利用生成式 AI 经济学超越竞争对手

不同于以往的任何技术，生成式 AI 正在迅速颠覆商业和社会形态，迫使企业领导者刻不容缓地重新思考其假设、计划和战略。

为了帮助 CEO 们掌握快速变化的形势，IBM 商业价值研究院 (IBM IBV) 发布了一系列有针对性、基于研究数据的生成式 AI 指南，涵盖数据安全、技术投资策略和客户体验等主题。

这是本指南的第二十二部分，重点关注“计算成本”

催化计算成本

计算成本曾被视为一个纯粹的 IT 问题，甚至在短短 24 个月前，情况也确实如此。然而，生成式 AI 的兴起正在将这一问题推至企业高管的议事日程。原因在于，如果不加以有效管理，支撑生成式 AI 所需的庞大计算资源可能会迅速引发意外成本的急剧上升，进而危及创新和业务转型。

深入理解生成式 AI 计算成本的关键驱动因素，CEO 能够制定更为明智的投资策略，明确战略重点，从而以更高效的成本推动创新与转型。

例如，企业需要进行大规模的资金投入或运营投资，以确保具备生成式 AI 所需的专用云计算能力。然而，计算能力和服务器仅仅是其中的一部分。此外，还需要考虑存储、数据中心、网络设备和服务，以及为生成式 AI 系统供电所需的能源消耗。当这些意外成本叠加时，可能会导致预算的急剧膨胀。

善于管理这些成本的 CEO 们，将能够像驾驭一台高性能引擎般运营组织，在运用尖端技术超越竞争对手的同时，最大限度地降低阻力。如此一来，计算成本不仅不再是负担，反而能成为一项竞争优势。当其他组织还在为生成式 AI 的预算分配焦头烂额时，那些精于成本控制的组织已经跨越财务障碍，率先迈向未来。

IBM 商业价值研究院甄别出了每位领导者都需要了解的三个要点：

1. 成本问题可能会让您精心规划的生成式 AI 蓝图偏离轨道。
2. 混合架构设计为生成式 AI 的扩展提供了经济高效的解决方案。
3. 生成式 AI 会大幅推高您的计算预算。

现在，每位领导者都需要采取以下三项行动：

1. 牢牢掌控您的计算成本。
2. 将生成式 AI 与混合云深度融合，形成强大合力。
3. 以更低成本实现快速突破。

1. 扩展 + 生成式 AI

需要了解的事项 →

成本问题可能会让您精心规划的生成式 AI 蓝图偏离轨道。

生成式 AI 正推动计算成本进入高速增长阶段。预计 2023 年至 2025 年间，平均计算成本将攀升 89%，而 70% 的高管认为，生成式 AI 是这一增长的主要驱动力。

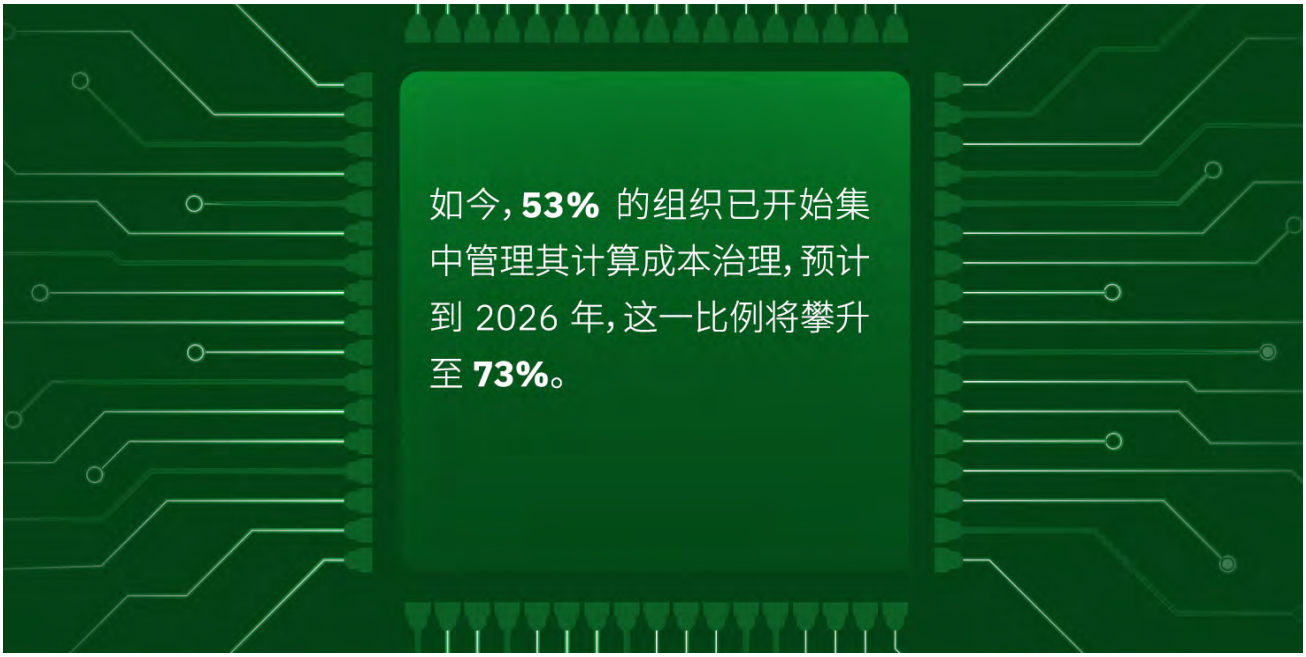
面对这一趋势，许多组织开始采取保守策略。受访高管均表示，由于对计算成本的担忧，他们的组织已取消或推迟了至少一项生成式 AI 项目。平均来看，15% 的项目被暂停，21% 的生成式 AI 计划因成本问题未能实现规模化扩展。

生成式 AI 的多个环节都会推高计算成本，如模型训练与微调、数据存储和处理支持，但这些成本大部分将通过云计算承担。目前，与部署生成式 AI 相关的云成本已达到模型本身成本的两倍，且随着云成为生成式 AI 构建和运行的核心平台，这一差距还在持续扩大。这形成了一个两难困境：如果缺乏有效监管，扩展生成式 AI 所需的云服务可能会成为规模化发展的最大成本瓶颈。

为了突破这一困境，CEO 需要为生成式 AI 项目制定清晰的成本目标，建立完善的成本治理架构，探索与合作伙伴协作降低成本的途径，并投资于更高效的架构以实现成本优化。

以混合云架构为例，它使组织能够将生成式 AI 直接与其必须结合的数据和应用程序对接，从而释放商业价值。采用配备统一控制面板和 FinOps 功能的混合云平台，领导者能够在成本最优的环境中运行数据、工作负载和应用程序。然而，尽管云平台潜力巨大，目前仅有 26% 的组织在较大程度上利用云平台和容器编排技术来降低计算成本。

生成式 AI 使成本问题愈发复杂，CEO 必须对支出进行细致分析，以实现更高效的管理。从数据标注到模型定制，一系列新的 AI 成本如果缺乏严格控制，可能会迅速推高预算。掌握模型优化的最佳实践，并精准配置计算资源，是提高生成式 AI 盈利能力的关键。



1. 扩展 + 生成式 AI

需要采取的行动 →

牢牢掌控您的计算成本。

精准锁定推高生成式 AI 成本的关键因素，并在项目扩展中保持前瞻性。在项目规划初期就设定明确的成本控制机制并评估计算需求，以避免未来出现高昂的意外支出。

剖析成本驱动因素。深入分析不同因素对生成式 AI 成本的影响（包括硬件、云服务、模型选择与训练、数据收集与清理、集成和维护），以及随着项目规模扩大，这些因素如何发生变化。制定清晰的成本控制标准，为每一项生成式 AI 决策提供指导，同时为团队配备必要的工具，以便在项目的每个阶段评估、监控和管理生成式 AI 对计算成本的影响。

优化计算资源配置。开展全生命周期成本评估，提前预测计算需求。投资更具成本效益的基础架构、定制化模型以及工作负载优化工具，确保在扩展过程中成本可控。与合作伙伴通力合作，降低训练、微调和开发成本。

借助 FinOps 和云优化降低生成式 AI 成本。将混合云平台作为计算成本的管理中枢。利用 Kubernetes 管理容器中的工作负载和服务，以最一致和高效的方式部署生成式 AI 应用。实时监控生成式 AI 带来的动态成本变化，涵盖数据存储、模型重新训练与微调、安全性及合规性等方面，避免造成财务上的突发损失。

2. 混合云 + 生成式 AI

需要了解的事项 →

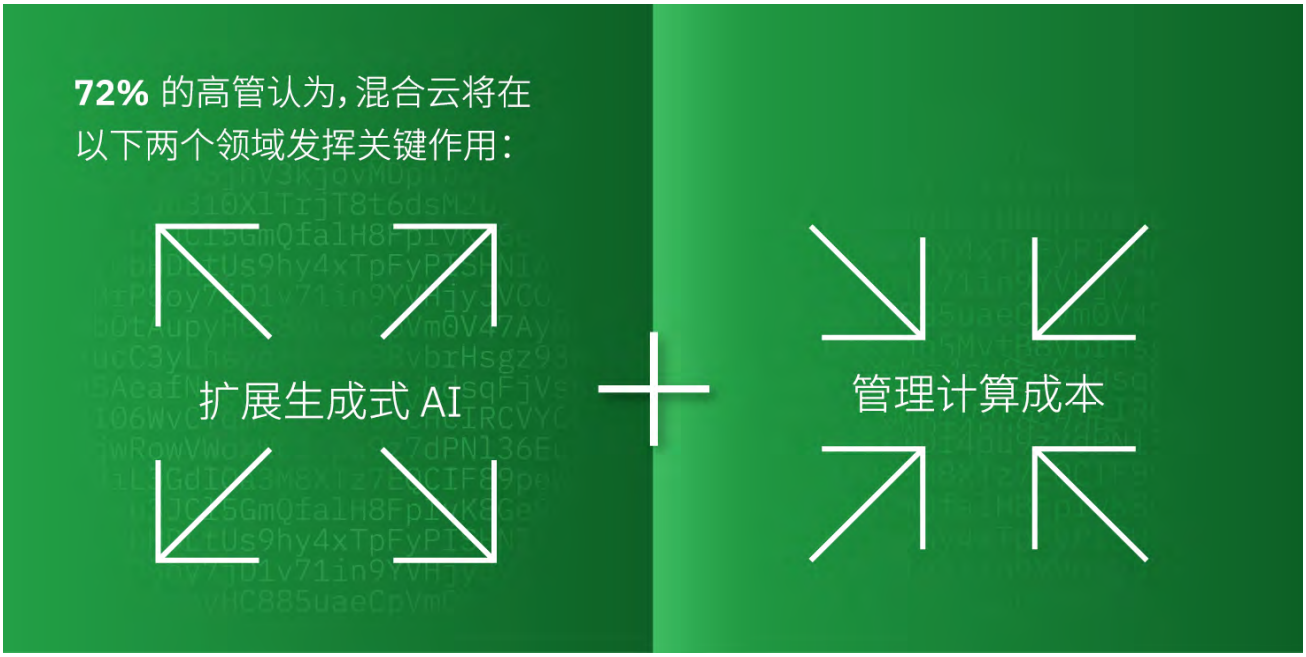
混合架构设计为生成式 AI 的扩展提供了经济高效的解决方案

生成式 AI 的应用并非千篇一律。每个用例都承载着独特的计算、数据和隐私需求。正因如此，混合云以其灵活性和经济性，迅速成为组织实现生成式 AI 规模化目标的首选架构。它使组织能够为每个工作负载量身定制最具成本效益的基础架构。

总体而言，72% 的高管认为，混合云是扩展生成式 AI 并有效管理计算成本的关键。对于那些已从试点阶段迈向全面部署生成式 AI 项目的组织，这一比例高达 85%。然而，要释放混合云在生成式 AI 中的全部潜力，必须将其核心理念贯穿于平台、安全、AI、云服务和数据管理等各个环节。

这种混合设计架构配备强大的引擎来提供原始处理能力，包括本地部署处理能力和云端敏捷性（用于支持快速扩展和数据访问）。通过巧妙的设计与有意识的整合，混合云架构将多种技术无缝结合，共同推动业务目标达成。

因此，使用统一、系统的混合设计方法，能够帮助组织更好地实现生成式 AI 项目的扩展，确保项目实施的效率和效果。如今，53% 的组织已开始集中管理其计算成本治理，预计到 2026 年，这一比例将攀升至 73%。在这一转型过程中，混合设计至关重要，其为领导者提供统一的视图，帮助他们监控、优化并控制计算资源成本。



2. 混合云 + 生成式 AI

需要采取的行动 →

将生成式 AI 与混合云深度融合，形成强大合力。

激发生成式 AI 与混合云的协同力量，实现明确的业务目标。采用精心规划的混合架构与容器化工作负载，优化资源配置，精准控制计算成本并提升运营效率。

打造您的核心中枢。在扩展生成式 AI 的过程中，清晰洞察计算资源需求增长的方向与方式。寻找更经济高效的方式调配资源，确保为每项任务精准分配所需的计算能力。在整个技术体系中推广混合云架构的设计理念。

有效抑制成本膨胀。采用模块化与灵活性的设计理念。遵循架构设计原则，使您的组织能够为每个 AI 应用场景和项目选择最优且最具成本效益的环境。

以清晰的指标定义成功。集中管理计算成本，制定以明确业务目标为驱动的企业级指导方针。构建治理框架，明确责任分配矩阵与绩效指标，确保每一步行动都有的放矢。

3. 优化 + 生成式 AI

需要了解的事项 →

生成式 AI 会大幅推高您的计算预算。

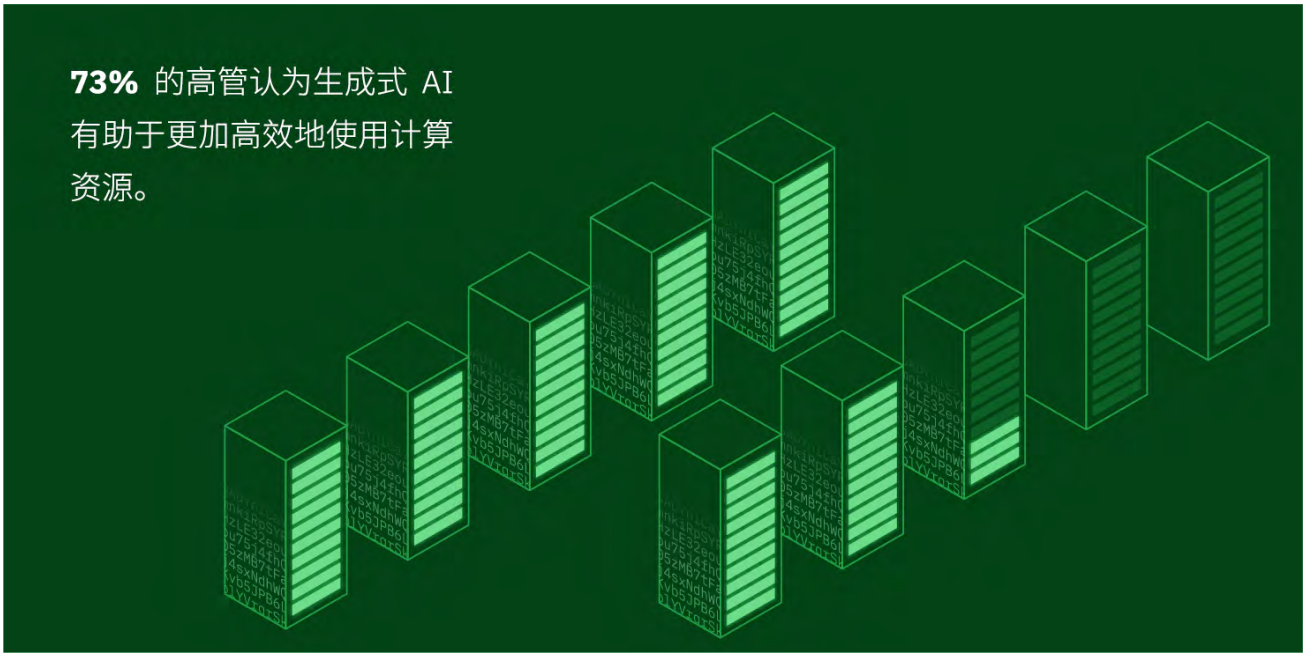
生成式 AI 或许是计算成本飙升的推手，但它同样可以成为破解难题的关键。73% 的高管认为，生成式 AI 能够大幅提升计算资源的使用效率，这一理论已在实践中得到验证。

例如，67% 的组织正利用生成式 AI 加速开发更高效的新模型、算法和应用程序。这不仅减少了开发资源所需的时间和精力，还帮助组织创建更高效的解决方案，从而减少对计算资源的需求。

此外，65% 的组织正利用生成式 AI 自动化任务，以此减少计算资源的消耗。这与传统自动化有何不同？一个独特之处在于，生成式 AI 模型能够并行处理数据，充分利用多个处理单元，从而减少完成自动化任务所需的总体处理时间和计算资源。

生成式 AI 的另一大潜力在于提升大型机的成本效益。大型机因高昂的管理成本和复杂的操作而备受诟病，但在许多关键场景中，它依然无可替代。例如，银行、保险公司和航空公司依赖大型机在危机时刻确保业务正常运转。其能够在系统故障时灵活转移工作负载，并以超快速度处理交易，这使得大型机在许多行业中依然具有强大的竞争力。

生成式 AI 则能将这种速度和弹性推向新的高度。它不仅可以优化系统利用率，还能通过 AI 驱动的自动化、预测分析和自我调节能力，重新定义大型机的运营模式。此外，借助生成式 AI 优化数据中心布局，组织能够显著降低能源消耗、减少成本并提升整体效率。2023 年，已有 25% 的组织将生成式 AI 应用于此领域，预计到 2024 年底，这一比例将增至 70%。



3. 优化 + 生成式 AI

需要采取的行动 →

以更低成本实现快速突破。

为管理者提供智能决策支持工具，显著降低计算成本并提升实时适应能力。自动化工作流程，优化模型，释放效率，降低成本，推动创新。

利用生成式 AI 驱动 IT 运营变革。为 IT 管理者提供生成式 AI 工具，助力自动化脚本生成、操作文档记录，并大幅减少合规性工作耗时。利用自动化故障检测与解决、预测性容量管理及实时性能监控，重塑大型机管理模式。

利用优化和自动化，推动效率提升。利用生成式 AI 生成合成数据、自动生成和优化代码，并实现动态资源调配，从而降低计算成本。

实时洞察市场动态，敏捷应对变化。利用生成式 AI 实时分析市场需求、市场趋势及竞争对手定价，优化定价策略，增加收入并减少因定价导致的损失。评估历史支出模式，精准预测预算需求，优化资源配置，减少浪费。

IBM 商业价值研究院

CEO 生成式 AI 行动指南

计算成本

本报告分析所依据的统计数据来自 IBM 商业价值研究院联合牛津经济研究院开展的两次专项调查。第一项调查于 2024 年 6 月至 7 月询问了 207 位美国高管关于计算成本和生成式 AI 的看法。第二项调查于 2023 年 12 月至 2024 年 4 月询问了全球 1,110 名高管关于可持续 IT 实践的看法。其他参考信息包括：《The Wall Street Journal》

IBM 商业价值研究院

IBM 商业价值研究院 (IBM IBV) 创立二十年来我们提供有研究支持和技术支持的战略洞察，帮助领导者做出更明智的业务决策。

凭借我们在商业、技术和社会交叉领域的独特地位，IBV 每年都会针对成千上万高管、消费者和专家展开调研、访谈和互动，将他们的观点综合成可信赖的、振奋人心和切实可行的洞察。

需要 IBV 最新研究成果，请在 ibm.com/ibv 上注册以接收 IBV 的电子邮件通讯。您可以通过 https://ibm.co/ibv-linkedin 在 LinkedIn 上联系我们。

访问 IBM 商业价值研究院中国网站，免费下载研究报告：https://www.ibm.com/ibv/cn

国际商业机器（中国）有限公司 IBM
北京市朝阳区金和东路 20 号院 3 号楼
正大中心南塔 12 层
邮编：100020

美国出品 | 2024 年 12 月

Q0JMKOAW-ZHCN-00

扫码关注 IBM 商业价值研究院

官网　　微博　　微信公众号