型集成到应用系统中。模型部署环节应该做到：简单 - 集成到新的应用程序不需要再更改代码、快捷 - 一键部署、自动化 - 自动化部署到各种位置、可扩展 - 跨多个区的自动扩展的部署能力。
- ▶ **模型治理：**持续进行模型的运行监测，对模型准确度、效果，甚至业务的价值提供持续的监测评估与反馈。进行负责任的模型治理：开发自动通知的功能，执行基于触发的治理，连接到人工智能资产，用于发布模型的负责任指标 / 透明度。



4.3.4 持续迭代

- 建立衡量应用是否成功的评估体系：针对生成式 AI 应用，尝试定义一套明确的关键绩效指标(KPI)，如客户满意度、员工生产力、提供服务的成本、用户信任度等。
- 在试点期间监控这些关键绩效指标，以决定哪些指标需要扩展。
- 评估数据、关键绩效指标及其他绩效和风险因素，以决定扩大哪些试点或开始其他试点。
- 评估生成式 AI 带来的业务效益。
- 从技能、流程和技术架构方面吸取经验教训，以便为新的应用场景提供参考。

展望未来,对于生成式 AI,有三点仍需特别关注

- 一、高管们需要革命性的思考,而不是渐进式的思考;
- 二、数据是关键,构建强大的数据价值链是企业差异化竞争优势的基石;
- 三、为未来而建,不局限于当前的需求,战略需要模块化、高度可扩展。

05

第五章 ▲

亚马逊云科技 助力企业解锁生成式 AI 潜力

亚马逊云科技致力于不断降低机器学习使用门槛，在人工智能和机器学习解决方案的三个层级都拥有至深至广的产品组合。在生成式 AI 时代，亚马逊云科技更是前所未有地加大投入，致力于让更多客户访问丰富的基础模型、为机器学习推理和训练提供基础设施、提高所有开发人员的编码效率，帮助客户更简单、更容易地在业务中使用生成式 AI。除了生成式 AI 应用的开发，亚马逊云科技也非常注重为用户提供端到端、全面、集成、可治理的数据底座，并帮助用户一起打造负责任的 AI 策略。

亚马逊云科技面向互联网行业、制造行业、金融行业、汽车行业以及出海群体提供了丰富先进的生成式 AI 云服务选择。全球已有超过 10 万家来自各行各业、不同规模的客户采用了亚马逊云科技机器学习服务，超过 1 万家客户利用 Amazon Bedrock 进行创新，如 Adidas 借助生成式 AI 使开发人员能够快速、准确地回答深层次的技术问题；Carrier 使用 Amazon Bedrock 将历史活动、预测分析、系统警报和数据趋势相结合，以提供帮助客户减少能源消耗和减少碳排放的建议；Nasdaq 利用 Amazon Bedrock 帮助实现可疑交易调查工作流程的自动化，并加强其反金融犯罪和监控能力。

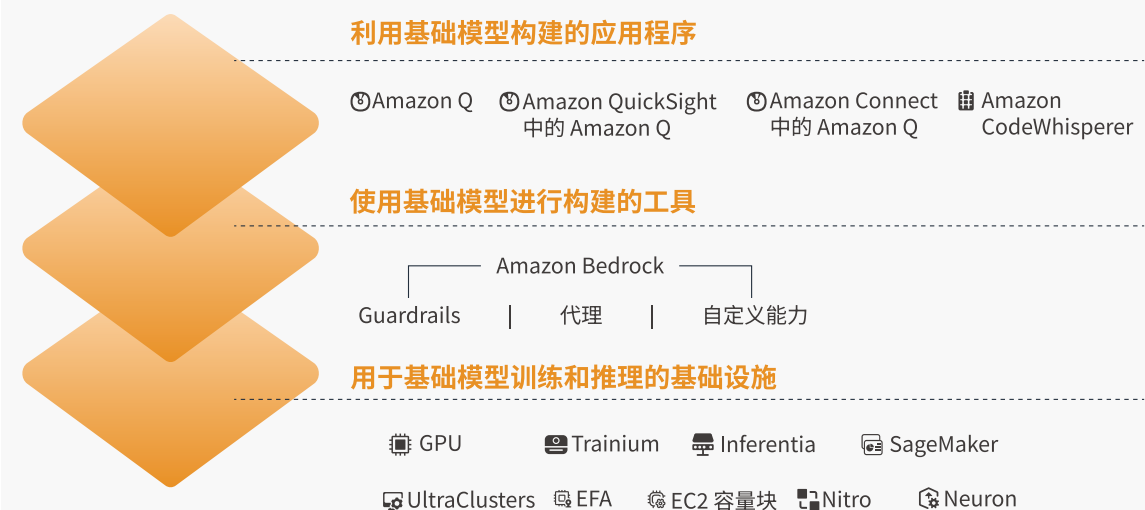
图 20 全球各行各业、不同规模的客户借助亚马逊云科技生成式 AI 持续创新



来源：亚马逊云科技，2024

面向客户，亚马逊云科技支持构建生成式 AI 应用的解决方案包括三层：用于基础模型训练和推理的基础设施、使用基础模型进行构建的工具、利用基础模型构建的应用程序。

图 21 亚马逊云科技生成式 AI 三层架构



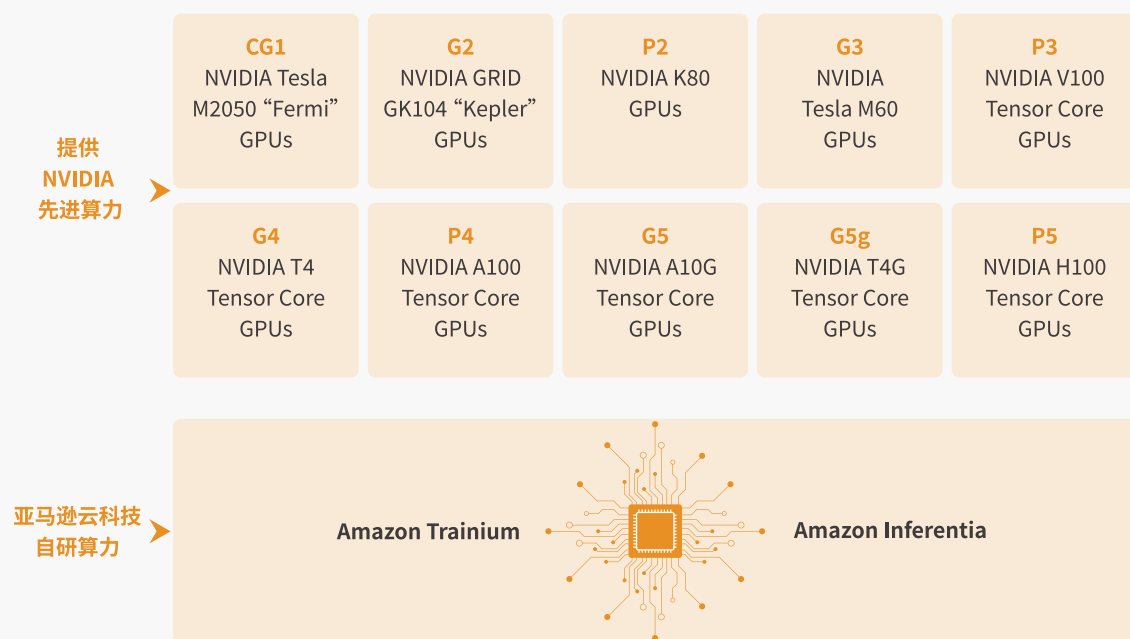
来源：亚马逊云科技，2024

5.1 高性价比且丰富的基础设施加速训练与推理

亚马逊云科技为用户提供丰富的算力选择。在 GPU 方面，NVIDIA 多款先进算力均率先在亚马逊云科技落地——亚马逊云科技将提供首款搭载 NVIDIA Grace Hopper 超级芯片的云 AI 超级计算机，以及首款采用 NVIDIA GH200 NVL32 的 NVIDIA DGX 云。此平台将在 Amazon EC2 实例上可用，由虚拟化系统 Amazon Nitro 及超大规模集群 Amazon EC2 UltraClusters 提供支持，让用户能轻松扩展至数千个 GH200 超级芯片。

同时，亚马逊云科技也在持续投入先进算力的研发以提升构建 AI 应用性价比，目前已发布支持千亿级甚至万亿级参数规模模型训练的 Trainium2，以及面向推理的 Inferentia2，为模型训练和推理提供更高算力、更快的加速比、更好的性价比。

图 22 亚马逊云科技提供丰富的算力选择

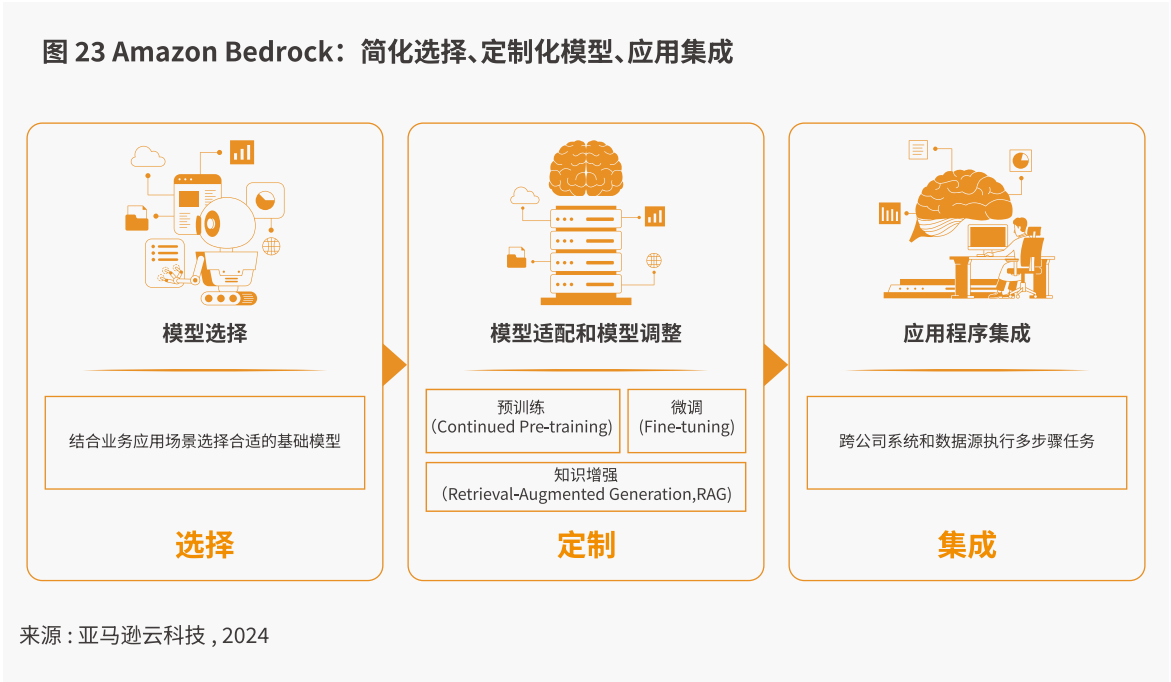


来源：亚马逊云科技，2024

Amazon SageMaker——训练与微调基础模型的一站式平台。MLops 编排：在统一的可视化界面中编码、构建、训练、部署和监控，提高生产力；模型实验管理：更高质量的模型训练，部署，通过配置自动调试训练错误、监控模型；直接获取算力资源：做到成本可控；可视化特征处理：连接 Amazon EMR、支持 Interactive Glue Session，Amazon S3 等语料数据源；大规模协作：多 domain，分享 Notebook、GitHub 集成。Amazon SageMaker 上提供了数百个预训练模型的访问权限，用户可以使用专门构建的工具来微调、实验、再训练和部署基础模型，构建自己的基础模型。Amazon SageMaker 产品面向分布式训练也做了专门优化，能够为客户提供更优的性价比。

5.2 丰富的模型可选,可基于私有数据差异化定制

Amazon Bedrock 使用基础模型构建 AI 应用：使一系列模型的构建和移动就像 API 调用一样简单，让所有开发人员都可以获得模型定制的新技术，并保护客户的安全和数据隐私。



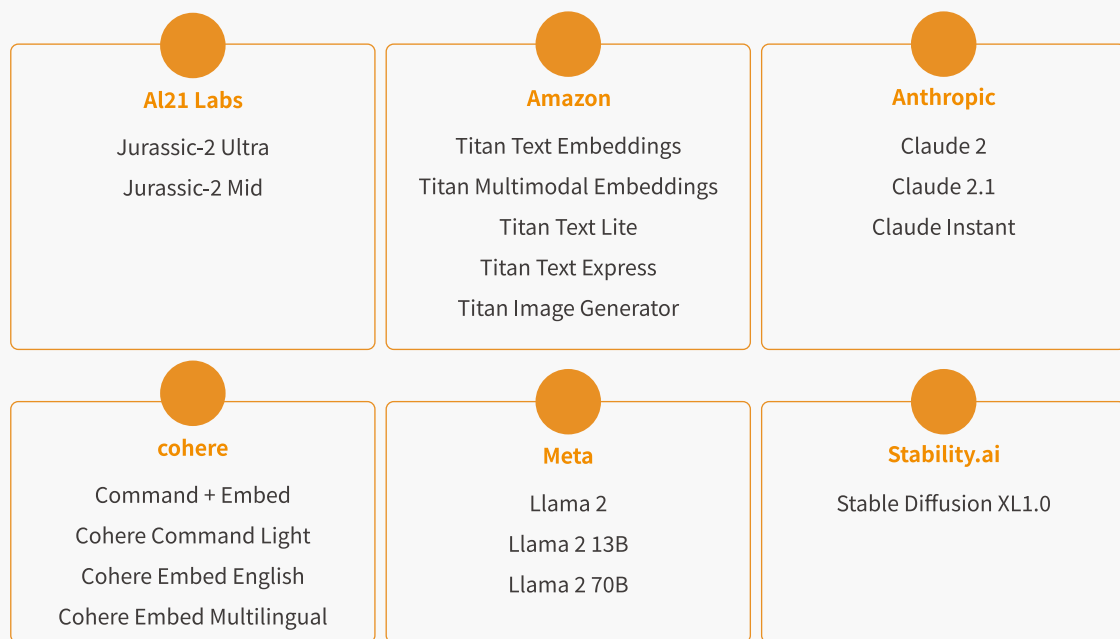
丰富的模型选择

企业面向场景构建每一个 AI 应用都要尝试多个模型，在 Amazon Bedrock 上可以支持用户轻松访问各种主流模型——多次尝试和切换模型，最终选出最匹配需求的模型。Amazon Bedrock 是一项完全托管的服务，通过单个 API 提供来自 AI21 Labs、Anthropic、Cohere、Meta、Stability AI 和 Amazon 等领先人工智能公司的高性能基础模型(FM)，以及通过安全性、隐私性和负责任的 AI 构建生成式人工智能应用程序所需的一系列广泛功能。

- Claude：通过负责任的 AI 进行训练的 FM，可进行深思熟虑的对话、内容创作、复杂推理、创意写作和编码。
- Stable Diffusion：图像生成模型可生成独特、逼真且高质量的视觉对象、艺术品、徽标和设计。
- Llama 2：经过微调的模型，非常适合对话用例和自然语言任务，例如问答和阅读理解。
- Command 和 Embed：文本生成和表示模型，用于生成文本、摘要、搜索、聚类、分类和利用 RAG。
- Jurassic：专为企业构建的指令遵循 FM，可执行一系列任务，包括文本生成、问答、摘要等。
-

除了业界领先的基础模型外，Amazon Bedrock 还提供了自研基础模型 Amazon Titan 供客户选择。Amazon Titan 包括一系列 FM，用于文本和图像生成、摘要、分类、开放式问答、信息提取以及文本或图像搜索等。

图 24 Amazon Bedrock 广泛的模型选择



来源：亚马逊云科技，2024

差异化定制

高效评估、比较和选择最适合其应用场景和业务需求的模型：支持用户自动或者人工评估、比较并选择模型。通过选择预定义的评估标准并上传自己的测试数据集或从内置的公开数据集中进行选择，可以实现自动评估。如需更加复杂的主观标准，用户可以通过定义特定的指标来设置，由 Amazon Bedrock 运行评估并生成报告，用户可以基于其结果做出最佳权衡。

通过 RAG、微调以及提示词工程实现模型定制功能：

- **RAG：**支持企业从多个数据源获取数据，使用专有数据补充模型，以便获得更准确的响应。Amazon Bedrock 知识库功能完全托管，会自动编排 RAG，通过模型获取相关文本来增强提示词，将提示词发给模型，最终返回响应。
- **微调：**是另外一项重要的模型定制技术，经常被用来控制模型去理解特定的输入格式（比如固定的问题问法），输出以及理解回答问题的角色。然后，Bedrock 会复

制一个客户专属的基础模型，并结合客户准备的 S3 上的数据，进行微调，并生成一个新的定制模型。在 Amazon Bedrock 上，Cohere Command Lite, Meta Llama 2, Amazon Titan Text Lite & Express 已经可以使用微调功能，Anthropic Claude 等即将推出微调功能。

- ▶ **对 Amazon Titan Text Lite 与 Express 持续预训练：**在 Amazon Bedrock 中，另一种自定义模型的方法是持续预训练。这是一种使用大量未标记数据，如内部报告、财务计划或研究成果的原始文本来提高大模型对特定领域的知识和推理能力的技术，这种方式是适合想要拥有“领域大模型”，或者说将领域知识以及技能嵌入大模型能力的场景。

在 Amazon Bedrock 上使用模型定制功能，用户数据在传输过程和静态时都经过加密。因此，所有数据都能保证安全私密。

应用程序集成

生成式 AI 应用通常需要集成公司内部系统、API 以及外部系统的数据。连接模型与数据源，开发人员需要配置和管理必要的基础设施，并制定数据安全和隐私策略。这些步骤耗时且具有一定的门槛，使用 Agents for Amazon Bedrock 可以计划并执行大多数业务任务，可以从专有数据源检索所需信息并提供准确且相关的响应。

5.3 生成式 AI 驱动的应用程序

未来企业的应用程序都将基于 AI 驱动，构建 AI 应用程序本身，亦需要生成式 AI 驱动。亚马逊云科技在云服务全栈产品中提供生成式 AI 工具的助力，目前在海外区域已有 Amazon Q 以及 Amazon CodeWhisperer 支持用户应用开发流程的自动化与智能化，未来将有更多服务引入生成式 AI 工具。

Amazon Q- 企业级生成式 AI 助手

Amazon Q 是专为您的业务量身定制的、生成式人工智能工作助手，可为员工提供信息和建议，帮助他们简化任务、加速决策和解决问题，并帮助激发工作中的创造力和企业创新。Amazon Q 为满足企业级客户严格的要求专门设计，可以根据企业既有的人员、角色和权限对每个用户交互进行个性化定制。此外，客户的内容绝不会用于训练 Amazon Q 的底层模型。Amazon Q 由亚马逊云科技 17 年来积累的知识和经验训练而成，改变了开发者和 IT 人员在亚马逊云科技上构建、部署和运维应用程序和工作负载的方式。客户可以通过亚马逊云科技管理控制台、文档页面、IDE、Slack 或其他第三方对话应用程序的聊天界面访问 Amazon Q。

图 25 亚马逊云科技企业级生成式 AI 助手 Amazon Q



来源：亚马逊云科技，2024

企业的业务专家：通过将 Amazon Q 连接到公司数据、信息和系统，可以根据您的业务定制 Amazon Q，并使用 40 多个内置连接器简化操作。商业用户，例如营销人员、项目和计划经理以及销售代表等，可以进行定制的对话、解决问题、生成内容、采取行动等等。Amazon Q 知道他们可以访问哪些系统，因此他们可以提出详细、细致的问题，并获得其中仅包含他们有权查看的信息的定制结果。

为充分发挥生成式 AI 的优势，客户需要针对其应用场景或行业专门构建的解决方案。因此，亚马逊云科技将 Amazon Q 引入多种服务和应用程序，包括：

- ▶ **Amazon QuickSight 中的 Amazon Q：**Amazon QuickSight 是专为云构建的统一 BI 服务，提供交互式仪表板、分页报告、嵌入式分析以及自然语言查询功能。借助 QuickSight 中的 Amazon Q，客户可以访问由生成式 AI 驱动的功能，由此来构建仪表板，并更轻松地利用数据简化决策、与业务利益相关者同步信息并获取洞察。借助新的故事生成功能，用户可以要求 Amazon Q“描述上个月业务发生的变化，用于向领导层汇报”。Amazon Q 能够在几秒钟内根据 Amazon QuickSight 中的可用数据创建一段数据驱动的、视觉效果良好的描述，用户可以进一步自定义描述并与整个企业共享。此外，通过仪表板和报告上的新的概要总结功能，Amazon Q 可以快速创建摘要，突出显示仪表板中需要注意的重要内容。业务用户还可以使用新的、简化的问答体验，他们可以提出开放性问题并获得相关答案，而不仅限于仪表板和报告中的视觉效果。
- ▶ **Amazon Connect 中的 Amazon Q：**Amazon Connect 是云联络中心，使各种规模的企业能够以更低的成本提供卓越的客户体验。联络中心客服能够通过一系列复杂的决策帮助客户，在企业建立客户信任和忠诚度方面发挥着关键作用，但招聘、培训和指导员工成为优秀的客服，确保他们能够快速、准确地响应客户需求是一项艰巨挑战。Amazon Connect 中的 Amazon Q 根据客户与客服之间的实时对话检测客户问题，并能够自动回复、给出建议以及提供相关资料。通过让客服人员能够在没有主管协助的情况下满足客户对各种问题的需求，Amazon Connect 中的 Amazon Q 提高了客户满意度，同时减少了客服人员培训、解决问题的时间并降低了成本。例如，Amazon Q 可以检测到客户正在联系租车公司更改预订。然后，Amazon Q 可以快速响应，发送公司的预定变更政策，并指导客服如何一步步地更新预订。

➤ **Amazon Supply Chain 中的 Amazon Q:** Amazon Supply Chain 是一款基于云的应用程序,通过将亚马逊近 30 年的供应链经验与亚马逊云科技的弹性、安全性和业务连续性相结合,让客户深入了解自身的供应链。许多客户正在寻找一种更直观的方式来了解上下游库存变化如何影响他们未来的运营。借助 Amazon Supply Chain 中的 Amazon Q,客户将能够提出有关供应链数据的一系列问题,如“是什么?”、“为什么?”和“如果.....会怎样?”,将复杂场景的结果可视化,还可以追问问题以权衡不同决定之间的优劣。例如,客户可能会问“是什么导致我的发货延迟以及如何加快速度”。Amazon Q 会分析客户的供应链,并标注大部分订单目前都在东海岸,风暴导致了延误,客户可以选择运往纽约而不是迈阿密来加快交货速度并降低成本

Amazon CodeWhisperer

Amazon CodeWhisperer 可以根据用户的注释和现有代码,在 IDE 中实时生成从代码片段到全函数的代码建议,也支持命令行中的 CLI 补全和自然语言到 bash 的转换。此外,在 Amazon CodeWhisperer 中,也可以使用 Amazon Q 通过自然语言的方式来解释代码、转换代码以及获取个性化的代码建议。

5.4 亚马逊云科技注重负责任的 AI

生成式 AI 一定要是安全、负责任的。亚马逊云科技强调帮助企业建立公正、可解释、稳健、透明、可治理、确保隐私与安全的“负责任的 AI”。在具体功能方面,亚马逊云科技确保用户用于自定义模型的数据将保留在客户自己的 VPC 内,任何数据均不会用于训练底层模型;在模型训练与推理的各个环节,所有数据在传输和静态时均经过加密,确保数据安全;支持 GDPR 以及 HIPAA 等标准。

全新的 Guardrails for Amazon Bedrock 也将助力落实“负责任的 AI”策略:企业希望生成式 AI 应用生成的内容能够确保“所答即所问”,少一些“幻觉”。

Guardrails for Amazon Bedrock 提供综合客户应用场景与根据“负责任的 AI”原则定制的功能，用户可以使用自然语言定义其应用程序中需要被拒绝的话题，也可以配置负面内容的阈值来过滤有害内容，还可以设定模型响应中的个人身份信息的变换方法，以保证安全、设置脏话过滤器，并提供自定义单词列表阻止用户和模型之间的交互。通过提供一致的用户体验并标准化生成型 AI 应用程序的安全和隐私控制，Guardrails for Amazon Bedrock 功能使客户能够安全地进行创新。

5.5 端到端数据基座打通壁垒,让所有应用可治理、可管控

端到端的数据服务能力是大模型成功应用的一个关键点——生成式 AI 不仅需要具备低成本高性能的基础设施、强大的模型与生成式 AI 工具，同时还需要构建一个稳固的数据基座，以实现不同环境、不同产品之间的无障碍连接，并确保所有应用的可治理、可管控。

亚马逊云科技提供了全面、集成、可治理的数据基座：

- **全面的数据服务：**通过大模型构建挖掘数据价值，离不开数据存储、查询、分析等全面数据服务能力的支持。用户可以利用 Amazon S3 构建数据湖，无论是结构化数据或非结构化数据，均可被用来分析，或者用于机器学习，并选择多种专用数据库（如 Amazon Aurora、Amazon DynamoDB）、数据仓库（如 Amazon Redshift）和大数据服务（如 Amazon EMR）；用户可以使用多种托管数据库引擎，包括 MySQL、PostgreSQL 等开源数据库引擎，还包括商业数据库引擎 Oracle、SQL Server；针对不同类型的数据处理需求，亚马逊云科技最新发布了 Amazon RDS for Db2、向量检索、Amazon OpenSearch Serverless、Amazon DocumentDB 和 Amazon DynamoDB 的全新向量搜索功能以及 Amazon MemoryDB for Redis 的向量检索功能预览版，使得用户能够在成本、性能和规模上取得平衡。

- ▶ **Zero-ETL 集成特性：**要想让大模型能够“更聪明”地为业务所用，底层数据支撑上需要全面的集成能力，需要将企业里散落在不同产品、不同业务环节的数据，进行采集、传输、存储、清洗、分析、展示等。在传统业务模式下，需要做很多 ETL 的工作，需要清洗、转化和加载，非常耗时费力。2022 年亚马逊云科技提出了 Zero-ETL 愿景，实现了 Amazon Aurora MySQL 到 Amazon Redshift 的 ETL 打通。2023 年，亚马逊云科技进一步扩展了 Zero-ETL 能力，包括发布 Amazon Aurora PostgreSQL、Amazon RDS for MySQL 和 Amazon DynamoDB 与 Amazon Redshift 的 Zero-ETL 集成预览版，以及 Amazon DynamoDB 与 Amazon OpenSearch Service、Amazon OpenSearch 与 Amazon S3 之间的 Zero-ETL 集成预览版。这些新功能使客户能够无缝地分析来自不同来源的数据，而无需自行构建和维护数据管道。
- ▶ **可治理：**大模型虽然发展潜力巨大，但访问和控制永远是一对不可妥协的矛盾。数据作为大模型的基座，必须安全、可靠、可治理，这是所有客户的“红线”。若授予用户过多的访问权限，带来整个生产力提升的同时，也有可能带来核心数据的泄露风险。若对于访问权限过于限制，毫无疑问会限制员工的创造力。所以，在保障数据安全的情况下，如何做到允许正确的员工来访问正确的数据，永远都是每个客户的共同诉求。为此，亚马逊云科技发布了 Amazon DataZone，从约束流程和操作方式开始，允许用户建立统一的数据目录，并建立严格的授权审批能力，让正确的人访问正确的数据。此外，还发布了 Amazon DataZone AI recommendations 和 Amazon Clean Rooms ML 预览版，为企业提供更易理解的业务描述和使用建议，以及在不共享基础数据的情况下与合作伙伴一起构建机器学习模型的产品。这些措施旨在平衡业务敏捷性和数据治理，描绘出负责任生成式 AI 的正确范式。

关于 IDC

国际数据公司 (IDC) 是在信息技术、电信行业和消费科技领域，全球领先的专业的市场调查、咨询服务及会展活动提供商。IDC 帮助 IT 专业人士、业务主管和投资机构制定以事实为基础的技术采购决策和业务发展战略。IDC 在全球拥有超过 1100 名分析师，他们针对 110 多个国家的技术和行业发展机遇和趋势，提供全球化、区域性和本地化的专业意见。在 IDC 超过 50 年的发展历史中，众多企业客户借助 IDC 的战略分析实现了其关键业务目标。IDC 是 IDG 旗下子公司，IDG 是全球领先的媒体出版，会展服务及研究咨询公司。

IDC China

IDC 中国 (北京)：中国北京市东城区北三环东路 36 号环球贸易中心 E 座 901 室

邮编：100013

+86.10.5889.1666

Twitter: @IDC

idc-community.com

www.idc.com

版权声明

凡是在广告、新闻发布稿或促销材料中使用 IDC 信息或提及 IDC 都需要预先获得 IDC 的书面许可。如需获取许可，请致信 gms@idc.com。翻译或本地化本文档需要 IDC 额外的许可。

获取更多信息请访问 www.idc.com，获取更多有关 IDC GMS 信息，

请访问 <https://www.idc.com/prodserv/custom-solutions>。

版权所有 2023IDC。未经许可，不得复制。保留所有权利。