Resources, Conservation & Recycling 226 (2026) 108670

Contents lists available at ScienceDirect

Resources, Conservation & Recycling

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

Full length article

Assessing the carbon footprint of language models: Towards sustainability in AI☆

Fleur Jeanquartier [a,f,1], Claire Jean-Quartier [a,c,1], Paul Rieder [d,1], Vedad Misirlić [f,1], Christian Pasero [d,1], Richard Hohensinner [e], Heimo Müller [b], Andreas Holzinger [a,f,b,*]

[a] Human-Centered AI Lab, Institute of Forest Engineering, Department of Ecosystem Management, Climate and Biodiversity, BOKU University Vienna, Austria
[b] Information Science and Machine Learning Group, Diagnostic and Research Institute of Pathology, Medical University Graz, Austria
[c] Research Data Management, Graz University of Technology, Austria
[d] Institute of Software Technology, Graz University of Technology, Austria
[e] Institute of Machine Learning and Neural Computation, Graz University of Technology, Austria
[f] Institute of Human-Centered Computing, Graz University of Technology, Austria

ARTICLE INFO

Dataset link: https://github.com/VedadTUG/HCI-SLM

Keywords:
Sustainability
Carbon footprint
Energy efficiency
Small language models
Generative AI

ABSTRACT

As language models gain prominence for their generative capabilities, their growing carbon footprint must be critically addressed in the context of the climate crisis. This paper aims to increase transparency by showcasing a comparison of emissions from training – particularly of small language models – and inference. We therefore scan existing benchmark data, investigate two representative models, TinyLlama and nanoGPT, evaluating energy consumption during training and task performance. We also reflect upon the specificity of use cases, model architecture, hardware choices, and their influence on efficiency and sustainability. Our findings indicate that existing benchmarks and publications rarely report energy consumption, creating a significant information gap, and urging for a harmonized evaluation framework that integrates standardized sustainability aspects. Despite this, small language models show potential for selected application scenarios where resource efficiency is key. To address the challenges of fair and sustainable AI, we emphasize the importance of ongoing documentation efforts. We also encourage model developers and providers to communicate energy usage data more openly. Transparent reporting supports responsible model selection and helps align AI development with climate-conscious technology practices.

1. Introduction

The next generation of language models may be small, both in industry (Kumar et al., 2025) and in science (Garg et al., 2025; Kim et al., 2025). With the digital transformation of various industries, even those that have been rather traditional industries are increasingly being transferred to Industry 5.0 (Turner et al., 2022; Holzinger et al., 2024b,a); these industries are more and more dependent on the success of artificial intelligence (AI) in business (Xiong et al., 2020).

Large language models (LLMs) are still growing in popularity across all industry verticals, owing to their unparalleled performance in diverse applications. As LLMs play an increasingly important role in industry and even in our daily lives, it is becoming more and more important to evaluate them not only at the task level but also at the societal level to gain a better understanding of their potential risks. In recent years, significant efforts have been made to study LLMs from a variety of diverse perspectives (Zhou et al., 2024; Chang et al., 2024; Ehrlich-Sommer et al., 2025; Kocic et al., 2025; Kraišniković et al., 2025). As the deployment of such LLMs accelerates globally, their significant environmental impact poses an urgent challenge in the context of the escalating climate crisis.

This is enormously intensifying the urgent need to understand and mitigate its carbon emissions and green innovation (Wang et al., 2023). Traditionally, sustainable operations management has grappled with inefficiencies in energy usage and resource allocation, compounded by rapidly evolving digital technologies and complex supply chain dynamics. Advanced technological developments, including AI, blockchain, cloud computing, big data, and IoT, have emerged as promising avenues to enhance sustainability across sectors by improving transparency, energy efficiency, and resource optimization. Within this

☆ This article is part of a Special issue entitled: 'ADV_TECH' published in Resources, Conservation & Recycling.

* Corresponding author at: Human-Centered AI Lab, Institute of Forest Engineering, Department of Ecosystem Management, Climate and Biodiversity, BOKU University Vienna, Austria.
E-mail address: andreas.holzinger@boku.ac.at (A. Holzinger).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.resconrec.2025.108670
Received 3 April 2025; Received in revised form 5 October 2025; Accepted 23 October 2025
Available online 31 October 2025

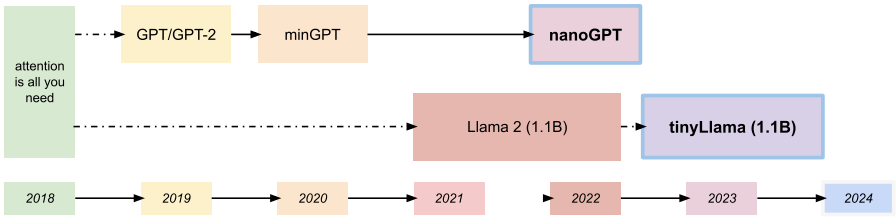

Fig. 1. Model version timeline (selected ones in bold font with lightblue border).

broader context, this paper specifically investigates the carbon footprint associated with language models, emphasizing smaller variants such as TinyLlama and nanoGPT. By analyzing energy usage during both training and inference phases, performance efficacy, and hardware dependencies, our research addresses critical gaps in standardized benchmarking and advocates for greater transparency. Our findings underline that while small language models offer valuable pathways towards environmentally responsible AI usage, comprehensive and transparent reporting standards are essential to achieve genuinely sustainable technological innovations.

Since the use of AI and its workhorse machine learning (ML) is rising at an almost exponential pace, it is imperative to keep up with the resultant challenges (Azhar, 2021). One such example is to understand how the training and the use of computationally expensive models impact the environment by the interrelated CO_2 release in the atmosphere (Gupta et al., 2021). Recent studies have targeted the topic of language models and their carbon footprint (Ren et al., 2024; Kleinig et al., 2024; Liu and Yin, 2024; Rehman et al., 2025).

In general, large language models (LLMs) like `ChatGPT` have been reflected upon in numerous recent studies and further have been employed not only for translation but alienated tasks such as data processing, hypothesis generation, and outreach, thereby attracting attention from academia, the economic and public sector as well as society in general (Ray, 2023; Raiaan et al., 2024) Ahn and Kim (2025). In contrast, this study focuses on the topic of small language models (SLMs) which are overlooked in the shadow of their large counterparts (Kim et al., 2024; Schick and Schütze, 2020; Lu et al., 2024). SLMs range from a few up to 5 billion (5B) parameters (Lu et al., 2024). Limitations of LLMs in regard to large parameters and computational costs, but also privacy concerns, have been proposed to be circumvented through SLMs (Wang et al., 2024c).

Representatively, we chose `NanoGPT` (nanoGPT, 2024) and `TinyLlama` (Zhang et al., 2024) to study energy consumption. Fig. 1 summarizes historic versions up to the selected two models, highlighted in bold.

NanoGPT is a rewrite of minGPT which is based on GPT-2 (Nogales Pérez, 2023; Ganescu and Passerat-Palmbach, 2024; Radford et al., 2019), TinyLlama is a reduced derivative of Llama2 LLM (Zhang et al., 2024; Touvron et al., 2023). There are also many other small language models, many of them suit very specific tasks and are fine-tuned on specific datasets, but are often build on famous models like Llama (f.i. MedAlpaca, Med-Pal etc.) or GPT (e.g. SmallDisMed etc.) correspondingly (Schick and Schütze, 2020; Lu et al., 2024).

Motivation

This study is intended to consider the future of SLMs in regard to energy consumption in the view of possible application scenarios. The paper puts a controversial focus on language models with regard to the current climate crisis and highlights associated gaps in current evaluation practices. Frameworks for measuring carbon emissions have been developed and used for evaluating AI models (Budennyy et al., 2022; Rodriguez et al., 2024; Jean-Quartier et al., 2023). These strategies need to be aligned and considered as a harmonized reference system for model sustainability on a policy-level (Appelle, 2023; Luccioni et al., 2024).

2. Background and related work

In this section, we present studies on the possibilities of measuring and estimating the energy consumption of language models and provide an overview of other aspects related to small modeling applications. We summarize frameworks applicable to evaluate SLMs in consideration of energy consumption.

2.1. Basics of energy consumption

With Language Models consuming large amounts of energy (Wells, 2023), concerns about the long-term feasibility of these models are rising (Luccioni et al., 2022). Watts are a measure of power required for a particular device to run. In contrast, watt-hours measure the energy consumption of a particular device over a defined period of time (Watts, 2024).

Services account for 31% of energy consumption within the EU (Eurostat, 2023). With the market for language models growing, data centers are already accounting for over 2% of annual energy consumption within the United States (Power, 2023). With language models connected to data centers, this number is only growing, showing that language models already account for a large portion of energy consumption.

2.2. Selection of language models

Considering that billions of parameters go into training a Language Model, the distinction between models is usually made based on the number of parameters considered when training the model. Language models are gaining increasing traction, and people use them for many different things. The main usages that we see right now are as virtual assistants (LLM, 2024), content-creation and Search Engine Optimization (SEO) (Vajrobol et al., 2024), translation of text (Kocmi and Federmann, 2023), (financial) sentiment analysis (Araci, 2019), as well as educational tools (Kasneci et al., 2023). SLMs compared to LLMs use fewer parameters but due to finetuning may be as performant as larger counterparts for specific tasks (Schick and Schütze, 2020).
The main difference between LLMs and SLMs is their size. Size relates to fewer parameters, faster training and fine-tuning time, narrower tasks, less resource requirements, eventually more accessibility and sustainability. Some use cases for SLM applications are outlined in Fig. 1. The exemplary application scenarios are categorized into three partly overlapping groups. One category summarizes the utilization of SLMs due to the availability of scarce data only for training and for few-shot learning, such as it is the case for low resource languages (Hasan et al., 2024; Tonja et al., 2024), task-oriented dialog (Budzianowski and Vulić, 2019), and expert language translation (Keles et al., 2024; Buhnila, 2025). SLMs can also be preferably applied in case of time dependencies, e.g. in the educational context (Wang et al., 2024c; Rashid et al., 2024), and domain adaption (Kang et al., 2023). Resource-constrained settings are given in the case of offline robots or embodied agents (Zhu et al., 2024; Erdogan et al., 2024; Islam et al., 2024), but could also be the case for e.g. medical devices, which have to be used offline due to data privacy (Wang et al., 2024a). Some examples partly overlap in this categorization and several factors would lead to the

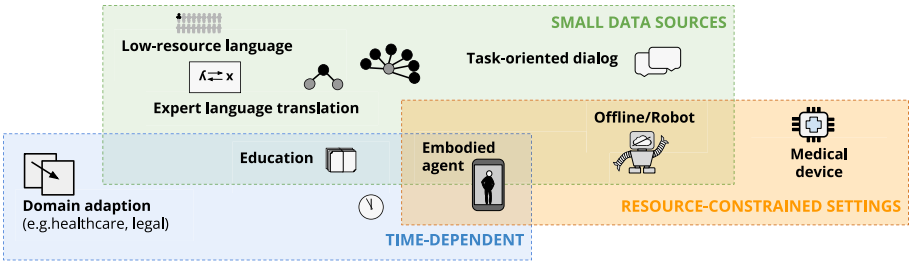

Fig. 2. SLM use cases for application scenarios: categorized into the group of small data sources (green), time-dependency (blue), and resource-constrained settings (orang), with overlapping examples.

choice of applying a SLM over a LLM. SLMs benefit from resource efficiency, being cost-effective, perform better with more straightforward tasks, and can also run locally on everyday devices. On the other hand, LLMs are superior in the context of complex diverse tasks and queries, handling complexity versus ambiguity, offer more versatility, as well as generally have a better response quality due to size and complexity of training data (Lu et al., 2024; Schick and Schütze, 2020) (see Fig. 2).

2.3. Tools to estimate the carbon footprint

Currently, only a few tools are available to predict the CO_2 consumption before training a LLM, namely `mlco2`, `LLMCarbon` (Faiz et al., 2023; Xu et al., 2025) and `OpenCarbonEval` (Yu et al., 2024).

`Mlco2` f.i. achieves this by taking your Hardware into account (GPU or CPU), the hours during which one wants to train its data, the provider where you train the data (e.g., Google Cloud Platform) and the region where one lives (e.g., europe-west1). Although the tool is not 100 percent accurate, researchers can get a good overview of how much energy consumption can be expected. Even though we did not use energy consumption prediction, `mlco2` also gives access to a Python library called `CodeCarbon` to measure energy consumption during training (Lacoste et al., 2019).

2.4. Estimating the carbon footprint of training LMs

As research and applications of AI are on the rise, it is important for both researchers as well as developers to be transparent about how much energy was consumed during training. While Luccioni et al. report on environment costs for training a model called BLOOM, they also elaborate on the challenges of measuring carbon footprints precisely (Luccioni et al., 2022). Another example that discusses the transparency and sustainability of model development is from Jean-Jean-Quartier et al. (2023). In this study, frameworks for measuring the carbon emissions while training models were tested, focusing algorithmic energy consumption during model optimization and reflecting on the incorporation of explainability methods. From another point of view, there is ongoing research on estimating the carbon footprint of model training in advance (Faiz et al., 2023). Such studies try to incorporate the implication of selected hardware settings used for model training. We will experimentally examine this aspect during training the selected models.

2.5. Evaluation of SLMs

A classification system for SLM capabilities has been recently presented (Sakib et al., 2025). Still, there is no comprehensive evaluation framework focusing on the tradeoff between SLM performance and energy efficiency in regard to application scenario (Ding et al., 2025). Ecological and social – including economic – aspects of language models could extend the comprehensible view on sustainability (Luccioni et al., 2024). Moreover, there are challenges in standardizing model energy metrics such as lifecycle (Miraghaei et al., 2025), per-model granularity, data movement energy, or platform variations (Chowdhury et al., 2025). Various evaluations frameworks have been introduced and used in SLM studies, involving general benchmarks such as Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020) (with its enhanced MMLU-Pro (Wang et al., 2024b)), or GSM8K (Cobbe et al., 2021), with varying emphasis on knowledge and reading-comprehension, reasoning and logic, common sense and social understanding, code generation, and multitask evaluation (Corradini et al., 2025).
Only very recently, SLMbench was proposed to take accuracy, computational efficiency, and sustainability metrics into account (Thanh Pham et al., 2025).

3. Methods

We used different setups for `nanoGPT` and `tinyLlama` described in Tables A.5 and A.6. Both SLMs, `TinyLlama` and `nanoGPT`, are compared to other SLMs considered energy efficient and still accurate (Lee, 2024; Lu et al., 2024; Ganescu and Passerat-Palmbach, 2024). For our research we also weighted good documentation and a suitable integration in the test bench. We selected the "tinyshakespeare" dataset for its manageable size, enabling reproducibility and clear benchmark results. The selected SLMs also had to run seamlessly on all the testing setups, without the need to change the entire codebase. There are fine-tuned versions of `TinyLlama` available called `medAlpaca` (Han et al., 2023), which focus on medical questions only, thus, they were not further tested. Additional models such as `Stable LM` (StableLM, 2024), `Mini CPM` (Hu et al., 2024), have been discarded of extended testing due to the given criteria above. Only `TinyLlama` and `nanoGPT` have been chosen for the experiment described in this manuscript. This manuscript further reports on tests on different devices, including different CPUs from `AMD`, `Intel`, and the `M1` processor. We also use Nvidia GPUs with the Cuda library to speed up the training process and observe the produced carbon footprint. We finally compare the measurements and observe which SLM produces less- and which produces more carbon emissions. Source code for our tests is available on https://github.com/VedadTUG/HCI-SLM.

3.1. Before the experiment

Three questions needed to be answered before we were able to start our experiment. First, we conducted the search for fitting SLMs for later evaluation. Second, we chose a framework to compute the energy consumption during training and querying. Third, we needed to find a suitable data set for the training procedure and an appropriate query for testing.

3.2. Model selection

Manifold of SLMs have been developed and are available in the web (Wang et al., 2024c; Nucci, 2024; Lu et al., 2024). We chose a smaller - `nanoGPT` - and a bigger model - `TinyLlama`, outlined beneath.

Model 1: nanoGPT. `nanoGPT` is one of the simplest and fastest repositories for training and fine-tuning medium-sized GPTs. It is based on `minGPT` (minGPT, 2024) but is far from finished because development is still ongoing. When executing the *train.py* file, `nanoGPT` reproduces GPT-2 with 124M parameters (nanoGPT, 2024). The version from Jun 3, 2024 was used for the experiment.

Model 2: TinyLlama. `TinyLlama` is one of the biggest SLMs as it aims to train 1.1B parameters on 3 trillion tokens. The architecture of `TinyLlama` is based on Llama 2, which is its LLM equivalent. Yet, because `TinyLlama` is effectively so small, it can run on nearly any hardware (Zhang et al., 2024). At the time when the experiment took place, the model version TinyLlama-1.1B-intermediate-step-240k-503b was selected.

3.2.1. Measuring energy consumption

The library `CodeCarbon` has been identified (Jean-Quartier et al., 2023) to be usable for measuring the energy consumption both during training and running a model. `CodeCarbon` is part of the `mlco2` package that can predict the carbon emissions of model training even before the start of the procedure. `CodeCarbon` is an open-source python package that can be seamlessly integrated into the codebase (CodeCarbon, 2024).

3.2.2. Data selection

As dataset for training the different models we chose a curated subset of Shakespeare's works, available on the nanoGPT Github project (nanoGPT, 2024), labeled as "tinyshakespeare".

3.3. Setup of models

In this section, we describe our model setup and the problems we faced during the setup.

`nanoGPT` is a breeze to set up. The documentation is easy to understand featuring code examples also on how to train the model. `CodeCarbon` could be integrated easily.

`TinyLlama` can be downloaded and accessed through the `HuggingFace` API. It runs in Linux environments only because of the bits and bytes library used to finetune its parameters. If operated on Windows, the WSL Linux subsystem has to be installed, which does not work on a MacBook. Still, `TinyLlama` can be run in cloud-based environments such as Google Colab. `CodeCarbon` could also easily be integrated thereafter.

`Codecarbon` features a clear and understandable documentation. We encountered only one minor issue when trying to write energy consumption output into a `.txt` file. This was achieved by using the Python logging library that `CodeCarbon` supports. Therefore, setting up a logger, formatting and finally handing over the logger to the `CodeCarbon` tracker during its' initialization. Subsequently, the terminal's output was saved into a `.log` file.

4. Results

The following section shows how much energy was consumed during the training and query execution and how long it took. For both SLMs, we created two tables, one for the training and one for the query execution.

Tables 1, 2, 3, 4 show training and query results for the respective model.

For completeness, as indicated in Table A.5, Tables 3 and 4 do not include an M1 column. This was due to setup issues for respective model on respective hardware. Each table presents energy consumption values taken from codecarbon's output. Both energy consumed for RAM, GPU and CPU as well as the total energy is given in kilowatthours (kWh). Total energy is the electricity used since the beginning of the respective computational task.

Table 1 presents measurements during training NanoGPT. Training on the M1 hardware performed most efficiently in the sense of least amount of energy consumed, on the other hand performed least in regard to time taken. Training with Nvidia 2060 Super needed most energy but was second-but-slowest with 47 min. According to expectations, fastest training was obtained with Nvidia 3060Ti, however, most climate friendly was training with M1.

Table 3 presents time and energy consumption for training TinyLlama. Interestingly, in contrast to Table 1, Nvidia 2080 superseded Nvidia 3060Ti with regard to least time taken, however, computation consumed the most energy in the comparison. Least energy consumption was again the slowest computation, in this case Nvidia 3060Ti.

Table 2 further shows computation time and energy consumption for querying NanoGPT. M1 hardware consumed the least amount of energy with only 0.000232 kWh in total, and needed most amount of time with 2 min 35 s, analogical to NanoGPT training time and energy costs. Also, querying with Nvidia 3060Ti was again, as during the training, fastest but not the most energy consuming hardware. The latter was with Nvidia 2080.

Finally, results from querying TinyLlama are shown in Table 4. Querying with Nvidia 2080 is fastest with 6 min 14 s response time, while consuming also most energy, namely 0.011466 kWh. Least energy but most time needed the TinyLlama example querying with Nividia 3060Ti.

NanoGPT performs faster than TinyLlama in both training as well as querying. NanoGPT Training time shows that training on a high performance GPU is faster than on low cost GPU and CPU but also needs more energy.

5. Discussion

Given the results from the initial experiments, presented in Section 4, fast training using a modern GPU does not imperatively provide the most climate friendly approach. However, future experiments should be aligned to a currently unavailable extended evaluation framework integrating sustainability benchmarks of AI models in general and SLMs in particular. The following subsections detail concomitant aspects of SLM experimentations.

5.1. Are SLMs more energy efficient than LLMs?

We used the larger TinyLlama model in comparison to the smaller nanoGPT model. Looking further ahead, comparing LLMs to SLMs, the latter are more efficient in the context of power consumption (Wells, 2023). However, this efficiency usually comes at the cost of reduced accuracy and relevance of the response. Still, the carbon footprint should be considered when opting for a solution to complete a given task. Moreover, we need to understand and weigh the advantages and disadvantages of available models. In the context of a language model as a tool to support a certain action, the computational effort of its development, its implementation, and its application has to be taken into account. In this regard, smaller tasks which do not require numerous tokens and parameters can be considered as the superior option in the context of effectiveness, sustainability and profitability. Use cases must be well-defined and specific. Energy consumption may also rise upon utilization of larger datasets and the increased iteration counts, but not necessarily provide better quality. The latter also depends on the dataset's representativeness and optimization strategies (Khan et al., 2025; Rusyn et al., 2024; Argerich and Patiño-Martínez, 2024).

Table 1
Training NanoGPT (Energy in kWh).

Training	Nvidia 2060 Super	Nvidia 2080	Nvidia 3060Ti	M1
Energy consumed RAM	0.001487	0.001190	0.001110	0.004202
Energy consumed GPU	0.085424	0.074114	0.032946	0.000828
Energy consumed CPU	0.016661	0.019540	0.006273	0.011631
Total energy	0.103572	0.094844	0.040329	0.016661
Time taken	47 min 31 s	24 min 26 s	11 min 20 s	2 h 23 min 34 s

Table 2
Query NanoGPT (Energy in kWh).

Query	Nvidia 2060 Super	Nvidia 2080	Nvidia 3060Ti	M1
Energy consumed RAM	0.000031	0.000028	0.000043	0.000083
Energy consumed GPU	0.000807	0.000853	0.000282	0.000109
Energy consumed CPU	0.000343	0.000465	0.000241	0.000041
Total energy	0.001181	0.001347	0.000566	0.000232
Time taken	23 s	20 s	12 s	2 min 35 s

Table 3
Training TinyLlama (Energy in kWh).

Training	Nvidia 2060 Super	Nvidia 2080	Nvidia 3060Ti
Energy consumed RAM	0.043221	0.026551	0.044313
Energy consumed GPU	2.553218	1.816966	1.951726
Energy consumed CPU	0.484433	0.435885	0.321111
Total energy	3.080872	2.279403	2.317150
Time taken	14 h 54 min 5 s	13 h 56 min 46 s	9 h, 52 min 35 s

Table 4
Query TinyLlama (Energy in kWh).

Query	Nvidia 2060 Super	Nvidia 2080	Nvidia 3060Ti
Energy consumed RAM	0.000375	0.000313	0.000190
Energy consumed GPU	0.006060	0.006021	0.002986
Energy consumed CPU	0.004200	0.005132	0.002221
Total energy	0.010634	0.011466	0.005398
Time taken	7 min 30 s	6 min 14 s	8 min 41 s

5.2. Which hardware settings are suitable for respective model?

Results only ambiguously point to the suitable HW with a focus on model training and application. The hardware landscape is diverse and rapidly evolving due to the performance requirements towards efficient utilization of the underlying hardware (Bavikadi et al., 2022; Zhao et al., 2022).

There is a debate on whether it is more profitable to train on a CPU, rather than on a GPU and the need of ML on CPUs (Li et al., 2016; Alizadeh and Castor, 2024; Daghaghi et al., 2021; Zhang et al., 2025). It is not a question CPU vs. GPU but rather which overall hardware infrastructure suits the specific use case. GPUs, especially next generation high performance GPUs, are faster at training and inference but come with an increased environmental impact due to energy costs, operating temperature, limited availability and other factors (Gyawali, 2023; Samsi et al., 2023; Štaka et al., 2025). Given the total energy consumed is far greater on a GPU than on a CPU for the same task, it could be assumed to better train on a CPU. However, at the same time, the training time is far greater if run on a CPU. With this in mind, it needs to be considered whether there are enough time resources to train the model on a CPU. Furthermore, it is also important to note that our training was done on a smaller iteration count (30 000) than the original model, so the time consumption was not lower. Moreover, it is important to find a balance between the energy consumption and the time consumption, as sometimes it is not feasible and realistic to train a SLM over the course of a several weeks instead of less than a day. The demand for expansion of multi-GPU solutions could be evaded to a certain limit, by focusing on smaller, more specialized and possibly smarter algorithms and solutions, that allow to reduce training and inference time.

5.3. What are feasible use cases for SLMs?

The preliminary comparison of smaller language models revealed specific application scenarios. SLMs are especially used in situations where computation power of the device is limited. They are being integrated into mobile devices to reduce battery drainage and improve data privacy (Sarhaddi et al., 2025). Furthermore, there are also projects where SLMs such as TinyLlama run on a Raspberry Pi (Laskaridis et al., 2024). Reports also describe deployment on microcontrollers (Scherer et al., 2024; Yang et al., 2024). Other studies have shown that SLMs perform better than LLMs at translating clinical texts (Han et al., 2024). These models are mainly trained for exactly this purpose and can therefore even outperform LLMs.

In summary, SLMs should preferably be chosen as applied model in certain scenarios that are based on a limited vocabulary for input parameters, limited computational resource availability, and/or require tailored or specialized training sets. Thereby, noise and bias could be reduced while being cost-effective enabling the model to run locally.

5.4. Limitations and future outlook

Main limitations of this study include the selection of only two models, a simple training as well as query set, and a reduced hardware configuration for the experiments. Yet, a simple setup can also underline the key message of communicating energy consumption transparently, and this is the intention of this work, rather than presenting a comprehensive overview of different SLMs. We point out as the missing central element in evaluating SLMs. There is a lack of standardized evaluation for SLMs integrating multiple metrics for performance, efficiency and sustainability. We thereby highlight the approach of Huggingface LLM leaderboard (Open LLM, 2025) and SLMbench (Thanh Pham et al., 2025) as starting points of harmonizing evaluation metrics for language models in general and SLMs in particular.

A promising direction for future research may be to further analyze SLMs for specialized use cases such as in the biomedical domain, to underline SLM's feasibility. Future work could also examine sustainability aspects of SLM training and maintenance resulting from variations based on the geographic location, latency and associated energy grid as shown to impact the carbon footprint of computing systems (Gupta et al., 2021; Jegham et al., 2025).

We call upon AI researchers and providers to clearly report energy consumption metrics, facilitating responsible and sustainable model choices. Small language models, paired with transparent benchmarks, offer a feasible pathway towards a sustainable AI ecosystem. Furthermore, we call upon language model providers to increase transparency means regarding energy consumption by providing standardized information per model via a dedicated table on model cards. Such a table could report energy consumption as average of kWh per amount of inferences, using a particular set of standardized prompts and model settings. Arising generalization problems like hardware capabilities can be overcome by dual reporting facts for consumer-grade and industrial-grade hardware. Similarly, model options can be generalized by the means of temperature settings and using a maximum limit for token generation to enable fair and standardized comparison between models of varying sizes.

6. Conclusion

So far, choosing the right model for a specific task and a specific hardware is not an easy task, especially when it comes to balance resource costs and application suitability. While the ML community already collects and reports on many models (Castaño et al., 2024b), the perspective of climate cost in a comprehensive manner is barely touched. Meanwhile, Huggingface LLM leaderboard already includes some CO_2 cost estimations on SLMs (Open LLM, 2025), and first efforts are made to harness these data fields (Castaño et al., 2024a). Though, our selected models were not part of this leaderboard until the manuscript's finalization.

SLMs have been postulated as the future of Agentic AI and with its rapidly expanding usage in the AI domain, often containing a plethora of individual agents, this work highlights the importance of further investigation of sustainability concerns regarding SLMs (Belcak et al., 2025). This paper points to the importance of transparently communicating energy consumption in a standardized way including use case suitability, and advocates thoughtful model selection from a more responsible and sustainable perspective.

CRediT authorship contribution statement

Fleur Jeanquartier: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Conceptualization. **Claire Jean-Quartier:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Conceptualization. **Paul Rieder:** Validation, Software, Data curation. **Vedad Misirlić:** Validation, Software, Data curation. **Christian Pasero:** Validation, Software, Data curation. **Richard Hohensinner:** Validation, Software, Data curation. **Heimo Müller:** Writing – review & editing. **Andreas Holzinger:** Writing – review & editing, Supervision, Funding acquisition.

Code and data availability

Source code for tests is available on https://github.com/VedadTUG/HCI-SLM and generated data is shown in the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Parts of this work have been funded by the Austrian Science Fund (FWF), Project: https://doi.org/10.55776/P32554, explainable Artificial Intelligence.

Table A.5
Devices setups for NanoGPT.

Spec	Device 1	Device 2	Device 3	Device 4
CPU	AMD Ryzen 5 3600	Intel i5	AMD Ryzen 7 5700G	M1
GPU	Nvidia 2060 Super	Nvidia 2080	Nvidia 3060Ti	M1
OS	Windows	Windows	Windows	MacOS
Platform	WSL	WSL	Windows	MacOS

Table A.6
Devices setups for TinyLlama.

Spec	Device 1	Device 2	Device 3
CPU	AMD Ryzen 5	Intel i5	AMD Ryzen 7 5700G
GPU	Nvidia 2060 Super	Nvidia 2080	Nvidia 3060Ti
OS	Windows	Windows	Windows
Platform	WSL	WSL	WSL

Appendix. Devices setups

See Tables A.5 and A.6.

Data availability

https://github.com/VedadTUG/HCI-SLM.

References

Ahn, Y., Kim, N.W., 2025. Understanding why ChatGPT outperforms humans in visualization design advice. arXiv:2508.01547, URL https://arxiv.org/abs/2508.01547.

Alizadeh, N., Castor, F., 2024. Green ai: A preliminary empirical study on energy consumption in dl models across different runtime infrastructures. In: Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI. pp. 134–139.

Appelle, A., 2023. Will businesses or laws and regulations ever prioritise environmental sustainability for AI systems?. (Accessed 01 October 2025).

Araci, D., 2019. FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063, URL https://arxiv.org/abs/1908.10063.

Argerich, M.F., Patiño-Martínez, M., 2024. Measuring and improving the energy efficiency of large language models inference. IEEE Access 12, 80194–80207.

Azhar, A., 2021. Exponential thinking.. Res.-Technol. Manag. 65, 11–17. http://dx.doi.org/10.1080/08956308.2022.1999641.

Bavikadi, S., Dhavlle, A., Ganguly, A., Haridass, A., Hendy, H., Merkel, C., Reddi, V.J., Sutradhar, P.R., Joseph, A., Pudukotai Dinakarrao, S.M., 2022. A survey on machine learning accelerators and evolutionary hardware platforms. IEEE Des. Test 39 (3), 91–116. http://dx.doi.org/10.1109/MDAT.2022.3161126.

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y.C., Molchanov, P., 2025. Small language models are the future of agentic AI. arXiv:2506.02153, URL https://arxiv.org/abs/2506.02153.

Budennyy, S.A., Lazarev, V.D., Zakharenko, N.N., Korovin, A.N., Plosskaya, O., Dimitrov, D.V., Akhripkin, V., Pavlov, I., Oseledets, I.V., Barsola, I.S., et al., 2022. Eco2ai: carbon emissions tracking of machine learning models as the first step towards sustainable ai. In: Doklady Mathematics. Vol. 106, Springer, pp. S118–S128. http://dx.doi.org/10.1134/S1064562422060230.

Budzianowski, P., Vulić, I., 2019. Hello, it's GPT-2 - how can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. In: Birch, A., Finch, A., Hayashi, H., Konstas, I., Luong, T., Neubig, G., Oda, Y., Sudoh, K. (Eds.), Proceedings of the 3rd Workshop on Neural Generation and Translation. Association for Computational Linguistics, Hong Kong, pp. 15–22. http://dx.doi.org/10.18653/v1/D19-5602, URL https://aclanthology.org/D19-5602/.

Buhnila, I., 2025. Explain this medical term in my language: A case study of small language models for medical paraphrase generation. In: 3rd UniDive Workshop. Hal, pp. 1–4, URL https://hal.science/hal-04927075v1.

Castaño, J., Martínez-Fernández, S., Franch, X., 2024a. Lessons learned from mining the hugging face repository. In: Proceedings of the 1st IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering. pp. 1–6.

Castaño, J., Martínez-Fernández, S., Franch, X., Bogner, J., 2024b. Analyzing the evolution and maintenance of ml models on hugging face. In: 2024 IEEE/ACM 21st International Conference on Mining Software Repositories. MSR, IEEE, pp. 607–618.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., 2024. A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol. 15 (3), 1–45. http://dx.doi.org/10.1145/3641289.

Chowdhury, M.N.U.R., Haque, A., Soliman, H., 2025. The hidden cost of AI: Unraveling the power-hungry nature of large language models. http://dx.doi.org/10.20944/preprints202502.1676.v1, Preprints, URL https://doi.org/10.20944/preprints202502.1676.v1.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al., 2021. Training verifiers to solve math word problems. http://dx.doi.org/10.48550/arXiv.2110.14168, arXiv preprint arXiv:2110.14168.

2024. CodeCarbon official. URL https://codecarbon.io/.

Corradini, F., Leonesi, M., Piangerelli, M., 2025. State of the art and future directions of small language models: A systematic review. Big Data Cogn. Comput. 9 (7), 189. http://dx.doi.org/10.3390/bdcc9070189.

Daghaghi, S., Meisburger, N., Zhao, M., Shrivastava, A., 2021. Accelerating slide deep learning on modern cpus: Vectorization, quantizations, memory optimizations, and more. Proc. Mach. Learn. Syst. 3, 156–166.

Ding, Z., Wang, J., Song, Y., Zheng, X., He, G., Chen, X., Zhang, T., Lee, W.-J., Song, J., 2025. Tracking the carbon footprint of global generative artificial intelligence. Innov. 6 (5), http://dx.doi.org/10.1016/j.xinn.2025.100866.

Ehrlich-Sommer, F., Eberhard, B., Holzinger, A., 2025. ForestGPT and beyond: A trustworthy domain-specific large language model paving the way to forestry 5.0. Electronics 14 (18), 3583. http://dx.doi.org/10.3390/electronics14183583.

Erdogan, L.E., Lee, N., Jha, S., Kim, S., Tabrizi, R., Moon, S., Hooper, C., Anumanchipalli, G., Keutzer, K., Gholami, A., 2024. Tinyagent: Function calling at the edge. http://dx.doi.org/10.48550/arXiv.2409.00608, arXiv preprint arXiv:2409.00608.

Eurostat, 2023. Final energy consumption by sector, EU, 2022. URL https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_statistics_-_an_overview#Final_energy_consumption.

Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., Jiang, L., 2023. Llmcarbon: Modeling the end-to-end carbon footprint of large language models. arXiv preprint arXiv:2309.14393.

Ganescu, B.-M., Passerat-Palmbach, J., 2024. Trust the process: Zero-knowledge machine learning to enhance trust in generative ai interactions. arXiv preprint arXiv:2402.06414.

Garg, M., Raza, S., Rayana, S., Liu, X., Sohn, S., 2025. The rise of small language models in healthcare: A comprehensive survey. arXiv preprint arXiv:2504.17119.

Gupta, U., Kim, Y.G., Lee, S., Tse, J., Lee, H.-H.S., Wei, G.-Y., Brooks, D., Wu, C.-J., 2021. Chasing carbon: The elusive environmental footprint of computing. In: 2021 IEEE International Symposium on High-Performance Computer Architecture. HPCA, IEEE, pp. 854–867.

Gyawali, D., 2023. Comparative analysis of cpu and gpu profiling for deep learning models. arXiv preprint arXiv:2309.02521.

Han, T., Adams, L.C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., Bressem, K.K., 2023. MedAlpaca–An open-source collection of medical conversational AI models and training data. arXiv preprint arXiv:2304.08247.

Han, L., Gladkoff, S., Erofeev, G., Sorokina, I., Galiano, B., Nenadic, G., 2024. Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning. Front. Digit. Heal. 6, 1211564.

Hasan, M.A., Tarannum, P., Dey, K., Razzak, I., Naseem, U., 2024. Do large language models speak all languages equally? A comparative study in low-resource settings. http://dx.doi.org/10.48550/arXiv.2408.02237, arXiv:2408.02237.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J., 2020. Measuring massive multitask language understanding. http://dx.doi.org/10.48550/arXiv.2009.03300, arXiv preprint arXiv:2009.03300.

Holzinger, A., Fister Jr., I., Fister, I., Kaul, H.-P., Asseng, S., 2024a. Human-centered AI in smart farming: Towards agriculture 5.0. IEEE Access 12, 62199–62214. http://dx.doi.org/10.1109/ACCESS.2024.3395532.

Holzinger, A., Schweier, J., Gollob, C., Nothdurft, A., Hasenauer, H., Kirisits, T., Häggström, C., Visser, R., Cavalli, R., Spinelli, R., Stampfer, K., 2024b. From industry 5.0 to forestry 5.0: Bridging the gap with human-centered artificial intelligence. Curr. For. Rep. 10 (6), 442–455. http://dx.doi.org/10.1007/s40725-024-00231-7.

Hu, S., Tu, Y., Han, X., He, C., Cui, G., Long, X., Zheng, Z., Fang, Y., Huang, Y., Zhao, W., et al., 2024. MiniCPM: Unveiling the potential of small language models with scalable training strategies. arXiv preprint arXiv:2404.06395.

Islam, M.R., Dhar, N., Deng, B., Nguyen, T.N., He, S., Suo, K., 2024. Characterizing and understanding the performance of small language models on edge devices. In: 2024 IEEE International Performance, Computing, and Communications Conference. IPCCC, pp. 1–10. http://dx.doi.org/10.1109/IPCCC59868.2024.10850044.

Jean-Quartier, C., Bein, K., Hejny, L., Hofer, E., Holzinger, A., Jeanquartier, F., 2023. The cost of understanding—XAI algorithms towards sustainable ML in the view of computational cost. Computation 11 (5), http://dx.doi.org/10.3390/computation11050092, URL https://www.mdpi.com/2079-3197/11/5/92.

Jegham, N., Abdelatti, M., Elmoubarki, L., Hendawi, A., 2025. How hungry is ai? Benchmarking energy, water, and carbon footprint of llm inference. arXiv preprint arXiv:2505.09598.

Kang, M., Lee, S., Baek, J., Kawaguchi, K., Hwang, S.J., 2023. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. Adv. Neural Inf. Process. Syst. 36, 48573–48602. http://dx.doi.org/10.48550/arXiv.2305.18395.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al., 2023. ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. 103, 102274.

Keles, B., Gunay, M., Caglar, S.I., 2024. LLMs-in-the-loop part-1: Expert small AI models for bio-medical text translation. http://dx.doi.org/10.48550/arXiv.2407.12126, arXiv preprint arXiv:2407.12126.

Khan, T., Motie, S., Kocak, S.A., Raza, S., 2025. Optimizing large language models: Metrics, energy efficiency, and case study insights. arXiv preprint arXiv:2504.06307.

Kim, H., Hwang, H., Lee, J., Park, S., Kim, D., Lee, T., Yoon, C., Sohn, J., Park, J., Reykhart, O., et al., 2025. Small language models learn enhanced reasoning skills from medical textbooks. NPJ Digit. Med. 8 (1), 240.

Kim, W., Spörer, J., Lee, C.L., Handschuh, S., 2024. Is small really beautiful for central bank communication? Evaluating language models for finance: Llama-3-70B, GPT-4, FinBERT-FOMC, FinBERT, and VADER. In: Proceedings of the 5th ACM International Conference on AI in Finance. ICAIF '24, Association for Computing Machinery, New York, NY, USA, pp. 626–633. http://dx.doi.org/10.1145/3677052.3698675.

Kleinig, O., Sinhal, S., Khurram, R., Gao, C., Spajic, L., Zannettino, A., Schnitzler, M., Guo, C., Zaman, S., Smallbone, H., et al., 2024. Environmental impact of large language models in medicine. Intern. Med. J. 54 (12), 2083–2086.

Kocic, V., Lukac, N., Rozajac, D., Schweng, S., Gollob, C., Nothdurft, A., Stampfer, K., Ser, J.D., Holzinger, A., 2025. LLM in the loop: A framework for contextualizing counterfactual segment perturbations in point clouds. IEEE Access 13, 85507–85525. http://dx.doi.org/10.1109/ACCESS.2025.3568052.

Kocmi, T., Federmann, C., 2023. Large language models are state-of-the-art evaluators of translation quality. arXiv preprint arXiv:2302.14520, URL https://arxiv.org/abs/2302.14520.

Kraišniković, C., Harb, R., Plass, M., Al Zoughbi, W., Holzinger, A., Müller, H., 2025. Fine-tuning language model embeddings to reveal domain knowledge: An explainable artificial intelligence perspective on medical decision making. Eng. Appl. Artif. Intell. 139, 109561. http://dx.doi.org/10.1016/j.engappai.2024.109561.

Kumar, A., Davenport, T., Bean, R., 2025. The case for using small language models. https://hbr.org/2025/09/the-case-for-using-small-language-models, (Accessed 01 October 2025).

Lacoste, Alexandre, Luccioni, Alexandra, Schmidt, Victor, Dandres, Thomas, 2019. Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700, URL https://mlco2.github.io/impact/#publish.

Laskaridis, S., Katevas, K., Minto, L., Haddadi, H., 2024. Mobile and edge evaluation of large language models. In: Workshop on Efficient Systems for Foundation Models II @ ICML2024. p. 20, URL https://openreview.net/forum?id=aAtCQnCsya.

Lee, Y.-S., 2024. Analysis of small large language models (LLMs). Int. J. Adv. Smart Converg. 13 (4), 155–160.

Li, D., Chen, X., Becchi, M., Zong, Z., 2016. Evaluating the energy efficiency of deep convolutional neural networks on CPUs and GPUs. In: 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom). BDCloud-SocialCom-SustainCom, pp. 477–484. http://dx.doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.76.

Liu, V., Yin, Y., 2024. Green AI: exploring carbon footprints, mitigation strategies, and trade offs in large language model training. Discov. Artif. Intell. 4 (1), 49.

2024. 10 LLM use cases to enhance your business. URL https://www.coursera.org/articles/llm-use-cases.

Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., Lane, N.D., Xu, M., 2024. Small language models: Survey, measurements, and insights. arXiv preprint arXiv:2409.15790.

Luccioni, S., Gamazaychikov, B., Hooker, S., Pierrard, R., Strubell, E., Jernite, Y., Wu, C.-J., 2024. Light bulbs have energy ratings—so why can't AI chatbots? Nature 632 (8026), 736–738. http://dx.doi.org/10.1038/s42256-025-00979-y.

Luccioni, A.S., Viguier, S., Ligozat, A.-L., 2022. Estimating the carbon footprint of BLOOM, a 176B parameter language model. http://dx.doi.org/10.48550/arXiv.2211.02001, ArXiv URL https://arxiv.org/abs/2211.02001.

2024. Mingpt. URL https://github.com/karpathy/minGPT.

Miraghaei, P., Moreschini, S., Kolehmainen, A., Hästbacka, D., 2025. Towards a small language model lifecycle framework. http://dx.doi.org/10.48550/arXiv.2506.07695, arXiv preprint arXiv:2506.07695.

2024. Nanogpt. URL https://github.com/karpathy/nanoGPT.

Nogales Pérez, D., 2023. A Deep Learning Based Tool For Ear Training B.S. thesis. Universitat Politècnica de Catalunya.

Nucci, A., 2024. Small language models: Trends and use cases. URL https://aisera.com/blog/small-language-models/#:~:text=Small%20Language%20Models%20offer%20a,speed%20is%20of%20the%20essence.

2025. Open LLM Leaderboard, https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/.

2023. Power-hungry AI: Researchers evaluate energy consumption across models. URL https://cse.engin.umich.edu/stories/power-hungry-ai-researchers-evaluate-energy-consumption-across-models.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., 2019. Language Models are Unsupervised Multitask Learners. Tech. rep., OpenAi, URL https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf, (Accessed 26 February 2025).

Raiaan, M.A.K., Mukta, M.S.H., Fatema, K., Fahad, N.M., Sakib, S., Mim, M.M.J., Ahmad, J., Ali, M.E., Azam, S., 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. IEEE Access 12, 26839–26874. http://dx.doi.org/10.1109/ACCESS.2024.3365742.

Rashid, S.F., Duong-Trung, N., Pinkwart, N., 2024. Generative AI in education: Technical foundations, applications, and challenges. In: Kadry, S. (Ed.), Artificial Intelligence and Education. IntechOpen, Rijeka, pp. 33–54. http://dx.doi.org/10.5772/intechopen.1005402.

Ray, P.P., 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Physical Syst. 3, 121–154. http://dx.doi.org/10.1016/j.iotcps.2023.04.003, URL https://www.sciencedirect.com/science/article/pii/S266734522300024X.

Rehman, T., Sanyal, D.K., Chattopadhyay, S., 2025. How green are neural language models? Analyzing energy consumption in text summarization fine-tuning. http://dx.doi.org/10.48550/arXiv.2501.15398, arXiv preprint arXiv:2501.15398.

Ren, S., Tomlinson, B., Black, R.W., Torrance, A.W., 2024. Reconciling the contrasting narratives on the environmental impact of large language models. Sci. Rep. 14 (1), 26310.

Rodriguez, C., Degioanni, L., Kameni, L., Vidal, R., Neglia, G., 2024. Evaluating the energy consumption of machine learning: Systematic literature review and experiments. http://dx.doi.org/10.48550/arXiv.2408.15128, arXiv preprint arXiv:2408.15128.

Rusyn, B., Lutsyk, O., Kosarevych, R., Kapshii, O., Karpin, O., Maksymyuk, T., Gazda, J., 2024. Rethinking deep CNN training: A novel approach for quality-aware dataset optimization. IEEE Access 12, 137427–137438.

Sakib, T.H., Hosain, M.T., Morol, M.K., 2025. Small language models: Architectures, techniques, evaluation, problems and future adaptation. http://dx.doi.org/10.48550/arXiv.2505.19529, arXiv preprint arXiv:2505.19529.

Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., Gadepally, V., 2023. From words to watts: Benchmarking the energy costs of large language model inference. In: 2023 IEEE High Performance Extreme Computing Conference. HPEC, IEEE, pp. 1–9.

Sarhaddi, F., Nguyen, N.T., Zuniga, A., Hui, P., Tarkoma, S., Flores, H., Nurmi, P., 2025. LLMs and IoT: A comprehensive survey on large language models and the internet of things. TechRxiv.

Scherer, M., Macan, L., Jung, V.J., Wiese, P., Bompani, L., Burrello, A., Conti, F., Benini, L., 2024. Deeploy: Enabling energy-efficient deployment of small language models on heterogeneous microcontrollers. IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 43 (11), 4009–4020.

Schick, T., Schütze, H., 2020. It's not just size that matters: Small language models are also few-shot learners. arXiv preprint arXiv:2009.07118.

2024. StableLM: Stability AI language models. URL https://github.com/Stability-AI/StableLM, (Accessed 18 March 2025).

Štaka, Z., Mišić, M., Tomašević, M., 2025. CPU vs. GPU: Performance evaluation of classical machine and deep learning algorithms. In: 2025 24th International Symposium INFOTEH-JAHORINA. INFOTEH, pp. 1–6. http://dx.doi.org/10.1109/INFOTEH64129.2025.10959248.

Thanh Pham, N., Kieu, T., Nguyen, D.-M., Xuan, S.H., Duong-Trung, N., Le-Phuoc, D., 2025. SLM-bench: A comprehensive benchmark of small language models on environmental impacts–extended version. pp. arXiv–2508. http://dx.doi.org/10.48550/arXiv.2508.15478, arXiv e-Prints.

Tonja, A.L., Dossou, B.F., Ojo, J., Rajab, J., Thior, F., Wairagala, E.P., Aremu, A., Moiloa, P., Abbott, J., Marivate, V., et al., 2024. Inkubalm: A small language model for low-resource african languages. http://dx.doi.org/10.48550/arXiv.2408.17024, arXiv preprint arXiv:2408.17024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G., 2023. LLaMA: Open and efficient foundation language models. arXiv:2302.13971, URL https://arxiv.org/abs/2302.13971.

Turner, C., Oyekan, J., Garn, W., Duggan, C., Abdou, K., 2022. Industry 5.0 and the circular economy: Utilizing LCA with intelligent products. Sustainability 14 (22), 14847. http://dx.doi.org/10.3390/su142214847.

Vajrobol, V., Aggarwal, N., Saxena, G.J., Singh, S., Pundir, A., 2024. Transforming SEO in the era of generative AI: Challenges, opportunities, and future prospects. Revolutionizing AI-Digital Landsc. 86–100.

Wang, X., Dang, T., Kostakos, V., Jia, H., 2024a. Efficient and personalized mobile health event prediction via small language models. In: Proceedings of the 30th Annual International Conference on Mobile Computing and Networking. pp. 2353–2358. http://dx.doi.org/10.1145/3636534.3698123.

Wang, L., Li, M., Wang, W., Gong, Y., Xiong, Y., 2023. Green innovation output in the supply chain network with environmental information disclosure: An empirical analysis of Chinese listed firms. Int. J. Prod. Econ. 256, 108745. http://dx.doi.org/10.1016/j.ijpe.2022.108745.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., Chen, W., 2024b. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (Eds.), Advances in Neural Information Processing Systems. Vol. 37, Curran Associates, Inc., pp. 95266–95290, URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., Wang, W., Li, R., Xu, J., Tang, X., et al., 2024c. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. http://dx.doi.org/10.48550/arXiv.2411.03350, arXiv preprint arXiv:2411.03350.

2024. Watts and watt-hours: what are they and how is energy consumption measured? URL https://www.enelgreenpower.com/learning-hub/gigawhat/search-articles/articles/2022/11/watts-watt-hours-energy-consumption.

Wells, S., 2023. Generative AI's energy problem today is foundational. URL https://spectrum.ieee.org/ai-energy-consumption.

Xiong, Y., Xia, S., Wang, X., 2020. Artificial intelligence and business applications, an introduction. Int. J. Technol. Manage. 84 (1–2), 1–7. http://dx.doi.org/10.1504/IJTM.2020.112615.

Xu, S., Jia, X., Yan, L., 2025. Research on carbon footprint in the whole process of LLM based on refined modeling. In: Proceedings of the 2024 3rd International Conference on Algorithms, Data Mining, and Information Technology. ADMIT '24, Association for Computing Machinery, New York, NY, USA, pp. 300–303. http://dx.doi.org/10.1145/3701100.3701162.

Yang, Z., Chen, R., Wu, T., Wong, N., Liang, Y., Wang, R., Huang, R., Li, M., 2024. MCUBERT: Memory-efficient BERT inference on commodity microcontrollers. arXiv preprint arXiv:2410.17957.

Yu, Z., Wu, Y., Deng, Z., Tang, Y., Zhang, X.-P., 2024. Opencarboneval: A unified carbon emission estimation framework in large-scale ai models. arXiv preprint arXiv:2405.12843.

Zhang, T., Yi, J., Yao, B., Xu, Z., Shrivastava, A., 2025. Nomad-attention: Efficient llm inference on cpus through multiply-add-free attention. Adv. Neural Inf. Process. Syst. 37, 112706–112730.

Zhang, P., Zeng, G., Wang, T., Lu, W., 2024. TinyLlama: An open-source small language model. URL https://paperswithcode.com/paper/tinyllama-an-open-source-small-language-model.

Zhao, Y., Gao, X., Shumailov, I., Fusi, N., Mullins, R., 2022. Rapid model architecture adaption for meta-learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.), Advances in Neural Information Processing Systems. Vol. 35, Curran Associates, Inc., pp. 18721–18732, URL https://proceedings.neurips.cc/paper_files/paper/2022/file/76df3f555683bc6c4b988bb81b930d5b-Paper-Conference.pdf.

Zhou, J., Müller, H., Holzinger, A., Chen, F., 2024. Ethical ChatGPT: Concerns, challenges, and commandments. Electronics 13 (17), 3417. http://dx.doi.org/10.3390/electronics13173417.

Zhu, Y., Zhu, M., Liu, N., Xu, Z., Peng, Y., 2024. LLaVA-phi: Efficient multi-modal assistant with small language model. In: Proceedings of the 1st International Workshop on Efficient Multimedia Computing under Limited. EMCLR '24, Association for Computing Machinery, New York, NY, USA, pp. 18–22. http://dx.doi.org/10.1145/3688863.3689575.