

accenture

赋能可持续AI



作者

stephanie ja
mison

全球可持续发展服务主管及全
球资源行业实践主席



桑贾伊·波德
尔

全球领先—技术可
持续发展创新



亚当·伯登

全球创新领军者



Bhaskar Gho
sh

首席战略与创新官



Senthil Ram
ani

全球领先—数据与
人工智能



沙拉布·库马尔·
辛格

首席总监— Accenture
研究



马修·罗宾逊

总经理—埃森哲研究



人工智能正在快速扩展，但代价是什么？我们的新分析介绍了可持续人工智能指数（SAIQ），这是一个衡量指标，追踪人工智能如何将金钱、能源和排放转化为可衡量的性能，使用代币作为标准化的绩效单位。它使企业能够根据其独特的组织优先事项平衡财务可行性、能源弹性和环境影响。基于专有模型和专家见解，我们确定了四个要点，以帮助领导者可持续地扩展人工智能：更智能的硅、更清洁的数据中心、更具战略性的人工智能使用和管理即代码。



我们相信那些会问“我们从投入到人工智能的资源中得到什么？”——并基于答案采取行动的公司，将定义未来。

人工智能不再是新兴趋势——它正在重塑从医疗诊断到零售供应链的一切。

随着人工智能的采用增长，支持大规模人工智能所需能源、排放和水的成本正以同样快的速度上升。如果放任不管，人工智能的环境足迹可能会威胁到企业的可持续发展目标，并超出我们地球的限度。但这不仅仅是可持续发展风险——它也是设计和创新的机会。解决这一问题可能是实现韧性和长期价值创造的途径。

我们希望将对话基于事实。与其依赖人工智能能源和资源需求的零散估计，我们产生了自己关于人工智能能源消耗、碳排放和水资源利用的全球估计。目标：基于真实的基础设施数据、建模场景和可持续性影响指标，确定领导者现在需要了解的内容。

我们从一个问题开始：我们是否以一种可持续的方式扩展人工智能？根据当前趋势，答案是否定的。推动人工智能革命的基础设施正在以堪比整个国家的速度消耗着前所未有的电力和水资源。

这一见解促使我们采取行动。我们联系了17位外部专家，包括来自计算机科学、人工智能和能源行业的学者、实践者和思想领袖。我们还汇集了来自安永可持续发展、技术创新、战略和数据与人工智能实践部门的专业人士。我们的目标是理解领导者应该采取哪些措施来确保长期绩效和韧性。

一个关键成果是新的指标：可持续人工智能商数（SAIQ）。它衡量公司将投资——美元、能源、水——转化为宝贵人工智能成果的有效程度。这个指标不仅仅是一个基准。它是一个用于决策的战略视角。

为了帮助公司最大化其人工智能投资的回报——不仅在财务方面，还包括能源、水 and 环境影响——我们围绕四个要点制定了一个实用框架：部署更智能的硅、去碳化数据中心、审慎使用人工智能以及将可持续性融入人工智能治理。这些措施共同使企业能够在符合新指标的情况下可持续地扩展人工智能，在增长与环境保护和合规准备之间取得平衡。

人工智能效率悖论



人工智能不再仅仅是一个创新的前沿——它正迅速成为全球商业的支柱。根据埃森哲的 [AI规模发展领跑者指南](#)，那些热衷于利用生成式人工智能重塑自身的组织，预计在18个月内，通过在全企业范围内部署和扩展生成式人工智能后，[生产力将提高13%，收入增长率将提高12%，客户体验将改善11%，成本将降低11%](#)。¹ 但它的发展付出了我们再也无法忽视的成本。支撑当今人工智能革命的设施正在消耗大量的电力和水资源，同时排放大量碳，造成的环境足迹正威胁要破坏人工智能本应带来的进步。

到2030年，为人工智能提供动力的数据中心预计每年将消耗612太瓦时能量——相当于加拿大的全年总电力消耗量（图1）。² 冷却这些数据中心可能需要消耗30.2亿立方米的新鲜水，这超过了挪威或瑞典等国家的年新鲜水总取用量。³

人工智能的碳排放可能占全球总排放量的3.4%——十年内增长了11倍。如果不受控制，这一趋势不仅会损害地球。它还会推高成本，加剧供应链压力，并使企业面临监管风险、碳税和利益相关者反弹。

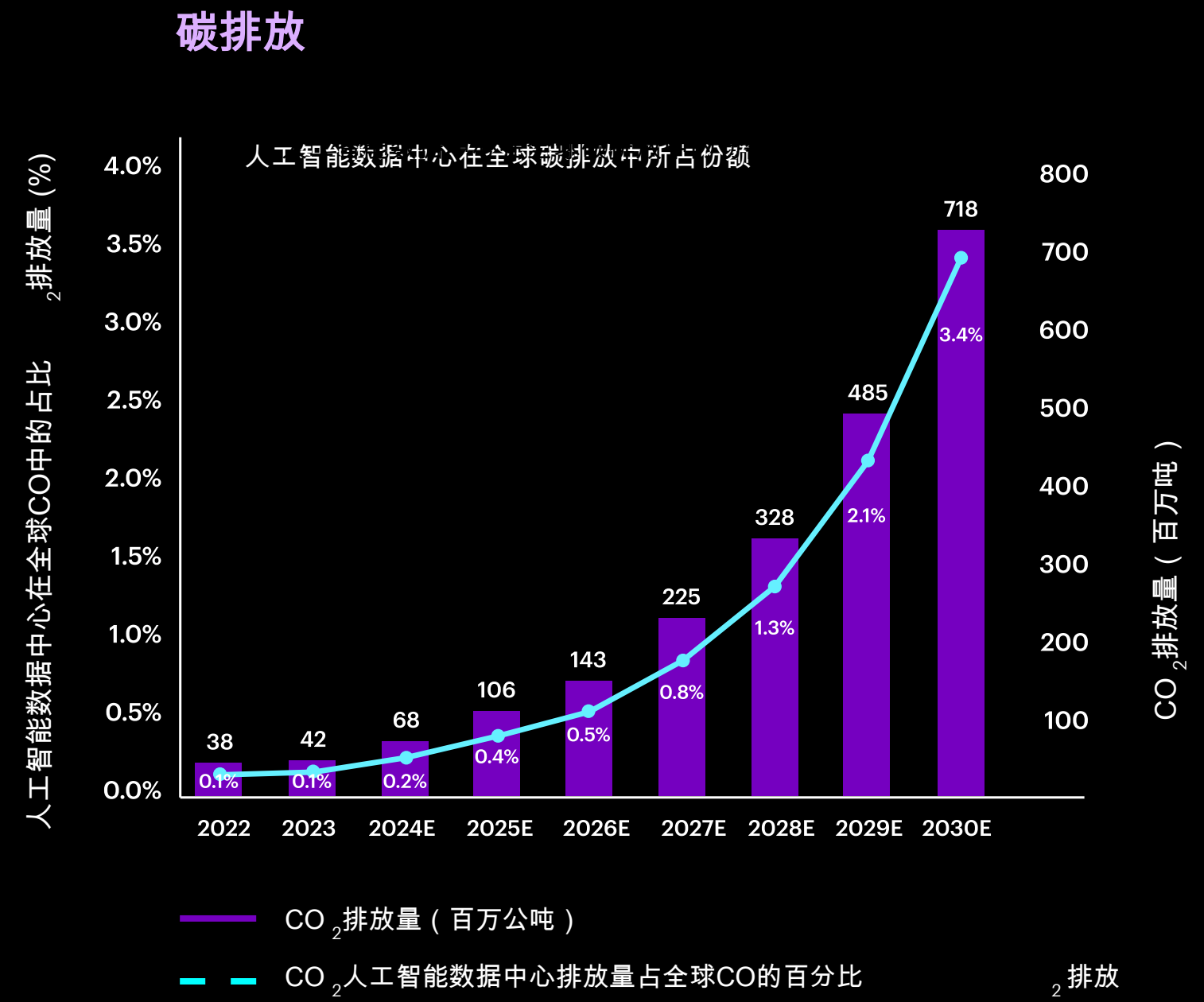
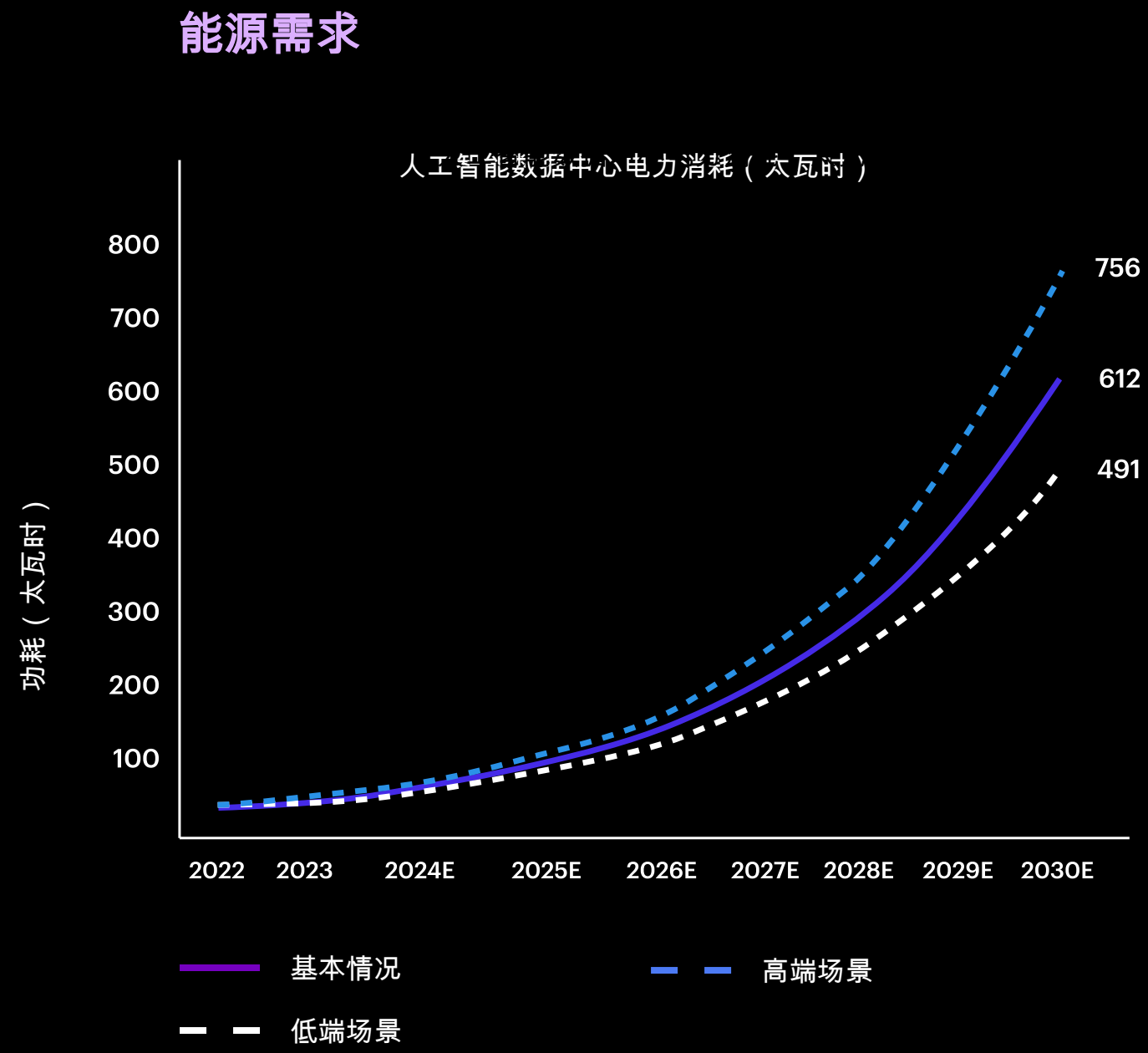
这是时代的讽刺。同样，代表企业在效率、生产力和能力方面的代际飞跃的技术，可能会破坏商业目标。

根据 Accenture 的 [目标净零排放](#) 报告显示，全球2000家最大公司中，只有16%的企业正在按计划实现其净零排放目标。⁴ 人工智能不断增长的能源足迹可能会进一步拉大差距，除非得到控制。然而，这里存在着悖论：虽然人工智能本身具有巨大潜力来减少大多数公司的碳足迹，但目前只有14%的公司使用人工智能来减少排放。

是时候将问题从“我们的AI有多强大？”重新定义为“我们为人工智能投入的资源能得到什么？”

预计到本世纪末，人工智能数据中心电力使用量将增长10倍以上，达到612太瓦时（TWh），相当于2022年加拿大的电力需求。

图1：
人工智能数据中心——不断增长的能源需求和碳排放



来源：埃森哲研究基于来自各个来源的数据估算

让我们算笔账

我们需要一个新的智能衡量标准——一个不仅反映AI的强大程度，还反映其构建的智慧和运行的负责任程度的指标，一个不仅衡量AI计算负载（通常用“token”数量，即处理信息的基本单位来衡量），还衡量其在经济压力下的弹性和效率的指标。

传统的效率衡量方法虽然有用，但无法全面反映情况。以电力使用效率（PUE）为例：它告诉我们在数据中心总能耗与IT设备能耗的相对基础上，数据中心的运行效率如何，但它并未说明这种效率衡量的是什么。它并非设计用来衡量人工智能模型将电力、美元、碳和水转化为智能、转化为影响的能力如何。它无法帮助企业回答这样的问题：对于我们投入的每一单位资源，我们得到了什么样的产出？

为弥补这一差距，我们提出了可持续性调整智商（SAIQ），这是一种衡量人工智能系统将资金、电力、水和碳转化为实际性能效率的多维度指标。SAIQ是一个复合效率得分，用于衡量人工智能系统的环境和经济性能，计算公式为\$/token、MWh/e/token和m³水/token的加权总和。随着时间的推移，token, tCO₂

一个组织的安全智能（SAIQ）可以通过衡量四个方面的效率来跟踪其人工智能的可持续性：成本、能源、碳排放和水资源使用。SAIQ越低，人工智能系统就越高效和负责任。该指标可以引入标准化的、数据驱动的方法来应对衡量人工智能相对于所用资源真实影响这一挑战。

可持续人工智能商数：

SAIQ = w1 . $\frac{\text{兆瓦时能量}}{\text{token}}$ + w2 . $\frac{\text{tCO}_2\text{e}}{\text{token}}$ + w3 . $\frac{\text{m}^3\text{水}}{\text{token}}$ + w4 . $\frac{\text{\$ 2}}{\text{token}}$

组织可以通过调整权重 (w1, w2, w3, w4) 来适应其战略重点，从而最小化其 SAIQ，提升 AI 性能并最有效地应用，而不是将其作为一刀切的基准。例如，一个成本敏感的企业可能会强调 w1 (每个 token 的美元)，而一个以可持续性为中心的组织可能会更重视 w3 (每个 token 的吨 e)，从而实现灵活的协调对齐。 二氧化碳²商业和ESG目标。

如图2所示，SAIQ使组织能够在三个相互关联的维度上保持平衡：

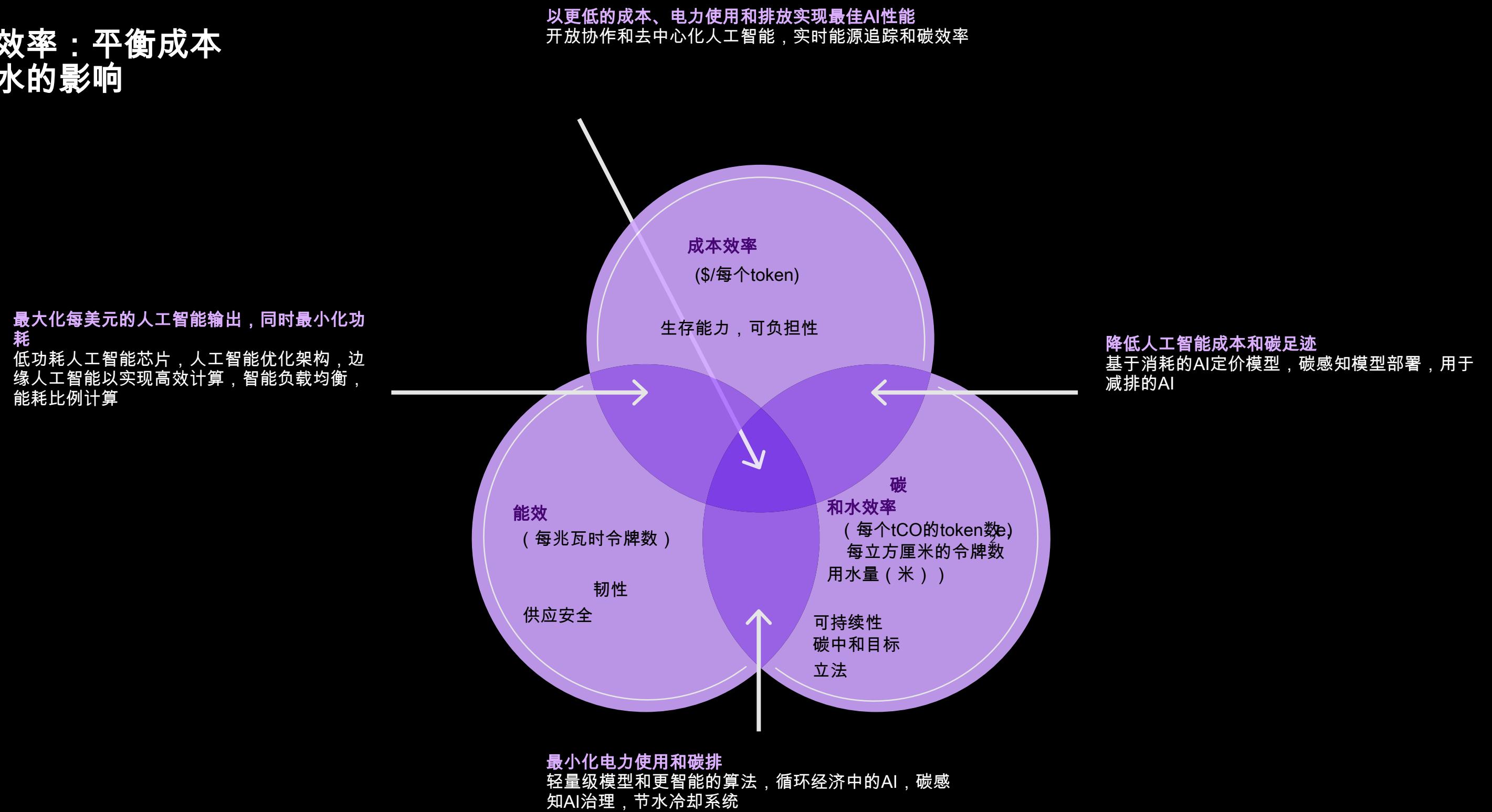
- 成本效益 (\$/token)：可行性、可负担性
- 能源效率 (兆瓦时/代币) 韧性，供应安全
- 碳和水效率 (tCO₂e/个令牌)
m³水/令牌): 可持续性，净零目标，立法

认识到人工智能的真实成本-输出比，仅仅是开始。随着其能源和资源需求的加速，组织必须将SAIQ透镜应用于人工智能堆栈的每一层——从硬件和基础设施到部署策略和治理。

传统指标（例如准确率、延迟）已无法全面反映AI的成本。诸如每美元、每兆瓦时、每吨二氧化碳当量以及每立方米水的多维度指标提供了更全面的视角——使利益相关者能够看到相对于其财务和环境成本，系统生产有价值输出的效率如何。

– 米哈埃拉·范德·舒尔教授，机器学习、人工智能和医学教授，剑桥大学医学人工智能中心主任，剑桥大学教授

图2：
优化人工智能效率：平衡成本、能源、碳和水的影响



人工智能的更广阔视角

我们提供一种平衡增长与环保责任和商业绩效的方式，实现人工智能的规模化发展。四种行动——使用智能硅、去碳化数据中心、审慎部署人工智能并嵌入人工智能治理为代码——可以帮助领导者提高其组织在三个维度上的 SAIQ：财务（每个代币的美元）、能源（每个代币的瓦特）和立方米。环境（吨CO₂每token的水量）。

每个行动都会以相互交织和增强的方式影响 SAIQ 指标的一个或多个组成部分，而非孤立地影响。例如，使用节能硅和芯片，模型和 IT 架构可以减少能源消耗，同时也能降低成本和碳排放。类似地，强大的 AI 治理可以减少模型浪费并提高利用率，同时也能改善成本和碳排放影响。

SAIQ为组织在可持续地扩展AI方面的有效性提供了一个整体衡量标准，捕捉了下面推荐的四个相互依赖行动的相互影响，它们共同构成了一份将可持续性转化为AI战略可衡量成果的实用手册。

人工智能在全球能源消耗中的占比预计将从2024年的0.2%上升至2030年的1.9%，以48%的复合年均增长率（CAGR）呈现异常迅速的增长速度，超过该时期整体电力需求增长预期的1.5%。



操作 1：

Put smart 硅 工作

大多数AI模型运行在老旧硬件上，这些硬件从未打算用于当今内存密集型计算需求。结果：巨大的能源浪费、高昂的运营成本和高排放。这些低效率不仅仅是可持续性方面的担忧。在日益增长的地缘政治和供应链不确定性中，它们是商业风险。更智能的计算不仅是一种气候要求，也是对波动的对冲。

优化GPU能效和数据中心利用率可以削减AI预计到2030年的能耗121万亿瓦时——相当于挪威的年用电量。

实现这一潜力需要在更智能、计算效率更高的硬件和模型上运行人工智能工作负载，以降低成本、减少排放并减少对大型云数据中心的依赖，同时不损害人工智能能力。企业可以更进一步，将其未充分利用的人工智能计算资源货币化，以创建一个不仅减少浪费而且创造新的收入来源的更高效的人工智能生态系统。

这里是一些需要关注的领域：

在更智能的基础设施上运行 AI

为了获得每兆瓦时的更多AI输出（“token”），应将AI工作负载转移到更有效地集成硬件和软件的架构上。计算存储（CIM）和处理存储（PIM）技术可以直接在数据存储处处理数据，从而大幅降低能耗。这最小化了内存和处理器之间昂贵的移动数据，这是传统系统中的关键瓶颈。例如，作为CIM领域的领导者，Mythic正通过将模拟计算与闪存集成，开创AI推理的转变。⁵

神话的方法最大限度地减少了功耗和热问题，已被证明可以将边缘设备的AI推理功耗降低20倍。类似地，三星电子已开发出PIM技术，该技术将AI专用半导体集成到高带宽内存中。⁶将部分数据计算工作从处理器卸载到内存，从而节省了85%的数据移动能量消耗。

其他硬件设置，例如类脑计算，为更有效地处理数据提供了一种有前景的方式。人脑运行与

难以置信的效率——使用大约 20 瓦的功率来执行一些估计相当于 1 屏算次（每秒一百亿亿次计算）⁷。相比之下，训练大型语言模型的单个处理单元可消耗超过400瓦特——而这项任务需要数千个此类单元。神经形态系统模拟人脑的神经结构和功能来处理信息，使用仅在活跃时才耗电脉冲神经网络（SNNs）。OpenAI近日签署了一封意向书，计划从初创公司Rain AI购买价值5100万美元的神经形态处理单元（NPU），旨在与传统硬件相比显著节省能源。⁸

使用轻量级模型和开放生态系统

要获得每美元更多的人工智能输出，请使用轻量级人工智能模型和低精度计算格式，如浮点8位（FP8）。通过仅使用8位，FP8减少了内存需求并允许更快的计算，使其成为快速训练和部署大型人工智能模型的理想选择。中国初创公司DeepSeek通过一种具有独特架构的大型语言模型（LLMs），采用了专家混合（MoE）方法，其中仅一小部分参数

在任何给定时间都处于活动状态。⁹ 使用fp8、moe和多词预测格式——允许模型一次性生成多个单词而不是一次生成一个单词——DeepSeek不仅降低了计算需求，还在技术查询和代码生成方面提供了有竞争力的性能。

在边缘部署人工智能可以减少云的使用，并通过降低延迟来提高性能，同时减少碳排放。边缘人工智能应用特别适用于依赖实时数据处理行业的制造、医疗保健、零售和金融服务等。通用电气医疗与英特尔合作，将边缘人工智能集成到其医疗成像设备中。通过在设备上直接处理扫描，临床医生可以更快、更准确地检测如气胸等病症——而无需将大量成像数据发送到云端。¹⁰

这种实时分析提高了诊断速度并减少了延迟，同时还减少了数据传输量和相关能耗，从而改善了患者预后并使临床运营更可持续。



脱碳化

动作 2:

数据中心

超大规模数据中心——那些支持云计算、大数据分析及其他如人工智能等大规模应用需求的庞大设施——每天消耗数百万加仑的水。许多数据中心由化石燃料供电。随着他们人工智能工作负载的增加，他们的能源消耗甚至增长得更快，而不是成比例地增长，随着时间的推移，增加了他们的碳足迹。

这个资源密集型的基础设施正在给电力系统和水资源带来压力，特别是在干旱地区，工业设施、农场和城市都在争夺有限的水资源。企业还面临着越来越严格的监管审查和投资者压力，要求他们展示在可持续性方面取得的实质性进展，而目前的基础设施可能难以应对这一挑战。

通过优化人工智能工作负载和数据中心运营，企业可以确保人工智能与企业的可持续发展目标保持一致地扩展，同时降低运营成本和环境影响。

这里是一些需要关注的领域：

采用动态伸缩和智能负载均衡

为减少功耗和排放，将能源使用与AI工作负载相匹配。例如，Meta正在设计考虑能源比例性的AI基础设施，借鉴动态效率扩展等原则来优化AI工作负载的功耗。¹¹ 通过提高模型效率和系统级设计，Meta旨在随着AI的发展减少运营和隐含的碳排放。谷歌也通过使用自适应调度将AI处理转移到电力最便宜和最清洁的时间来减少峰值能源需求。¹²

变现数据中心

人工智能计算市场和GPU共享平台为变现闲置资源以及分摊人工智能成本提供了机会。去中心化市场和企业云转售模式在减少碳足迹的同时也开辟了新的收入来源。一个例子是NodeGo.AI，它允许个人和企业通过将其转化为人工智能和其他应用的有价值资产来变现其未使用的计算资源。¹³ NodeGo.AI使用区块链技术，在点对点生态系统内购买和销售计算能力时，进行安全高效的交易。

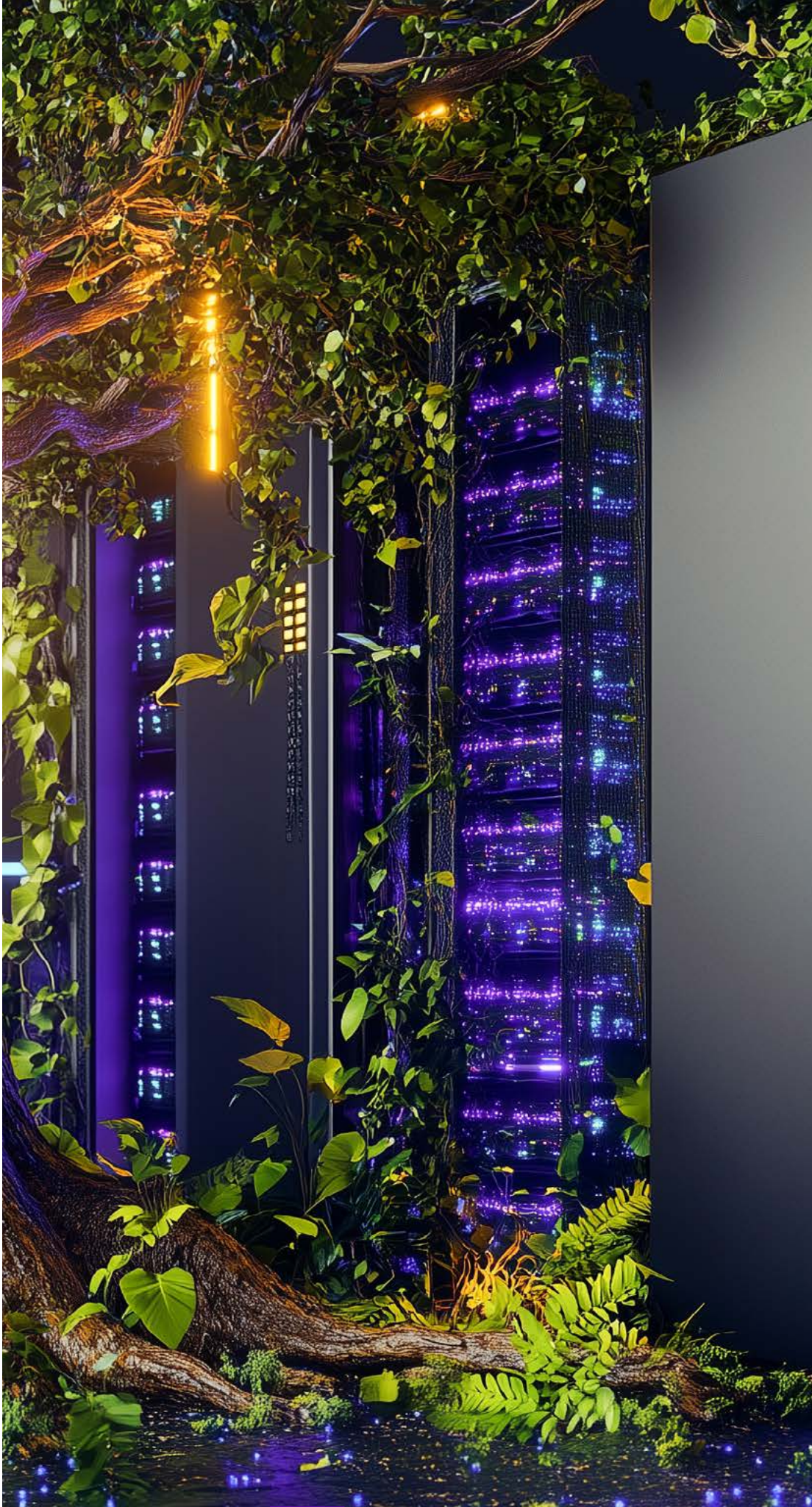
优化数据中心位置

选择战略位置以最小化成本和碳排放。电力碳强度——即每单位电力产生的二氧化碳量——因地区和时间不同而有显著差异。这意味着AI工作负载可以安排在能获得清洁能源的位置。像谷歌、微软和亚马逊这样的科技巨头正专注于数据中心位置和可再生能源作为其长期可持续发展战略的一部分。¹⁴ 并且像Lonestar Data Holdings这样的初创公司正在探索在月球表面利用太阳能和自然冷却等创新解决方案。¹⁵

微软已采取措施，优化其人工智能处理，以服务于该国可再生能源水平较高的地区。¹⁶ 谷歌与能源公司Intersect Power和TPG合作，将数据中心扩展与清洁能源发展同步进行。¹⁷

位置也会影响可持续性。例如，挪威等北欧国家提供了自然冷却的机会。¹⁸ 另一方面，某些地区容易受到洪水、野火和极端高温等气候风险的影响。海拔、空气质量和能源供应韧性等因素对于建设低碳、气候适应基础设施日益重要。^{19, 20}





整合低碳能源选项

核能，尤其是小型模块化反应堆（SMRs），为支持人工智能数据中心提供了一种可靠的方案，在确保持续电力供应的同时减少对化石燃料的依赖。最近的合作，例如康斯特拉尔能源与微软之间的合作，以及塔伦能源与亚马逊网络服务之间的合作，凸显了向核能人工智能基础设施发展的势头日益增强。事实上，谷歌已投资建设SMRs，预计到2035年将生产500兆瓦的核能。²¹ 除了环保之外，这些选项还有可能降低每单位AI输出的成本。全天候零碳联盟展示了如何通过小时级匹配数据中心需求与低碳能源来实施这些策略（参见“全天候零碳联盟”）。

先进的冷却技术，如直冷芯片液体冷却、无蒸发冷却和热重用系统，均可用于最小化水和电力消耗。微软对两相浸没冷却的实验（参见“让每一滴都发挥作用”）以及谷歌在特定数据中心使用再生水的应用，展示了人工智能基础设施如何在保持性能的同时减少对淡水依赖。^{22, 23}

循环水解决方案也正成为一种关键策略，以实现数据中心可持续增长。²⁴ 例如，一些设施现在正在采用现场水处理系统，将废水回收用于冷却过程中的非饮用用途。像封闭式冷却系统和水回用这样的做法，在实施的地方已经证明可以节省50%到70%的水量。微软通过推出一种使用环境空气和封闭式冷却系统的新型数据中心设计，证明了这一点，从而消除了对新鲜水的需求。²⁵



案例研究

让每一滴都发挥作用

微软在其数据中心采用了两相液体浸没冷却技术，以解决人工智能工作负载产生的过多热量。这种方法将服务器机架浸没在一种在50摄氏度沸腾的专业液体中，远低于水的沸点。当服务器运行并产生热量时，液体沸腾，产生的蒸汽上升到罐盖上的冷却冷凝器处。这种相互作用使蒸汽凝结回液体，然后液体滴落到服务器上，建立了一个闭环冷却系统。通过实施两相浸没冷却，微软每台服务器的能耗降低了5%至15%。这种方法不仅提高了能源效率，还支持了公司积极可持续发展的目标。²⁶

24/7无碳联盟

由气候集团发起的24/7零碳联盟召集采购商，以实现每小时每分钟消耗的每兆瓦时电力都与当地零碳电力相匹配。谷歌、阿斯利康、英国沃达丰、铁山等公司在2024年承诺开发联盟试点，并计划在2025年全面全球推广。除其他多项活动外，会员还分享采购手册，倡导诸如基于时间的能源属性证书等政策，并释放新的储能和灵活性技术，以帮助公司降低清洁电力采购风险，包括数据中心增长。²⁷



操作 3：

使用 AI 体贴地

人工智能的采用往往是由炒作驱动的，而非战略匹配，导致在成本和能源使用方面产生可避免的低效。许多组织即使使用更简单、任务特定的模型也能胜任时，也默认使用大型通用人工智能模型，如大型语言模型。此外，人工智能定价模型通常按固定费率收费，这并不能鼓励高效使用。

虽然人工智能不可否认是一个强大的工具，但将其视为万能的解决方案——哪里有钉子哪里就有锤子——却可能适得其反。大型模型在训练和推理过程中都需要消耗大量能源，从而导致沉重的碳足迹。在基于规则的场景中，依赖人工智能进行实时执行也可能会减慢应用程序并引入意外错误，尤其是在需要确定性结果的情况下。让传统系统处理此类任务，并选择性地使用人工智能来监控、改进或调整规则，要高效得多也更具可持续性。周到的、基于上下文的部署——而不是全面采用——才是最终在最小化成本、复杂性和环境影响的同时驱动有意义的性能提升的关键。

但低效只是问题的一部分。被忽视的是一个更大的机遇。与其说AI是能源消耗的倍增器，不如说AI的超级能力用于可持续发展会怎样？AI通过预测分析、流程优化和智能自动化，有潜力使工业脱碳。但这需要思维方式的转变：AI应作为以目标为导向的创新的工具来部署，为组织创造最大价值并带来最高投资回报。除真正的创新之外，它可以更选择性地使用，以控制其成本和影响。

那些战略性地部署人工智能、将合适的模型匹配到合适任务的企业，能够最小化能源使用，推动可持续性，更具竞争力，并使其产业更智能、更盈利且碳密集度更低。

选择合适的 AI 模型

不要默认使用大型LLM，而是部署针对特定任务的AI模型。Accenture Model Switchboard等平台通过基于性能、成本和上下文无缝协调多个基础模型，帮助组织为每项任务选择合适的模型。² 像检索增强生成（RAG）这样的技术可以通过仅在需要时访问数据来降低推理成本，而混合式AI规则方法可以进一步提高效率。摩根士丹利在其基于GPT-4的AI助手使用了RAG，该助手帮助财务顾问从其庞大的知识库中快速检索信息。现在，该公司超过98%的财富管理顾问使用该AI代理在几秒钟内检索内容，与约20%准确性的较慢手动方法相比，内容相关性达到80%。²⁹

激励效率

从固定费用 AI 定价模式过渡到基于使用量或效率驱动的定价模式，以优化 AI 资源的消耗方式和时间。Hugging Face 为 AI 模型推理提供按请求计价的方案，企业只需为实际使用的 AI 调用量付费。这有助于抑制低效的过度使用，并促进对模型进行精细调整以实现成本效益。³⁰

实现净负排放

AI 应该被用来驱动各行业的脱碳化。例如包括 AI 驱动的供暖、通风和空调（HVAC）优化（例如 BrainBox AI，它通过减少能源使用来降低排放，请参见“用 AI 减少排放”），预测性 25%，物流削减运输排放，以及人工智能驱动的智能电网优化电力分配。³¹ 佳能的低碳制造之旅展示了企业级人工智能如何将碳中和愿景转化为可衡量的减排成果，从供应链到数据中心的工作负载（参见“佳能的能效人工智能战略”）。

案例研究

用人工智能削减排放

加拿大初创公司脑盒人工智能正引领使用人工智能优化商业暖通空调系统。通过利用深度学习于连续的传感器数据流、天气预报和占用模式，脑盒人工智能自主调整暖通空调设置以达最低可能能耗，同时维持室内舒适度。建筑暖通空调能耗排放已减少 使用高达25%并减少与暖通空调相关的CO₂

作为40%——所有无需昂贵的硬件升级。BrainBox AI已部署在数百万平方英尺的商业地产中，包括办公楼、酒店、购物中心和零售连锁店。一个例子是Dollar Tree，它在600家门店中实施了BrainBox AI，并共同节省了798万千瓦时的电力²，相当于103万美元的成本削减和排放减少。防止 5,632 公吨二氧化碳³²

Jabil的AI战略以提高能源效率

Jabil是一家价值280亿美元的全球制造领导者，正在开创利用人工智能和生成式人工智能推动可持续制造。目标是在2045年实现碳中和，Jabil已实施一个多层次的绿色战略，专注于优化其人工智能基础设施的能源消耗。

公司的AI驱动供应链平台提高了物流和采购效率，最大限度地减少不必要的燃油使用和排放。通过节能AI解决方案，捷普优化硬件和算法，在保持高性能的同时降低功耗。它还集成了智能绿色技术来管理太阳能和风能等可再生能源的间歇性，确保AI系统实现最佳负载平衡。为此，捷普已投资储能系统，在峰值生产期间捕获可再生能源以便后续使用——尤其是为了在需求高峰期为AI供电。通过将本地工作负载转移到云服务提供商并采用低代码/无代码平台，捷普进一步减少了其数字碳足迹。³³

操作 4:

嵌入AI 治理- as-code

随着人工智能在各行业和应用加速，它正获得全球监管机构的更多关注。目前，这些关注大多集中在负责任的AI这一主题上，该主题涉及诸如权利侵犯、非预期的偏见和歧视以及虚假信息等问题。这些风险是真实的，必须得到治理，就像环境和资源风险必须得到解决一样。这些都是当今存在重大缺陷的领域。³⁴

人工智能的快速采用已经超越了治理框架，导致了低效和

高碳部署。许多组织对其人工智能工作负载的碳足迹缺乏可见性，最终将可持续性视为次要绩效指标而非核心绩效指标。此外，人工智能部署很少基于碳感知决策，因此错失了减少碳排放的机会——例如，通过根据实时能源组合选择基础设施。当今的现实是，大多数企业在塑造人工智能法规方面是被动而非主动的，导致合规挑战和僵化政策，从而可能扼杀创新。

未来，公司可以将人工智能可持续性嵌入其运营中，以应对新的法规。这包括将人工智能治理系统嵌入业务运营中，以确保人工智能模型在各个职能中得到部署。人工智能能源使用的实时追踪也应成为标准做法，企业应积极为其工作负载选择碳感知的基础设施，以保持合规。

随着人工智能能耗不断增加，将可持续性嵌入自动化合规系统的人工智能治理即代码——对于降低风险、降低成本和构建弹性的、面向未来的人工智能生态系统至关重要。理解并执行到位的企业将获得显著的竞争优势，并确保人工智能是商业价值的驱动者，而不是成本中心或合规负担。



这里是一些需要关注的领域：

将AI可持续性融入治理

管理始于测量，因此应与传统关键绩效指标（KPIs）并行跟踪AI碳强度和能源效率。实时能源监控确保AI部署与可持续性目标一致。组织可以使用OpenTelemetry和Prometheus等云原生可观测性工具以及Open Policy Agent等政策即代码框架来动态执行可持续性阈值。^{35, 36, 37}

微软在这个领域是领导者，从多个角度应对人工智能资源可持续性问题，将其作为更广泛可持续性计划的一部分，投资于优化数据中心并提高人工智能模型效率。通过微软排放影响控制面板等工具，治理被嵌入到他们的人工智能运营中，这些工具帮助组织衡量、监控并减少他们的人工智能工作负载的碳足迹。

微软也开发了自动化的AI政策合规机制，确保AI部署遵循碳感知原则，并遵守欧盟AI法案等监管框架。³⁸

类似地，谷歌已经对其张量处理单元（TPU）进行了全面的生命周期评估，这是一种专门用于人工智能工作负载的硬件加速器。³⁹ 这导致了它们的发展

计算碳强度 (CCI) 指标，衡量每单位计算所产生的碳排放。⁴⁰ 谷歌还实施了一种碳感知测量策略，该策略始于将计算任务转移到太阳能和风能等可再生资源丰富的时段。随后，他们将灵活的计算任务迁移到其全球数据中心，使运营与碳免费能源更易获取的地区保持一致。这种动态方法利用电网碳强度的提前一天预测，使谷歌能够智能地在清洁能源地区安排任务，从而在不影响流媒体或导航等用户服务的情况下减少其碳足迹。

自动化治理策略

人工智能驱动的自动化可以帮助落实可持续性政策，并在实时中管理环境风险。自动化还可以更容易地为每个模型部署选择最可持续的基础设施。

云原生计算基金会（CNCf）项目如开普勒（基于Kubernetes的高效电力等级导出器）支持节能型AI调度，而像Karmada（基于Kubernetes的多云、多集群编排器）这样的智能工作负载编排工具则根据碳排放强度优化跨区域AI工作负载。^{41, 42}

缺乏统一的全球标准为公司提供了一个机会，使其能够积极塑造人工智能可持续性标准和最佳实践，理想情况下是与生态合作伙伴合作。几个关键的行业同侪团体倡议已经启动。IEEE P7100工作组正在制定首个衡量人工智能环境影响的国际标准，绿色软件基金会正在推广节能型人工智能开发，经济合作与发展组织人工智能计算和气候专家组正在推动共享碳核算方法。^{43, 44, 45}

与其等待僵化的合规要求，公司应与监管机构合作，制定平衡创新与可持续发展的政策，帮助政策制定者弥合设计与执行之间的差距。包括SSI联盟和CNCf的TAG环境可持续性在内的行业团体，正在推动开发开放的AI能效基准和与可持续发展相一致的治理模型。^{46, 47}



一个愿景 可持续 大规模AI

人工智能的未来取决于调和指数增长与有限资源的能力。在平衡之术取得成功的组织将把人工智能从高耗能的雄心转变为可持续价值创造的催化剂。最近的全球事件，从气候压力到供应冲击，使这种平衡之术更加紧迫。随着地缘政治紧张局势加剧可能进一步导致人工智能基础设施成本更高和可用性受限，建设弹性行之有效的资源节约型系统对于在不妥协的情况下保持势头至关重要。

通过致力于硅的智能使用、数据中心脱碳、人工智能应用的精心使用以及强大的治理框架，企业可以在不损害地球福祉或其利润底线的情况下，拥抱人工智能的全部潜力。

通过在硬件、软件和政策方面的持续创新，人工智能可以可持续地扩展，确保明天的最智能系统运行在可再生能源、透明标准和明确指南上。通过这样做，组织可以引领向气候韧性经济的转型，同时解锁一个真正服务于人类和地球的人工智能驱动突破的新时代。

实施可持续的人工智能始于将SAIQ整合到您的更广泛的治理和优化框架中。该指标既可用于诊断工具，也可用作决策视角——但只有在其得到可靠数据、系统监控和协调激励的支持时才如此。

1. 在人工智能生命周期中建立一个强大的测量基础。

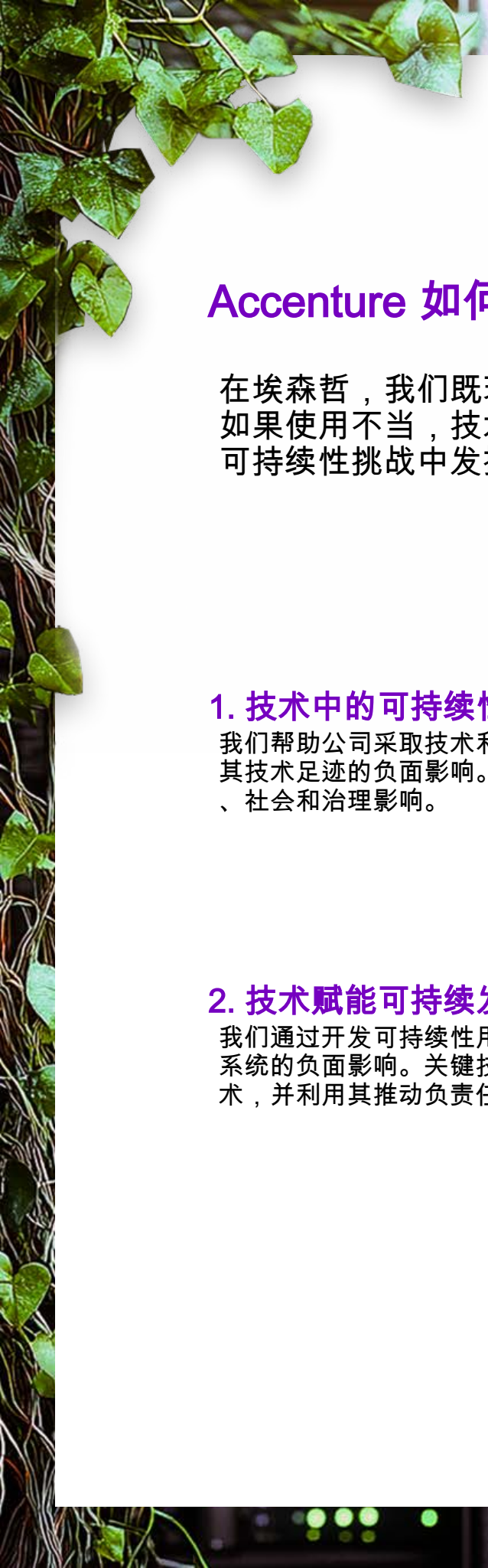
首先在整个人工智能生命周期中收集细粒度的实时遥测数据。在模型复杂度、硬件配置和部署地域的背景下解释这些数据。使用软件碳强度 (SCI) 等其他指标来增强基线，并将范围扩展到包括其他环境指标，如用水量。

2. 将效率嵌入治理和控制系统中。

通过策略即代码方法将指标嵌入人工智能治理平台。Watsonx.governance等平台可实现基础设施、数据和模型层级的可追溯性、实时监控和基于阈值的警报。定义碳风险级别，对低效的训练运行执行自动终止，并优化模型路由和压缩策略，以确保每次部署都符合环境基准。

3. 建立反馈机制和问责机制。

在高管评审和人工智能操作模型中使SAIQ可见，与历史和行业基准跟踪绩效，激励工程团队在可持续性KPI方面取得进展，并利用跨学科洞察支持决策。建立自适应反馈循环以反映不断发展的技术和标准。



Accenture 如何提供帮助

在埃森哲，我们既理解技术的承诺，也理解它带来的挑战。确实，如果使用不当，技术可能是有害的。但我们也相信，技术将在解决可持续性挑战中发挥关键作用。我们专注于两个使命：

1. 技术中的可持续性

我们帮助公司采取技术和管理措施，从基础设施和设备到运营和软件，从根本上减少其技术足迹的负面影响。关键能力正在被引入，以改善核心IT和云运营足迹内的环境、社会 and 治理影响。

2. 技术赋能可持续发展

我们通过开发可持续性用例和先进组织能力来创造数字技术，从而减少商业生态系统的负面影响。关键技术产品和服务正在开发中，以帮助公司可持续地部署技术，并利用其推动负责任的治理和更好的客户成果。

我们的目标是改善技术的社会和环境
影响，并利用以技术引领的创新
来创造一个可持续的未来。作为全
球服务提供商，埃森哲的独特优势
在于确保可持续性被设计到为全球
业务提供动力的技术解决方案中。

[联系我们了解更多](#)

研究负责人

吉朱·马修

贡献者

马尼莎·达什 米歇拉·科波拉 莫尼克·德·里特 拉马尼·莫斯 塞布兰苏·萨霍

方法论

我们根据预期安装的GPU数量，并考虑利用率调整和数据中心额外的能源需求来建模能源消耗，假设相关创新轨迹没有变化。我们通过应用指数移动平均法来估算缺失值。通过基于正在使用的设施直接估算能源消耗，该模型涵盖了训练和推理工作负载。我们按Jefferies对各类数据中心的所有预测比例分配地区，预计到2030年，人工智能数据中心将占有数据中心能源需求的超过一半。随后，我们将IEA各地区的电力生产、能源组合以及IEA各能源来源的CO排放数据结合起来，以估算CO排放。^{2 2} 每个地区的人工智能数据中心排放。

根据人工智能数据中心能耗和已公布的水强度系数，评估了两个范围——直接（冷却）和间接（发电）——的用水情况，得出总运营用水量。

专家访谈

我们采访了17位人工智能和可持续性领域的专家，涵盖行业、角色和地区——从云服务提供商到跨国公司再到学术界、初创企业和行业团体。我们使用定性主题方法分析了这些访谈记录。每份访谈记录都进行了编号，并手动筛选，仅包含专家的回应。我们采用了归纳编码方法来识别反复出现的主题，并辅以基于关键词的内容分析以确保一致性。主题类别通过迭代细化和通过审查代表性引言进行验证。

关于埃森哲

埃森哲是一家领先的全球专业服务机构，帮助世界领先的企业、政府和其他组织构建其数字核心、优化运营、加速收入增长和提升民服务——以速度和规模创造切实价值。我们是一家以人才和创新为导向的公司，拥有大约791,000名员工，服务于120多个国家的客户。如今，技术是变革的核心，我们凭借强大的生态系统关系，是推动这一变革的世界领导者之一。我们将我们在技术方面的优势以及云、数据、AI领域的领导力与无与伦比的行业经验、职能专业知识和全球交付能力相结合。我们涵盖战略与咨询、技术、运营、行业X和宋的广泛服务、解决方案和资产，以及我们共享成功的文化和对创造360°价值的承诺，使我们能够帮助客户重塑并建立信任、持久的合作关系。我们通过为客户、彼此、股东、合作伙伴和社区创造的360°价值来衡量我们的成功。访问我们的 [accenture.com](https://www.accenture.com)

埃森哲研究

埃森哲研究就组织面临的最紧迫的商业问题创建思想领导力。结合创新研究技术，如数据科学主导的分析，以及对行业和技术深刻理解，我们20个国家的300名研究人员每年发布数百份报告、文章和观点。我们与世界领先组织共同开发的有启发性的研究帮助客户拥抱变革、创造价值并实现技术和人类智慧的力量。欲了解更多信息，请访问www.accenture.com/research上的埃森哲研究。

本文件涉及第三方拥有的标志。所有此类第三方标志均为其各自所有者的财产。本文件无意表示或暗示此类标志的所有者对本文内容提供赞助、认可或批准。本文仅供参考，并非取代咨询我们专业顾问。

版权所有 © 2025 Accenture。保留所有权利。Accenture及其标志是Accenture的注册商标。