nature computational science

Brief Communication

https://doi.org/10.1038/s43588-024-00712-6

E-waste challenges of generative artificial intelligence

Received: 22 February 2024

Accepted: 24 September 2024

Published online: 28 October 2024

Check for updates

Peng Wang[1,2,5], Ling-Yu Zhang[1,5], Asaf Tzachor[3,4] & Wei-Qiang Chen[1,2]

Generative artificial intelligence (GAI) requires substantial computational resources for model training and inference, but the electronic-waste (e-waste) implications of GAI and its management strategies remain underexplored. Here we introduce a computational power-driven material flow analysis framework to quantify and explore ways of managing the e-waste generated by GAI, with a particular focus on large language models. Our findings indicate that this e-waste stream could increase, potentially reaching a total accumulation of 1.2–5.0 million tons during 2020–2030, under different future GAI development settings. This may be intensified in the context of geopolitical restrictions on semiconductor imports and the rapid server turnover for operational cost savings. Meanwhile, we show that the implementation of circular economy strategies along the GAI value chain could reduce e-waste generation by 16–86%. This underscores the importance of proactive e-waste management in the face of advancing GAI technologies.

Generative artificial intelligence (GAI) represents a pivotal advancement in the field of artificial intelligence (AI) by generating text, images, videos or other content types on the basis of input prompts[1]. Large language models (LLMs), a form of GAI that leverages natural language processing, are often trained on vast datasets and can be fine-tuned to offer expert-level insights in specialized fields[2,3]. However, LLMs demand considerable computational resources for training and inference, which require extensive computing hardware and infrastructure[4]. This necessity raises critical sustainability issues, including the energy consumption and carbon footprint associated with these operations[5,6]. The development of LLMs such as GPT-4 and DeBERTa, along with GAI applications in image and video generation such as Sora, highlights the growing trend of global hardware expansion, and emphasizes the timeliness and importance of sustainable computing.

Previous studies on sustainable computing have primarily focused on the energy use and carbon emissions of AI models[1,5–7]. However, the physical materials involved in their life cycle, and the waste stream of obsolete electronic equipment—known as electronic waste (e-waste)—have received less attention. For example, the weight of Nvidia's latest Blackwell platform (designed for intensive LLM inference, training and data processing tasks) stands at around 1.36 tons (~3,000 pounds), positioning GAI as a substantial material-intensive sector. Additionally, predictions indicate that AI's installed computational capacity could increase approximately 500-fold from 2020 to 2030[8]. This rapid growth in hardware installations, driven by swift advancements in chip technology, may result in a substantial increase in e-waste and the consequent environmental and health impacts during its final treatment[9,10]. In light of this, the International Energy Agency[11] and some leading tech companies have noted the importance of circular economy strategies, focusing on reducing, reusing, repairing and recycling obsolete equipment from data centers (Supplementary Table 4). Despite this recognition, there remains a lack of thorough quantification method and analysis of these strategies.

In response, we introduce a computational power-driven material flow analysis framework (Fig. 1a) designed to quantify the inflow, in-use (operating) stock and end-of-service (EoS) volume of GAI servers in data

[1]Key Lab of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]School of Sustainability, Reichman University, Herzliya, Israel. [4]Centre for the Study of Existential Risk, University of Cambridge, Cambridge, UK. [5]These authors contributed equally: Peng Wang, Ling-Yu Zhang. e-mail: pwang@iue.ac.cn; atzachor@runi.ac.il; wqchen@iue.ac.cn

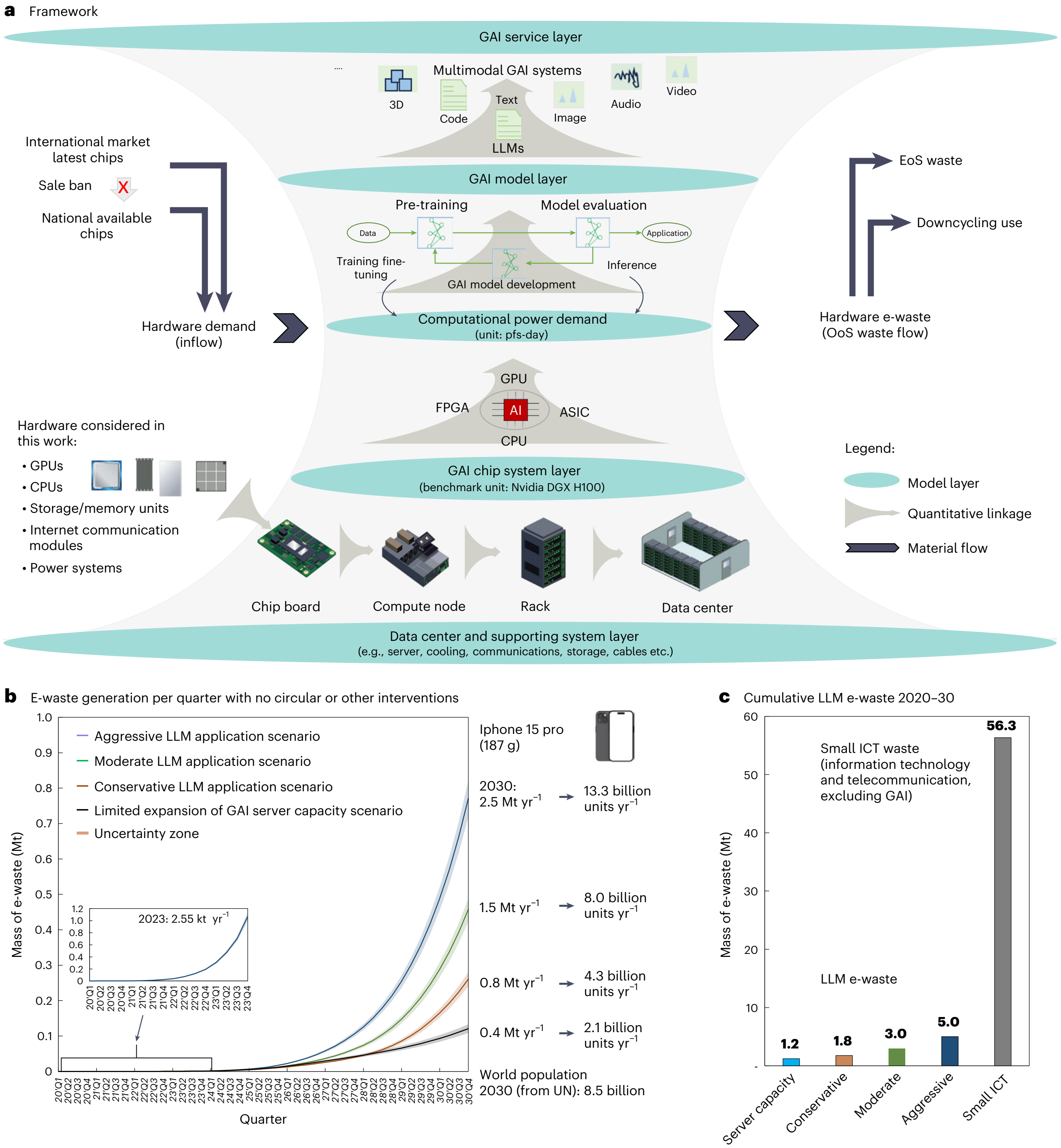

Fig. 1 | Hierarchical framework of our computational power-driven material flow analysis model and the corresponding scenario results regarding LLM-related waste generation without interventions. a, Framework containing five layers to link the public GAI service needs with final in-use hardware demand through the service layer, model layer, computational power layer, chip system layer and data center layer. Meanwhile, different strategies can be applied between two linked layers to reduce hardware demand while providing higher service, such as GAI model innovations and high-performance chip innovations. FPGA, field-programmable gate array; ASIC, application-specific integrated circuit. **b**,**c**, The e-waste generation per quarter when no further treatment is applied (**b**) and the cumulative amount from different LLM development scenarios during 2020–2030 when no further treatment such as refurbishment or lifespan extension is considered (**c**).

centers that support the computational needs of LLMs. Our quarterly assessment spans quarterly from 2020 to 2030 under different future GAI scenario settings. The analysis focuses on AI servers that include graphics processing units (GPUs), central processing units (CPUs), storage, memory units, internet communication modules and power systems. Ancillary machinery such as cooling and communication units is excluded from this study.

As illustrated in Fig. 1a, we link the demand for computational power from LLMs to a widely adopted benchmark server—specifically, the most commonly used eight-unit GPU server, the Nvidia DGX H100 system, introduced in 2023. This server serves as an initial proxy to determine the conversion factor of computational power to physical infrastructure components. Anticipating future advancements in digital electronics, we quantify this benchmark server's development by modeling the computational power per server as exponentially increasing according to Moore's law.

Our study aims not to precisely forecast the quantity of AI servers and their associated e-waste, but rather to provide initial gross estimates that highlight the potential scales of the forthcoming challenge, and to explore potential circular economy solutions. To this end, we develop four future scenarios: (1) limited expansion of GAI chip and server manufacturing based on historical trend (~41% during 2022–2023), and three scenarios based on varying levels of global application of LLMs—(2) an aggressive scenario (widespread applications with compound annual growth rate—CAGR—of computer power demand 136%), (3) a moderate scenario (limited applications with a CAGR of 115%) and (4) a conservative scenario (specific applications with CAGR of 85%). Additionally, we conducted Monte Carlo simulations to present an uncertainty analysis of our findings.

Figure 1b depicts the global potential increase in the studied e-waste generation on a quarterly basis under the four scenarios. Our results indicate potential for rapid growth of e-waste from 2.6 thousand tons (kt) yr^{-1} in 2023 to around 0.4–2.5 million tons (Mt) yr^{-1} in 2030, when no waste reduction measures are considered. For context, this total annual mass would be equivalent to discarding 2.1–13.3 billion units of the iPhone 15 Pro (187 g per unit, Fig. 1b) in 2030, which translates to 0.2–1.6 units for every person on the planet that year. Specifically, these values refer to LLM-related waste defined as EoS e-waste generation in our model, without further treatments such as refurbishing or lifespan extension. As shown in Fig. 1c, the results indicate that cumulatively the untreated EoS e-waste stream from designated data centers during 2023–2030 could total 5.0 Mt, 3.0 Mt, 1.8 Mt and 1.2 Mt in the aggressive, moderate, conservative and limited server manufacturing capacity growth scenarios, respectively.

For context, the most recent Global E-waste Monitor report indicates that annual e-waste related to small information technology equipment such as personal computers totaled 4.6 Mt in 2022, and will sum up to 43.2 Mt by 2030, meaning that AI servers could increase this quantity by 3% to 12% (Fig. 1c). In our scenarios, the CAGR of LLM-related e-waste mass ranges from 129% to 167% during 2023–2030, compared with 3.6% for global conventional e-waste mass growth[9]. Given that AI data centers are highly geographically clustered, these untreated waste streams would be mainly located in Europe (14%), North America (58%) and East Asia (25%) (Fig. 2b). This calls for stringent regulation and careful monitoring of e-waste from data center operations in those regions.

Figure 2 presents the effects of potential circular economy strategies and other related factors on the cumulative out-of-system (OoS) e-waste flow. Among these strategies, three levers are developed, each targeting different life-cycle stages of servers. The first lever (C1) examines the effects of immediate upgrades to the latest servers to improve the performance of data centers. The second lever (C2) examines lifespan extension via improved maintenance in the use phase. The third lever (C3) explores key module reuse in the (re)manufacturing phase. To assess the maximum effectiveness of each proposed strategy, we use the aggressive LLM proliferation scenario as our baseline. These strategies aim to mitigate the environmental impacts of OoS e-waste by incorporating circular processes at the upstream EoS waste stage, as illustrated in Fig. 2a. Accordingly, we explored two additional levers regarding early chip and AI model design stages: increasing computing efficiency (C4)[12] and introducing an advanced computing algorithm, such as sparsity[13] (C5).

We find that different circular economy strategies have varied impacts on the cumulative OoS e-waste between 2023 and 2030 (Fig. 2c). The most effective strategy is lifespan extension (C2): around 3.1 Mt (62%) of obsolete AI servers can be avoided if an extra 1 yr downcycling usage is applied. Similarly, the module reuse strategy (C3) reduces e-waste by 42% (2.1 Mt). This measure refers to dismantling, renovation and reassembly of obsolete critical modules (GPU, CPU, battery and so on), so that they can be reused in downcycled computing. These strategies (C2 and C3) are expected to have positive results, while the presumed effect of immediate upgrading (C1) is potentially countereffective (with 2.3 Mt more cumulative e-waste), despite reducing the need for operating servers by 15% when compared with the baseline. This scenario indicates the purchase and upgrade of new servers in data centers when the latest version is available, rather than retirement on a fixed-period schedule.

Our findings reveal that stakeholder involvement in GAI model and chip innovation substantially influences e-waste reduction (Fig. 2b). Innovations in GAI models can decrease the demand for computational power to deliver the same GAI service (Fig. 2a). For example, introducing 2× sparsity in strategy C4 can cut expected e-waste generation by 50% (2.5 Mt). Similarly, improvements in chip computing efficiency (C5) lead to a 16% reduction in e-waste (0.8 Mt). It is important to note that these strategies focusing on chip and model optimization could also trigger a rebound effect—increased service demand leading to higher server-related e-waste. Implementing strategies C2 through C5 could reduce e-waste by 86% when compared with the baseline scenario. Therefore, coordinated efforts across the entire GAI value chain—including model development, chip manufacturing, data center operations and waste management—are crucial.

Considering the uneven development of data centers in the global GAI industry and existing semiconductor export bans, we further investigate how potential technical barriers can impact the scale of OoS e-waste flows. Currently, major chip suppliers such as the United States have restricted the sale of advanced GPUs to certain countries, including China[14]. As a result, data centers there are forced to use outdated server models. We have developed several geopolitical scenarios to assess the impact of these restrictions: (T1) countries subject to bans without taking any countermeasures, (T2) countries subject to bans that achieve technological breakthroughs within 10 yr (ref. 14) and (T3) banned countries implementing circular economy strategies C2–C5.

Our results indicate that technical barriers can impact e-waste management, although the severity depends on circularity practices adopted by countries subject to bans. In the absence of trade restrictions, data centers worldwide can freely purchase the latest model. However, geopolitical factors leading to the concentrated supply of AI server components, such as GPU chips, can result in the loss of computational efficiency in countries that do not have access to such chips, resulting in higher physical server demand. For instance, the Nvidia H800's bandwidth efficiency is half that of the H100, necessitating double the number to achieve equivalent performance. Our analysis indicates that a 1 yr delay in obtaining the latest chips could result in a 14% increase in EoS e-waste, cumulatively totaling 5.7 Mt from 2023 to 2030, higher than the global quantity of small ICT waste in 2022[9]. In a more optimistic case, where the regional semiconductor industry rapidly advances (strategy T2), our results indicate a 12% increase in e-waste when compared with the baseline. Nonetheless, implementing circularity measures can mitigate the additional e-waste generated by these technical barriers (T3), underscoring the importance of server utilization optimization explored in levers C2 and C3.

https://doi.org/10.1038/s43588-024-00712-6

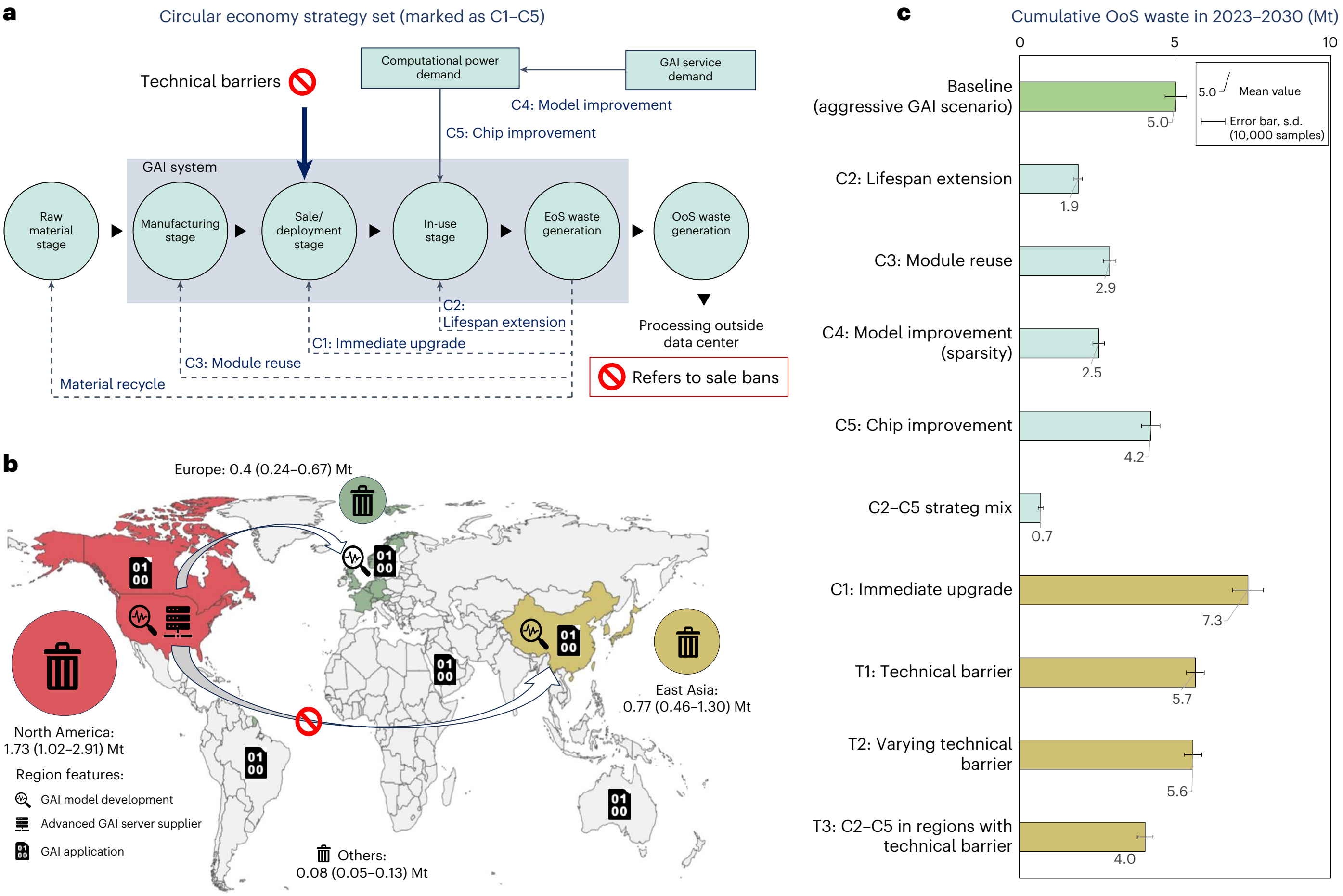

Fig. 2 | Circular economy strategies and their potential impacts on GAI-related e-waste generation. a, Circular economy strategies in different life-cycle stages of the AI server. **b**, International landscape of GAI model development, server manufacturing, application and potential e-waste generation under the conservative, moderate and aggressive scenarios. **c**, Cumulative OoS e-waste amount for different improvement levers, compared with the aggressive scenario results. Colors refer to scenarios with negative effect (brown) and positive effect (blue) and the baseline scenario (green).

We further quantify the valuable and hazardous material associated with obsolete LLM-related servers, mainly in three component categories: printed circuit boards with mounted semiconductor units, batteries and structural parts. They contain toxic metals including lead and chromium, and valuable metals such as gold, silver, platinum, nickel and palladium. A detailed estimation of these materials is presented in Supplementary Fig. 5. At the upper bound, we find that the aggressive scenario (with 1.5 Mt of printed circuit boards—mainly epoxy and polyamide—and 0.5 Mt of lead batteries—mainly lead, acrylonitrile butadiene styrene and polycarbonate) could generate substantial amounts of toxic substances (Supplementary Section 1.3). Conversely, it could create great economic gains if recycled properly, with an estimated value of around US$14–28 billion (2020–2030 cumulative value calculated at 2023 fixed prices). Given these facts, the potential toxic emissions and the recycling technologies call for further study[15].

In the realm of data centers, several prominent operators, such as the Microsoft Azure data center, have committed to sustainable practices and outlined zero-waste-series strategies (Supplementary Section 4.2). Nonetheless, similar to the current pattern of e-waste trade flows, GAI-related e-waste from environmentally aware nations may be exported to low- and middle-income countries, harming the environment and health. Thus, it is vital to boost circular economy strategies, track cross-border e-waste and encourage data center sustainability self-reporting. Sustainability certification or the digital life-cycle passport from battery waste management can serve as a ref.16.

Ultimately, our findings highlight the pressing need to anticipate future surges in GAI-related e-waste and to proactively adopt circular design and management strategies to mitigate its impact. Despite incorporating uncertainty analysis, our study has limitations. These include assuming constant computational power intensity for GPU servers, rough estimations of parameter configurations, and overlooking regional and inter-data-center variations. These factors could lead to both underestimation (for example, the usage scenarios and scope of GAI are underestimated) and overestimation (especially when chips use more advanced manufacturing processes in the future, increasing computational power intensity per unit mass of material) of the e-waste impact. Future research will need to address these limitations and refine assumptions to explore and develop more effective circular strategies. Nonetheless, our findings underscore the need to acknowledge the potential for future swells of GAI-related e-waste and to proactively implement circular, design and management strategies to avoid them.

Methods

We develop a dynamic model to estimate future amounts of global GAI-related e-waste under different scenarios. The logistic diagram is shown in Supplementary Fig. 1. We consider only the training and inference servers used for LLM computing inside data centers, ignoring servers for other purposes and accessory modules (Fig. 1a). The estimation of LLM proliferation is the basis of the model, and is determined by the number of models, number of parameters (in both training and inferring), training time, daily active users and queries per user per day. Unlike the simple exponential estimation, we use the limitation of training data size (that is, the sentences for training the model are finite) as the constraint to set limits for its development. Then, the computational power demand, which is derived from LLM proliferation, can be transformed into server demand by considering the computation efficiency and computational power per chip, which evolves according to Moore's law. To estimate the number of servers being discarded, we first assume the global new server deployment and stock amount. Then, the e-waste amount can be quantified on the hypothesis that the servers have a fixed lifespan and are discarded at the end of this period. In the baseline scenario, a 3 yr lifespan is selected on the basis of the historical general lifespan of computing devices, which is between 2 and 5 yr (Supplementary Section 3.3). We quantify the e-waste amount generated in 2030 and the cumulative amount between 2020 and 2030 at quarterly intervals. A brief description of the model components is given here with a detailed explanation of each parameter provided in Supplementary Section 1.

Estimation model of LLM-related GPU server flows

It is hard to gauge the accurate number of servers in data centers, as this is regarded as proprietary information by operators. Here we use the dynamic demands of computational power and the performance of GPU servers to approximate the number of servers, by evaluating the evolution of LLM parameter scales, training dataset scale, number of LLMs worldwide, number of daily active users and computational power of GPU servers. The configuration of these factors is derived from complementary research and recognized theorems such as Moore's law. 'Dynamic' means that the estimation is not carried out by simply dividing the total computational power demands by the individual computational power per GPU server. The stock computational power should be considered using this principle: number of new servers = (new computational power demands − current computational power stockage)/computational power per new server. The computational power is measured in pfs-day (24 h computing at 1 petaflop s^{-1}). This unit is widely used to describe the scale of AI computing tasks by OpenAI, Google and so on. For full details, see Supplementary Section 1.1.

Regional distribution analysis of LLM development

AI data centers are currently geographically clustered. Here, we suppose that the training and inference of LLMs will be undertaken in the same region as model development as a simplifying assumption. In this regard, three major LLM regions are selected: North America (United States and Canada), East Asia (China, South Korea and Japan) and Europe (European Union and United Kingdom). The regional distribution proportion is calculated by counting the number of existing LLMs in the three regions. Details of accounting are available in Supplementary Section 2.

Scenario development and settings

In this study, we develop four LLM development scenarios to explore possible LLM-related e-waste generation trends, namely a limited expansion of GAI chip capacity scenario, an aggressive scenario, a moderate scenario and a conservative scenario. The limited expansion of GAI chip capacity scenario is based on the hypothesis that the development of GAI scope is constrained by the manufacturing capacity of chip and server companies. The aggressive scenario follows the radical adoption of LLMs for daily usage such as some search engines and social platforms (for example, Google, Bing, Baidu and Facebook). Then, moderate and conservative scenarios are modeled with the assumption that LLMs have a specific yet wide-range target user (for example, TikTok), and that LLMs serve only those who become accustomed to this interaction (for example, voice assistants on smartphones). The detailed settings of different key parameters for these scenarios are given in Supplementary Sections 3.1 and 3.2 and listed in Supplementary Data 1. We further made comparisons with existing projections to indicate how our assumptions align with those of other authors. We then explored six scenarios to examine the extent to which the circular economy strategies (Supplementary Sections 3.3 and 3.4) and technical barriers (Supplementary Section 3.5) might affect the estimated e-waste generation amounts. The hypotheses of these scenarios are listed below, and detailed value configurations, potential impacts and practical applications of each strategy are discussed in Supplementary Section 3.

(1) C1: immediate upgrade. This is realized by changing the training computational power and inference computational power with the hypothesis that data center operators decide to substitute all the servers as soon as there are major upgrades of on-sale GPU (to reduce energy or maintenance costs, for example).

(2) C2: lifespan extension. This refers to deployment of servers at the end of their *L* yr lifespan (generally 3 yr) to downcycled server applications, such as less intensive AI computation or non-AI computation (for an extra 1 yr).

(3) C3: module reuse. This refers to the dismantling, renovation and remanufacturing of key modules of an obsolete GPU server, for example, GPU modules, CPU modules, memory modules or communication modules. The ultimate goal of this strategy is similar to that of the lifespan extension strategy: to extend the usage phase of a certain computational power. However, it differs from lifespan extension in two aspects. First, remanufacturing requires extra material; therefore extra e-waste is bound to reused servers, which does not occur in the lifespan extension case. Second, the renovated and remanufactured servers are re-endowed with a full *L* yr lifespan, rather than merely a 1 yr addition life in the lifespan extension case.

(4) C4: introduce advanced computing algorithm to models. We introduce sparsity features in this scenario. To conduct the calculation, we halve the value of computational power demand.

(5) C5: increase chip's computing efficiency. Only the computational power efficiency is changed in this scenario.

(6) T1–T3: three scenarios for technical barriers. A technical barrier refers to export regulations of certain GPU servers to certain countries or regions. Under these circumstances, the countries subject to the barrier will have to use servers with weaker computational power to conduct LLM computing tasks. For instance, the Nvidia H800 is a specifically adapted version of the H100 for the Chinese market due to the technical barrier implemented in August 2022. The difference between them is the drop in interconnect bandwidth, which leads to longer computing time. This is equivalent to a lag in computational power of pfs-day. In the three technical barrier scenarios (T1, T2 and T3), we reconfigure the training and inference computational power for the countries subject to the barrier and redo the calculation.

Data availability

This paper analyzes existing and publicly available data. All data sources used in this research are referenced in the main text or in Supplementary Information[17]. Source data for Figs. 1b,c and 2b,c are available with this paper.

Code availability

The main code of our approach (as well as datasets to run the code) is available[17].

References

1. Crawford, K. Generative AI's environmental costs are soaring—and mostly secret. *Nature* **626**, 693 (2024).

2. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
3. Grossmann, I. et al. AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
4. Jia, Z. et al. The importance of resource awareness in artificial intelligence for healthcare. *Nat. Mach. Intell.* **5**, 687–698 (2023).
5. Lannelongue, L. et al. GREENER principles for environmentally sustainable computational science. *Nat. Comput. Sci.* **3**, 514–521 (2023).
6. Mytton, D. & Ashtine, M. Sources of data center energy estimates: a comprehensive review. *Joule* **6**, 2032–2056 (2022).
7. Masanet, E., Shehabi, A. & Koomey, J. Characteristics of low-carbon data centres. *Nat. Clim. Change* **3**, 627–630 (2013).
8. *Computing 2030: Building a Fully Connected, Intelligent World* (Huawei, 2021).
9. Baldé, C. P., et al. *Global E-waste Monitor 2024* (ITU/UNITAR, 2024); https://ewastemonitor.info/the-global-e-waste-monitor-2024/
10. Parvez, S. M. et al. Health consequences of exposure to e-waste: an updated systematic review. *Lancet Planet. Health* **5**, e905–e920 (2021).
11. *Data Centres and Data Transmission Networks* https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks (IEA, 2023).
12. Ambrogio, S. et al. An analog-AI chip for energy-efficient speech recognition and transcription. *Nature* **620**, 768–775 (2023).
13. Wen, W. et al. Learning structured sparsity in deep neural networks. In *Proc. 30th International Conference on Neural Information Processing Systems* 2082–2090 (Curran, 2016).
14. Jonathan, O. Who's making chips for AI? Chinese manufacturers lag behind US tech giants. *Nature* https://doi.org/10.1038/d41586-024-01292-1 (2024).
15. Nuss, P. & Eckelman, M. J. Life cycle assessment of metals: a scientific synthesis. *PLoS ONE* **9**, e101298 (2014).
16. Walden, J., Angelika, S. & Maroye, M. Digital product passports as enabler of the circular economy. *Chem. Ing. Tech.* **93**, 1717–1727 (2021).
17. Johnly233. E-waste-Challenges-of-Generative-Artificial-Intelligence: revised version (V1.1). *Zenodo* https://doi.org/10.5281/zenodo.13790035 (2024).

Acknowledgements

This research was financially supported by the National Natural Science Foundation of China (72274187 to P.W., 71961147003 to W.-Q.C.) and CAS IUE Research Program (IUE-JBGS-202202 to P.W.). We thank E. Masanet and our other colleagues for their contributions, which have improved this study.

Author contributions

P.W. and L.-Y.Z. designed the research; L.-Y.Z., P.W. and A.T. led the drafting of the manuscript. P.W., L.-Y.Z. and W.-Q.C. contributed to the methodology; L.-Y.Z., P.W. and A.T. interpreted the results. All authors contributed to the final writing of the article.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-024-00712-6.

Correspondence and requests for materials should be addressed to Peng Wang, Asaf Tzachor or Wei-Qiang Chen.

Peer review information *Nature Computational Science* thanks Loïc Lannelongue, Mengmeng Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Kaitlin McCardle, in collaboration with the *Nature Computational Science* team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.