Engineering 40 (2024) 202–210

Contents lists available at ScienceDirect

Engineering

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

Research
Energy Systems Engineering—Perspective

Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots

Peng Jiang [a], Christian Sonne [b], Wangliang Li [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, Roskilde DK-4000, Denmark
[c] Chinese Academy of Sciences 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

Keywords:
Large language models
Intelligent chatbots
Carbon emissions
Energy and environmental footprints
Life-cycle assessment
Global cooperation

ABSTRACT

Intelligent chatbots powered by large language models (LLMs) have recently been sweeping the world, with potential for a wide variety of industrial applications. Global frontier technology companies are feverishly participating in LLM-powered chatbot design and development, providing several alternatives beyond the famous ChatGPT. However, training, fine-tuning, and updating such intelligent chatbots consume substantial amounts of electricity, resulting in significant carbon emissions. The research and development of all intelligent LLMs and software, hardware manufacturing (e.g., graphics processing units and supercomputers), related data/operations management, and material recycling supporting chatbot services are associated with carbon emissions to varying extents. Attention should therefore be paid to the entire life-cycle energy and carbon footprints of LLM-powered intelligent chatbots in both the present and future in order to mitigate their climate change impact. In this work, we clarify and highlight the energy consumption and carbon emission implications of eight main phases throughout the life cycle of the development of such intelligent chatbots. Based on a life-cycle and interaction analysis of these phases, we propose a system-level solution with three strategic pathways to optimize the management of this industry and mitigate the related footprints. While anticipating the enormous potential of this advanced technology and its products, we make an appeal for a rethinking of the mitigation pathways and strategies of the life-cycle energy usage and carbon emissions of the LLM-powered intelligent chatbot industry and a reshaping of their energy and environmental implications at this early stage of development.

1. Introduction

Intelligent chatbots powered by large language models (LLMs) have become extremely popular worldwide, especially since the end of 2022 [1–4]. Just two months following its release as an advanced generative pretrained transformer (GPT) tool, ChatGPT embraced 100 million users, with 590 million visits in January 2023 [5]. This trend increased significantly as of the end of 2023, with more than 1.5 billion visits per month [6]. LLM-powered intelligent chatbots are part of the development of generative artificial intelligence (AI). Even considering the impacts of the coronavirus disease 2019 (COVID-19) pandemic and regional wars, generative AI has been projected to maintain an annual growth rate of 24.4% from 2023 to 2030, corresponding to a market volume increase from 44.9 to 207.0 billion USD [7]. Within this frenzied competition, similar intelligent chatbots such as Google Bard, Bing Chat, My AI, LLaMA, PaLM, PaLM-E, Copilot X, Ernie, Qwen-72B, and Alpaca [8] have appeared, along with global frontier technology companies and research institutions, including Microsoft OpenAI, Google AI, Meta AI, DeepMind, Amazon, Huawei, GitHub, Anthropic, Baidu, NVIDIA, Ali Cloud, and more, resulting in a very large number of different choices regarding AI assistance. LLMs are accelerating and refining the world's development from art [1,9] to science [2,10–12].

Before ChatGPT was embedded with LLMs, conversational chatbots had already been widely applied in various sectors, including healthcare, hospitality, public service, entertainment, education,

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

https://doi.org/10.1016/j.eng.2024.04.002

and manufacturing [13]. There are three main types of conversational chatbots: rule-based chatbots, live chatbots, and basic AI-powered intelligent chatbots [13,14]. The first two types of chatbots communicate by following predetermined rules and respectively incorporate chatbot software and human conversations to provide customer service. The third type, basic AI-powered chatbots, promotes communications outside predefined commands without human interactions, laying the foundation for advanced intelligent chatbots. However, the three types of conversational chatbots require much less model training and fewer fine-tuning parameters and utilize less powerful hardware in comparison with LLM-powered intelligent chatbots, including the recently developed ChatGPT with GPT-3, GPT-3.5, and GPT-4, which feature a massive number of model parameters. Fig. 1 [15] shows the growth trend in the number of model parameters in 12 representative LLMs released from 2019 to 2021, starting from DistilBERT, with 66 million parameters, to Switch-C, with 1570 billion parameters—a remarkable increase of 23 788 times within three years.

Behind the popularity of the intelligent tools and services supported by LLMs, the related energy and environmental impacts have been concerning [3,16]. LLM-powered intelligent chatbots consume substantial quantities of electricity and require hardware such as graphics processing units (GPUs), tensor processing units (TPUs), and supercomputers [16]. From a whole-system perspective, beyond software with intelligent LLMs and hardware with GPUs, TPUs, and supercomputers, an intelligent chatbot service requires additional support such as advanced manufacturing, logistics and facility operations management, massive data collection, data center management, and waste management. Mitigation of the energy and environmental impacts of the LLM-powered intelligent chatbots industry calls for consideration of the industry's energy and resource consumption and its environmental implications throughout the entire life cycle of its development. This work identifies major energy usage and carbon emission phases throughout the entire life cycle of the development and application of LLM-powered intelligent chatbots. A system-level solution targeted at associated stakeholders is proposed to assist in optimizing the management and development of this industry. Finally, the work concludes by projecting insights into and outlooks on new research directions.

2. Life-cycle energy and carbon footprints

2.1. Identification of life-cycle phases

Based on a life-cycle assessment (LCA) approach and following a "cradle-to-grave" analysis method, the environmental impacts of intelligent chatbot development and application include those from resource extraction, production, use, and disposal [17]. According to the International Electrotechnical Commission (IEC) 62890:2020 standard [18], all partners including product producers, suppliers, service providers, and users are involved in an LCA, and design, planning, development, operations, and maintenance should be considered. Accordingly, eight main phases are expected to be relevant to the life-cycle energy usage and carbon emission management related to LLM-powered intelligent chatbots (Fig. 2). More specifically, these include chatbot research and development (R&D; phase 1); hardware manufacturing (e.g., GPUs, TPUs, supercomputers, and service devices/robots; phase 2); global commercial logistics (phase 3); facility operations and maintenance (O&M; phase 4); massive data collection and management (e.g., source data management via energy-intensive large data centers; phase 5); LLM training and fine-tuning (phase 6); online/offline chatbot services (phase 7); and hardware material recycling and waste disposal at the end (phase 8).

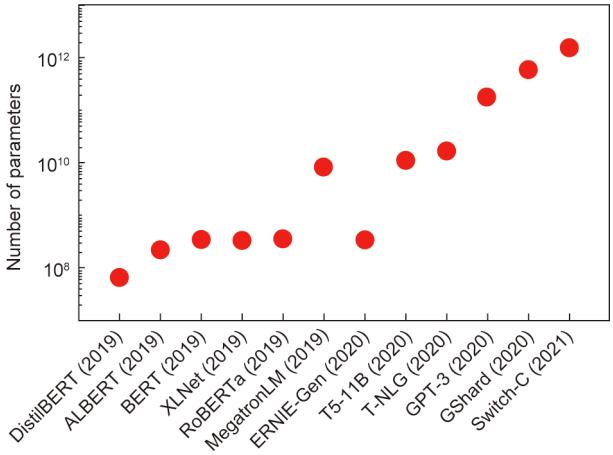

Fig. 1. Growth trend in the number of parameters in LLMs from 2019 to 2021 (based on data retrieved from Ref. [15]).

2.2. Energy and carbon footprints

Thus far, estimation of the life-cycle energy consumption and carbon emissions of the LLM-powered intelligent chatbot industry has been an open question, with limited data and information available about the overall energy consumption and carbon emissions of this industry. Moreover, embodied energy consumption and carbon emissions are closely related to global trading, LLM training, and chatbot online services, the tracking of which is challenging. Carbon footprint quantification results for the autoregressive LLM named BLOOM indicated that its embodied emissions, idle consumption-related emissions, and dynamic consumption-related emissions accounted for 22.2%, 28.9%, and 48.9% of its carbon footprint, respectively [19]. In this section, we utilize the available data to explore and analyze the possible energy and carbon footprints of the eight main phases of the life-cycle energy usage and carbon emission management of LLM-powered intelligent chatbots.

2.2.1. Phase 1: Chatbot R&D

The R&D of intelligent chatbots involves both LLMs and hardware (e.g., GPUs, TPUs, and supercomputers). Although the years

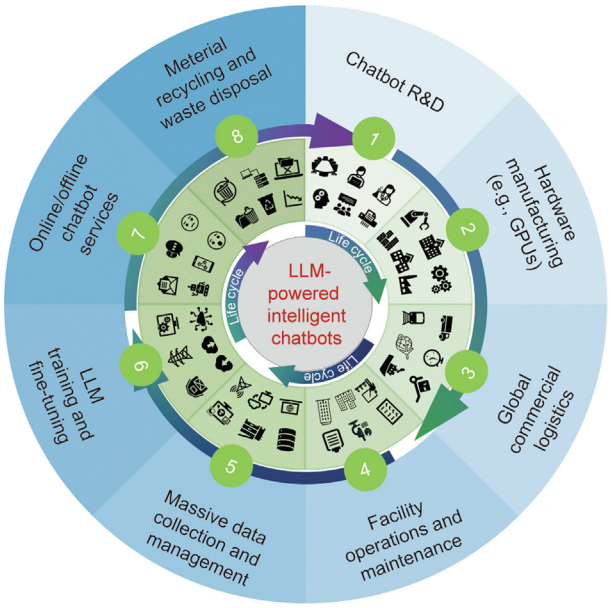

Fig. 2. An illustration of the life-cycle energy usage and carbon emission management related to LLM-powered intelligent chatbots, from research and development (R&D) and hardware manufacturing to services and waste management.

2022 and 2023 are regarded as the breakthrough period for the emergence and broad applications of ChatGPT and LLMs, the related R&D progress started much earlier. For example, OpenAI was created in 2015, and the first-generation GPT-1 with 117 million parameters was introduced in June 2018, followed by the release of GPT-2, with over one billion parameters, in February 2019. Similar trials have been witnessed for other global frontier technology companies and research institutions in the field. For a life-cycle perspective, the R&D-related energy and carbon emissions should be traced back to earlier years, as well as including the associated materials and effort inputs since then.

2.2.2. Phase 2: Hardware manufacturing

The hardware manufacturing related to LLM-powered intelligent chatbots is both resource- and energy-intensive. Hardware requirements include a significant number of high-performance GPUs, TPUs, supercomputers, large data center facilities, and offline service devices/robots. For example, due to the fast adoption and application of generative AI in the field, the company NVIDIA was expected to ship 550 000 pieces of the latest flagship H100 GPUs globally within 2023 [20]. Supercomputers with H100 NVL GPUs can elevate the performance of GPT-4 by approximately 12 times in comparison with those equipped with NVIDIA DGX A100 GPUs [21]. However, while the cutting-edge RTX 4000-series GPU can be mass-produced via the Taiwan Semiconductor Manufacturing Company Ltd. (TSMC) 5 nm process, high-performance H100 GPUs (each piece being loaded with 5120-bit memory, 80 GB of HBM3 capacity, and 14 592 Compute Unified Device Architecture (CUDA) cores) are manufactured via high-end and high-energy-consumption equipment such as that used in the TSMC 4 nm process [20].

Another aspect is the manufacturing of integrated circuits. An existing LCA study showed that the cumulative energy used in the manufacturing of integrated circuits ranged from 9 to 38 $MJ{\cdot}cm^{-2}$ [22]. Moreover, materials must be extracted for use in hardware manufacturing, and the mining and production of metals is energy-intensive, globally accounting for about 38% of industrial energy usage and approximately 15% of global electricity usage [23]. The energy and environmental footprints of rare earth metals mining and production for the manufacturing of GPUs and TPUs chips are very high [24]. In summary, for the necessary high-end equipment manufacturing, which includes various advanced and intelligent manufacturing technologies [25] and resource- and energy-intensive raw material production, the associated energy consumption and carbon emissions are non-trivial.

2.2.3. Phase 3: Global commercial logistics

The global commercial logistics associated with intelligent chatbot development and application include traditional logistics with product delivery and reverse logistics involving the waste equipment and materials associated with phase 8. Due to the spatiotemporal heterogeneity of the demand and supply of GPUs, TPUs, data servers, and offline service chatbots, a considerable amount of global shipping is involved in hardware distribution via multiple coordinated modes of transportation. For example, GPUs are essential components for intelligent chatbot development: training the lightweight BloombergGPT with 50 billion parameters and 363 billion token data required 512 NVIDIA A100 GPUs [26]; GPT-4 was trained on 10 000–25 000 A100-type high-performance GPUs, while the quantity for the coming GPT-5 could be 30 000–50 000 [27]. Existing data reveals that the specialized GPUs shipped in 2022 could consume about 9500 GW·h of electricity in a single year [16].

2.2.4. Phase 4: Facility O&M

Facility O&M are part of intelligent manufacturing and supply chain management and involve overseeing all the assets, participants, and processes related to facilities, as well as future maintenance execution covering both facility maintenance and related equipment maintenance. In the LLM-powered intelligent chatbot industry, energy consumption and carbon emissions related to O&M are present in multiple sectors, such as manufacturing facility management, smart warehouse management, and hardware equipment maintenance management such as large data centers [28]. Currently, very limited information is available on the O&M energy and carbon footprints of the LLM-powered intelligent chatbot industry, although the O&M of other industries such as solar and wind power suggested that they are likely to be significant. For example, O&M and decommissioning can account for about 10% of the overall carbon footprint for wind power generation [29].

2.2.5. Phase 5: Massive data collection and management

Massive data must be collected and managed to support LLMs and intelligent chatbot services. The multiple data sources used for training GPT-3 include the filtered Common Crawl (410 billion tokens), WebText2 (19 billion tokens), Books1 (12 billion tokens), Books2 (55 billion tokens), and Wikipedia (3 billion tokens) [30]. For GPT-4, to avoid refusing valid requests by the model, OpenAI collected a diverse dataset with labeled production data, model-produced prompts, and human red-teaming [31]. Although the overall energy usage and carbon emission information on LLM-related data management is still unclear, the potential resource requirement should be of concern. It has been shown that large data centers and their management are energy- and resource-intensive. For example, it was estimated that the total energy consumption of global data centers was 205 TW·h (i.e., approximately 1% of global electricity usage) in 2018 and that computing instances using data centers were elevated by 550% compared with those in 2010 [32]. Data centers in the United States contributed a total of 3.15×10^7 tCO_2-equivalent (CO_2-eq) greenhouse gas (GHG) emissions, making up approximately 0.5% of overall emissions in the United States in 2018 [33]. Beyond energy and carbon footprints, it was also estimated that 5.4 million litre of water was used to cool Microsoft's high-performance data servers when training GPT-3, and that 500 mL of water was consumed to respond to 10–50 queries in ChatGPT services [34].

2.2.6. Phase 6: LLM training and fine-tuning

As another key energy-intensive phase in terms of software and algorithms, training AI-driven LLMs consumes large amounts of energy and carry significant carbon footprints [3,35]. For example, training a single transformer architecture with 213 million parameters produces around 300 tCO_2-eq (Fig. 3), which is equivalent to the emissions from 125 round-trip flights between Beijing and New York City [35]. This does not account for the substantial resources used for training trials before the final training run [36]. For example, approximately 5000 additional models must be pretrained to gain a final model with better performance [37]. Even with technology development improving its energy efficiency, the final training run of GPT-3 with 175 billion parameters consumes 1287 MW·h of electricity with a carbon footprint of 552 tCO_2-eq [36,38], which does not even count the energy usage and emissions for model fine-tuning and updating (Fig. 3). This is similar to the case of Meta's LLaMA model, which needs more than 100 A100 high-performance GPUs for model fine-tuning [39]. Overall, this implies that 5000 times of pretraining (as assumed) of GPT-3 LLMs could cause up to 2.76 $MtCO_2$-eq emissions, which is nearly equivalent to the life-cycle carbon footprint of 8.36 billion COVID-19 vaccines fully covering one dose for the earth's population [40].

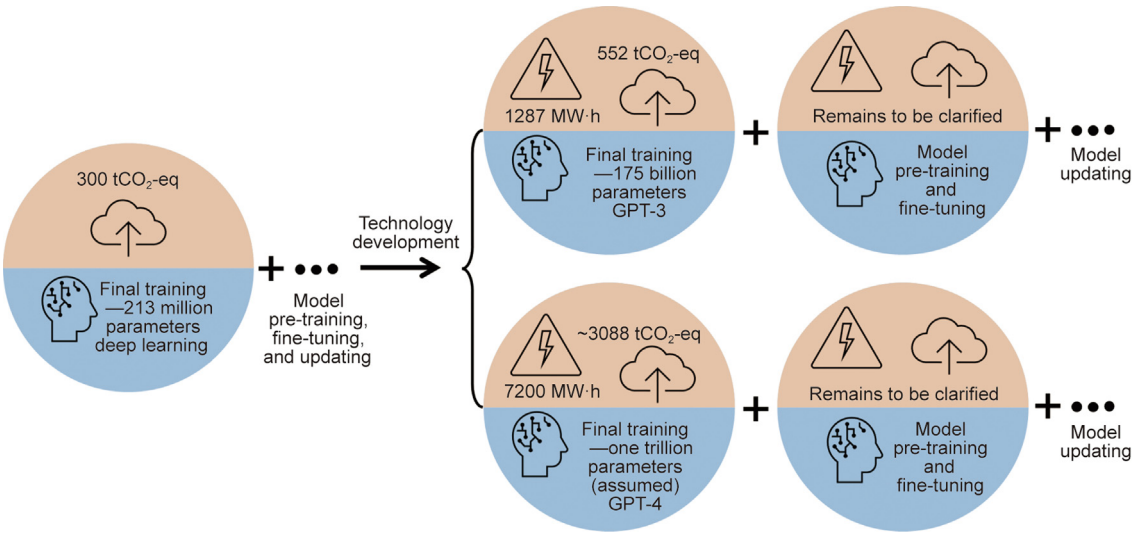

Fig. 3. Electricity consumption and carbon emissions from LLM training and fine-tuning, along with model development. New algorithms and technologies have the potential to improve the energy efficiency of LLM training, so significant research is required to clarify the additional energy consumption and carbon footprints of the computation beyond the final training run of LLMs.

In a humorous and light-hearted way, upon asking ChatGPT the question "What is the electricity consumption of training ChatGPT?" we received the response "...training the GPT-3 model could consume as much as 355 000 kW·h of energy, which is roughly equivalent to the energy consumption of an average American household for 32 years." It is expected that the energy consumption of GPT-3 from a systems perspective will be much higher when considering at least eight main phases throughout the life cycle, as shown in Fig. 2. Moreover, if the LLM training algorithms were to be significantly improved, energy consumption and emissions would skyrocket, as LLMs continue to increase in terms of model size and input data volume—for example, from GPT-1's 117 million parameters and 5 GB of data, to GPT-2's 1.2 billion parameters and 40 GB of data, to GPT-3's 175 billion parameters and 45 TB of data [41]. According to the known settings of GPT-3 and the assumed one trillion parameters of GPT-4, it was estimated that it might take approximately 26 and 150 days to train GPT-3 and GPT-4, respectively, using 10 000 V100 GPUs [42]. For the final training run of LLMs, compared with the electricity usage of GPT-3, the corresponding number for GPT-4 is 7200 MW·h [42]. This corresponds to estimated carbon emissions of about 3088 tCO_2-eq, based on the calculation for GPT-3 [36]. The data and information are compared in Fig. 3.

2.2.7. Phase 7: Online/offline chatbot services

Compared with LLM training and fine-tuning, end users are more familiar with the human–device interactions that occur as they enjoy intelligent chatbots' web services (e.g., computer and phone) or offline service devices such as robots. Based on the trained and fine-tuned LLMs hosted in the cloud end, rapid inference and response in the process of feedbacking to users' prompts consume energy and incur emissions. For each new prompt, the inference requires a series of GPUs/TPUs to run the built LLMs and implement a rapid computation. For example, the inference, fine-tuning, and training processes of Meta's LLaMA model typically use 16, 100+, and 2000 A100 high-performance GPUs, respectively [39].

Although the energy usage and emissions related to a single inference are not significant, the long-term accumulative carbon emissions for global chatbot online/offline chatbot services are obvious and non-trivial. The related energy and carbon footprints per month for online services can be comparable to those from the final training run of LLMs described in phase 6. Based on assumptions and estimated data from the TRG Datacenters [42] and the 590 million visits in January 2023 [5], the overall monthly electricity consumption of ChatGPT services under nine different scenarios was estimated and is shown in Fig. 4. The estimation indicates that five scenarios and eight scenarios have electricity consumption higher than that of the final training runs of GPT-4 and GPT-3, respectively. The worst-case scenario (i.e., scenario 3) incurs 23 364 MW·h of electricity usage per month. If based on the peak data of an average of 1.5 billion visits per month from August to October 2023 [6], the annual electricity consumption of ChatGPT services via Google search under the worst case would be up to 7.128 TW·h.

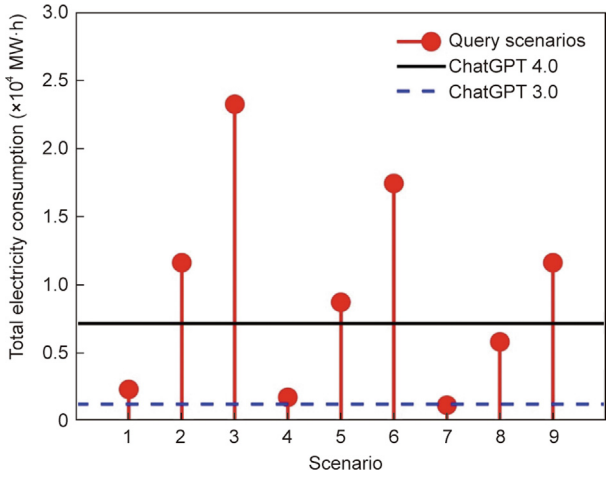

Fig. 4. Total electricity consumption of monthly queries related to overall Google search-based ChatGPT services under different scenarios. Scenarios 1, 4, and 7 are based on one query per visit; scenarios 2, 5, and 8 are based on five queries/visit; and scenarios 3, 6, and 9 are based on ten queries/visit; simultaneously, scenarios 1–3 denote low efficiency with an electricity usage per query of 0.00396 kW·h; scenarios 4–6 denote medium efficiency with 0.00297 kW·h per query; and scenarios 7–9 denote high efficiency with 0.00198 kW·h per query [42]. Such figures on electricity usage per query are similar to those reported in Refs. [43,44]. The estimated electricity consumption of the final training run of the GPT-3 and GPT-4 models is shown in the figure for comparison. ChatGPT 4.0: ChatGPT with GPT-4; ChatGPT 3.0: ChatGPT with GPT-3.

2.2.8. Phase 8: Material recycling and waste disposal

In the LLM-powered intelligent chatbot industry, material recycling and disposal mainly occur in the end-of-life handling of hardware equipment, as stated in phase 2. Effective material recycling in this industry has great relevance to the development of an intelligent circular economy concept. The recycling and reuse of robotic materials [45] and electronic components from printed circuit boards [46] hold global influence to mitigate economic costs and energy and environmental footprints. Moreover, the heavy energy footprints of rare earth metals mining and production [24] suggest that material recovery and recycling—especially for critical raw metals—are crucial pathways toward resource circularity and can lead to substantial indirect energy and carbon savings. However, the fact remains that the resource recovery efficiency of critical rare materials is low, and relevant policy intervention is still premature [47]. This situation calls for extensive R&D in product design, social behavior analysis, the thermodynamics of separation, and recycling technologies to overcome the issues of low extractability and lengthy purifications [48,49]. Despite the rapid development of the LLM-powered intelligent chatbot industry, the overall energy and carbon footprints can be mitigated if materials can be effectively recovered and recycled.

2.3. How the different phases interrelate and influence each other?

The visualized relationships and interactions among the different life-cycle phases are shown in Fig. 5(a), which presents a directed network structure with the eight phases. An analysis of these complex interactions requires knowledge of the systems-engineering discipline and must take several inherent factors into consideration.

First, within the eight phases under analysis, there is a wide range of participants worldwide, including R&D teams (phase 1), manufacturing workers (phase 2), logistics staff and e-commerce operators (phase 3), facility management workers (phase 4), data engineers (phase 5), algorithm engineers (phase 6), software engineers and operators (phase 7), and recycling workers, disposal workers, and logistics staff (phase 8). This implies the relevance of substantial energy and environmental footprints due to participants' activities and interactions. The possible information flows and connections among the participants within such an interrelated system suggest the importance of synergizing their activities toward optimal overall energy and environmental performance.

Second, the arrows in Fig. 5(a) represent the interactions among different phases. For example, the interactions between phase 1 and phases 2 and 5–7 indicate that the R&D ① should align with existing intelligent manufacturing technologies, ② can improve the data requirements and data management, ③ is key in elevating the efficiency of LLM training and fine-tuning, and ④ always aims to enhance the service quality of intelligent chatbots, which indirectly affects the overall energy consumption and carbon emissions. Hardware manufacturing in phase 2 has been identified as a resource- and energy-intensive phase that interrelates with all the other phases. Phases 3–6 and their possible interactions with other phases were preliminarily discussed in Section 2.2. The online/offline chatbot services in phase 7 involve the collection of new data from global users, which will be managed in data centers in phase 5 and will then be further leveraged for LLM training and fine-tuning in phase 6. For example, OpenAI fine-tuned GPT-4 with the feedback data from GPT-3 users and 50 AI experts that were stored in its data centers [31]. In phase 8, the recovery, recycling, and reuse of materials and electronic equipment serve to reduce the manufacturing activities in phase 2 and the global commercial logistics in phase 3, indirectly decreasing the energy and carbon footprints of the whole system via these two phases.

3. The way forward: A system-level solution to optimize management

Based on the analysis of the life-cycle energy and carbon footprints of the LLM-powered intelligent chatbot industry (Section 2.2) and the interactions among various phases (Section 2.3), a system-level solution with few pathways to mitigate the footprints is proposed and illustrated in this section.

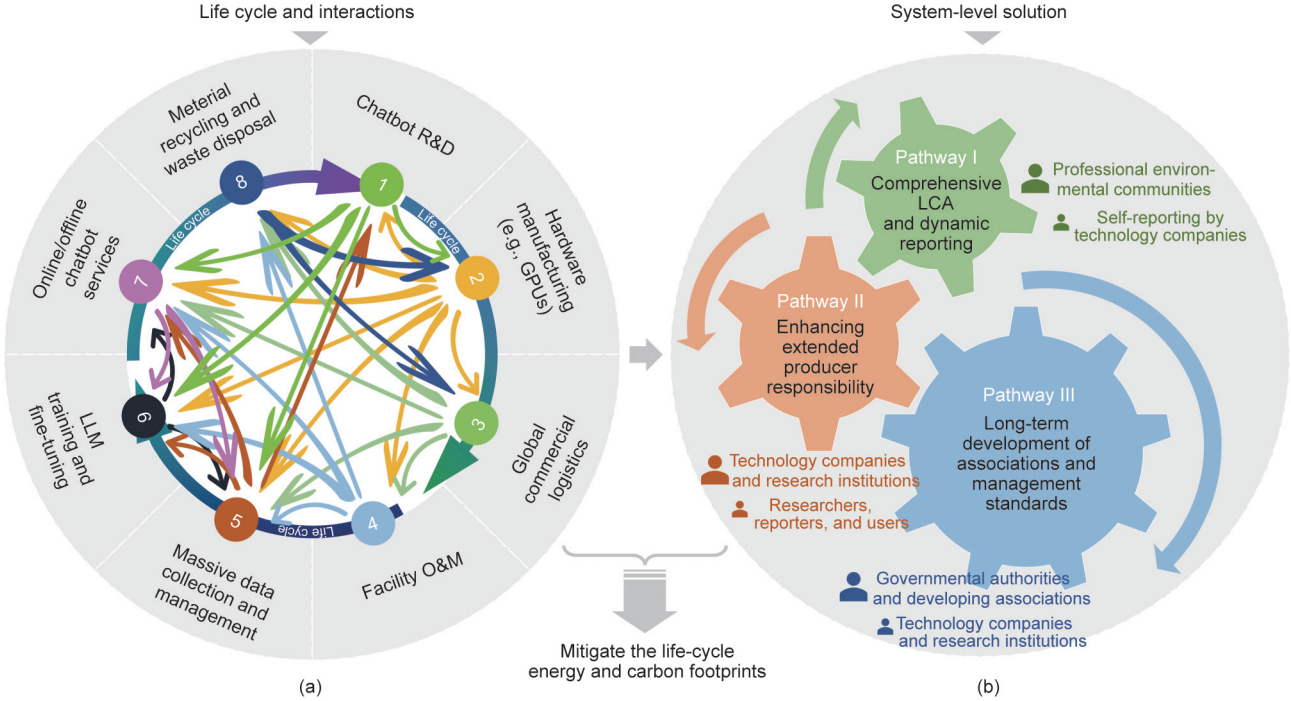

Fig. 5. Pathways to mitigate the life-cycle energy and carbon footprints of the LLM-powered intelligent chatbot industry based on the analysis in Section 2. (a) Relationships and interactions among the different life-cycle phases. Each arrow denotes how a phase (i.e., a source) interrelates and/or influences another phase (i.e., a sink). (b) Diagram of the system-level solution for optimizing the management of the LLM-powered intelligent chatbot industry based on the life-cycle energy usage and carbon emission analysis.

Similar to LLM-powered intelligent chatbots, there is another emerging technology that requires the support of hardware such as GPUs, TPUs, supercomputers, and large data servers, as well as software: namely, blockchain technology-based cryptocurrency mining, which has significant energy consumption and carbon emissions [50,51]. For example, under a scenario of no policy interventions, the carbon emissions and energy consumption of Bitcoin mining per year in China are estimated to peak in 2024 at 130.5 $MtCO_2$-eq and 296.6 TW·h, respectively [50]. However, if the transaction fees are insufficient to incentivize cryptocurrency mining based on blockchain, the energy consumption of mining activities would peak in 2140, due to the number limit of Bitcoins [52]. In contrast, there are limited mechanisms as yet to regulate the development and global competition of LLM-powered intelligent chatbots toward energy and environmental sustainability.

It is believed that LLM-powered intelligent chatbots will become extremely common and will profoundly transform the world by incorporating multimodal data-processing capabilities [8,53,54]. However, while anticipating the enormous potential of this advanced technology, concrete solutions are needed to optimize the management of this industry and mitigate its energy and environmental footprints through dedicated action pathways and collaboration among wider communities and stakeholders. A few preliminary ideas are under development. For example, a conceptual framework based on employing energy-efficient hardware driven by green energy such as solar, hydrogen, and wind energy and tailoring the model structure of LLMs has been initially suggested [16]. These ideas require key breakthroughs in technology development, such as ① the revolution of hardware manufacturing technologies under the popular Industry 4.0 [25] and human-centric Industry 5.0 [55]; ② the enhancement of energy-efficient computing power boosted by algorithms, software, and hardware architecture improvements [36,56]; ③ the optimization of LLM performance dynamics [57]; and ④ the technical utilization of the assistance of human feedback [58].

Accordingly, beyond technological innovation, we believe that there is a need for substantial global collaboration efforts among the multiple stakeholders in this industry. According to the analysis in Section 2, the main stakeholders in the energy and environmental management in this industry include governmental authorities/agencies, newly developing associations, professional environmental communities, technology companies and research institutions, technology reporters, and worldwide chatbot users. Based on the life-cycle and interaction analysis shown in Fig. 5(a) and considering that the related improvement involves scientific problems and mechanism exploration including multi-stakeholder influence/interaction, multi-objective optimization, and multi-agent decision-making, we propose the system-level solution shown in Fig. 5(b), with three strategic pathways that cover all these main stakeholders. The specific stakeholders that are well-positioned for the implementation of each strategic pathway are explained below.

3.1. Pathway I: Comprehensive LCA and dynamic reporting

In pathway I, the relevant stakeholders are mainly professional environmental communities, including authoritative environment organizations and institutions, non-governmental institutions, researchers in the environmental field, and environmentalists with professional backgrounds. The assessment of the cumulative energy demand and carbon emissions related to LLM-powered intelligent chatbots should cover phases 1 to 8, as outlined in Fig. 2. Such assessments should adopt a whole-system approach (e.g., a systematic reporting approach for the carbon and energy footprints of machine learning [59]), and associated reports from authoritative environment organizations and institutions, such as the United Nations Environment Programme (UNEP), should be constantly promoted to facilitate the accurate characterization of related environmental issues, in line with the United Nation (UN)'s sustainable development goals (SDGs). Based on a life-cycle perspective, an analysis of the various scenarios and dynamic trends of intelligent chatbot development and application should be carried out in order to project consequences and impacts. The estimated results shown in Figs. 3 and 4 and the life cycle and interactions in Fig. 5(a) may suggest an overall trend of the LLM-related energy footprint in the future, which could be significantly larger than that of the LLM training and fine-tuning phase. Dynamic reporting that accounts for changes in the different phases, assurance of data quality and accuracy, and related implications will have the potential to facilitate effective policy formulation and action planning to optimize the development of the intelligent chatbot industry from a coherent socio-economic, energy, and environmental perspective.

However, it is worth noting that the limited data reliability and transparency in various commercial situations are barriers that must be tackled properly before comprehensive assessment and dynamic reporting can be possible. For example, there are no existing portals and there is limited motivation for private organizations such as OpenAI to share and report their latest relevant data, due to concerns about commercial secrets and intellectual property. This is part of the reason why the energy and carbon footprints related to GPT-4 are still secrets. From the perspective of information disclosure, this barrier could be addressed by enhancing producer responsibility (pathway II) and developing relevant associations and standards (pathway III), as shown in Fig. 5(b). Moreover, from the perspective of data management, the integration of advanced emerging technologies, such as blockchain, Internet of Things (IoTs), and machine learning models, could serve to improve data reliability and transparency for accurate environmental accounting [60].

3.2. Pathway II: Enhancing extended producer responsibility (EPR) with incentives

The lead stakeholders in pathway II are mainly frontier technology companies and research institutions, especially those who are leading the R&D revolution of the intelligent chatbot industry, while other participating stakeholders, including small and medium-sized technology enterprises, environment-community researchers, technology reporters, and global chatbot users, also contribute to this pathway. The concept of EPR stresses that all life-cycle environmental impacts should be counted for a product.

First, under an ideal situation, intelligent chatbot designers should understand and choose greener pathways to train and fine-tune LLMs. For example, rather than constructing its whole loop with the involvement of at least phases 3–6 and 8, Bloomberg assigned approximately 1.3 million hours of training based on distributed shared resources from Amazon's cloud services [26]. This route is especially practical for small- and medium-technology enterprises designing LLM-based platforms and applications.

Second, suitable incentive mechanisms play a key role in the implementation of EPR. For the end-of-life intelligent chatbot management in phase 8, supervision of EPR can incentivize the best practices in material recycling management. This in turn helps to simultaneously reduce the end-of-life costs and energy and carbon footprints, considering the strategic importance of the macro governance of the production and recovery of key rare metals [47]. There are also benefits in incentivizing green O&M management in phase 4, sustainable data center management in phase 5, and energy-efficient LLM training and fine-tuning in phase 6.

Third, in an ideal case of implementing EPR in the LLM-powered intelligent chatbot industry, the proactive reporting of carbon

footprints, energy consumption, and resource usage could assist in avoiding the issues of lag management that have been experienced in Bitcoin mining [61]. For example, various open-source alternatives to Bard and GPT-4 are providing technical reports and/or codes for building LLMs and systems and estimating their energy and carbon footprints; these include LLaMA and LLaMA-2 by Meta, BLOOM by BigScience, Qwen-14B and Qwen-72B by Ali Cloud, Alpaca by Stanford University, and ChatGLM-6B by Tsinghua University. However, due to the absence of clear policy legislation and incentives, the outcome may end up as a non-cooperative game, in which leading technology companies and research institutions have insufficient motivations to continuously implement EPR. This issue could not be solved easily in a top-down manner (i.e., by leading companies and institutions to consciously and proactively improve the industry) before the appearance of a major non-profit organization or association (as discussed in pathway III, later). On the other hand, it is likely that the issue can be addressed via a bottom-up approach—that is, through revolutionary actions from niche-market companies, emission facts revealed by environment-community researchers, and the spontaneous voice of global users. All these actions from the "bottom" can emerge as motivations for the leaders of R&D technology companies and research institutions. Concrete options for the three kinds of participating stakeholders are provided as follows:

(1) From the perspective of a niche market, for small- and medium-technology enterprises that act as competitors at the grassroots level, relatively lightweight intelligent chatbots like mini-GPT-4 [62] and Alpaca [8] can be designed and made to be open-access sources. Such tools are still useful for some users without the requirements of advanced ChatGPT applications. Lightweight intelligent chatbots are also applicable in various industries/sectors with the addition of domain-specific knowledge. For example, the Bloomberg company developed the lightweight BloombergGPT with many fewer parameters and data requirements for financial domain services [26]. Another example is Socratic by Google for the education domain, which consumes only 50–300 MW·h of electricity for model training [42]. Additional examples in the drug discovery domain include small ChatGPT variants with high accuracy in medical applications [63].

(2) From the perspective of environment-community researchers, the costs and benefits of LLMs and ChatGPT have previously been discussed [53,54]; however, further in-depth and quantitative discussions on the pros and cons of LLM-powered intelligent chatbots are desirable based on the reliable and standardized presentation of data. It is expected that significant energy consumption and carbon emissions will come from several phases such as hardware manufacturing (phase 2), LLM training and fine-tuning (phase 6), and online/offline chatbot services (phase 7). Researchers should work with relevant stakeholders, such as non-profit organizations and authoritative environmental institutions, to accurately map emission hotspots and explore possible means for data-sharing partnerships and the proposal of mitigation measures.

(3) From the perspective of technology reporters and end users (especially environmentalists), feedback on not only technology development and user experiences but also concerns about energy and environmental issues related to intelligent chatbots has the potential to stimulate the responsible development of intelligent chatbot enterprises. For example, if many reporters and users were to speak out via traditional media platforms (e.g., newspapers and magazines) and social media platforms (e.g., Twitter, Facebook, and TikTok), respectively, the spontaneous voice of media and users could emerge as a kind of mechanism to stimulate and incentivize improvements in technology companies and research institutions in phases 1 and 2. The public voice could also strongly support the incentivization of green O&M management in phase 4 and end-of-life intelligent chatbot management in phase 8.

3.3. Pathway III: Long-term development of associations and management standards

The stakeholders of pathway III are mainly governmental authorities/agencies and newly developing associations that set standards for the industry, on the one hand, and intelligent chatbot R&D companies and research institutions that need to be regulated and incentivized, on the other. In the long run, for the sustainable development of this industry, it is essential to develop consensus and collaboration by means of the formulation of relevant associations and management standards among technology companies, academics, organizations, and agencies. This can be in the form of the development of relevant international non-profit organizations, similar to the International Federation of Robotics (IFR), the International Air Transport Association (IATA), the International Automotive Task Force (IATF), and so forth. The action route of the automotive industry could be taken as a reference to define a specific route for the LLM-powered intelligent chatbot industry in which at least three-dimensional constraints are leveraged to regulate and support the development of the industry. These constraints could include: ① the implementation of emission legislation in different regions, such as the Euro VI, China VI, and American standards (i.e., LEVIII and US Environmental Protection Agency (EPA) Tier3) [64]; ② developing IATF 16949:2016 based on International Organization for Standardization (ISO) 9001 to increase operational efficiency and reduce operating costs and waste in the supply chain [65]; and ③ the UN pushing for net-zero emissions transport by 2050 to further boost the economy and reduce inequality [66]. Compared with IATA's ambitious plan to achieve net-zero carbon emissions by 2050, how the industry of LLM-powered intelligent chatbots will respond to the carbon neutrality ambition remains uncertain, especially considering the possible impacts of ongoing and future global crises [67].

Transboundary collaboration involving different stakeholders at different scales is needed to develop relevant associations and management standards as massive capital enters the intelligent chatbot industry and promising expectations spawn a new era. However, it must be acknowledged that it will take substantial effort and time to establish well-functioning associations in this industry. For example, after the foundation of the first association under a zero-waste concept, namely, the Zero Waste International Alliance (ZWIA) in 2002, it took about nine years and about 13 years, respectively, to promote Zero Waste Europe and the resolution "Support of Municipal Zero Waste Principles" in the United States [68]. As of 2018, only 23 global cities among the C40 Cities had signed the commitment "Regions Advance Towards Zero Waste" [69]. For the LLM-powered intelligent chatbot industry, there is still a long way to go, and it is necessary to simultaneously strengthen pathways I and II during the construction process of pathway III, as shown in Fig. 5(b).

4. Conclusions and outlook

ChatGPT and other LLM-powered intelligent chatbot products are providing beneficial services and insights in various industries ranging from social services (e.g., entertainment, business, and education) to science and engineering applications in fields such as chemistry, mathematics, medicine, renewable energy transition, intelligent manufacturing, and the Industrial IoTs (IIoT). While embracing the benefits of these emerging products and services, we call for a rethinking of mitigation pathways for the life-cycle energy usage and carbon emissions of LLM-powered intelligent chatbots and a reshaping of the development of their environmental and climate change implications at this early stage of intelligent chatbot industry development. Taking the life-cycle energy and

carbon footprint analysis provided here as a starting point, mitigation practices such as process and system efficiency improvement should be promoted and implemented, based on improved analyses incorporating comprehensive LCA and dynamic result reporting, the enhancement of EPR, and the long-term development of associations and management standards. These require immediate action and collaborative efforts among the different stakeholders at different scales and transboundary cooperation for the development of a framework with wide consensus and benefits.

The key takeaways of this work can be summarized in four points:

(1) The energy and carbon footprints related to the LLM-powered intelligent chatbot industry are non-trivial, let alone those projected from the industry's rapid development trend.

(2) Eight main life-cycle phases and their interactions are identified for this emerging industry, among which two phases (i.e., hardware manufacturing and LLM training) are expected to be the most energy-intensive. In addition, the estimated data imply that the chatbot service phase could emerge as a new footprint hotspot in the future.

(3) A system-level solution comprising three specific pathways with associated stakeholders is proposed to optimize the management of the industry.

(4) A whole-system life-cycle view and a system-level solution supported by interdisciplinary experiences are crucial for understanding and tackling the emerging energy and environmental issues of this industry.

Society has been concerned about the privacy and security related to LLM-powered intelligent chatbot applications and services [70,71]. Such issues are being addressed by leading technology companies and institutions such as OpenAI [31]. Future research can focus on several aspects, as follows: ① Efforts should always be made to develop data that are more accurate and highly granular timewise and geographically. When more useful data are available, including information on long-tail marketing with small- and medium-sized technology enterprises, dynamic environmental accounting based on LCA, for example, should be carried out while considering all the relevant life-cycle phases. ② Issues related to the spatiotemporal heterogeneity of emissions and energy usage (especially the infrastructure with renewable energy sources) are worth further investigation, considering that the associated embodied energy consumption and emissions are closely related to several life-cycle phases. The results of the analysis can clarify the hotspots in energy and carbon footprints and are key to guiding the industry's development. ③ LLMs can be applied to support the environment, climate, and sustainability development [72]. This might include the application of LLMs to optimize the design, development, and management of LLMs themselves, further contributing to practical applications and thereby forming a closed loop among different industries. ④ Cloud services and the industrial internet could be utilized to enhance energy and carbon conservation, and emerging technologies such as the IIoT and blockchain could be incorporated to improve the tracking, collection, and management of data toward greater credibility of analysis. ⑤ The interactions among the major stakeholders and the various detailed pathways within or beyond the proposed system-level solution should be investigated and designed, acknowledging that there is a long way to go to improve the energy and carbon management of the LLM-powered intelligent chatbot industry.

Acknowledgments

Peng Jiang was supported by the National Natural Science Foundation of China (72061127004 and 72104164) and the System Science and Enterprise Development Research Center (Xq22B04). Siming You would like to acknowledge the financial support from the Engineering and Physical Sciences Research Council (EPSRC) Programme (EP/V030515/1). Wangliang Li would like to acknowledge the financial support from the Science and Technology Support Project of Guizhou Province ([2019]2839). We would like to thank editors and anonymous reviewers their constructive comments. All data supporting this study are provided in full in the paper.

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] ChatGPT reaches 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] Generative AI—worldwide. 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, a single GPT query consumes 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.