UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Bridging the gap between industry and academia

Sustainability in LLM-assisted software engineering

Kirkeby, Maja; de Reus, Pepijn; Oprescu, Ana-Maria; Pronk, Kalle; Zhao, Qin; Castor, Fernando; Fernandes, João Paulo

DOI
10.1145/3786148.3788638

Publication date
2026

Document Version
Final published version

Published in
GREENS 2026 : 2026 IEEE/ACM 10th International Workshop on Green and Sustainable Software

License
CC BY

Link to publication

Citation for published version (APA):
Kirkeby, M., de Reus, P., Oprescu, A.-M., Pronk, K., Zhao, Q., Castor, F., & Fernandes, J. P. (2026). Bridging the gap between industry and academia: Sustainability in LLM-assisted software engineering. In *GREENS 2026 : 2026 IEEE/ACM 10th International Workshop on Green and Sustainable Software: 12-18 April 2026, Rio de Janeiro, Brazil : proceedings* (pp. 153-157). Association for Computing Machinery. https://doi.org/10.1145/3786148.3788638

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.
UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:16 Sept 2026

Bridging the gap between industry and academia: sustainability in LLM-assisted software engineering

Maja Kirkeby
Roskilde University
Roskilde, Denmark
majaht@ruc.dk

Pepijn de Reus
University of Amsterdam
Amsterdam, Netherlands
p.dereus@uva.nl

Ana-Maria Oprescu
Universiteit van Amsterdam
Amsterdam, Netherlands
a.m.oprescu@uva.nl

Kalle Pronk
Fontys University of Applied Science
Eindhoven, Netherlands
kallepronk@proton.me

Qin Zhao
Fontys University of Applied Science
Eindhoven, Netherlands
q.zhao@fontys.nl

Fernando Castor
University of Twente
Enschede, Netherlands
f.castor@utwente.nl

João Paulo Fernandes
New York University Abu Dhabi
Abu Dhabi, United Arab Emirates
jpf9731@nyu.edu

Abstract

The advent of Large Language Models in software engineering brings both academic research and industry to the cutting edge of uncharted territory, breaking the typical division of roles, and inviting a new type of symbiosis between academia and industry. This paper draws insights from the SILAS workshop, collocated with ICT4S 2025, which aimed to examine sustainability in the context of software development supported by Large Language Models (LLMs). This paper identifies key gaps between academic and industrial approaches to methods and metrics, accessibility of systems, and efficiency trade-offs. Examples include differences in technology maturity levels, metrics, and assumptions about operational environments in each context. We outline key considerations for shaping research agendas that emphasize standardized, outcome-oriented studies linking reproducibility with real-world sustainability, aiming at achieving greener AI.

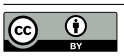
CCS Concepts

• **General and reference** → *Metrics*; **Empirical studies**; • **Hardware** → **Impact on the environment**.

Keywords

Sustainability, Software Engineering, LLMs, Industry, Green AI

ACM Reference Format:
Maja Kirkeby, Pepijn de Reus, Ana-Maria Oprescu, Kalle Pronk, Qin Zhao, Fernando Castor, and João Paulo Fernandes. 2026. Bridging the gap between industry and academia: sustainability in LLM-assisted software engineering. In *10th International Workshop on Green and Sustainable Software (GREENS '26), April 12–18, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3786148.3788638



This work is licensed under a Creative Commons Attribution 4.0 International License.
GREENS '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2381-0/26/04
https://doi.org/10.1145/3786148.3788638

1 Introduction

Artificial Intelligence (AI) has become a central technology in recent years, proving valuable across numerous fields by uncovering patterns in vast datasets [14]. With the introduction of Large Language Models (LLMs) such as GPT-3.5[1], AI quickly reached a broad audience of hundreds of millions of users [10]. Alongside these benefits, AI's rapid advancement has raised societal challenges, particularly regarding the growing energy demands of large-scale models. Calls to reduce AI's environmental impact have intensified, as training and deploying these models contributes significantly to global energy consumption [6, 21]. Companies such as Microsoft and Alphabet have reported steep increases in both energy use and related emissions as a result of AI development [12, 19]. In response, the software engineering community has increasingly focused on defining research agendas, metrics, and practices for environmentally sustainable AI [4].

For software development, specialised LLMs known as code LLMs aid developers with programming, with GitHub Copilot being one of the most popular [16]. These models are fine-tuned with code training data [23] or trained from scratch with a mix of natural and programming language text [13]. As code LLMs become integrated into programming practice, questions about their long-term sustainability gain importance. To address these challenges, we organised a hybrid workshop at the 11^{th} ICT for Sustainability (ICT4S) conference in 2025, focusing on Sustainability in LLM-assisted Software Development (SILAS). During the workshop, participants collaboratively identified challenges through structured discussions, working with Padlet (Figure 1) to capture, group, and refine themes into potential research directions. This paper examines the landscape of LLM-assisted software development from a broad perspective. It presents a research agenda that emerged from the workshop, highlighting the misalignment between academia and industry, and directions for future research.

When considering industry's needs, we should not confuse usefulness with immediacy. Industry prioritises latency, reliability, and

[1] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt

Figure 1: Padlet capturing the discussions of the workshop.

cost, as these are operational bottlenecks. Academia, by contrast, stretches the horizon of what might count as important: fairness across deployments, sustainability metrics, standardised reporting. Taking into account industry needs is not abandoning this horizon, but defining our contributions in ways that industry can imagine adopting. That means thinking about comparability, simplicity, and scale from the outset, even if the first versions of our metrics remain abstract. In this sense, usefulness is not measured only by immediate uptake, but by whether our ideas can mature into a perspective that the industry will one day recognise as indispensable.

We identify three main focus areas, each with a gap between academic research and industry needs: (i) methods & metrics, (ii) accesibility and relevance of systems, and (iii) costs, trade-offs and efficiency. These focus areas structure this paper and capture both the breadth and depth of our workshop discussions. For each area, we examine academic literature, industry needs and priorities, and the tensions that arise between them. By documenting the discussions involing participants from industry and academia during SILAS, we aim to bridge the gap between these two camps and outline research directions that can better align their interests.

2 Methods & Metrics

Academic strength. Academic research has made important progress in measuring the energy footprint of AI systems. At the hardware level, socket-based measurements remain the ground truth, complemented by hardware-supported software estimators such as Intel's RAPL and NVIDIA's NVML. Several studies in software engineering and systems research have highlighted both their value and their limitations, showing that estimator accuracy varies significantly across workloads and hardware platforms [2]. This demonstrates the need for transparency and reproducibility in reporting.

Beyond measurement tools, serving infrastructure also affects efficiency. Serverless inference frameworks use optimizations such as PagedAttention [17] to boost throughput while maintaining latency, showing that energy per token depends on the model and also on its backend implementation [9]. At the algorithmic level, inference strategies such as early exit improve latency–throughput trade-offs and support energy-aware optimization [5].

Benchmarking studies are increasingly evaluating energy efficiency alongside accuracy and latency. Argerich et al. [2] systematically analyze how architecture, batch size, and quantization affect LLM energy efficiency, showing that efficiency is inseparable from deployment design choices. A recent study [1] examines how diverse models perform software development tasks—such as code generation, bug fixing, and documentation—considering both accuracy and energy efficiency. They find that models performing the same task vary in energy efficiency and accuracy, and that individual models behave differently across tasks. More broadly, user-defined benchmarking frameworks are emerging that treat efficiency as one axis of trade-off in practical decision contexts [11].

Industry needs/priority. From the industry perspective, metrics must be more than technically precise: they must also be *relevant* and *comparable* to operational contexts. Current metrics such as Joules per token (or request) rarely capture the end-user experience. A developer using Copilot does not care about the efficiency of a single token; they care about the total energy required to obtain a *useful answer*. If multiple prompt–response cycles are needed before code compiles or a bug is fixed, the real footprint may be two or three times higher than reported by per-token measures. Furthermore, Copilot users do not care about energy specifically, since Copilot is not locally hosted. Instead, they are concerned with the financial cost of using the service. Conversely, for locally-hosted models, they may worry about energy, but only on the long run, never for joules per token or similar metrics. These low level metrics, although important and meaningful to evaluate efficiency, are not relevant for most practitioners.

This mirrors the regulatory logic of the EU's energy labels, which require metrics to be *accurate, relevant, and comparable*. Accuracy ensures measurements are trustworthy; comparability allows distinguishing among alternatives; relevance guarantees that metrics reflects actual use. For LLMs, accuracy and comparability are progressing (validated estimators, benchmark frameworks), but relevance is largely missing.

Considerations. Bridging the gap requires **developing metrics** that connect technical measurement to end-user outcomes. One promising direction is to **extend benchmarks** to capture *energy per useful outcome* such as successful code completion, accepted suggestion, or resolved bug. This would align academic methods with industry needs, where efficiency must be assessed for tasks rather than tokens. Even better would be to **report cost, e.g., in euro, per useful outcome**. This is feasible and, at scale, a more meaningful measure than joules or Wh.

Another direction is **standardization**. Today, energy studies use heterogeneous tools and protocols, making comparisons difficult. Establishing **reproducible evaluation protocols**—specifying how to measure, report, and normalize results across hardware—would provide common ground for academia and industry. Research should expand beyond single-inference benchmarks to include **multi-turn interactions** and workflow components (e.g., retrieval in RAG pipelines), which are central to how LLMs are used in practice.

Evidence of practitioner interest is also visible in **community initiatives** such as Hugging Face's "AI Energy Score," which explicitly aims to provide comparable reporting across models[2]. While not peer-reviewed, such efforts showcase that the demand for standardized, cross-model comparisons exists within the practitioner and open-source communities.

Taken together, these considerations point a research agenda that treats sustainability metrics for LLMs with the same rigor that

[2]https://huggingface.co/ai-energy-score, Accessed: 2025-09-30

policy frameworks such as EU energy labels apply in other domains: **metrics must be accurate, relevant, and comparable**.

3 Accessibility and Relevance of Systems

Academic strength. Most academic work on efficiency assumes an open-system setting, where model weights and architectures are accessible. Within this context, pruning, quantization, and distillation have been shown to substantially reduce computational demands while preserving accuracy [8]. Automated quantization pipelines further demonstrate that such improvements can be applied at scale without extensive manual intervention [3].

Recent advances include post-training pruning methods such as Plug-and-Play, which achieves efficiency gains without retraining [28], and system-level strategies such as energy-aware ensemble model selection, which dynamically balances accuracy against energy costs in production [20]. These methods generally assume full access to model weights and architectures, making them effective for open or research-grade models yet limiting their applicability to the closed, API-based systems that dominate industry use.

Industry needs/priority. In industry, the most widely deployed LLMs (e.g., GPT, Gemini and Claude) are closed systems accessed only through APIs. This makes weight-level optimizations such as pruning or distillation impossible to apply directly, since practitioners have no access to model internals.

At the same time, enterprises rarely use models in isolation. Real deployments embed LLMs into larger workflows: retrieval-augmented generation (RAG) requires indexing, embedding, and vector search; orchestration frameworks chain multiple calls and external tools; user-facing applications such as Copilot or Cursor integrate models into interactive systems. Each of these components consumes additional compute resources and contributes to the overall energy footprint.

Most current evaluations, however, still focus on the bare model in isolation. While such benchmarks provide useful lower bounds, they systematically underestimate the footprint of production systems. To reflect operational reality, evaluation must extend beyond single-model inference and capture the costs of retrieval, orchestration, and repeated calls across full workflows. Such a shift would complement existing academic approaches them with system-level perspectives reflecting how LLMs are actually used in practice.

Considerations. **Bridging the divide** between the assumption of **open models** made by academia and the **closed models** often used in industry requires methods that remain effective without access to **model internals** and that capture the complexity of deployed systems. Promising directions include input–output or system-level optimizations such as **context pruning, efficient retrieval, and orchestration** strategies. Rather than replacing existing model-level work, these approaches extend the efficiency agenda to settings that more closely mirror production.

Signs of convergence are already emerging. Frameworks such as BERGEN [22] benchmark retrieval-augmented generation with efficiency alongside accuracy, reflecting practitioner demand for workflow-relevant evaluation. The next step is to establish **shared, reproducible protocols** that move beyond open-model assumptions and capture the **full system costs of deployment**. By complementing model-level insights with workflow-aware approaches, the community can develop evaluation practices that are both scientifically rigorous and operationally meaningful.

Finally, research explores **how RAG can support software development** tasks [27], and other studies address **its energy efficiency** [25]. Notwithstanding, we are not aware of any paper examining the **intersection** of these two aspects.

4 Costs, Trade-offs, and Efficiency

Academic strength. Academic research has established a solid foundation for understanding and improving AI energy efficiency. For example, Alizadeh et al. [1] show that in software development tasks some models can produce best-of-class results while also being energy-efficient. This suggests that these two attributes do not always need to be treated as a trade-off, as long as the ML task at hand is taken into account.

Recent work broadens this perspective by introducing more holistic sustainability frameworks. Wright et al. [26] argue that efficiency alone is insufficient, noting that improvements in energy or compute efficiency can mask broader environmental trade-offs, including increased demand, resource intensification, and shifts in embodied carbon. They advocate for life cycle assessments that extend beyond training-time energy measurements to include upstream and downstream impacts.

This broader framing is supported by Malmodin et al. [18], who provide empirical data on global ICT energy use, including AI. They show that end-user devices account for more than half of total consumption during operation. Although their analysis is not specific to AI workloads, it demonstrates the importance of considering client-side energy use when evaluating system-wide efficiency. These results challenge the current academic focus on server-side or training-centric assessments and emphasize the need for system-level and usage-aware evaluation methods.

To keep academic studies relevant to deployment realities, benchmarks should emulate production-scale conditions. Real-world inference involves sustained, high-throughput workloads where concurrent requests, GPU saturation, and memory scheduling strongly affect power and latency. Frameworks such as vLLM [17] offer an important step toward reproducible, yet realistic evaluation by enabling efficient request handling and hardware use.

Industry needs/priorities. Industry evaluations of AI energy use increasingly emphasize operational performance and system-level efficiency. In a recent large-scale study, Elsworth et al. [7] present a methodology for measuring the environmental impact of delivering AI services at Google scale. Their framework uses full-stack measurement boundaries that include accelerator energy use as well as host systems, idle infrastructure, and datacenter overhead. This marks a shift toward assessing the environmental cost of inference at the point of use rather than focusing solely on training.

To capture output utility, Elsworth et al. [7] introduce the Arena score—a large-scale, Elo-style metric comparing generated responses. This enables comparison of energy and emissions per unit of output quality. The authors report a 33-fold reduction in energy use for the median Gemini prompt within one year, achieved through

integrated software, hardware, and infrastructure optimizations reflecting a commitment to sustainability-focused design.

The focus on quality-per-prompt efficiency reflects a shift toward outcome-oriented metrics that emphasize deployment-phase costs and user-perceived performance. However, the study explicitly delimits the system boundary to exclude end-user energy consumption, which—as noted in academic work—can constitute a substantial portion of the overall footprint. Aligning the measurement boundaries used in academic research with those adopted in industry remains essential for comparability and to ensure that efficiency gains translate meaningfully to system-level sustainability.

Considerations. While industry has advanced system-level measurement and deployment-oriented optimization, the conceptual basis for meaningful evaluation still stems from academic and interdisciplinary research. Contributions such as Wright et al. [26] and Malmodin et al. [18] provide essential framing for interpreting energy efficiency beyond narrow hardware or training scopes. They highlight the need for **metrics that encompass full life-cycle** impact, including **client-side energy** and **rebound effects**. However, these perspectives have yet to be fully operationalized.

In contrast, recent industry efforts have implemented pragmatic measurement boundaries, such as energy per prompt, tied to output quality via metrics like the Arena score [7]. These metrics represent a concrete, **quality-of-outcome-oriented framework** that aligns with deployment realities. To ensure that energy optimization strategies are effective and comparable, academic work should consider fair quality and energy mixing score. Research can thus contribute to improvement strategies grounded in **realistic performance baselines** while retaining the critical perspective necessary to evaluate their broader environmental and societal implications.

To ensure that energy-performance evaluations **generalise beyond lab conditions**, academic studies must emulate the operational environments of industry deployments. This includes running benchmarks **under high-throughput scenarios**, where **shared infrastructure**, **GPU saturation**, and **memory scheduling** significantly affect energy consumption. To improve representativeness, inference servers such as vLLM [17], which replicate **production-like request handling**, are preferable to single-request setups using libraries like HuggingFace Transformers.

5 Discussion

Academic and industrial communities approach AI sustainability from complementary perspectives: academia develops foundational, reproducible methods, while industry optimizes large-scale systems under real-world constraints. We synthesize the resulting tensions across the workshop's three focus areas.

Challenge 1: TRL-driven divergence in goals and scope. These differences in focus and context can be viewed through the lens of Technology Readiness Levels (TRLs)[3]. Academic work typically falls within TRL 3–5, where ideas are developed and tested in controlled settings, whereas industry operates at TRL 7–9, integrating methods into large-scale, production-grade systems. This variation helps explain differences in metrics, optimization strategies, and system boundaries seen across research and practice.

[3]https://ec.europa.eu/research/participants/data/ref/h2020/other/wp/2016-2017/annexes/h2020-wp1617-annex-ga_en.pdf

Challenge 2: Bridging model-level methods and deployment. Academic research contributes mainly at the conceptual and methodological levels. Techniques such as pruning, quantization, and compression [24] yield generalisable improvements in model-level efficiency, typically validated under controlled conditions. This places them at TRL 3–5: effective in constrained scenarios but not yet deployed in real-world systems. Likewise, recent calls for holistic sustainability assessments [18, 26] expand the scope of inquiry but remain early in practical adoption. These limitations are particularly salient when dominant LLM deployments are closed and accessed via APIs, which restricts access to model internals (Section 3). In industry, inference efficiency is pursued under production constraints at scale. For example, Elsworth et al. [7] report energy and emissions per user prompt within a full-stack measurement boundary. Their use of the *Arena score* to relate output quality to resource consumption reflects TRL 9 deployment maturity: outcome-oriented, data-driven, and embedded in operational systems.

Challenge 3: Bridging evaluation frameworks. The disconnect between academic and industrial efforts, e.g., Sections 2 and 3, does not appear to stem from a classic "valley of death" scenario [15], where promising research fails to translate due to a lack of resources or infrastructure. On the contrary, both communities have continued to advance independently—industry by deploying outcome-driven efficiency metrics and optimization pipelines at scale [7], and academia by deepening methodological understanding and developing reproducible frameworks [24, 26]. This suggests that the divergence is not one of inaction, but rather of differing priorities and scopes of evaluation. The key challenge is therefore not failed translation from lower to higher TRLs, but enabling shared frameworks that connect scientific rigor with real-world utility and support principled evaluation of energy efficiency across the AI lifecycle. Between these extremes, we observe an emerging middle ground. Approaches such as realistic benchmarking with vLLM [17] or input-level optimizations like prompt restructuring and context pruning reflect an upward trajectory from mid-TRL academic work. These methods retain the transparency and replicability valued in research while addressing production constraints like concurrency, latency, and GPU saturation. They offer promising avenues for advancing research toward deployable solutions without requiring privileged access to closed-source models.

Challenge 4: Completeness vs. feasibility of measurement boundaries. This tension is already visible in the trade-offs discussed in Section 4. Elsworth et al. [7] report energy and emissions per user prompt within a full-stack measurement boundary, but deliberately exclude end-user devices—highlighting the tension between completeness and feasibility in high-TRL environments. Such boundary choices shape what is counted as "efficiency": they affect comparability across studies and can shift the apparent location of energy consumption to components outside the reported system. Because boundary choices determine where impacts are accounted for, they also raise a fairness question: where and by whom is energy consumed? A substantial share of ICT energy use occurs on the client side [18], beyond the boundaries typically considered in industrial reporting. When efficiency metrics exclude end-user devices, the environmental burden risks being displaced rather than reduced, e.g., externalized to regions with higher carbon intensity.

In light of these findings, academic work remains essential—not only in safeguarding the validity and reproducibility of efficiency claims, but in ensuring that what is optimised truly reflects a fair and sustainable distribution of environmental impact.

6 Conclusion

This paper has explored the multifaceted landscape of sustainable AI, focusing on the measurement, optimization, and evaluation of energy efficiency across both academic research and industrial deployment. We have shown that while academia and industry operate under different priorities and constraints, they contribute in complementary ways to the sustainability of AI systems. Academic research excels at developing foundational methods and reproducible frameworks for model-level efficiency, while industry leads in deploying full-stack solutions that optimize performance and environmental cost under real-world usage conditions.

These efforts often differ in scope and assumptions, particularly in how system boundaries are drawn and outcomes are evaluated. Bridging this gap requires metrics that make sense to both sides, consideration of relevant trade-offs, and rigor and transparency in resource estimation, so as to ensure accountability and fairness.

Discussions during the workshop made it evident that this topic is broader than software development and reaches towards other aspects of AI sustainability. To advance the field, we call for stronger collaboration between communities and the development of shared frameworks that bridge scientific insight with operational relevance. In 2026, we are planning to organize a follow-up workshop with this broader scope in mind. Only by aligning methods, metrics, and goals across the full lifecycle of AI—from model design to user interaction—can we ensure that the drive for efficiency results in long-term environmental and societal benefit.

Acknowledgments

This work was partially supported by GreenDIGIT, an European Union project funded under the grant agreement 101131207 and by the Independent Research Fund Denmark Project no. 2102-00281B.

References

[1] Negar Alizadeh, Boris Belchev, Nishant Saurabh, Patricia Kelbert, and Fernando Castor. 2025. Language models in software development tasks: An experimental analysis of energy and accuracy. In *2025 IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR)*. IEEE, 725–736.

[2] Mauricio Fadel Argerich and Marta Patiño-Martínez. 2024. Measuring and Improving the Energy Efficiency of Large Language Models Inference. *IEEE Access* 12 (2024), 80194–80207. doi:10.1109/ACCESS.2024.3409745

[3] Yuwei Cai et al. 2020. ZeroQ: A Novel Zero-Shot Quantization Framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13169–13178. doi:10.1109/CVPR42600.2020.01319

[4] Luís Cruz, João Paulo Fernandes, Maja H Kirkeby, Silverio Martínez-Fernández, June Sallou, Hina Anwar, Enrique Barba Roque, Justus Bogner, Joel Castaño, Fernando Castor, et al. 2025. Greening ai-enabled systems with software engineering: A research agenda for environmentally sustainable ai practices. *ACM SIGSOFT Software Engineering Notes* 50, 3 (2025), 14–23.

[5] Yinwei Dai, Rui Pan, Anand Iyer, Kai Li, and Ravi Netravali. 2024. Apparate: Rethinking Early Exits to Tame Latency-Throughput Tensions in ML Serving. In *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles* (Austin, TX, USA) *(SOSP '24)*. ACM, New York, NY, USA, 607–623. doi:10.1145/3694715.3695963

[6] Alex de Vries. 2023. The growing energy footprint of artificial intelligence. *Joule* 7, 10 (Oct. 2023), 2191–2194. doi:10.1016/j.joule.2023.09.004

[7] Cooper Elsworth, Keguo Huang, David Patterson, Ian Schneider, Robert Sedivy, Savannah Goodman, Ben Townsend, Parthasarathy Ranganathan, Jeff Dean, Amin Vahdat, Ben Gomes, and James Manyika. 2025. Measuring the environmental impact of delivering AI at Google Scale. https://arxiv.org/abs/2508.15734

[8] Elias Frantar and Dan Alistarh. 2023. SparseGPT: massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. Article 414.

[9] Yao Fu, Leyang Xue, Yeqi Huang, Andrei-Octavian Brabete, Dmitrii Ustiugov, Yuvraj Patel, and Luo Mai. 2024. ServerlessLLM: low-latency serverless inference for large language models. In *Proceedings of the 18th USENIX Conference on Operating Systems Design and Implementation* (Santa Clara, CA, USA) *(OSDI'24)*. USENIX Association, USA, Article 8.

[10] Marzyeh Ghassemi, Abeba Birhane, Mushtaq Bilal, Siddharth Kankaria, Claire Malone, Ethan Mollick, and Francisco Tustumi. 2023. ChatGPT one year on: who is using it, how and why? *Nature* 624, 7990 (Dec. 2023), 39–41. doi:10.1038/d41586-023-03798-6

[11] Ana Gjorgjevikj, Ana Nikolikj, Barbara Koroušić Seljak, and Tome Eftimov. 2025. User-defined trade-offs in LLM benchmarking: balancing accuracy, scale, and sustainability. *Knowledge-Based Systems* 330 (2025), 114405.

[12] Google Sustainability. 2024. Environmental Report. https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf

[13] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. DeepSeek-Coder: When the Large Language Model Meets Programming – The Rise of Code Intelligence. https://arxiv.org/abs/2401.14196

[14] María Gutiérrez, Coral Calero, Félix García, and Mª Ángeles Moraga. 2024. The Effects of Class Balance on the Training Energy Consumption of Logistic Regression Models. In *Research Challenges in Information Science*. Springer Nature Switzerland, Cham, 324–337. doi:10.1007/978-3-031-59465-6_20

[15] Mihály Héder. 2017. From NASA to EU: the evolution of the TRL scale in Public Sector Innovation. *The Innovation Journal* 22, 2 (2017), 1–23.

[16] Eirini Kalliamvakou. 2022. Quantifying GitHub Copilot's impact on developer productivity and happiness. https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness/

[17] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C.H. Yu, J.E. Gonzalez, H. Zhang, and I. Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*. doi:10.1145/3600006.3613165

[18] Jens Malmodin, Nina Lövehagen, Pernilla Bergmark, and Dag Lundén. 2024. ICT sector electricity consumption and greenhouse gas emissions – 2020 outcome. *Telecommunications Policy* 48, 3 (2024), 102701.

[19] Microsoft. 2024. *Environmental Sustainability Report*. Sustainability Report. Microsoft. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1lhhu

[20] Nienke Nijkamp, June Sallou, Niels van der Heijden, and Luís Cruz. 2024. Green AI in Action: Strategic Model Selection for Ensembles in Production. In *Proceedings of the 1st ACM International Conference on AI-Powered Software (AIware 2024)*. ACM, New York, NY, USA, 50–58. doi:10.1145/3664646.3664763

[21] The Shift Project. 2023. Energy, climate: which virtual words for which real world? (2023), 73. https://theshiftproject.org/wp-content/uploads/2023/12/The-Shift-Project-Quels-mondes-virtuels-pour-quel-monde-reel-Rapport-intermediaire-2023-002.pdf

[22] David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. 2024. BERGEN: A Benchmarking Library for Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 7640–7663. doi:10.18653/v1/2024.findings-emnlp.449

[23] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code Llama: Open Foundation Models for Code. arXiv:2308.12950 [cs.CL] https://arxiv.org/abs/2308.12950

[24] Roberto Verdecchia, June Sallou, and Luís Cruz. 2023. A systematic review of Green AI. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 13, 4 (2023), e1507.

[25] Konstantinos Vrettos and Michail E. Klontzas. 2025. Accurate and Energy Efficient: Local Retrieval-Augmented Generation Models Outperform Commercial Large Language Models in Medical Tasks. https://arxiv.org/abs/2506.20009

[26] Dustin Wright, Christian Igel, Gabrielle Samuel, and Raghavendra Selvan. 2025. Efficiency Is Not Enough: A Critical Perspective on Environmentally Sustainable AI. *Commun. ACM* 68, 7 (June 2025), 62–69. doi:10.1145/3724500

[27] Zezhou Yang, Sirong Chen, Cuiyun Gao, Zhenhao Li, Xing Hu, Kui Liu, and Xin Xia. 2025. An Empirical Study of Retrieval-Augmented Code Generation: Challenges and Opportunities. *ACM Trans. Softw. Eng. Methodol.* 34, 7, Article 188 (Aug. 2025), 28 pages. doi:10.1145/3717061

[28] Yingtao Zhang, Haoli Bai, Haokun Lin, Jialin Zhao, Lu Hou, and Carlo Vittorio Cannistraci. 2024. Plug-and-Play: An Efficient Post-training Pruning Method for Large Language Models. In *The Twelfth International Conference on Learning Representations*.