

From computation to environmental cost the resource burden of artificial intelligence

Corresponding Author: Ms Sophia Falk

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

**** Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors ****

Dear Ms Falk,

Your manuscript titled "From FLOPs to Footprints: The Resource Cost of Artificial Intelligence" has now been seen by 3 reviewers, whose comments are appended below. You will see that they find your work of some potential interest. However, they have raised substantial concerns that must be addressed. In light of these comments, extensive revisions will be required before we can further consider the manuscript for publication. We would, however, be interested in considering a revised version that fully addresses these serious concerns.

We hope you will find the reviewers' comments useful as you decide how to proceed. If additional work allows you to either incorporate or refute these criticisms, we will be happy to look at a substantially revised manuscript. If you choose to take up this option, please either highlight all changes in the manuscript text file, or provide a list of the changes to the manuscript with your responses to the reviewers.

To be publishable in Communications Earth & Environment your study must:

****provide compelling novel insights into the model's performance taking into account the nonlinear effects as well.**

****transparently report the limitations associated with the modelling framework along with inclusion of a simple sensitivity analysis.**

When resubmitting, please provide a point-by-point response to the reviewers' comments. Please submit your responses as a separate file, distinct from your cover letter where you can add responses to the Editors' comments that you do not want to be made available to the reviewers. Word files are preferred. We recommend that any figures, tables or graphs that are included in the response to reviewers are also included in the main article or Supplementary Information.

Please bear in mind that we will be reluctant to approach the reviewers again in the absence of substantial revisions.

If the revision process takes significantly longer than three months, we will be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Earth & Environment or published elsewhere in the meantime.

We are committed to providing a fair and constructive peer-review process. Please do not hesitate to contact us if you wish to discuss the revision in more detail.

Please use the following link to submit your revised manuscript, point-by-point response to the reviewers' comments with a list of your changes to the manuscript text (which should be in a separate document to any cover letter), a tracked-changes version of the manuscript (as a PDF file) and any completed checklist:

Link Redacted

**** This url links to your confidential home page and associated information about manuscripts you may have submitted or be**

reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

Please do not hesitate to contact us if you have any questions or would like to discuss the required revisions further. Thank you for the opportunity to review your work.

Best regards,

Nandita Basu, PhD
Consulting Editor, Communications Earth & Environment
Associate Editor, Communications Sustainability
Nature Portfolio

EDITORIAL POLICIES AND FORMAT

If you decide to resubmit your paper, please ensure that your manuscript complies with our editorial policies and complete and upload the checklist below as a Related Manuscript file type with the revised article:

For Manuscripts that fall into the following fields:

- Behavioural and social science
- Ecological, evolutionary & environmental sciences
- Life sciences

An updated and completed version of our Reporting Summary must be uploaded with the revised manuscript

You can download the form here:

<https://www.nature.com/documents/nr-reporting-summary.pdf>

For your information, you can find some guidance regarding format requirements summarized on the following checklist: (<https://www.nature.com/documents/commsj-phys-style-formatting-checklist-article.pdf>) and formatting guide (<https://www.nature.com/documents/commsj-phys-style-formatting-guide-accept.pdf>).

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

The paper introduces a link between AI training compute (FLOPs) and the material footprint of specialised hardware, using the Nvidia A100 GPU as the reference device. It combines empirical elemental composition measurements with scenario-based modelling of GPU computational throughput, assumed hardware lifetimes, and Model FLOPs Utilisation (MFU) to estimate the elemental mass (including toxic metals) associated with training several large models, including GPT-4.

The topic is timely and important. Extending AI sustainability analysis beyond energy/water to material resources is valuable and underdeveloped in the literature. The manuscript's principal contribution is the translation from hardware to materials in the compute-to-materials layer and scenario analysis.

I enjoyed reading this paper, which is well written and opens another dimension on the sustainability of the AI endeavour, beyond energy and water consumption. Unfortunately, the paper carries some issues in conceptual modelling, uncertainty treatment, and scope that currently prevent the results from supporting the paper's broader conclusions.

Pros

- Valuable attempt to connect compute metrics (FLOPs) to physical resource demand; policy relevance is very high.
- Empirical elemental characterisation of a flagship AI GPU is rare and useful.
- Clear stepwise methodology and transparent equations; scenario approach is appropriate given uncertainty.
- The performance vs. resource framing is potentially impactful, provided interpretability issues are addressed.

Cons

- Novelty vs. incremental contribution is not sufficiently explicit. A core empirical input (A100 elemental composition) is imported from prior work by the authors. This is acceptable, but the manuscript has to better describe what is reused vs. new, and how this paper advances beyond the earlier LCA-oriented publication. Currently, the novelty claim is somewhat blurred, which complicates editorial evaluation.

- Sequential training abstraction is misframed and risks reader confusion. The manuscript assumes sequential training on one GPU until end-of-life as a simplifying device. While the Discussion acknowledges this does not reflect real training. In reality, the efficiency of training (wrt execution on a single GPU) is known, despite being dependent on specific training settings (but this is also true for one GPU). Moving from sequential training to parallel (e.g. fully sharded) would not add further uncertainty and would avoid the framing that can still mislead non-ML readers into interpreting GPU counts as

operational provisioning rather than an equivalence class of hardware depreciation/lifetime utilisation.

- Uncertainty ranges are so wide that interpretability suffers. Varying MFU (20–50%), lifespan (1–3 years), and GPT-4 architectural assumptions produce order-of-magnitude ranges. Scenario analysis is appropriate given poor transparency, but the paper currently does not help the reader identify the most plausible region or what parameters dominate outcomes. The result risks being read as “illustrative only.” A sensitivity/elasticity analysis would help.

- Inference is not considered, but inference is by far the dominant AI workload. The paper focuses on training, but draws broader claims about AI scalability/sustainability. For deployed frontier systems, inference frequently dominates lifetime compute and hardware utilisation. Without at least bounding inference, the manuscript cannot support general statements about the material footprint of “AI” beyond training.

- Toxicity framing. Aggregating “toxic elements” by mass is rhetorically strong but can conflate mass with realised harm. Impacts depend on chemical forms, exposure pathways, governance, and disposal conditions. The paper should avoid language that suggests direct health impacts from mass totals without exposure modelling.

Reviewer #2 (Remarks to the Author):

This article presents important and novel results looking in details at the composition of GPUs widely used for modern AI training. They provide a rather comprehensive assessment of the material requirements (at element level) of the needs of training a range of popular AI models, and analysed different cost/performance trade-offs.

This is a really important and novel piece. It is overall well written, transparent on its methodology and limitations, and I commend the authors for the quality of their research. I don't have much comments at this stage, I have listed some below.

Section 0.2.1. Although I can appreciate the authors' argument for only looking at 3 years lifespan, I believe the first part of this section makes the case for at least mentioning what the results would be with a 5 year lifespan.

On the same note, while assuming that just 1 GPU is used and replaced is indeed a convenient framework for this work, it could then be seen as unfair to also allocate shorter lifespan to this GPU because it would be used continuously. It would be good to provide some assessment (even if accompanied by large uncertainty intervals) in a more realistic setting of parallel training, if feasible. Or at least explain more how parallel training can be expected to impact lifespan.

****Minor comments:****

(l.64) “Capable of delivering up to 312 teraFLOPs (in BF16 precision) for deep learning workloads, the A100 provided a significant performance increase over its predecessor, the Volta series, with an approximately threefold increase in FP16 and sixfold in FP32” - This would likely needs some explanations as it's significantly more technical than the surrounding text.

Overall, I believe the flow of the results section is complicated by switching between very granular element-level statistics and more generic results. It may be useful, depending on the expected audience, to either streamline element-level results (in the text at least), or maybe just put results interested in a general audience at the start.

Reviewer #3 (Remarks to the Author):

Summary

The paper explores the environmental and material costs associated with training large-scale artificial intelligence models, moving beyond traditional concerns related to energy and water consumption. By disassembling an Nvidia A100 GPU, the authors identify thirty-two distinct elements, showing that these devices are composed primarily of heavy metals such as copper and iron, as well as toxic materials including lead and arsenic. The study introduces a framework that links computational workloads to physical hardware consumption and estimates that training a model such as GPT-4 may result in the extraction and eventual disposal of up to several tons of toxic elements. While software optimizations and extended hardware lifespans can significantly mitigate these impacts, the results suggest that current AI scaling trends yield diminishing performance returns relative to rapidly increasing resource demands. Overall, the paper argues for a shift toward more resource-conscious AI development in light of the environmental risks associated with mineral extraction and electronic waste.

General comments

- From a high-performance computing perspective, the computational model employed in the manuscript is intentionally simple. It effectively reduces to a linear scaling exercise in which peak theoretical GPU throughput is integrated over time

and adjusted by two scalar parameters: Model FLOPs Utilization (MFU) and assumed hardware lifespan. This approach is transparent and easy to interpret, and is appropriate for a first-order analysis. At the same time, it does not constitute a performance or system-level model, as non-linear effects, scaling inefficiencies, performance characteristics of the computational process, and interactions between utilization, parallelism, and hardware degradation are not captured. GPU efficiency is therefore condensed into a single scalar (MFU), which in practice encapsulates a wide range of system- and algorithm-level factors.

- The choice to focus on GPT-3.5 and GPT-4 is understandable given their prominence and relevance. However, these models are also associated with substantial uncertainty regarding architectural details, parameter counts, token counts, and training configurations. In addition to emphasizing the environmental implications, it could be valuable if the authors explicitly highlighted the need for greater transparency from model providers regarding these quantities, for example in the conclusions.

- The statement that GPUs operate "at or near maximum power draw" even at low MFU (lines 203-205) requires a more precise justification. Low MFU is often caused by non-compute delays, such as communication overheads (network bottlenecks) or memory-bound operations where tensor cores remain idle. In these wait states, average power consumption (and consequently thermal stress) may drop below the peak TDP sustained during dense matrix multiplication. If lower MFU correlates with reduced thermal degradation, the worst-case material impact scenarios might be overestimated.

- The analysis focuses exclusively on GPUs when estimating material consumption. While GPUs are clearly the computational core of AI workloads, it is not obvious that they dominate material waste compared to other components required for large-scale AI deployment, such as networking, storage, power delivery, and cooling infrastructure. This point is already addressed by the authors in the conclusions and does not detract from the main contribution, but becomes more relevant in the later sections (from Section 0.8 onward), where broader generalizations are attempted.

Major comment:

The comparative analysis presented in Section 0.7, and in particular in Figure 6, links benchmark performance improvements to estimated GPU resource requirements in a way that is central to the manuscript's broader narrative on diminishing returns from AI scaling. While the qualitative conclusion (that performance gains are heterogeneous across benchmarks and do not scale uniformly with computational investment) is convincing, the quantitative ratios reported (e.g., statements such as "~30x more resources") deserve more careful treatment. These ratios are derived from point estimates of GPU counts that scale linearly with the assumed Model FLOPs Utilization (MFU), a parameter that is not measured in this study and is known to vary substantially across models, training runs, and software employed. As a result, relatively modest changes in the assumed MFU for different models (e.g., GPT-3.5 versus GPT-4) can shift the reported GPU ratios by factors of several, potentially altering the quantitative interpretation of the results while leaving the qualitative conclusion unchanged. In its current form, the presentation of single-valued GPU ratios risks conveying a level of precision and robustness that is not fully supported by the underlying first-order model. To mitigate this risk and align the quantitative presentation more closely with the acknowledged limitations of the framework, I recommend incorporating a simple sensitivity analysis with respect to MFU. For example, Figure 6 could display ranges or error bars corresponding to the MFU interval already considered elsewhere in the manuscript (e.g., 20-60%), rather than single point estimates. This modification would not require additional experiments, but would substantially improve the interpretability and robustness of the comparative analysis by making explicit that the reported GPU resource estimates should be understood as ranges rather than precise values.

Minor comments

- Line 159: "80GB" → "80 GB".

- Table 2: please standardize the number of significant digits across entries.

- Table 4: the caption states that values are rounded up to the nearest integer; this does not appear to be true for LLaMa 2 with lifespan = 2 and MFU = 50%.

- Table 5: consider representing the elemental composition as a color-coded heat map to allow readers to more immediately identify which elements dominate each GPU component.

- Table 5: the statement "Elements marked with (*) are either inherently toxic or present significant health hazards" would benefit from supporting references. Please clarify how toxicity is defined or assessed in this context and provide appropriate citations.

- Section 0.6: the manuscript reports many absolute quantities of materials required for training. Providing contextual references (e.g., copper required for training relative to global copper extraction or annual production) would help readers better interpret the magnitude of these numbers.

- The manuscript cites a non-negligible number of arXiv preprints, which is understandable given the rapidly evolving AI/HPC literature and the limited availability of peer-reviewed sources for some training details. However, I recommend checking whether peer-reviewed versions exist and citing them where available. For key assumptions supported primarily by preprints, a brief clarification that these sources function as technical reports or preliminary studies would help contextualize their use.

**** Visit Nature Portfolio's author and referees' website at www.nature.com/authors for information about policies, services and author benefits****

Communications Earth & Environment is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on 'Modify my Springer Nature account' and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

For more information please visit <http://www.springernature.com/orcid>

If you experience problems in linking your ORCID, please contact the Platform Support Helpdesk.

Version 1:

Decision Letter:

**** Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors ****

Dear Ms Falk,

Your manuscript titled "From FLOPs to Footprints: The Resource Cost of Artificial Intelligence" has now been seen by our reviewers, whose comments appear below. In light of their advice we are delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Earth & Environment.

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table".

*******Please take care to match our formatting and policy requirements. We will check revised manuscript and return manuscripts that do not comply. Such requests will lead to delays. *******

Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed at the end of the Editorial Requests Table; the list of required files is also available at <https://www.nature.com/documents/commsj-file-checklist.pdf>.

OPEN ACCESS:

Communications Earth & Environment is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Portfolio, please visit <https://www.nature.com/commsenv/open-access>

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

Please use the following link to submit the above items:
Link Redacted

**** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first ****

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Nandita Basu, PhD
Consulting Editor, Communications Earth & Environment
Associate Editor, Communications Sustainability
Nature Portfolio

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

The authors have addressed my previous comments effectively. As already written, the work is interesting and original enough to be published.

Reviewer #2 (Remarks to the Author):

Thank you for the detailed responses and clear changes made to the manuscript. I believe all my comments have been suitably addressed.

Reviewer #3 (Remarks to the Author):

All recommendations from the previous review have been taken into account and I have no further comments.

**** Visit Nature Portfolio's author and referees' website at <http://www.nature.com/authors> for information about policies, services and author benefits****

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were

made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

**Revisions for the manuscript titled
"From FLOPs to Footprints: The Resource Cost of Artificial Intelligence"
(COMMSENV-25-6190)**

Dear Editor,

Thank you for the opportunity to revise the manuscript "From FLOPs to Footprints: The Resource Cost of Artificial Intelligence" (COMMSENV-25-6190) and for coordinating the review process. We are grateful to the reviewers for their thorough and constructive feedback, which has substantially improved the quality and rigor of the manuscript.

We have addressed all reviewer comments in full. Most notably, we have added a dedicated section titled "Model Uncertainty and Limitations," which extends the baseline computational model in two ways: a simple sensitivity analysis quantifying the influence of MFU and hardware lifespan assumptions on GPU requirement estimates, and a parallelization efficiency analysis introducing a non-linear scaling factor η .

In addition, we have highlighted what the novel insights of this manuscript are and expanded the discussion of modelling framework limitations.

All changes are documented in detail in the response to the reviewers, which is set out as follows: The response to the reviewer is **bold** and example text passage to highlight revisions are given in *red italics*. Furthermore, all changes are visible via the "track changes" function in the revised manuscript.

The following documents are provided: the point-by-point response to reviewers, the clean revised manuscript and the revised manuscript with track changes.

We hope the revised manuscript meets the standards of the journal Communications Earth & Environment and look forward to your decision.

Best regards,

Sophia Falk
on behalf of all co-authors

Response to Reviewer #1

We thank Reviewer 1 for their careful and constructive reading of the manuscript. Their comments were substantive and have led to meaningful improvements across several dimensions of the paper. In response to their feedback, we have: (1) explicitly delineated the novel contributions of this study from the prior empirical dataset reported in Falk et al. (2025); (2) reframed the sequential training abstraction as an equivalent GPU lifetime/depreciation metric, removing the potential for misinterpretation by non-ML readers; (3) added a dedicated sensitivity analysis and identified throughout the manuscript the most plausible scenario; (4) narrowed the scope of all sustainability claims to AI training specifically, with inference noted as an important avenue for future work; and (5) revised all toxicity-related language to clearly distinguish between hazard classification and realized risk, framing mass totals as indicators of hazardous material flows rather than direct proxies for health impact. We believe these revisions substantially strengthen the manuscript's conceptual clarity, interpretability, and scientific rigor.

Reviewer #1 (Remarks to the Author):

The paper introduces a link between AI training compute (FLOPs) and the material footprint of specialised hardware, using the Nvidia A100 GPU as the reference device. It combines empirical elemental composition measurements with scenario-based modelling of GPU computational throughput, assumed hardware lifetimes, and Model FLOPs Utilisation (MFU) to estimate the elemental mass (including toxic metals) associated with training several large models, including GPT-4.

The topic is timely and important. Extending AI sustainability analysis beyond energy/water to material resources is valuable and underdeveloped in the literature. The manuscript's principal contribution is the translation from hardware to materials in the compute-to-materials layer and scenario analysis.

I enjoyed reading this paper, which is well written and opens another dimension on the sustainability of the AI endeavour, beyond energy and water consumption. Unfortunately, the paper carries some issues in conceptual modelling, uncertainty treatment, and scope that currently prevent the results from supporting the paper's broader conclusions.

Pros

- Valuable attempt to connect compute metrics (FLOPs) to physical resource demand; policy relevance is very high.
- Empirical elemental characterisation of a flagship AI GPU is rare and useful.
- Clear stepwise methodology and transparent equations; scenario approach is appropriate given uncertainty.
- The performance vs. resource framing is potentially impactful, provided interpretability issues are addressed.

Cons

- Novelty vs. incremental contribution is not sufficiently explicit. A core empirical input (A100 elemental composition) is imported from prior work by the authors. This is acceptable, but the manuscript has to better describe what is reused vs. new, and how this paper advances beyond the earlier LCA-oriented publication. Currently, the novelty claim is somewhat blurred, which complicates editorial evaluation.

Thank you for this comment. We agree that the distinction between reused and novel contributions needed to be made more explicit. To address this, we have added the following clarifying text to the manuscript:

While the elemental composition dataset originates from the same empirical analysis reported in Falk et al. (2025) \cite{falk2025morethan}, that earlier work used the data solely as background input for LCA software modeling, without reporting the elemental composition results themselves. In this study, we shift the emphasis from the LCA lens to the analytical procedures, reporting the elemental analysis results in detail and providing methodological context and interpretation not included in the prior publication. These results are then used to develop a novel framework linking computational demand (measured in FLOPs) to material consumption - a connection that has not been established in prior work.

This addition explicitly delineates what is reused (the underlying empirical dataset) from what is new (the full reporting of results, methodological detail, and analytical interpretation), and clarifies how this paper advances beyond the prior publication.

- Sequential training abstraction is misframed and risks reader confusion. The manuscript assumes sequential training on one GPU until end-of-life as a simplifying device. While the Discussion acknowledges this does not reflect real training. In reality, the efficiency of training (wrt execution on a single GPU) is known, despite being dependent on specific training settings (but this is also true for one GPU). Moving from sequential training to parallel (e.g. fully sharded) would not add further uncertainty and would avoid the framing that can still mislead non-ML readers into interpreting GPU counts as operational provisioning rather than an equivalence class of hardware depreciation/lifetime utilisation.

Thank you for this comment! Indeed, the sequential framing was unnecessarily misleading. We have reframed the metric as a hardware lifetime/depreciation equivalent, which is both more accurate and more robust to parallelization assumptions. We changed the phrasing to the following:

To link GPU throughput to the demands of training a given AI model, we express computational demand in terms of equivalent GPU lifetime consumed. This is representing the total hardware depreciation attributable to training, independent of the actual parallelization architecture employed. This framing treats GPU count as an equivalence class of hardware lifetime utilization: the number of GPUs whose full productive lifespan would be exhausted by the cumulative compute required to train a given model.

We also changed the discussion section that related to the sequential training:

GPU demand is expressed as equivalent hardware lifetime consumed, translating cumulative computational workloads into fractional depletion of GPU productive capacity regardless of the parallelization strategy employed. This allows standardized comparison across models with heterogeneous training strategies. To illustrate: GPT-4 was reportedly trained on approximately 25,000 A100 GPUs for 90–100 days \cite{katerinaptrvGPT4Leaked}. Under a two-year lifespan assumption, such a training run corresponds to roughly 13% wear of each GPU's productive capacity.

- Uncertainty ranges are so wide that interpretability suffers. Varying MFU (20–50%), lifespan (1–3 years), and GPT-4 architectural assumptions produce order-of-magnitude ranges. Scenario analysis is appropriate given poor transparency, but the paper currently does not help the reader identify the most plausible region or what parameters dominate outcomes. The result risks being read as “illustrative only.” A sensitivity/elasticity analysis would help.

Thank you for this constructive comment. We agree that the manuscript needed to more clearly identify the most plausible region of outcomes and communicate which parameters dominate uncertainty. We have addressed this on three levels. First, we have revised the text accompanying Figure 4 to explicitly designate the 2-year hardware lifespan at 35% MFU as the most plausible baseline scenario for GPT-4, framing the 1-year and 3-year lifespan scenarios as conservative upper and lower bounds respectively. Second, we have added a clarifying statement that the GPU counts presented in Figure 6 represent the most plausible central estimates for each model, derived from reported MFU values where available and from independent technical assessments for remaining models, combined with the 2-year lifespan baseline. Third, and most directly addressing the request for a sensitivity analysis, we have added a dedicated section on model uncertainty and limitations which includes a one-at-a-time sensitivity analysis varying MFU (20-60%) and hardware lifespan (1-3 years) independently around the baseline. For GPT-4, architectural uncertainty is additionally quantified across the full range of MoE scenarios.

- Inference is not considered, but inference is by far the dominant AI workload. The paper focuses on training, but draws broader claims about AI scalability/sustainability. For deployed frontier systems, inference frequently dominates lifetime compute and hardware utilisation. Without at least bounding inference, the manuscript cannot support general statements about the material footprint of “AI” beyond training.

Thank you for this comment. We agree that inference represents a significant and often dominant workload for deployed AI systems. However, this study is explicitly scoped to training compute and does not aim to characterize the full lifecycle footprint of deployed AI systems. We have revised the manuscript language throughout to ensure that broad claims about “AI” sustainability are qualified as pertaining specifically to AI training, avoiding overgeneralization beyond the scope of this study. We acknowledge inference as an important direction for future work.

- Toxicity framing. Aggregating “toxic elements” by mass is rhetorically strong but can conflate mass with realised harm. Impacts depend on chemical forms, exposure pathways, governance, and disposal conditions. The paper should avoid language that suggests direct health impacts from mass totals without exposure modelling.

Thank you for this important methodological point. We have revised the language throughout the manuscript to distinguish between intrinsic hazard classification and realized risk, which depends on chemical form, exposure pathway, and disposal conditions. Mass totals are now framed as indicators of hazardous material flows rather than direct proxies for health impact. Here are two examples of those framings:

Electronic components contain numerous heavy metals, including arsenic (As), mercury (Hg), lead (Pb), cadmium (Cd), chromium (Cr), zinc (Zn), copper (Cu), nickel (Ni), antimony (Sb), cobalt (Co) and beryllium (Be), that are classified as hazardous and have well-documented toxic properties if released during mining, manufacturing or disposal at the end of their life \cite{wang2024analysis, wu2022antimonyBeCO, WHO2010toxic}.

This corresponds to a material footprint of 19,940 kg of extracted resources, 18,545 kg of which are hazardous metals with recognised toxic properties under relevant exposure conditions.

Response Reviewer #2

We thank Reviewer 2 for their generous and encouraging assessment of the manuscript, and for their thoughtful suggestions for improvement.

In response to their feedback, we have: (1) added the results for a four- and five-year lifespan to Table 4, and we have added a paragraph to the Discussion section explaining how the lifespan relates to parallelization in our framework; (2) extended the baseline computational model with a parallelization efficiency factor accounting for communication overhead, memory bandwidth bottlenecks, and load imbalance at scale, reported in a new table, which directly addresses the concern that continuous single-GPU operation may not reflect realistic distributed training conditions. We additionally clarify that the linear baseline should be interpreted as a conservative lower bound on GPU requirements; (3) added clarifying parentheticals and a follow-up sentence to the technical description of precision formats (BF16, FP16, FP32) to make this passage more accessible to a general scientific audience; and (4) restructured the Results section so that headline findings and big-picture takeaways precede element-level detail, improving narrative flow for readers with varying levels of technical background.

We are grateful for the reviewer's recognition of the novelty and policy relevance of this work, and hope that these revisions further improve its clarity and accessibility.

Reviewer #2 (Remarks to the Author):

This article presents important and novel results looking in details at the composition of GPUs widely used for modern AI training. They provide a rather comprehensive assessment of the material requirements (at element level) of the needs of training a range of popular AI models, and analyzed different cost/performance trade-offs.

This is a really important and novel piece. It is overall well written, transparent on its methodology and limitations, and I commend the authors for the quality of their research. I don't have much comments at this stage, I have listed some below.

Section 0.2.1. Although I can appreciate the authors' argument for only looking at 3 years lifespan, I believe the first part of this section makes the case for at least mentioning what the results would be with a 5 year lifespan.

That is true, thank you for this comment. We have added the results for a 4 and 5-year lifespan in table 4.

On the same note, while assuming that just 1 GPU is used and replaced is indeed a convenient framework for this work, it could then be seen as unfair to also allocate shorter lifespan to this GPU because it would be used continuously. It would be good to provide some assessment (even if accompanied by large uncertainty intervals) in a more realistic setting of parallel training, if feasible. Or at least explain more how parallel training can be expected to impact lifespan.

Thank you for this thoughtful comment, this is a very interesting point. We have discussed this internally and agree that the relationship between parallelization architecture and hardware lifespan deserves explicit clarification. Unfortunately, we cannot empirically proof how parallelization can be expected to impact lifespan. While such an experiment would be very interesting and relevant for future work, it is outside the scope of this paper.

However, we can explain why, within our study, we believe that parallelization does not affect lifespan. Our framework is based on equivalent operational hardware lifetime consumed, that is, the fraction of each GPU's total productive throughput depleted by a given training run, rather than calendar time. Since we model our calculation for the training partition of a supercomputer used in industry context, idle periods are excluded from the calculation. Whether a GPU's productive capacity is consumed continuously over two years or distributed across a longer calendar period with intermittent use, the total wear attributable to the training workload remains identical. Parallel training redistributes hardware depreciation across more devices simultaneously but does not reduce the aggregate material footprint of training. The reported estimates therefore hold regardless of the parallelization architecture employed. We have added a clarifying paragraph to the manuscript making this reasoning explicit:

One might expect that distributed parallel training, by spreading the computational workload across many GPUs simultaneously, would reduce wear on individual devices and thus extend their operational lifespan. However, this does not affect the total hardware depreciation attributed to a training run in this framework. GPU requirements

are expressed as equivalent hardware lifetime consumed, that is, the share of each device's total productive capacity depleted by training. Since idle time is excluded from this calculation, it does not matter whether a GPU's capacity is used continuously over two years or spread across a longer period with intermittent use: the total wear assigned to the training workload remains the same. Parallel training redistributes this depreciation across more devices, but does not reduce the aggregate material footprint, meaning the reported estimates hold regardless of the parallelization strategy employed.

****Minor comments:****

(l.64) “Capable of delivering up to 312 teraFLOPs (in BF16 precision) for deep learning workloads, the A100 provided a significant performance increase over its predecessor, the Volta series, with an approximately threefold increase in FP16 and sixfold in FP32” - This would likely need some explanations as it's significantly more technical than the surrounding text.

Thank you for this comment. We have added clarifying parentheticals to define each precision format and added a follow-up sentence explaining the underlying precision-speed tradeoff:

Capable of delivering up to 312 teraFLOPs in BF16 precision - a 16-bit numerical format optimized for deep learning workloads, the A100 provided a significant performance increase over its predecessor, the Volta series, with an approximately threefold increase in FP16 (standard 16-bit precision) and sixfold in FP32 (32-bit precision) \cite{nvidia2020ampere}. Higher precision formats use more bits to represent numerical values, yielding greater accuracy but at the cost of increased memory and computational demand, making the choice of precision a key factor in training efficiency.

Overall, I believe the flow of the results section is complicated by switching between very granular element-level statistics and more generic results. It may be useful, depending on the expected audience, to either streamline element-level results (in the text at least), or maybe just put results interested in a general audience at the start.

Thank you for this structural suggestion. We agree that the alternation between granular element-level statistics and higher-level findings disrupted the narrative flow of the results section. We have restructured both the “Elemental Composition” and “Resource Cost of AI Training” subsections to lead with headline findings and big-picture takeaways before drilling into element-level detail. In addition, the discussion of die scaling limitations, including area-wall constraints and die yield, has been relocated from the Elemental Composition subsection to the Resource Cost of AI Training subsection, where its implications for aggregate material consumption are more directly relevant.

Response to Reviewer #3

We thank Reviewer 3 for their thorough and technically rigorous reading of the manuscript. Their comments have led to substantive improvements across both the computational modelling and the presentation of results.

In response to their feedback, we have: (1) extended the baseline computational model with a parallelization efficiency factor that accounts for communication overhead, memory bandwidth bottlenecks, and load imbalance at scale, reported in a new table, and clarified that the linear baseline estimates should be interpreted as conservative lower bounds; (2) added an explicit call in the conclusions for AI developers and regulators to treat training configuration disclosure as part of broader sustainability reporting frameworks; (3) added a careful qualification to distinguish the GPU power draw during sustained training workloads from inference contexts; (4) more prominently flagged the GPU-only scope of the analysis in the sections where broader generalizations are attempted; (5) added context to Figure 6 to highlight most plausible MFU assumptions, while adding a dedicated sensitivity analysis section varying MFU across the 20-60% interval and lifespan to make the sensitivity of reported GPU ratios fully transparent; (6) addressed all minor comments, including standardizing significant digits, correcting a rounding discrepancy in Table 4, adding a color-coded highlight to the dominant elements in Table 5, updating toxicity language with appropriate citations, contextualizing the single-run vs. global extraction comparison, and replacing where possible arXiv preprints with their peer-reviewed equivalents. We are grateful for the care and precision of this review, which has meaningfully strengthened the manuscript's scientific rigor.

Reviewer #3 (Remarks to the Author):

Summary

The paper explores the environmental and material costs associated with training large-scale artificial intelligence models, moving beyond traditional concerns related to energy and water consumption. By disassembling an Nvidia A100 GPU, the authors identify thirty-two distinct elements, showing that these devices are composed primarily of heavy metals such as copper and iron, as well as toxic materials including lead and arsenic. The study introduces a framework that links computational workloads to physical hardware consumption and estimates that training a model such as GPT-4 may result in the extraction and eventual disposal of up to several tons of toxic elements. While software optimizations and extended hardware lifespans can significantly mitigate these impacts, the results suggest that current AI scaling trends yield diminishing performance returns relative to rapidly increasing resource demands. Overall, the paper argues for a shift toward more resource-conscious AI development in light of the environmental risks associated with mineral extraction and electronic waste.

General comments

- From a high-performance computing perspective, the computational model employed in the manuscript is intentionally simple. It effectively reduces to a linear scaling exercise in which peak theoretical GPU throughput is integrated over time and adjusted by two scalar parameters: Model FLOPs Utilization (MFU) and assumed hardware lifespan. This approach is transparent and easy to interpret, and is appropriate for a first-order analysis. At the same time, it does not constitute a performance or system-level model, as non-linear effects, scaling inefficiencies, performance characteristics of the computational process, and interactions between utilization, parallelism, and hardware degradation are not captured.

GPU efficiency is therefore condensed into a single scalar (MFU), which in practice encapsulates a wide range of system- and algorithm-level factors.

Thank you for this observation. We agree that the linear scaling assumption represents a simplification that does not capture the full range of system-level inefficiencies present in large-scale distributed training. We have added a dedicated section extending the baseline model by introducing a non-linear parallelization efficiency factor, which accounts for communication overhead, memory bandwidth bottlenecks, and load imbalance at scale. Table 8 reports GPU requirements across efficiency factor values ranging from 1.0 to 0.2, with MFU fixed at 35% and lifespan at 2 years.

We additionally note that the non-linear scaling factor can only increase GPU requirements relative to the linear baseline, and the estimates reported in the main analysis should therefore be interpreted as conservative lower bounds.

- The choice to focus on GPT-3.5 and GPT-4 is understandable given their prominence and relevance. However, these models are also associated with substantial uncertainty regarding architectural details, parameter counts, token counts, and training configurations. In addition to emphasizing the environmental implications, it could be valuable if the authors explicitly highlighted the need for greater transparency from model providers regarding these quantities, for example in the conclusions.

Thank you for this constructive suggestion. We agree that the uncertainty surrounding architectural and training details of closed models such as GPT-3.5 and GPT-4 represents a significant limitation, and that calling for greater transparency from model providers is both warranted and relevant to the broader contribution of this paper. In addition to our interactive tool (<https://huggingface.co/spaces/sophia-falk/flops-2-footprints>) to accommodate alternative parameter assumptions, we have expanded this passage to explicitly call on model providers, developers, and regulators to treat training configuration disclosure as part of broader AI sustainability reporting frameworks:

The estimates presented in this study are inherently constrained by the limited transparency of frontier model providers regarding architectural details, parameter counts, token counts, and training configurations. For closed models such as GPT-3.5 and GPT-4, key inputs must be inferred from leaked data, third-party analyses, or architectural analogies, introducing unavoidable uncertainty into material footprint estimates. Alternative configurations, such as different parameter counts, token counts, expert counts, activation ratios, and training configurations would yield different computational and material requirements. To accommodate this uncertainty, we provide an interactive tool ^{footnote}{The tool can be accessed via the following URL: <https://huggingface.co/spaces/sophia-falk/flops-2-footprints>}. that enables users to input alternative parameters and compute the corresponding results themselves.

Greater transparency from model providers regarding these quantities would not only improve the accuracy of environmental assessments but would also enable more meaningful comparisons across models and organizations. We therefore call on AI developers and regulators to treat training configuration disclosure as a component of broader AI sustainability reporting frameworks.

- The statement that GPUs operate "at or near maximum power draw" even at low MFU (lines 203-205) requires a more precise justification. Low MFU is often caused by non-compute delays, such as communication overheads (network bottlenecks) or memory-bound operations where tensor cores remain idle. In these wait states, average power consumption (and consequently thermal stress) may drop below the peak TDP sustained during dense matrix multiplication. If lower MFU correlates with reduced thermal degradation, the worst-case material impact scenarios might be overestimated.

Thank you for this precise technical observation. We agree that our original statement required more careful qualification. We have added a clarification distinguishing training from inference workloads. That is, while instantaneous power draw fluctuates during training due to communication overhead and memory-bound phases, average GPU power consumption during sustained training workloads remains close to peak TDP. This contrasts with inference, where GPUs may sit genuinely idle between requests and power draw can drop more substantially. We highlight the difference between training and inference utilization and maintain that reduced MFU does not straightforwardly extend GPU lifespan in the context of AI training.

Note that a lower MFU does not necessarily imply proportionally lower power consumption. While instantaneous power draw fluctuates during training due to communication overhead and memory-bound phases, average GPU power consumption during sustained training workloads remains close to peak TDP. Technical reports indicate that GPUs often operate at or near maximum power draw even at 20%-50% MFU ^{cite{behindthekeyboardLlama, chowdhery2022palm}}.

This distinguishes training from inference, where GPUs may sit genuinely idle between requests during which power draw and thermal stress fall below peak TDP.

Since power draw and associated thermal dissipation are key drivers of component degradation and catastrophic failure ^{cite{tang2024brief, Ycui_2021_flexible}}, reduced MFU during training workloads does not straightforwardly extend GPU lifespan.

- The analysis focuses exclusively on GPUs when estimating material consumption. While GPUs are clearly the computational core of AI workloads, it is not obvious that they dominate material waste compared to other components required for large-scale AI deployment, such as networking, storage, power delivery, and cooling infrastructure. This point is already addressed by the authors in the conclusions and does not detract from the main contribution, but becomes more relevant in the later sections (from Section 0.8 onward), where broader generalizations are attempted.

Thank you for this comment. We flagged the limitation of our research to GPUs more clearly in section 0.8 and onward, ensuring that readers are aware that the material footprint estimates refer exclusively to GPU hardware in a training environment and that total infrastructure-level material consumption would be substantially higher.

Major comment:

The comparative analysis presented in Section 0.7, and in particular in Figure 6, links benchmark performance improvements to estimated GPU resource requirements in a way that is central to the manuscript's broader narrative on diminishing returns from AI scaling. While the qualitative conclusion (that performance gains are heterogeneous across benchmarks and do not scale uniformly with computational investment) is convincing, the quantitative ratios reported (e.g., statements such as "~30x more resources") deserve more careful treatment. These ratios are derived from point estimates of GPU counts that scale linearly with the assumed Model FLOPs Utilization (MFU), a parameter that is not measured in this study and is known to vary substantially across models, training runs, and software employed.

As a result, relatively modest changes in the assumed MFU for different models (e.g., GPT-3.5 versus GPT-4) can shift the reported GPU ratios by factors of several, potentially altering the quantitative interpretation of the results while leaving the qualitative conclusion unchanged.

Thank you for this careful and fair observation. We agree that the quantitative GPU ratios reported in Section 0.7 deserve more cautious treatment given their dependence on MFU assumptions. We have added an explicit paragraph in the limitations section (very closely to your phrasing) acknowledging that MFU is not directly measured in this study, that modest changes in assumed MFU can shift reported ratios by factors of several, and that the quantitative comparisons should therefore be interpreted as indicative estimates. We note that the qualitative conclusion regarding heterogeneous and non-uniform performance scaling remains robust to these uncertainties.

In its current form, the presentation of single-valued GPU ratios risks conveying a level of precision and robustness that is not fully supported by the underlying first-order model. To mitigate this risk and align the quantitative presentation more closely with the acknowledged limitations of the framework, I recommend incorporating a simple sensitivity analysis with respect to MFU. For example, Figure 6 could display ranges or error bars corresponding to the MFU interval already considered elsewhere in the manuscript (e.g., 20-60%), rather than single point estimates. This modification would not require additional experiments, but would substantially improve the interpretability and robustness of the comparative analysis by making explicit that the reported GPU resource estimates should be understood as ranges rather than precise values.

Thank you for this constructive recommendation. We agree that presenting single-valued GPU ratios without uncertainty bounds risks conveying a level of precision not fully supported by the underlying model. However, incorporating error bars or ranges directly into Figure 6 conflicts with feedback from another reviewer requesting that the analysis foreground the most likely scenarios rather than expanding the range of presented estimates. To reconcile these concerns, we have taken the following approach: Figure 6 is retained with point estimates corresponding to the reported or most plausible MFU assumptions for each model, with a clarifying statement that these represent the most likely scenario rather than precise measurements.

However, to also address your comment, we have added a dedicated sensitivity analysis section that explicitly varies MFU across the 20-60% interval and in a separate one-at-a-time sensitivity analysis varies lifespan while keeping MFU values constant. The Sensitivity Analysis is demonstrating how the reported GPU ratios shift across this range. This approach preserves the interpretability of the main figure while making the sensitivity of the quantitative comparisons to MFU and lifespan assumptions fully transparent, directly addressing the concern that reported ratios should be understood as ranges rather than precise values.

Minor comments

- Line 159: "80GB" → "80 GB".

Thank you, done!

- Table 2: please standardize the number of significant digits across entries.

Thank you, done!

- Table 4: the caption states that values are rounded up to the nearest integer; this does not appear to be true for LLaMa 2 with lifespan = 2 and MFU = 50%.

Thank you, done!

- Table 5: consider representing the elemental composition as a color-coded heat map to allow readers to more immediately identify which elements dominate each GPU component.

Good idea! I changed the color of the 5 most dominant elements per component in the table.

- Table 5: the statement "Elements marked with (*) are either inherently toxic or present significant health hazards" would benefit from supporting references. Please clarify how toxicity is defined or assessed in this context and provide appropriate citations.

Good point! I replaced "inherently toxic" with "classified as hazardous based on well-documented toxic properties," and added citations from the respective section.

- Section 0.6: the manuscript reports many absolute quantities of materials required for training. Providing contextual references (e.g., copper required for training relative to global copper extraction or annual production) would help readers better interpret the magnitude of these numbers.

Thank you for this suggestion. We have thought about this before, and we agree that contextual references can help readers interpret the magnitude of reported material quantities. However, we note that comparing the material footprint of a single training run to global annual copper or metal extraction would yield negligible ratios that could inadvertently minimize the significance of the findings. The more meaningful comparison would be at the industry scale, aggregating across all AI training runs globally, but as noted in the manuscript, this is not possible due to the lack of publicly disclosed GPU shipment data from Nvidia and other hardware providers. We have added a brief sentence making this reasoning explicit, clarifying that single-run comparisons to global extraction figures would not constitute a meaningful benchmark given the scale difference involved.

- The manuscript cites a non-negligible number of arXiv preprints, which is understandable given the rapidly evolving AI/HPC literature and the limited availability of peer-reviewed sources for some training details. However, I recommend checking whether peer-reviewed versions exist and citing them where available. For key assumptions supported primarily by preprints, a brief clarification that these sources function as technical reports or preliminary studies would help contextualize their use.

Thank you for this comment. We have reviewed all arXiv preprint citations and replaced five with their corresponding peer-reviewed publications. The remaining preprints predominantly comprise AI benchmark papers, for which arXiv represents the primary and widely accepted publication venue in the field. A practice driven by the rapid pace of model development that makes traditional peer-review cycles impractical for timely benchmark dissemination. We have added a brief clarifying statement to this effect in the Performance vs. Resource Consumption section.