Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Energy Systems Engineering—Perspective

防控大语言模型赋能的智能聊天机器人全生命周期能源与碳足迹剧增

蒋鹏[a], Christian Sonne[b], 李望良[c], Fengqi You[d], Siming You[e,*]

[a] Department of Industrial Engineering and Management, Business School, Sichuan University, Chengdu, 610064, China
[b] Department of Ecoscience, Aarhus University, Frederiksborgvej 399, Roskilde, DK-4000, Denmark
[c] CAS Key Laboratory of Green Process and Engineering, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, 100190 China
[d] Systems Engineering, Cornell University, Ithaca, NY 14853, USA
[e] James Watt School of Engineering, University of Glasgow, Glasgow, G12 8QQ, UK

ARTICLE INFO

Article history:
Received 20 August 2023
Revised 6 April 2024
Accepted 7 April 2024
Available online 17 April 2024

关键词
大语言模型
智能聊天机器人
碳排放
能源与环境足迹
全生命周期评估
全球合作

摘要

近年来，大语言模型（large language model, LLM）赋能的智能聊天机器人在全球范围内迅速发展，并展现出广泛的行业应用潜力。全球前沿科技公司正积极参与基于LLM的聊天机器人设计与开发，除了广为人知的ChatGPT之外，市场上也涌现了多种替代方案。然而，训练、微调及更新此类智能聊天机器人需要消耗大量电力，从而导致显著的碳排放。所有LLM及其软件和硬件（如图形处理单元及超级计算机）的研发制造、相关数据及运营管理，以及支持聊天机器人服务的材料回收等环节，均不同程度地伴随能源消耗及碳排放。因此，当前及未来应关注基于LLM的智能聊天机器人全生命周期的能耗及碳足迹，以缓解其对气候变化的影响。本研究明确并强调了在此类智能聊天机器人开发的生命周期中八个主要阶段的能源消耗及碳排放影响。基于这些阶段的生命周期及相互作用分析，我们提出一种系统级解决方案，并通过三条战略路径优化该行业管理，从而降低相关碳足迹。在预见这一先进技术及其产品巨大潜力的同时，我们呼吁重新思考大语言模型赋能的智能聊天机器人行业全生命周期的能耗与碳排放的减排路径与策略，并在其早期发展阶段重塑其能源与环境影响。

© 2024 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. 引言

自2022年年底以来，大语言模型（large language models, LLM）赋能的智能聊天机器人在全球范围内迅速流行[1–4]。作为一种先进的生成型预训练变换器（generative pretrained transformer, GPT）工具，ChatGPT在发布仅两个月内便吸引了1亿用户，并在2023年1月实现了5.9亿次访问量[5]。这一趋势在2023年年底显著加速，每月访问量超过15亿次[6]。基于LLM的智能聊天机器人是生成式人工智能（artificial intelligence, AI）发展的重要组成部分。即使考虑到新冠病毒肺炎疫情和地区战争的影响，生成式AI预计在2023—2030年间仍将保持24.4%的年增长率，市场规模将从449亿美元增长至2070亿美元[7]。在这场激烈的竞争中，类似的智能聊天机器人如Google Bard、Bing Chat、My AI、LLaMA、PaLM、PaLM-E、Copilot X、Ernie、Qwen-72B和Alpaca等[8]相继涌现。同时，OpenAI、Google AI、Meta AI、Deepmind、Amazon、华为、GitHub、Anthropic、百度、Nvid-

* Corresponding author.
E-mail address: Siming.You@glasgow.ac.uk (S. You).

2095-8099/© 2024 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
英文原文：Engineering 2024, 40(9): 202–210
引用本文：Peng Jiang, Christian Sonne, Wangliang Li, Fengqi You, Siming You. Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots. *Engineering*, https://doi.org/10.1016/j.eng.2024.04.002

ia、阿里云等全球前沿科技公司和研究机构也纷纷参与其中，提供了多样化的人工智能辅助选择。LLM正在加速并优化从艺术[1,9]到科学[2,10–12]等各领域的全球发展。

在融入LLM的ChatGPT之前，会话型聊天机器人已广泛应用于多个领域，包括医疗、酒店服务、公共服务、娱乐、教育和制造业[13]。会话型聊天机器人主要分为三类：基于规则的聊天机器人、实时聊天机器人和基础AI赋能的聊天机器人[13–14]。前两类聊天机器人通过预设规则进行交流，分别结合聊天机器人软件和人工对话来提供客户服务。第三类基础AI赋能的聊天机器人能够在无需人工干预的情况下超越预定义命令开展对话，为更高级的智能聊天机器人奠定了基础。然而，与最近开发的基于LLM的智能聊天机器人（如搭载GPT-3、GPT-3.5和GPT-4的ChatGPT）相比，这三类会话型聊天机器人的模型训练量、微调参数数量较少，使用的硬件也相对较差。这些LLM赋能的智能聊天机器人具有庞大的模型参数量。图1展示了从2019年到2021年发布的12种代表性LLM的模型参数数量增长趋势，从拥有0.66亿参数的DistilBERT到拥有1.57万亿参数的Switch-C，参数量在三年内显著增加了23 788倍[15]。

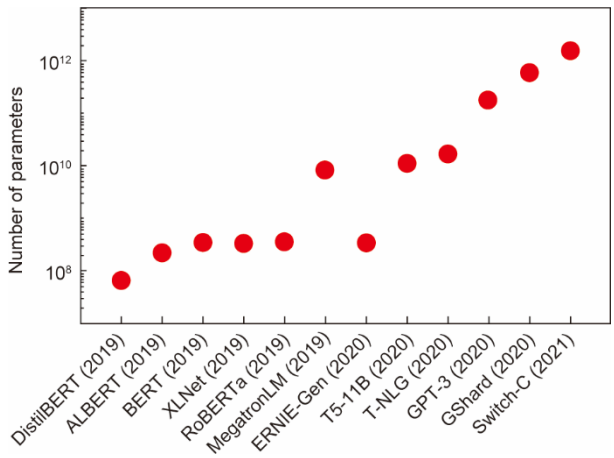

图1. 2019—2021年LLM参数数量的增长趋势（基于文献[15]的相关数据）。

在LLM赋能的智能工具和服务广受欢迎的背后，相关的能耗与环境影响引发了广泛关注[3,16]。LLM赋能的智能聊天机器人消耗大量电力，并依赖于图形处理单元（graphics processing units, GPUs）、张量处理单元（tensor processing units, TPUs）和超级计算机等硬件[16]。从整个系统的角度来看，除了包含智能LLM的软件及GPU、TPU和超级计算机等硬件外，智能聊天机器人服务还需要诸如先进制造、物流和设施运营管理、海量数据采集、数据中心管理及废弃物管理等额外的支持。为减轻LLM赋能的智能聊天机器人行业对能源和环境的影响，必须综合考虑该行业在整个生命周期内的能源与资源消耗及其环境影响。本文识别了LLM赋能的智能聊天机器人在开发与应用全过程中的主要能耗和碳排放阶段，并提出一种针对多利益相关方的系统级解决方案，旨在协助优化该行业的管理与发展。最后，本文展望了新的研究方向并提出一些具有前瞻性的见解。

2. 全生命周期能源与碳足迹

2.1. 生命周期阶段的识别

基于生命周期评估（life cycle assessment, LCA）方法，并按照“从摇篮到坟墓”的分析方法，智能聊天机器人的开发与应用在资源开采、生产、使用和处置过程中产生的环境影响均需考虑[17]。根据IEC 62890:2020标准[18]，所有合作方，包括产品生产商、供应商、服务提供商和用户，都应参与到LCA中，且设计、规划、开发、运营和维护阶段均需纳入考量。因此，与LLM赋能的智能聊天机器人相关的生命周期能耗和碳排放管理预计涵盖八个主要阶段（图2）。具体而言，这些阶段包括聊天机器人的研发（research and development, R&D）（阶段1）、硬件制造（如GPU、TPU、超级计算机及服务设备/机器人）（阶段2）、全球商业物流（阶段3）、设施运营和维护（阶段4）、海量数据采集与管理（如通过高能耗的大型数据中心进行源数据管理）（阶段5）、LLM的训练和微调（阶段6）、线上/线下聊天机器人服务（阶段7），以及硬件材料的回收和废弃物处理（阶段8）。

2.2. 能源与碳足迹

截至目前，LLM赋能的智能聊天机器人行业生命周期能耗和碳排放的估算仍是一个未解决的问题，关于该行业整体能耗和碳排放的数据和信息有限。此外，隐含的能耗和碳排放与全球贸易、LLM训练及聊天机器人在线服务密切相关，而对这些环节的追踪具有较大挑战性。自回归类型BLOOM的碳足迹量化结果显示，其隐含碳排放、与空闲消耗相关的碳排放以及与动态消耗相关的碳排放分别占其碳足迹的22.2%、28.9%和48.9% [19]。本节中，我们利用现有数据，探讨并分析LLM赋能的智能聊天机器人生命周期能耗和碳排放管理八个主要阶段的潜在能耗和碳足迹。

2.2.1. 阶段1：聊天机器人的研发

智能聊天机器人的研发涉及LLM及硬件（如GPU、

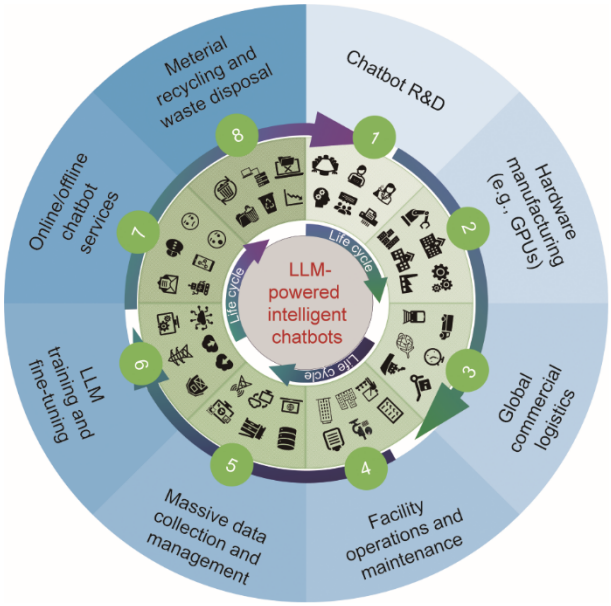

图2. LLM赋能的智能聊天机器人生命周期能耗和碳排放管理示意图，涵盖从研发、硬件制造到服务及废弃物管理全过程。

TPU和超级计算机）。尽管2022年和2023年被视为ChatGPT和LLM广泛应用的突破期，相关的研发进展却始于更早时期。例如，OpenAI于2015年成立，首代GPT-1于2018年6月推出，拥有1.17亿参数，随后在2019年2月发布的GPT-2参数超过10亿。类似的研发尝试也在全球前沿科技公司和研究机构中出现。从生命周期的角度来看，研发相关的能源消耗和碳排放应追溯到更早的年份，并包括自那时以来的相关人力及物力等的投入。

2.2.2. 阶段2：硬件制造

LLM赋能的智能聊天机器人硬件制造属于资源与能源双密集型过程。所需硬件包括大量高性能GPU、TPU、超级计算机、大型数据中心设施以及线下服务设备/机器人。例如，由于生成式AI在该领域的快速应用，预计英伟达（Nvidia）公司将在2023年内全球交付55万片最新旗舰H100 GPU [20]。与配备英伟达DGX A100 GPU的超级计算机相比，配备H100 NVL GPU的超级计算机可以将GPT-4的性能提升约12倍[21]。然而，尽管先进的RTX 4000系列GPU可以通过台湾积体电路制造股份有限公司（以下简称台积电）5 nm工艺进行量产，高性能的H100 GPU（每片配备5120位内存、80 GB HBM3容量和14 592个CUDA核心）则需要通过台积电4 nm工艺等高端且高能耗的设备制造[20]。

另一个方面是集成电路的制造。现有的生命周期评估研究表明，集成电路制造过程中累计的能耗范围为9~38 MJ·cm^{-2} [22]。此外，硬件制造所需材料须经过金属开采与生产过程，金属开采与生产是一个高能耗过程，全球约38%的工业能耗和约15%的电力消耗用于金属生产[23]。用于GPU和TPU芯片制造的稀土金属开采与生产的能耗和环境足迹非常高[24]。总之，智能聊天机器人所必需的高端设备制造涉及多种先进和智能制造技术[25]，再加上资源和能源密集型的原材料生产，其相关的能耗和碳排放不可忽视。

2.2.3. 阶段3：全球商业物流

与智能聊天机器人开发和应用相关的全球商业物流包括传统物流（产品交付）和逆向物流（处理阶段8产生的废弃设备和材料）。由于GPU、TPU、数据服务器和线下服务聊天机器人在供需方面的时空异质性，硬件分配过程中涉及大量通过多种协调运输方式进行的全球运输。例如，GPU是智能聊天机器人开发中的关键组件：训练拥有500亿参数和3630亿数据词元的轻量级BloombergGPT需要512块英伟达A100 GPU [26]；GPT-4的训练使用了1万~2.5万块A100高性能GPU，而未来的GPT-5可能需要3万~5万块[27]。现有数据显示，2022年出货的专用GPU在一年内可能消耗约9500 GW·h的电力[16]。

2.2.4. 阶段4：设施运营和维护

设施运营与维护（operations & maintenance, O&M）是智能制造和供应链管理的重要组成部分，涉及监督所有与设施相关的资产、参与者和流程，以及涵盖设施维护和相关设备的未来维护执行。在LLM赋能的智能聊天机器人行业中，与运维相关的能耗和碳排放涉及多个领域，如制造设施管理、智能仓库管理以及大型数据中心等硬件设备的维护管理[28]。目前，关于LLM赋能的智能聊天机器人行业的运维能耗和碳足迹信息非常有限，尽管其他行业（如太阳能和风能行业）的运维数据表明这些碳足迹可能较为显著。例如，运维和退役阶段的能耗和碳排放可能占到风力发电总体碳足迹的约10% [29]。

2.2.5. 阶段5：海量数据采集与管理

为支持LLM和智能聊天机器人服务，需要收集和管理大量数据。用于训练GPT-3的多种数据源包括过滤后的Common Crawl（4100亿词元）、WebText2（190亿词元）、Books1（120亿词元）、Books2（550亿词元）和Wikipedia（30亿词元）[30]。为了避免模型拒绝有效请求，OpenAI在GPT-4的训练中收集了多样化的数据集，其中包括标记的生产数据、模型生成的提示数据以及人工对抗类型的数据[31]。尽管目前关于LLM相关数据管理的总体能耗和碳排放信息尚不明确，但潜在的资源需求应引起关注。

已有研究表明，大型数据中心及其管理过程是资源和能源密集型的。例如，据估算全球数据中心在2018年的总能耗为205 TW·h（约占全球电力消耗的1%），而数据中心的计算实例较2010年增加了550% [32]。2018年，美国数据中心的温室气体（greenhouse gas, GHG）排放总量为3.15×10^7 t CO_2-eq，占该国总体排放量的约0.5% [33]。除了能耗和碳足迹外，用于训练GPT-3的微软高性能数据服务器的冷却水消耗量为540万升，同时ChatGPT服务在处理10~50次查询时会消耗约500 mL水[34]。

2.2.6. 阶段6：LLM的训练和微调

作为在软件和算法方面另一个关键的高能耗阶段，训练AI赋能的LLM消耗大量能源并具有显著的碳足迹[3, 35]。例如，训练一个拥有2.13亿参数的单一Transformer架构会产生约300 t CO_2-eq的碳排放（图3），相当于北京和纽约之间125次往返航班的排放量[35]。这还未包括在最终训练之前为进行训练试验而消耗的大量资源[36]。例如，大约需要预训练5000个额外的模型，才能获得性能更优的最终模型[37]。即使随着技术发展提高了能效，GPT-3的最终训练（拥有1750亿参数）仍消耗了1287 MW·h的电力，其碳足迹为552 t CO_2-eq [36,38]，这还不包括模型微调和更新的能耗与排放（图3）。类似的情况也发生在Meta的LLaMA模型上，微调该模型需要超过100块A100高性能GPU [39]。总体而言，这意味着假设GPT-3的预训练进行5000次，其碳排放可能高达2.76×10^6 t CO_2-eq，相当于83.6亿剂新冠疫苗的生命周期碳足迹，后者足以覆盖全球人口的一剂疫苗需求[40]。

在一个幽默轻松的情境中，当我们询问ChatGPT“训练ChatGPT的电力消耗是多少？”时，收到的回答是：“……训练GPT-3模型可能消耗高达355 MW·h的能源，这大致相当于一个普通美国家庭32年的能源消耗。”然而，从系统的角度来看，考虑到生命周期中所包含的至少八个主要阶段（图2），GPT-3的能耗预计会更高。此外，随着LLM训练算法的显著改进，能耗和碳排放将急剧上升，因为LLM的模型规模和输入数据量不断增加。例如，从GPT-1的1.17亿参数和5 GB数据，到GPT-2的12亿参数和40 GB数据，再到GPT-3的1750亿参数和45 TB数据[41]。根据已知的GPT-3设置以及假设GPT-4拥有1万亿参数，估算表明，使用1万块V100 GPU训练GPT-3和GPT-4分别需要大约26天和150天[42]。相比GPT-3的电力使用情况，GPT-4的最终训练耗电量为7200 MW·h [42]。根据GPT-3的碳排放计算，GPT-4的碳排放约为3088 t CO_2-eq [36]。相关数据和信息如图3所示。

2.2.7. 阶段7：线上/线下聊天机器人服务

与LLM训练和微调相比，终端用户更熟悉的是他们在使用智能聊天机器人的网络服务（如计算机和手机）或线下服务设备（如机器人）时的人机交互。基于云端托管的已训练和微调的LLM，在向用户反馈提示时进行快速推理和响应，这一过程消耗能源并产生碳排放。对于每个新提示，推理需要一系列GPU/TPU运行已构建的LLM并进行快速计算。例如，Meta的LLaMA模型在推理、微调和训练过程中通常分别使用16块、100多块和2000块A100高性能GPU [39]。

虽然单次推理的能耗和碳排放并不显著，但全球范围内的长期累积碳排放对于线上/线下聊天机器人服务而言

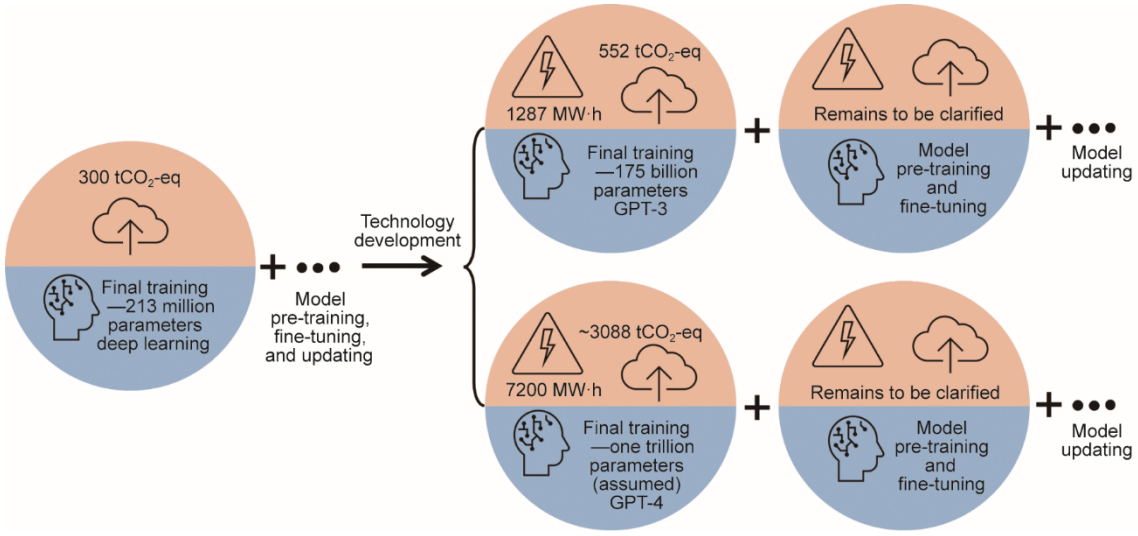

图3. LLM训练与微调以及模型开发过程中的电力消耗与碳排放。新算法和技术有望提高LLM训练的能效。需要进行大量研究，以明确在LLM最终训练之外的计算过程中额外的能耗与碳足迹。

是显而易见且不可忽视的。每月在线服务的相关能耗和碳足迹可与阶段6描述的LLM最终训练过程的能耗相当。根据TRG数据中心的假设和估算数据[42]，以及2023年1月的5.9亿次访问量[5]，在九种不同情景下，本文估算了ChatGPT服务的总体月度电力消耗，如图4所示。估算表明，有五种情景和八种情景的电力消耗分别超过了GPT-4和GPT-3的最终训练能耗。在最坏情况下（即情景3），每月电力消耗达到23.364 GW·h。如果基于2023年8—10月期间月均15亿次访问的峰值数据[6]，按此发展趋势估计，在最坏情况下未来通过Google搜索使用ChatGPT服务的年电力消耗将高达7.128 TW·h。

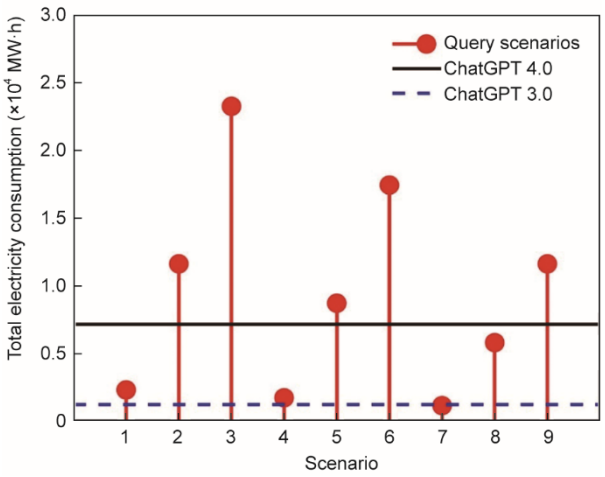

图4. 不同情景下，基于Google搜索的ChatGPT服务相关的每月查询总耗电量。情景1、4和7基于每次访问一次查询；情景2、5、8基于每次访问五次查询；情景3、6和9基于每次访问十次查询；同时，情景1、2、3代表效率较低情况：每次查询用电量为0.00396 kW·h；情景4、5和6表示中等效率：每次查询用电量为0.00297 kW·h；情景7、8和9表示高效率：每次查询用电量为0.00198 kW·h [42]。每次查询的用电量数据与参考文献[43–44]中报道的数据类似。GPT-3和GPT-4模型最终训练运行的预计耗电量对比如图所示（ChatGPT 4.0对应ChatGPT的GPT-4版本，ChatGPT 3.0对应ChatGPT的GPT-3版本）。

2.2.8. 阶段8：硬件材料的回收和废弃物处理

在LLM赋能的智能聊天机器人行业中，材料回收和处置主要发生在硬件设备的生命周期终结阶段（如阶段2所述）。该行业中有效的材料回收与智能循环经济概念的发展密切相关。机器人材料[45]和印刷电路板电子元件[46]的回收及再利用在全球范围内具有重要影响，有助于降低经济成本及能源和环境足迹。此外，稀土金属开采和生产的高能耗[24]表明，材料回收，尤其是关键原材料的回收，是实现资源循环的重要途径，并可以带来显著的间接能源节约和碳减排效果。然而，事实是关键稀有材料的资源回收效率较低，相关政策干预仍处于初期阶段[47]。要解决提取率低和纯化过程漫长等问题，亟须在产品设计、社会行为分析、分离热力学及回收技术方面进行广泛的研发[48–49]。尽管LLM赋能的智能聊天机器人行业发展迅速，但如果能够有效回收和再利用材料，则可以有效地降低整体的能耗和碳足迹。

2.3. 不同阶段间的交互和影响

不同生命周期阶段之间的可视化关系和交互如图5（a）所示，其中呈现一个包含八个阶段的有向网络结构。对这些复杂交互的分析需要具备系统工程学的知识，并且需考虑若干内在因素：

首先，在分析的八个阶段中，全球范围内涉及广泛的参与者，包括研发团队（阶段1）、制造工人（阶段2）、物流人员和电子商务运营商（阶段3）、设施管理人员（阶段4）、数据工程师（阶段5）、算法工程师（阶段6）、软件工程师和运营人员（阶段7）以及回收工人、处置工人和物流人员（阶段8）。这意味着参与者的活动和交互会导致显著的能耗和环境足迹。在这样一个相互关联的系统中，参与者之间可能的信息流和连接表明协同其活动以实现最佳的整体能源和环境绩效至关重要。

其次，图5（a）中的箭头代表不同阶段之间的相互作用。例如，阶段1与阶段2、5、6、7之间的相互作用表明：①研发应与现有的智能制造技术保持一致；②研发可以改进数据需求和数据管理；③研发是提升LLM训练和微调效率的关键；④研发应始终旨在提高智能聊天机器人的服务质量，进而间接影响整体能耗和碳排放。阶段2的硬件制造被认为是一个资源和能源密集的阶段，与所有其他阶段相互关联。在第2.2小节中我们初步讨论了阶段3、4、5、6及其与其他阶段可能的相互作用。阶段7的线上/线下聊天机器人服务涉及从全球用户收集新数据，这些数据将在阶段5涉及的数据中心进行管理，然后将在阶段6进一步用于LLM训练和微调。例如，OpenAI使用来自GPT-3用户和50位AI专家的反馈数据对GPT-4进行微调，这些数据存储在其数据中心[31]。在阶段8，材料和电子设备的回收、再利用有助于减少阶段2的制造活动和阶段3的全球商业物流，进而通过这两个阶段间接减少整个系统的能耗和碳足迹。

3. 前进之路：通往优化管理的系统级解决方案

基于对LLM赋能的智能聊天机器人行业生命周期能耗与碳足迹的分析（第2.2节）以及各阶段之间相互作用的探讨（第2.3节），本节提出并展示了一个包含若干路径的系统级解决方案，以降低碳足迹。

类似于LLM赋能的智能聊天机器人，另一项新兴技

术也依赖于GPU、TPU、超级计算机和大型数据服务器等硬件以及软件的支持，即基于区块链技术的加密货币挖矿，该技术具有显著的能耗和碳排放[50–51]。例如，在无政策干预的情景下，预计中国的比特币挖矿年碳排放和能耗将在2024年达到峰值，分别为1.305×10^8 t CO_2-eq和296.6 TW·h [50]。然而由于比特币数量的限制，如果交易费用不足以激励基于区块链的加密货币挖矿，挖矿活动的能耗峰值可能会推迟至2140年[52]。相比之下，目前针对LLM赋能的智能聊天机器人的发展和全球竞争，还缺乏有效的机制来促进其朝能源和环境可持续性方向发展。

人们认为，LLM赋能的智能聊天机器人将变得极为普遍，并通过整合多模态数据处理能力深刻改变世界[8, 53–54]。然而，在期待这一先进技术巨大潜力的同时，需要制定切实的解决方案，通过专门的行动路径和更广泛的社区与利益相关方的合作，来优化该行业的管理并降低其能耗和环境足迹。对此，现有文献提出了一些初步想法。例如，已有学者初步提出一个概念框架，建议使用由绿色能源（如太阳能、氢能和风能）驱动的高能效硬件，并调整LLM的模型结构[16]。这些构想需要在技术研发方面取得关键突破，包括：①在工业4.0和以人为中心的工业5.0 [55]背景下革新硬件制造技术[25]；②通过算法、软件和硬件架构的改进来提升高能效的计算能力[36,56]；③优化LLM的性能动态[57]；④人类反馈辅助的技术应用[58]。

因此，除了技术创新外，我们认为该行业的多个利益相关方之间也需要开展大规模的全球合作。根据第2节的分析，参与该行业能源和环境管理的主要利益相关方包括政府当局/机构、新兴行业协会、专业环保团体、科技公司和研究机构、技术报道者以及全球聊天机器人用户。基于图5（a）所示的生命周期和交互分析，并考虑到相关改进涉及科学问题和机制探索，包括多方利益相关者的影响/交互、多目标优化以及多主体决策，我们提出了图5（b）所示的系统级解决方案，该方案包括涵盖所有主要利益相关方的三条战略路径。下文将详细说明适合执行每条战略路径的特定利益相关方及其关联情况。

3.1. 战略路径I：全面生命周期评估和动态报告

在路径I中，利益相关方主要是专业环境团体，包括权威环境组织和机构、非政府机构、环境领域的研究人员，以及具有专业背景的环保人士。LLM赋能的智能聊天机器人相关的累积能耗和碳排放评估应涵盖图2中列出的第1至第8阶段。这类评估应采用全系统方法（如机器学习碳排放和能耗足迹的系统报告方法[59]），并应不断推广来自权威环境组织和机构的相关报告，如联合国环境规划署（United Nations Environment Programme, UNEP）的报告，以便促进相关环境问题的准确表征并符合联合国

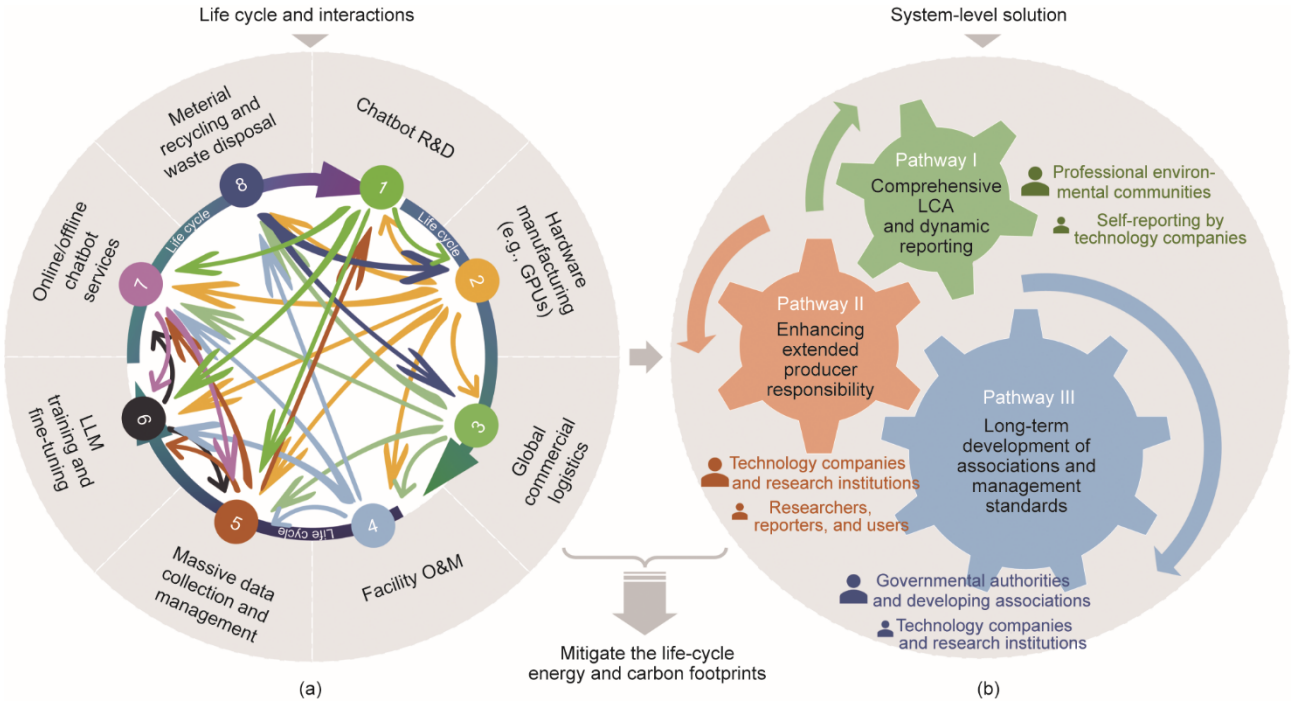

图5. 降低LLM赋能的智能聊天机器人行业全生命周期能源消耗与碳足迹的途径（基于第2节分析）。（a）全生命周期中八个主要阶段之间的关系和相互作用。每个箭头表示一个阶段（源）如何相互关联和（或）影响另一个阶段（汇）；（b）基于生命周期能源使用和碳排放分析，优化LLM赋能的智能聊天机器人行业管理的系统级解决方案示意图。

可持续发展目标（sustainable development goals, SDGs）。基于生命周期的视角，应对智能聊天机器人开发和应用的各种情景及动态趋势进行分析，以预测其后果和影响。图3和图4中的估算结果以及图5（a）中的生命周期和相互作用表明，未来LLM相关的总体能源及碳足迹将显著超过LLM训练和微调阶段的足迹总和。针对不同阶段变化的动态报告、数据质量和准确性保证及其相关影响，有望促进有效的政策制定和行动计划，从社会经济、能源和环境的整体角度优化智能聊天机器人行业的发展。

然而，值得注意的是，在各种商业环境中数据可靠性和透明度的有限性是必须妥善解决的障碍，否则无法实现全面评估和动态报告。例如，出于对商业机密和知识产权的担忧，诸如OpenAI等私营机构缺乏动力共享和报告其最新相关数据，且目前没有现成的门户用于数据共享。这也是GPT-4相关能耗和碳足迹仍然保密的部分原因。从信息披露的角度看，可以通过激励措施加强生产者责任延伸（路径II）及发展行业协会与管理标准（路径III）来解决这一障碍，如图5（b）所示。此外，从数据管理的角度看，整合区块链、物联网（internet of things, IoT）和机器学习模型等先进的新兴技术，有助于提高数据的可靠性和透明度，以实现精确的环境核算[60]。

3.2. 战略路径II：通过激励措施加强生产者责任延伸

路径II的利益相关方主要是前沿科技公司和研究机构，特别是那些引领智能聊天机器人行业研发革命的企业，而其他参与者如中小型科技企业、环境社区研究人员、技术记者和全球聊天机器人用户，也能为这一路径做出贡献。延伸生产者责任（extended producer responsibility, EPR）概念强调，产品所有生命周期的环境影响都应被计入。

首先，在理想情况下，智能聊天机器人的设计者应了解并选择更绿色的途径来训练和微调LLM。例如，Bloomberg公司采用了基于亚马逊云服务的分布式共享资源，分配了大约130万小时的训练，而不是涉及至少第3、4、5、6和8阶段的完整循环[26]。这一路径对于设计基于LLM平台和应用程序的中小型科技企业尤其具有现实可行性。

其次，合适的激励机制在EPR的实施中起着关键作用。在第8阶段的智能聊天机器人生命周期终结管理中，EPR的监督可以激励材料回收管理中的最佳实践。鉴于关键稀有金属生产和回收的宏观治理具有战略重要性，这有助于同时降低生命周期终结的成本以及能耗和碳足迹[47]。此外，在阶段4的绿色运维管理、阶段5的可持续数据中心管理以及阶段6的高能效LLM训练与微调中引入激励机制也具有积极作用。

再次，在LLM赋能的智能聊天机器人行业实施EPR的理想情况下，主动报告碳足迹、能耗和资源使用情况有助于避免比特币挖矿中出现的滞后管理问题[61]。例如，多个开源替代方案，如Meta的LLaMA和LLaMA-2、BigScience的BLOOM、阿里云的Qwen-14B和Qwen-72B、斯坦福大学的Alpaca以及清华大学的ChatGLM-6B，正在提供技术报告和（或）代码，用于构建LLM系统并估算其能耗和碳足迹。然而，由于缺乏明确的政策法规和激励措施，最终可能导致一种非合作博弈局面，领先的科技公司和研究机构缺乏足够动力来持续实施EPR。在出现主要的非营利组织或协会（路径III中讨论）之前，这一问题无法轻易通过自上而下的方式解决（即依靠领先公司和机构自觉主动地推动行业改进）。另一方面，该问题可能通过自下而上的方式得到解决，即通过利基市场企业的革命性行动、环境社区研究人员揭示的碳排放事实以及全球用户的自发声音。这些“底层”的行动可以成为激励研发科技公司和研究机构领导者的动力。以下为三类参与利益相关方的具体行动选项：

（1）从利基市场的角度来看，对于作为基层竞争者的中小型科技企业，可以设计和开发相对轻量级的智能聊天机器人，如mini-GPT4 [62]和Alpaca [8]，使其成为开放访问的资源。这类工具对于不需要高级ChatGPT应用的某些用户仍然有用。轻量级智能聊天机器人还可以通过加入领域特定的知识应用于多个行业/领域。例如，Bloomberg公司开发了轻量级的BloombergGPT，其参数和数据需求远少于其他模型，专用于金融领域服务[26]。另一个例子是Google的Socratic应用于教育领域，其模型训练的电力消耗仅为50~300 MW·h [42]。在药物发现领域，还有一些小型ChatGPT变体，在医疗应用中具有较高的准确性[63]。

（2）从环境社区研究人员的角度来看，已经讨论过LLM和ChatGPT的成本和效益[53–54]，但仍需基于可靠和标准化的数据展示来进一步深入量化讨论LLM赋能的智能聊天机器人的利弊。预计在硬件制造（阶段2）、LLM的训练和微调（阶段6）、线上/线下聊天机器人服务（阶段7）等几个主要的生命周期阶段会出现较大的能耗和碳排放。研究人员应与利益相关方（如非营利组织和权威环境机构）合作，精确绘制碳排放热点，并探索数据共享伙伴关系的可能性，提出减排措施的建议。

（3）从科技记者和终端用户（尤其是环保人士）的角度来看，针对技术发展和用户体验的反馈，尤其是对智能

聊天机器人相关能耗和环境问题的关注，有可能推动智能聊天机器人企业负责任地发展。例如，如果众多记者和用户分别通过传统媒体平台（如报纸和杂志）及社交媒体平台（如Twitter、Facebook和TikTok）发声，媒体和用户的自发声音有可能成为一种机制，推动激励科技公司和研究机构在阶段1、2进行改进。公众的声音也可以成为大力支持阶段4的绿色运维管理以及阶段8的智能聊天机器人生命周期终结管理的激励措施。

3.3. 战略路径III：长远发展行业协会与管理标准

路径III的利益相关方一方面是设立行业标准的政府当局/机构和新兴协会，另一方面是需要被监管和激励的智能聊天机器人研发公司和研究机构。从长远来看，为了该行业的可持续发展，通过科技公司、学术界、组织和机构之间的协作和共识制定相关协会和管理标准至关重要。这可以通过发展类似于国际机器人联合会（International Federation of Robotics, IFR）、国际航空运输协会（International Air Transport Association, IATA）和国际汽车工作组（International Automotive Task Force, IATF）等国际非营利组织的形式来实现。以汽车行业的行动路径作为参考，定义LLM赋能的智能聊天机器人行业的具体路线，利用至少三个维度的约束来规范和支持该行业的发展。这些约束可能包括：①在不同地区实施排放立法，如欧六标准、国六标准和美国标准（即LEVIII和EPA Tier3）[64]；②基于国际标准ISO 9001制定IATF 16949:2016，以提高运营效率，减少供应链中的运营成本和浪费[65]；③联合国推动到2050年实现净零排放交通，以进一步促进经济发展并减少不平等现象[66]。相比IATA计划到2050年实现净零碳排放的雄心，特别是考虑到当前和未来全球危机的可能影响，LLM赋能的智能聊天机器人行业如何应对碳中和目标仍不确定[67]。

随着大量资本进入智能聊天机器人行业并引发广泛的前景期待，涉及不同规模的利益相关方的跨境合作是制定相关协会和管理标准的必要条件。然而，必须承认在该行业建立运作良好的协会将需要大量的努力和时间。例如，自2002年在“零废弃”理念下成立首个协会——国际零废弃联盟（Zero Waste International Alliance, ZWIA）以来，分别用了约9年和13年时间推动欧洲零废弃和美国通过“支持市政零废弃原则”决议[68]。截至2018年，C40城市中仅有23个城市签署了“地区迈向零废弃”承诺[69]。对于LLM赋能的智能聊天机器人行业来说仍然任重道远，在建设路径III的过程中，有必要同时加强路径I和路径II的推进，如图5（b）所示。

4. 结论和展望

ChatGPT及其他基于LLM的智能聊天机器人产品正在为多个行业提供有益的服务和见解，覆盖从社会服务（如娱乐、商业和教育）到科学与工程应用领域，如化学、数学、医学、可再生能源转型、智能制造和工业物联网（industrial internet of things, IIoT）。在拥抱这些新兴产品和服务带来的益处的同时，我们呼吁重新思考LLM赋能的智能聊天机器人全生命周期能耗与碳排放的减缓路径，并在该行业发展的早期阶段重塑其对环境和气候变化的影响。以本文提供的生命周期能耗与碳足迹分析为起点，应推动并实施诸如过程与系统效率改进等减排实践。基于改进的分析方法并结合全面的LCA与动态结果报告，加强延伸生产者责任，推动协会与管理标准的长期发展。这些措施需要立即行动，并且要求不同规模的利益相关方之间的合作与跨境协作，以建立具有广泛共识的框架。

本文的主要启示可归纳为以下四点：

（1）与LLM赋能的智能聊天机器人行业相关的整体能耗和碳足迹不容忽视，尤其是在该行业快速发展趋势下的预期能耗和碳排放更为显著。

（2）识别了该新兴行业的生命周期八个主要阶段及其相互作用关系，其中硬件制造和LLM训练两个阶段目前是能耗密集的环节。此外本文的估算数据显示，未来聊天机器人服务阶段可能成为能耗和碳足迹热点。

（3）提出了一个包含三个具体行动路径及其多个利益相关方的系统级解决方案，以改善该行业的能耗和碳排放管理。

（4）全生命周期视角和跨学科经验支持的系统级解决方案对解决该行业能源和环境影响的新问题至关重要。

社会层面一直在关注由LLM赋能的智能聊天机器人应用和服务造成的隐私安全问题[70–71]。前沿科技公司和研发机构（如OpenAI）也正持续着手解决类似问题[31]。除此以外，在工程管理领域，以下几个方面仍值得持续研究：①应始终致力于收集更加精确且具有高度时间和空间颗粒度的数据，当获得更全面的数据（涵盖中小型科技企业信息），即可在统筹考量所有生命周期阶段的基础上，开展基于生命周期评估（LCA）的动态环境核算；②鉴于相关的隐含能耗和碳排放与多个生命周期阶段密切相关，值得进一步研究碳排放和能耗的时空异质性问题（特别是可再生能源基础设施），分析结果可以明确能耗和碳足迹热点，对于指导行业发展具有关键作用；③LLM可用于支持环境、气候和可持续发展[72]，这也意味着LLM能被用来优化其自身的设计、开发和管理，进一步

利用改善后的LLM来促进其他领域的实践，从而在不同行业之间形成管理闭环；④利用云服务和工业互联来促进节能减碳，并结合IIoT和区块链等新兴信息技术改进数据的追踪、收集和管理，从而提高量化分析的可信度；⑤需要研究和设计主要利益相关方之间的相互作用以及拟定系统级解决方案内外的各种详细行动路径，同时意识到LLM赋能的智能聊天机器人行业在能耗和碳管理方面的改进还有很大的空间。

致谢

蒋鹏感谢国家自然科学基金项目（72061127004、72474149和72104164）和系统科学与企业发展研究中心项目（Xq22B04）的资助。Siming You感谢英国工程与物理科学研究委员会（EPSRC）计划项目（EP/V030515/1）的资助。李望良感谢贵州省科技支撑项目（[2019]2839）的资助。所有作者感谢期刊匿名审稿人、主编和语言编辑的建设性意见。支持本研究的所有数据均在论文中完整提供。感谢四川大学商学院工业工程与管理系张磊、杜德龙、张轶博等研究生对本文校订工作的协助。

Compliance with ethics guidelines

Peng Jiang, Christian Sonne, Wangliang Li, Fengqi You, and Siming You declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Adam D. The muse in the machine. Proc Natl Acad Sci USA 2023; 120(19): e2306000120.
[2] Grossmann I, Feinberg M, Parker DC, Christakis NA, Tetlock PE, Cunningham WA. AI and the transformation of social science research. Science 2023; 380(6650):1108–9.
[3] Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. Nature 2023;614(7947):214–6.
[4] Wang H, Li J, Wu H, Hovy E, Sun Y. Pre-trained language models and their applications. Engineering 2023;25:51–65.
[5] reachesChatGPT 100 million users two months after launch. Report. Belize City: The Guardian; 2023 Feb.
[6] Chat. openai. com traffic & engagement analysis. Report. New York City: Similarweb; 2023.
[7] AI—worldwideGenerative. Report. New York City: Statista; 2023.
[8] What's the next word in large language models? Nat Mach Intell 2023; 5(4): 331–2.
[9] Epstein Z, Hertzmann A; the Investigators of Human Creativity. Art and the science of generative AI. Science 2023;380(6650):1110–1.
[10] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023; 379(6637):1123–30.
[11] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80.
[12] White AD. The future of chemistry is language. Nat Rev Chem 2023; 7(7): 457–8.
[13] Luo B, Lau RY, Li C, Si YW. A critical review of state-of-the-art chatbot designs and applications. Wiley Interdiscip Rev Data Min Knowl Discov 2022; 12(1): e1434.
[14] McTear M. Conversational AI: dialogue systems, conversational agents, and chatbots. Heidelberg: Springer Nature; 2022.
[15] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 Mar 3–10; Canada. New York City: ACM; 2021. p. 610–23.
[16] An J, Ding W, Lin C. ChatGPT: tackle the growing carbon footprint of generative AI. Nature 2023;615(7953):586.
[17] Hellweg S, Milà i Canals L. Emerging approaches, challenges and opportunities in life cycle assessment. Science 2014;344(6188):1109–13.
[18] IEC 62890: 2020. Industrial-process measurement, control and automation life-cycle-management for systems and components. Report. Geneva: International Electrotechnical Commission; 2020.
[19] Luccioni AS, Viguier S, Ligozat AL. Estimating the carbon footprint of BLOOM, a 176b parameter language model. 2022. arXiv: 2211.02001.
[20] Eadline D. NVIDIA H100: are 550000 GPUs enough for this year? Report. San Diego: HPC Wire; 2023 Aug.
[21] NVIDIA H100 tensor core GPU. Report. Santa Clara: NVIDIA; 2023.
[22] Nagapurkar P, Das S. Economic and embodied energy analysis of integrated circuit manufacturing processes. Sustainable Comput Infor Syst 2022; 35: 100771.
[23] Torrubia J, Valero A, Valero A. Energy and carbon footprint of metals through physical allocation: implications for energy transition. Resour Conserv Recycl 2023;199:107281.
[24] Vahidi E, Zhao F. Assessing the environmental footprint of the production of rare earth metals and alloys via molten salt electrolysis. Resour Conserv Recycl 2018;139:178–87.
[25] Zhong RY, Xu X, Klotz E, Newman ST. Intelligent manufacturing in the context of Industry 4.0: a review. Engineering 2017;3(5):616–30.
[26] Shah A. Bloomberg uses 1.3 million hours of GPU time for homegrown large language model. Report. San Diego: HPC Wire; 2023 Apr.
[27] NVIDIA accused of needing 50000 H100 graphics cards to train GPT-5 for a starting price of 250000 GPT-5. Report. Warsaw: Kuai Technology; 2023.
[28] Wiboonrat M. Energy management in data centers from design operations and maintenance. In: Proceedings of 2020 International Conference and Utility Exhibition on Energy, Environment and Climate Change; 2020 Oct 20–22; Pattaya, Thailand. IEEE; 2020.
[29] Kumar I, Tyner WE, Sinha KC. Input-output life cycle environmental assessment of greenhouse gas emissions from utility scale wind energy in the United States. Energy Policy 2016;89:294–301.
[30] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst 2020; 33: 1877–901.
[31] GPT-4. Report. San Francisco: OpenAI; 2023.
[32] Masanet E, Shehabi A, Lei N, Smith S, Koomey J. Recalibrating global data center energy-use estimates. Science 2020;367(6481):984–6.
[33] Inventory of U. S. greenhouse gas emissions and sinks 1990 – 2018. Report. Washington: U.S. Environmental Protection Agency; 2020.
[34] Li P, Yang J, Islam MA, Ren S. Making AI less "thirsty": uncovering and addressing the secret water footprint of AI models. 2023. arXiv: 2304.03271.
[35] Dhar P. The carbon impact of artificial intelligence. Nat Mach Intell 2020;2(8): 423–5.
[36] Patterson D, Gonzalez J, Le Q, Liang C, Munguia LM, Rothchild D, et al. Carbon emissions and large neural network training. 2021. arXiv: 2104.10350.
[37] Hao K. Training a single AI model can emit as much carbon as five cars in their lifetimes. MITS Technol Rev 2019;75:103.
[38] Power consumption when training artificial intelligence (AI) based large language models (LLMs) in 2023. Report. New York City: Statista; 2023.
[39] Yeluri S. Large language models—the hardware connection. Report. South Brisbane: APNIC; 2023 Aug.
[40] Klemeš JJ, Jiang P, Fan YV, Bokhari A, Wang XC. COVID-19 pandemics stage II—energy and environmental impacts of vaccination. Renew Sustain Energy Rev 2021;150:111400.
[41] Zhang M, Li L. A commentary of GPT-3 in MIT technology review 2021. Fundam Res 2021;1(6):831–3.
[42] AI chatbots: energy usage of 2023's most popular chatbots (so far). Report.

Spring: TRG Datacenters; 2023.
[43] According to ChatGPT, single GPT query consumesa 1567% (15x) more energy than a Google search query. Report. Reddit; 2023.
[44] ChatGPT's energy use per query. Report. Towards Data Science; 2023.
[45] Akl J, Alladkani F, Calli B. Feature-driven next view planning for cutting path generation in robotic metal scrap recycling. IEEE Trans Autom Sci Eng. In press.
[46] Zhao W, Xu J, Fei W, Liu Z, He W, Li G. The reuse of electronic components from waste printed circuit boards: a critical review. Environ Scie Adv 2023; 2(2):196–214.
[47] Charles RG, Douglas P, Dowling M, Liversage G, Davies ML. Towards increased recovery of critical raw materials from WEEE-evaluation of CRMs at a component level and pre-processing methods for interface optimization with recovery processes. Resour Conserv Recycling 2020;161:104923.
[48] Deng B, Wang X, Luong DX, Carter RA, Wang Z, Tomson MB, et al. Rare earth elements from waste. Sci Adv 2022;8(6):eabm3132.
[49] Reck BK, Graedel TE. Challenges in metal recycling. Science 2012;337(6095): 690–5.
[50] Jiang S, Li Y, Lu Q, Hong Y, Guan D, Xiong Y, et al. Policy assessments for the carbon emission flows and sustainability of Bitcoin blockchain operation in China. Nat Commun 2021;12(1):1938.
[51] Krause MJ, Tolaymat T. Quantification of energy and carbon costs for mining cryptocurrencies. Nat Sustain 2018;1(11):711–8.
[52] Vranken H. Sustainability of bitcoin and blockchains. Curr Opin Environ Sustain 2017;28:1–9.
[53] Rillig MC, Ågerstrand M, Bi M, Gould KA, Sauerland U. Risks and benefits of large language models for the environment. Environ Sci Technol 2023; 57(9): 3464–6.
[54] Zhu JJ, Jiang J, Yang M, Ren ZJ. ChatGPT and environmental research. Environ Sci Technol 2023;57(46):17667–70.
[55] Ordieres-Meré J, Gutierrez M, Villalba-Díez J. Toward the Industry 5.0 paradigm: increasing value creation through the robust integration of humans and machines. Comput Ind 2023;150:103947.
[56] Leiserson CE, Thompson NC, Emer JS, Kuszmaul BC, Lampson BW, Sanchez D, et al. There's plenty of room at the top: what will drive computer performance after Moore's law? Science 2020;368(6495):eaam9744.
[57] Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: a survey. ACM Comput Surv 2023;56(2):1–40.
[58] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst 2022;35:27730–44.
[59] Henderson P, Hu J, Romoff J, Brunskill E, Jurafsky D, Pineau J. Towards the systematic reporting of the energy and carbon footprints of machine learning. J Mach Learn Res 2020;21(1):10039–81.
[60] Jiang P, Zhang L, You S, Fan YV, Tan RR, Klemeš JJ, et al. Blockchain technology applications in waste management: overview, challenges and opportunities. J Clean Prod 2023;421:138466.
[61] Niaz H, Shams MH, Liu JJ, You F. Mining bitcoins with carbon capture and renewable energy for carbon neutrality across states in the USA. Energy Environ Sci 2022;15(9):3551–70.
[62] Zhu D, Chen J, Shen X, Li X, Elhoseiny M. MiniGPT-4: enhancing vision language understanding with advanced large language models. 2023. arXiv: 2304.10592.
[63] Savage N. Drug discovery companies are customizing ChatGPT: here's how. Nat Biotechnol 2023;41(5):585–6.
[64] Farrauto RJ, Deeba M, Alerasool S. Gasoline automobile catalysis and its historical journey to cleaner air. Nat Catal 2019;2(7):603–13.
[65] Laskurain-Iturbe I, Arana-Landín G, Heras-Saizarbitoria I, Boiral O. How does IATF 16949 add value to ISO 9001? An empirical study. Total Qual Manage Bus Excell 2021;32(11–12):1341–58.
[66] Decarbonizing all means of transport key for sustainable growth, achieving net-zero emissions by 2050, secretary-general tells Beijing conference. Report. United Nations; 2021.
[67] Jiang P, Sonne C, You S. Dynamic carbon-neutrality assessment needed to tackle the impacts of global crises. Environ Sci Technol 2022;56(14): 9851–3.
[68] Zero waste international alliance. Report. San Diego: ZWIA; 2023.
[69] 23 global cities and regions advance towards zero waste. Report. C40 Cities; 2018.
[70] Rastogi A. Moving towards better communication. Nature Comput Sci 2023; 3(10):808–9.
[71] Topol EJ. As artificial intelligence goes multimodal, medical applications multiply. Science 2023;381(6663):eadk6139.
[72] Larosa F, Hoyas S, García-Martínez J, Conejero JA, Fuso Nerini F, Vinuesa R. Halting generative AI advancements may slow down progress in climate research. Nat Clim Chang 2023;13(6):497–9.