Applied Computing and Intelligence

https://www.aimspress.com/journal/aci

Applied Computing and Intelligence
6(2): 157–172.
DOI: 10.3934/aci.2026009
Received: 01 March 2026
Revised: 24 June 2026
Accepted: 30 July 2026
Published: 25 August 2026

*Review*

# From carbon-intensive to carbon-neutral: sustainability and energy efficiency in small language models: bibliometric analysis (2020–2026)

**Avinash Shrivastava**[1], **Anamika Gupta**[1,*] **and Sarabjeet Kaur Kochhar**[2]

[1] Department of Computer Science, Shaheed Sukhdev College of Business Studies, University of Delhi, India

[2] Department of Computer Science, IP College for Women, University of Delhi, India

* **Correspondence:** Email: anamikargupta@sscbsdu.ac.in.

Academic Editor: Md Abdullah Al Hafiz Khan

**Abstract:** Large language models (LLMs) have achieved unprecedented performance in various tasks, such as classification, summarization, reasoning, and complex decision-making. But, as the size and parameter count of these models has increased exponentially, their computational demand, electricity consumption, and carbon footprint has also risen dramatically. This rapid scaling has put the global research community in front of a serious environmental concern. In response to this growing concern, researchers are gradually shifting their focus from red AI (compute-intensive, resource-heavy systems) to green AI. Particularly, small language models and optimization techniques like quantization, pruning, and knowledge distillation are being explored as sustainable alternatives to LLMs for domain-specific tasks. Covering the literature from 2020 to 2026, this review paper presents an extensive, deep, and rigorous bibliometric analysis. Using an advanced rule-based keyword preprocessing pipeline and Biblioshiny network mapping, we have mapped the annual publication growth, global research hubs, author influence networks, and dominant thematic clusters. Further, this review paper highlights a major research gap: Although many papers have already been written on energy optimization, publications on standardized carbon emission reporting frameworks are still limited.

**Keywords:** small language models; green AI; quantization; sustainability; knowledge distillation

## 1. Introduction

The recent history of AI reflects a highly dynamic and volatile progression. After 2020, the AI world was primarily driven by the idea of bigger is better. Every tech giant and research lab was trying to increase the parameter count of their new language models to either set or break a record. However,

behind this, a price was being paid which was initially ignored: the environmental cost. In this section, we will look at this transition step by step and see how the journey began from red AI to green AI.

### *1.1. The rise of red AI and the massive compute crisis*

One of the foundational pillars of AI's modern era is deep learning and neural networks. Initially, when Krizhevsky et al. (2012) [4] introduced AlexNet (trained on the ImageNet [2] dataset), compute power was not a big problem. But, when in 2020 Brown et al. [1] introduced GPT-3, which showed that language models are few-shot learners, this revolutionized the entire industry.

But, there was a catch. To train these models, thousands of GPUs needed to be running continuously at full power for months. Samsi et al. (2023) [10] in their empirical studies highlighted that the energy consumption and carbon footprints of these models had reached a dangerously high level. This approach, where a model's accuracy is the sole property of compute power and their data size, was termed "red AI" in academia. Red AI models are indeed very accurate. However, this incurs huge electricity demands and cooling systems that cause significant damage to the environment. The works of Schick and Schütze (2021) [8] and Devlin et al. (2019) [3] (who created BERT) brought a new level in natural language understanding. However, all these led to a multiple-fold increase in computational overhead.

### *1.2. Green AI*

As the energy demands of red AI increased exponentially, hardware constraints and global warming concerns gave the AI community a reality check. Scientists and researchers realized that they cannot always improve accuracy by increasing the number of GPUs. This required a new paradigm, which we call green AI.

The core philosophy of green AI is simple: Make AI smarter, not bigger. Here, the goal is not just to improve the accuracy but also to minimize the inference latency, carbon footprints, and carbon emission. Argerich and Patino-Martinez (2024) [21] have focused on energy-efficient inference metrics. Now, studies like Chen et al. (2025) [26] have begun focusing on energy-aware offloading by using digital twins and edge computing. A digital twin is a real-time replica of a virtual system. It is updated continuously through sensors and data feeds to reflect its current state. This enables simulation, monitoring, and optimization without physical intervention. The energy consumption of small language models (SLMs) on edge devices can be predicted without running them on actual hardware, and smart offloading (cloud vs edge) decisions can be made.

### *1.3. SLMs and model optimization*

There are two pillars that are driving the green AI revolution: SLMs and model compression techniques. SLMs are relatively newer than LLMs. Researchers began using this term in 2023 to refer to those models that are smaller in size and parameter count as compared to LLMs (usually, less than 15 billion parameters). However, these are trained on highly curated, high-quality data, and beat LLMs on multiple domain-specific tasks.

It is important for us to understand the techniques that make SLMs efficient:

- **Quantization:** This is a mathematical technique in which the precision of weights and attentions of neural networks is reduced (e.g., from 32-bit floating point to 8-bit or 4-bit integers). Shao et al.

(2024) [14] showed through OmniQuant, and Xiao et al. (2023) [5] proved through SmoothQuant that quantization can be used to reduce the model size by more than half without compromising accuracy.

- **Knowledge distillation:** In this process, the knowledge and decision-making of a heavy model (teacher model) is transferred to a smaller model (student model). Increasing work in this domain (like Gholami and Omar (2024) [17]'s student-teacher distillation research) is proof that open and efficient foundational models are the future.

Recent academic research is no longer focused solely on improving accuracy. Attention has been shifting toward prioritizing both aspects: accuracy and efficiency, as well as optimization and computational sustainability [11,13,28]. Even large technological companies are adopting architectural strategies to control inference cost and computational overhead [21].

### *1.4. Objective*

Although a lot of technical research is being conducted on green AI and SLMs, a comprehensive bibliometric study of this entire academic landscape on a macro level is still limited. In any emerging research domain, it is important to understand which thematic areas have received the most attention in the past 5–6 years, which countries and institutions are leading this development, and where meaningful research gaps still exist. The primary objective of this review paper is to address this gap by presenting a structured and data-driven overview of the field, thereby identifying major trends, collaborative patterns, and potential future research directions. We have five main research questions (RQs) that we aim to answer in this review paper:

- **RQ1:** What is the pattern of annual publication and citation in this field?
- **RQ2:** Which countries, institutions, and authors are the core hubs of this green AI revolution?
- **RQ3:** What are the dominant thematic clusters, and around which core topics is the research primarily centered?
- **RQ4:** What are the trending topics and how did they evolve?
- **RQ5:** What are the major research gaps in the current literature which should be the focus of researchers in the future?

## 2. Materials and methods

This study is based on a high-quality preprocessed dataset. We have followed a strict and reproducible methodology so that our results are accurate and reliable.

### *2.1. Data collection strategy*

To extract the literature corpus for this study, we used the Scopus databases. Scopus is one of the largest and a most highly reliable peer-reviewed abstract and citation database. We have designed our research query in such a way that it covers most variations of SLM, LLM optimization, green AI, and energy efficiency. We restricted our timespan to 2020 to 2026, with some exceptions.

*Note: The dataset was retrieved in February 2026 and therefore represents a partial year; 2026 figures reflect only the first two months of the year and should not be compared directly with full-year data from prior years.*

**Primary search keywords:**

```
TITLE-ABS-KEY(("Small Language Model*" OR "TinyLLM" OR
"Compact Language Model*" OR "Lightweight Language Model*"
OR ("Large Language Model*" AND ("small" OR "efficient")))
AND ("Sustainability" OR "Sustainable AI" OR "Green AI"
OR "Carbon footprint" OR "CO2 emission"
OR "Energy efficiency" OR "Energy consumption"
OR "Power consumption" OR "Low power"))
OR
TITLE-ABS-KEY(("Large Language Model*"
AND ("Quantization" OR "Pruning" OR "Distillation"
OR "Model compression")
AND ("Energy" OR "Efficiency" OR "Sustainability")))
AND PUBYEAR > 2019
AND NOT TITLE-ABS-KEY("Selective Laser Melting"
OR "Spatial Light Modulator" OR "Sintering")
```

The search strategy that we chose was based on Scopus. It was selected for its broad coverage of peer-reviewed literature in computer science, engineering, and interdisciplinary domains. The query was structured into complementary Boolean blocks. The idea was to maximize the conceptual coverage while maintaining precision. Block one targeted publications explicitly, discussing small, compact, lightweight, or efficient language models in conjunction with sustainability, energy, or environmental impacts. Block two included LLM compression techniques like quantization, pruning, and distillation, along with energy and sustainability terms. In order to remove the noisy and irrelevant results, some terms were explicitly excluded, like "selective laser melting" and "spatial light modulator" (which happens to be the expanded form of SLM in many other contexts).

There are also some limitations to the query which need to be stated. Scopus does not index preprint repositories like arXiv. Since foundational works on efficient and sustainable models mostly appeared as preprints, there is a possibility that the study overlooks those papers. Conceptually relevant publications which do not explicitly mention energy efficiency might have been excluded. The terminology small language model (SLM) itself was not formally established until 2023, which inherently limits the volume of indexed publications retrievable for the 2020–2022 period.

We obtained **1139 unique documents** as the result of the search query. These include articles (340), conference papers (695), reviews (85), and books and book chapters (19). We did not find any relevant publications in the year 2020. Since the query searched this span, we included the year in the study, and hence in the title. However, the charts drawn henceforth will be addressed as (2021–2026) rather than (2020–2026). The annual growth rate (AGR) of the dataset was observed to be **109.82%**, which is a large number in itself. This shows the urgency of this topic in academia. In our dataset, we had a total of **3204 unique authors** and **6492 unique cited references** (see Figure 1). Wang et al. (2026) [6], Wang and Yin (2026) [7], and Fernandez et al. (2025) [31] have published highly relevant works in this dataset boundary.


**Figure 1.** Overview of the dataset (retrieved February 2026).

### *2.2. The Python-based keyword preprocessing pipeline*

Bibliometric data, when obtained in its raw form, is very messy. The keyword columns include different variants of the same words. To address this issue, we develop a rule-based preprocessing pipeline in Python. The main purpose of the script was to systematically clean and standardize the authors' keywords and index keywords:

(1) **Standardizing LLM variants:** Different authors have used "LLM", "large language model", "large language models", and "large-language-model" to address the same entity. We grouped these variants under a master keyword: **"large language models"**.
(2) **Standardization for quantization:** The difference in British English and American English is quite common in the literature. "quantisation" and "quantization" were detected and mapped to a single term **"quantization"**. Shen et al. (2024) [14] and Xiao (2023) [5] have used the same standard terminology.
(3) **Merging optimization-related concepts:** Highly overlapping concepts like "fine tuning", "fine-tuning", "parameter efficient fine-tuning (PEFT)", and "low-rank adaptation (LoRA)" were mapped to broader architectural categories to ensure accurate thematic mapping.
(4) **Sustainability terminology:** Terms like "green AI", "eco-friendly AI", and "sustainable AI" were standardized based on their context without changing the title and the original abstract.

It is important to note that the preprocessing pipeline only modified the keyword columns of the dataset. Other metadata (like author's name, publication year, total citations, source title, etc.) were kept intact and untouched to maintain integrity.

### *2.3. Bibliometric analysis tools (Biblioshiny)*

After cleaning the data, the final csv was imported to an R-environment. We used Biblioshiny, a web interface of an open-source R package "bibliometric" to:

- plot the growth models of annual scientific production and average citations.
- generate co-word networks and thematic maps (using the Louvain clustering algorithm).
- verify Lotka's law (author productivity) and Bradford's law (journal scattering).

## 3. Results

In this section, we will discuss the results of our analysis and answer our RQs factually.

### *3.1. RQ1: What is the annual publication and citation growth?*

When we mapped the growth trajectory of the field, we found a classical exponential adoption curve. If a field grows by more than double year-on-year (109.82% AGR), then it means that it is undergoing a major paradigm shift.

**The exponential surge in production:** Annual scientific production data (Figure 2) not only show the increase in the number of publications, but also reflect the developing environmental concerns in academia. According to our dataset, there were no publications in the year 2020 addressing this issue. It was only in 2021 that studies such as Cui et al. (2021) [9] and Schick (2021) [8] began questioning the sustainability aspects of LLMs, resulting in three publications that year. Henceforth, we refer to this analysis as spanning 2021 to 2026 rather than 2020 to 2026. The following year (2022) had no publication on this domain based on our dataset, indicating that it was not a mainstream concern back then.

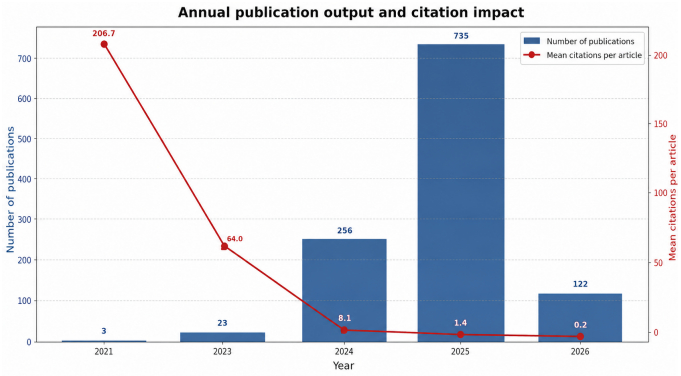


**Figure 2.** Annual scientific production of SLM and sustainability research. The bars show publication volume while the line indicates mean citation per article (retrieved February 2026).

When in 2023 LLMs were rapidly being adopted, discussions around their computational requirements and energy consumption began to intensify, shifting the academic focus from purely performance-driven benchmarks to concerns about carbon footprint, scalability, and long-term sustainability. This growing reflection is evident in the publication count, which rose sharply to 23 in 2023.

An increase in the number of publications from 256 in 2024 to 735 in 2025 indicates an increase in focus on sustainability in the AI community. Several researchers [21, 31] emphasized the shift from performance-driven AI to energy-conscious AI.

**The citation lag effect (quality vs. volume):** The analysis of the annual citations per year (Figure 2) reveals that the mean total citations (TC) per article was highest (206.67) in 2021, whereas in 2025, the mean TC dropped to only 1.44 (Table 1). This phenomenon is known as "citation lag". Older publications have had more time to accumulate citations, while recent papers have not yet had sufficient exposure within the community. To account for this time bias, the mean TC per year is a fairer measure, and even this normalized figure drops from 34.44 (2021) to 0.72 (2025), which is better explained by the rapid growth in publication volume, from just 3 papers in 2021 to 735 in 2025, spreading citations across a much larger pool of papers. Separately, a clear content shift is also visible: 2021–2023 served as a foundational period where basic theories, baseline metrics, and early quantization frameworks were developed [5, 14, 17], while 2024–2026 represents a period of

"mass implementation", where these foundational frameworks were applied to specific hardware and edge scenarios [18, 22]. These two trends together explain the observed citation pattern.

**Table 1.** Annual scientific production and citation impact growth (2021–2026). Data illustrates exponential volume growth alongside early foundational high-citation impact (retrieved February 2026).

| Year | Number of articles | Mean TC per article | Mean TC per year |
|---|---|---|---|
| 2021 | 3 | 206.67 | 34.44 |
| 2023 | 23 | 64.00 | 16.00 |
| 2024 | 256 | 8.14 | 2.71 |
| 2025 | 735 | 1.44 | 0.72 |
| 2026 | 122 | 0.21 | 0.21 |

### *3.2. RQ2: Which are the core hubs and influential actors?*

The literature found in this area is studied to find the dominating institutions and the authors.

#### 3.2.1. Country-wise influence: a bipolar geopolitical landscape

The cross-analysis of countries' scientific production and most cited countries (Figure 3) reveals an interesting conclusion. Global research has become a bipolar competition.

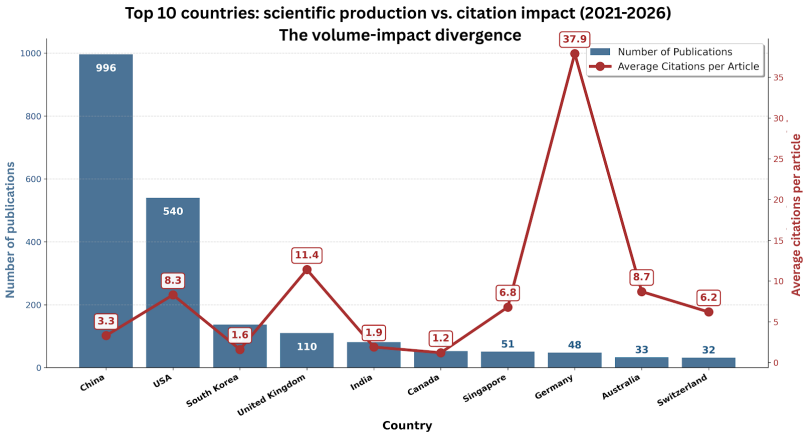


**Figure 3.** Countries' scientific production vs citation impact (retrieved February 2026).

*Note: Country publication counts reflect total output including co-authored papers, meaning one paper may be counted across multiple countries. Total counts therefore exceed the corpus size of 1139 documents.*

- **China—the volume leader:** China has taken a major lead in this domain with more than **990 publications**. Authors like Zeng et al. (2024) [15], Chen et al. (2024) [18], and Park et al. (2024) [20] have worked extensively on heavy hardware-software integration. China's below-average citation rate (3.3 per article) despite leading in volume (996 publications) can be explained by three factors. First, the bulk of China's output is concentrated in 2024–2025, meaning most papers are too recent to have accumulated significant citations yet, the same citation lag effect discussed earlier. Second, China has the highest single-country publication (SCP) rate among top producers (282 out of 374 corresponding author papers, or 75%), meaning most

Chinese papers are published without international collaboration; internationally co-authored papers tend to reach wider audiences and attract more citations. Third, Germany's contrasting profile (48 publications, 37.9 average citations) illustrates this point: smaller volume, older publications, and higher international collaboration rates produce stronger citation impact. In short, China's low citation average reflects the cost of rapid, recent, and largely domestic publication growth rather than a reflection of research quality.

- **USA—the early adopter:** The USA stands at the second position with 540 publications. Though the USA is lagging behind in volume, the foundational frameworks (like the GPT series [1, 21]) mostly come from US-based labs.
- **Germany—the quality hub:** The most interesting insight is from Europe. Germany has fewer publications (48 papers), but the average article citations are 37.9, which is greater than China (3.3) and the USA (8.3). This means that German institutions are publishing highly specific and high-quality fundamental research that is being cited by researchers around the globe. Pimenow et al. (2024) [19] and other European scholars are providing rigorous standards for edge AI and sustainability metrics.

While China leads overall publication output (996 papers), this dominance is a recent phenomenon rather than a long-standing trend. China produced only 1 paper in 2021 and 9 in 2023, before surging to 199 in 2024 and 860 in 2025, meaning over 86% of China's entire output appeared in 2025 alone. The USA follows a similar trajectory (2 papers in 2021, rising to 503 in 2025), suggesting that both countries entered this research space around the same time, with China simply scaling faster. This confirms that the volume gap between China and other nations is not a reflection of longer research experience, but of more aggressive recent scaling, likely driven by national AI investment priorities post-2023.

Chinese universities are clearly dominating the top ranks (Figure 4). **Tsinghua University (50 publications)** and **Shanghai Jiao Tong University (46 publications)** are two leading institutions. The focus of these universities are on model compression, FPGAs and accelerators, and other hardware-centric methodologies which make SLMs practically deployable. Zeng et al. (2024) [15]'s hardware micro-architecture research is an excellent example in this direction.

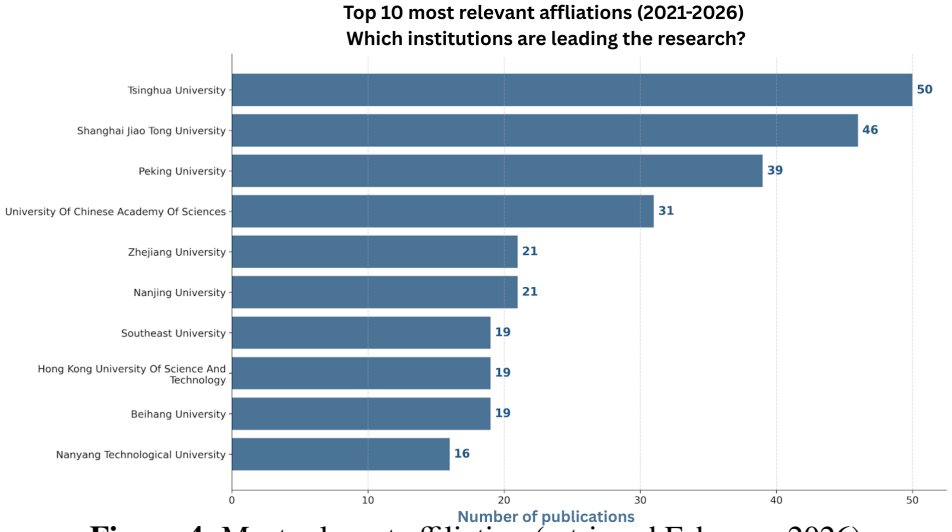


**Figure 4.** Most relevant affiliation (retrieved February 2026).

### 3.2.2. Authors and source dynamics

This section identifies the core contributors in the domain and differentiates between transient (single-publication) authors and continuing authors using Lotka's law.

**Lotka's law (author productivity):** The law states that the majority of publications in literature are written by a select few core authors, whereas the majority of authors have just one publication. According to our dataset, 77.03% of authors have a single publication on this subject (Figure 5). This is a classic signature of a trending, hype-driven scientific field. Because SLMs and green AI are trending topics, researchers across domains (like Aralimatti et al. (2025) [32] in mechanics, and Zong M. (2025)) are applying them in their domain. However, a core group of authors (like Wang Y., with 46 publications [6, 7]) are driving the true trajectory and foundational theories in this domain.

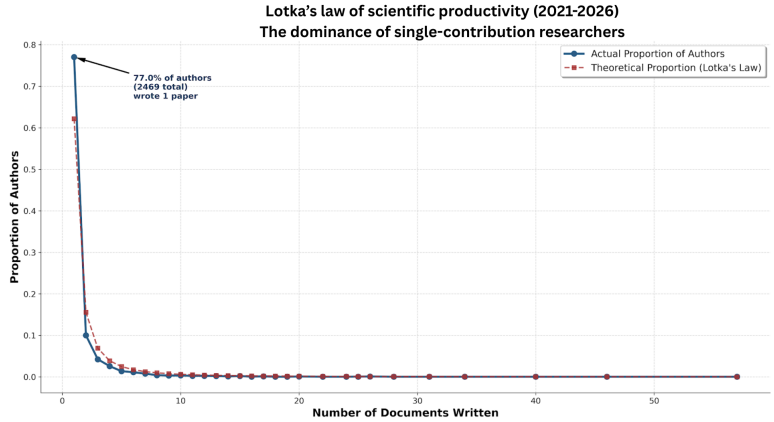


**Figure 5.** Lotka's law (retrieved February 2026).

Further, identifying the core of highly productive journals that contain the bulk of the relevant topics can be done using Bradford's law, which is described below.

**Bradford's law (journal scattering):** This law divides the literature into three zones based on the journals (sources). Our dataset clearly shows that the core zone (Zone 1) includes only a handful of venues where important research is concentrated; (1) *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* (41 publications); (2) *Proceedings of Machine Learning Research* (37 publications); (3) *Lecture Notes in Computer Science* (35 publications). Table 2 shows the distribution in some detail.

**Table 2.** Bradford's law of journal scattering: distribution of core, related, and peripheral sources (2021–2026) (retrieved February 2026).

| Bradford zone | Number of sources | Number of publications |
|---|---|---|
| Zone 1 (Core) | 20 | 377 |
| Zone 2 (Related) | 116 | 387 |
| Zone 3 (Peripheral) | 368 | 375 |
| **Total** | **504** | **1139** |

### *3.3. RQ3: Dominant thematic clusters (thematic mapping)*

To understand what the field is actually researching and how its topics relate to each other, we used thematic cluster analysis. This method groups the keywords that frequently appeared across documents, revealing underlying research themes. In this study, co-word analysis was applied to 250 author keywords from 1139 documents using the Biblioshiny interface. The underlying Louvain algorithm identified seven thematic clusters with a minimum keyword frequency of five.

Results are visualized on the thematic map as shown below (Figure 6). The horizontal axis shows centrality, or how well a theme connects with other themes in the field. The vertical axis shows density, or how well-developed a theme is internally. The dividing lines are drawn at the mean centrality (0.247) and mean density (20.23) across all seven clusters identified. Clusters above both thresholds are motor themes (central and dense); above density only are niche themes (dense but peripheral); above centrality only are basic themes (central but still developing); and below both are emerging or declining themes (neither central nor developed).

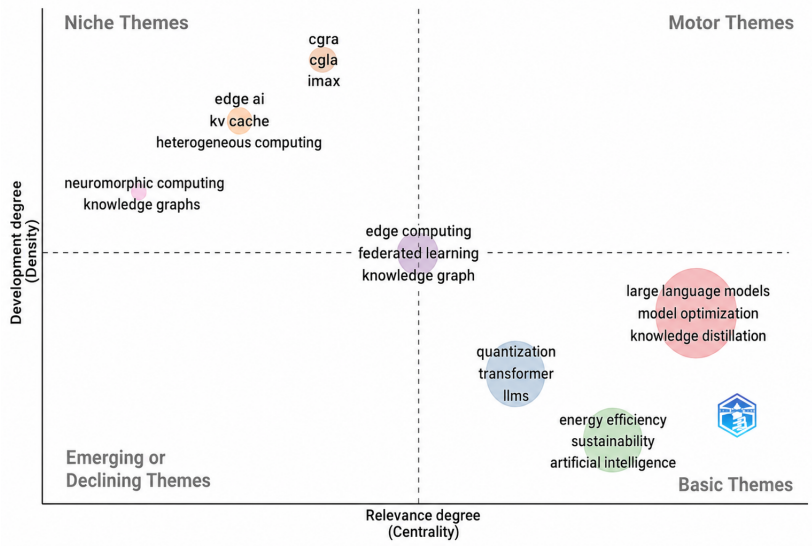


**Figure 6.** Dominant thematic clusters (retrieved February 2026).

- **Motor themes:** The algorithm identified that no cluster in this field crosses both thresholds simultaneously. The most internally coherent clusters (CGRA at density 38.17, edge AI at 24.35, neuromorphic computing at 20.00) all have centrality scores well below the mean (0.113, 0.061, and 0.040, respectively). This suggests that, although these topics are well-developed internally, they have limited influence across the broader research landscape. On the other hand, the most relevant clusters score low on density. Its emptiness clearly indicates that the carbon-intensive to carbon-neutral transition is in its early stage.
- **Niche themes:** Clusters like neuromorphic computing (brain-inspired energy-efficient hardware), coarse-grained reconfigurable architecture (CGRA) for parallel processing, and edge AI (on-device artificial intelligence without cloud reliance) can be observed in this quadrant. These are highly specified hardware-focused areas. CGRA has the highest density of all seven clusters (38.17), but the third-lowest centrality (0.113), and accounts for only 21 documents. This reflects a tightly connected hardware-architecture community publishing on reconfigurable computing for

AI acceleration, largely in isolation from the broader LLM-driven literature. Similarly, the edge AI cluster (density 24.35, centrality 0.061, 27 documents) focuses mainly on topics like edge AI, KV cache, and heterogeneous computing. Researchers in this cluster work closely on on-device AI inference, but their work is not strongly connected to the larger NLP research community. This highlights a significant hardware-software divide, meaning software solutions (quantization) are mainstream, but Green hardware solutions are still in a niche. The reason could be attributed to the fact that software solutions like quantization can be implemented and published rapidly on existing GPU infrastructure, while CGRA and neuromorphic computing require physical chip design and new toolchains that take years to develop. Our corpus also supports this idea. Quantization produces 249 documents with strong field connections (centrality 0.377), while CGRA yields only 21 documents with minimal cross-field influence (centrality 0.113). There is also no common benchmark for green hardware in the way that LLM research has standardized evaluation tools, which means these hardware communities mostly cite each other rather than engaging with the broader field.

- **Basic themes:** LLMs, energy efficiency, and quantization can be observed in this quadrant. The LLM cluster is the most connected of all (centrality 0.503, 985 documents), covering model optimization, knowledge distillation, fine-tuning, and NLP, making it the central thread running through the entire corpus. Energy efficiency (centrality 0.474, 340 documents) and quantization (centrality 0.377, 249 documents) are similarly well-referenced, linking sustainability, hardware efficiency, and model compression to the broader field [16, 24, 30]. However, all three score below the mean density (14.37, 11.65, and 13.65 respectively), meaning that while researchers are actively writing about these topics, the work is still scattered; there is no tight, agreed-upon core within each cluster yet, suggesting these themes are still evolving rather than settled.
- **Emerging or declining themes:** This quadrant includes edge computing cluster. The edge computing cluster (centrality 0.159, density 19.43, 106 documents) includes topics such as federated learning, knowledge graphs, IoT, and mobile edge computing. It lies very close to the average values on both axes, making it a transitional cluster rather than a clearly declining one. This suggests the area may continue to grow, especially as LLMs move toward on-device deployment [27].

### *3.4. RQ4: What are the trending topics?*

The AI world is changing at an unprecedented rate. We use the trending topic plot (Figure 7) to trace the word frequency based on year. Looking at the data deeply, one can observe a clear paradigm shift:

- **Pre-2023 era:** When the field was in its nascent stage, the focus of research was mainly in specific applications. Terms like sentiment analysis and natural language processing were prominent. Researchers applied large language models to various tasks.
- **2024–2025 (The efficiency era):** There was a marked shift in the focus of research in this era. The top trending terms became model optimization and quantization, which are now aggressively being applied to transformers and low-rank adaptation (LoRA) fine-tuning.

This shift effectively proves that academia has realized that the accuracy metric has hit a saturation point. The race ahead is based on efficiency, speed, latency, and compression. Tang et al. (2023) [12]

showed the importance of hardware-friendly low-precision quantization, which is now aggressively being applied to transformers.

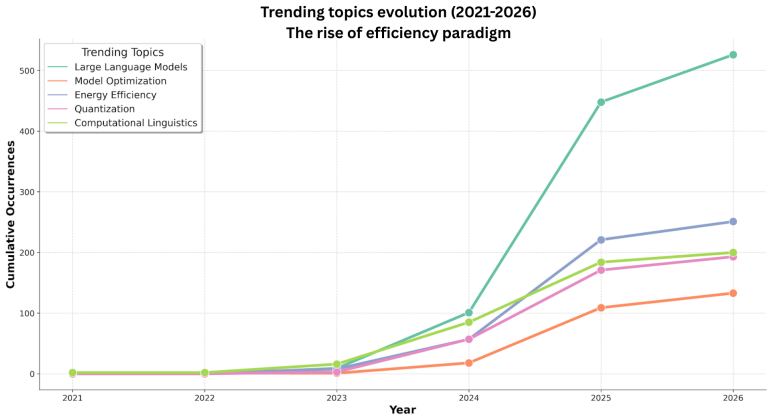


**Figure 7.** Trending topics (retrieved February 2026).

## 4. Discussion: the ground reality, research gaps, and future directions

Bibliometric analysis presents the data along with the research gaps in the available literature. The future trajectory depends on gaps discovered in the analysis. In this analysis, we mapped the keyword co-occurrence network and factorial analysis graphs to discover those gaps, which are listed below.

### *4.1. The carbon standardization*

The analysis revealed that the literature is focused on keywords such as green AI, sustainability, and energy efficiency. Studies show that hardware metrics, such as *floating point operations per second (FLOPs)*, *inference latency (milliseconds)*, and *joules consumed per token* are measured for defining efficiency.

While explicit carbon footprint or standardized carbon emission reporting frameworks are actively being explored, its presence is still limited in the literature. Models are indeed being compressed, but there is no globally accepted ISO standard metric that actually measures the carbon saved after model compression. Unless carbon metrics are explicitly benchmarked, efforts to mitigate the environmental impact of red AI risk remain merely theoretical or confined to paper [23, 33].

### *4.2. Edge deployment*

One of the most important conceptual advantages of SLMs are their localized computing capabilities. It is impractical to locally run an LLM [15] on smartphones or IoT devices. The data needs to be sent to the cloud. This incurs privacy risk and transmission energy.

In our analysis, edge computing (28 occurrences) is a growing field. Studies like Añón (2023) [26] discuss digital twins and edge offloading, and Zhang et al. (2025) [25] proposed collaborative edge partitioning for LLM inference. Sharshar et al. (2025) [29] gave a comprehensive survey of vision-language models on edge networks. However, the link between edge computing and sustainable AI is relatively weak. Future research should focus specifically on developing edge-native SLMs, which can

bring natural language understanding capability to IoT sensors and mobile devices consuming minimal battery.

### *4.3. The global south disparity*

In this research area, there is also an implicit geopolitical gap, as we have seen. Most major productions originate from China, the USA, and Europe (India is marginally present). For high-end AI research, thousands of GPUs are required per cluster (even if we are training SLMs, as it requires distillation of knowledge from LLMs).

The Global South (developing nations in Africa, South America, parts of Asia) is heavily lagging behind in this transition. Those tech giants that created the red AI problem are now selling green AI solutions. Future collaborative networks should grant open-source models such as llama and compute to researchers in the global south to develop locally relevant, diverse, and resource-constrained SLMs [27].

## 5. Conclusions

In this review paper, we have traced an exhaustive bibliometric journey of literature from 2020 to 2026. From Brown et al. (2020) [1]'s massive compute approach to Xiao (2023) [5]'s efficient quantization techniques, it is evident that the ecosystem of artificial intelligence is maturing.

There are three main takeaways of the entire analysis:

(1) **The red AI era is peaking:** The AI community has realized that there is a physical and environmental limit to raw compute and power scale. Exponential publication growth (109.82%) is not just proof of the interest in model optimization and quantization, but also evidence that the future is efficient models.
(2) **SLMs are the future:** SLMs have not just remained an alternative, but have become a fundamental necessity. Hardware accelerators, FPGAs, and software optimizations (knowledge distillation, PEFT, LoRA) have now become central themes in core CS literature.
(3) **A need for standardized carbon accounting:** Our research gap analysis highlights that green AI is still far from proper metrics. In the future, academia and industry should establish a globally verifiable metric like kgCO2eq per inference.

AI's future will not merely depend on how smart the model is, but will depend on how green, sustainable, and efficient it is. The path towards truly carbon neutral AI will not be achieved merely by increasing performance, but by efficiently compressing and intelligently optimizing AI models.

## Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

## Conflict of interest

The authors declare no conflict of interest.

## References


1. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., Language models are few-shot learners, *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, 1877–1901.
2. J. Deng, W. Dong, R. Socher, L. Li, K. Li, F. Li, ImageNet: a large-scale hierarchical image database, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, 248–255. https://doi.org/10.1109/CVPR.2009.5206848
3. J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423
4. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Commun. ACM*, **60** (2017), 84–90. https://doi.org/10.1145/3065386
5. G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, S. Han, SmoothQuant: accurate and efficient post-training quantization for large language models, *Proceedings of the 40th International Conference on Machine Learning*, 2023, 38087–38099.
6. Y. Liang, Y. Wang, Y. Li, Y. Zeng, Matrix-transformation based low-rank adaptation (MTLORA): a brain-inspired method for parameter-efficient fine-tuning, *Neural Networks*, **199** (2026), 108642. https://doi.org/10.1016/j.neunet.2026.108642
7. Y. Wang, B. Yin, GCL: group-shared continual learning fine-tuning for sparse LLMs, *Neurocomputing*, **675** (2026), 132918. https://doi.org/10.1016/j.neucom.2026.132918
8. T. Schick, H. Schütze, It's not just size that matters: small language models are also few-shot learners, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, 2339–2352. https://doi.org/10.18653/v1/2021.naacl-main.185
9. B. Cui, Y. Li, Z. Zhang, Joint structured pruning and dense knowledge distillation for efficient transformer model compression, *Neurocomputing*, **458** (2021), 56–69. https://doi.org/10.1016/j.neucom.2021.05.084
10. S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, et al., From words to watts: benchmarking the energy costs of large language model inference, *Proceedings of IEEE High Performance Extreme Computing Conference (HPEC)*, 2023, 1–9. https://doi.org/10.1109/HPEC58863.2023.10363447
11. Y. Li, B. Dong, F. Guerin, C. Lin, Compressing context to enhance inference efficiency of large language models, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, 6342–6353. https://doi.org/10.18653/v1/2023.emnlp-main.391
12. C. Guo, J. Tang, W. Hu, J. Leng, C. Zhang, F. Yang, et al., Olive: accelerating large language models via hardware-friendly outlier-victim pair quantization, *Proceedings of the 50th Annual International Symposium on Computer Architecture*, 2023, 1–15. https://doi.org/10.1145/3579371.3589038

13. X. Zhu, J. Li, Y. Liu, C. Ma, W. Wang, A survey on model compression for large language models, *Transactions of the Association for Computational Linguistics*, **12** (2024), 1556–1577 https://doi.org/10.1162/tacl_a_00704

14. W. Shao, M. Chen, Z. Zhang, P. Xu, L. Zhao, Z. Li, et al., Omniquant: omnidirectionally calibrated quantization for large language models, *Proceedings of International Conference on Learning Representations 2024*, 2024, 1–25.

15. S. Zeng, J. Liu, G. Dai, X. Yang, T. Fu, H. Wang, et al., FlightLLM: efficient large language model inference with a complete mapping flow on FPGAs, *Proceedings of the 2024 ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, 2024, 223–234. https://doi.org/10.1145/3626202.3637562

16. C. Lee, J. Jin, T. Kim, H. Kim, E. Park, Owq: outlier-aware weight quantization for efficient fine-tuning and inference of large language models, *Proceedings of the AAAI Conference on Artificial Intelligence*, **38** (2024), 13355–13364. https://doi.org/10.1609/aaai.v38i12.29237

17. S. Gholami, M. Omar, Can a student large language model perform as well as its teacher? In: *Innovations, securities, and case studies across healthcare, business, and technology*, Hershey: IGI Global Scientific Publishing, 2024, 122–139. https://doi.org/10.4018/979-8-3693-1906-2.ch007

18. H. Chen, J. Zhang, Y. Du, S. Xiang, Z. Yue, N. Zhang, et al., Understanding the potential of FPGA-based spatial acceleration for large language model inference, *ACM Trans. Reconfig. Techn.*, **18** (2024), 1–29. https://doi.org/10.1145/3656177

19. S. Pimenow, O. Pimenowa, P. Prus, Challenges of artificial intelligence development in the context of energy consumption and impact on climate change, *Energies*, **17** (2024), 5965. https://doi.org/10.3390/en17235965

20. S. Park, K. Kim, J. So, J. Jung, J. Lee, K. Woo, et al., An LPDDR-based CXL-PNM platform for TCO-efficient inference of transformer-based large language models, *Proceedings of IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2024, 970–982. https://doi.org/10.1109/HPCA57654.2024.00078

21. M. Argerich, M. Patiño-Martínez, Measuring and improving the energy efficiency of large language models inference, *IEEE Access*, **12** (2024), 80194–80207. https://doi.org/10.1109/ACCESS.2024.3409745

22. X. Shen, P. Dong, L. Lu, Z. Kong, Z. Li, M. Lin, et al., Agile-Quant: activation-guided quantization for faster inference of LLMs on the edge, *Proceedings of the AAAI Conference on Artificial Intelligence*, **38** (2024), 18944–18951. https://doi.org/10.1609/aaai.v38i17.29860

23. S. Iftikhar, S. Davy, Reducing carbon footprint in AI: a framework for sustainable training of large language models, *Proceedings of the Future Technologies Conference (FTC) 2024, Volume 1*, 2024, 325–336. https://doi.org/10.1007/978-3-031-73110-5_22

24. M. Kim, S. Lee, W. Sung, J. Choi, RA-LoRA: rank-adaptive parameter-efficient fine-tuning for accurate 2-bit quantized large language models, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024, 15773–15786. https://doi.org/10.18653/v1/2024.findings-acl.933

25. M. Zhang, X. Shen, J. Cao, Z. Cui, S. Jiang, Edgeshard: efficient llm inference via collaborative edge computing, *IEEE Internet Things J.*, **12** (2025), 13119–13131. https://doi.org/10.1109/JIOT.2024.3524255

26. C. Chen, K. Zhao, J. Leng, C. Liu, J. Fan, P. Zheng, Integrating large language model and digital twins in the context of industry 5.0: Framework, challenges and opportunities, *Robot. Comput.-Int. Manuf.*, **94** (2025), 102982. https://doi.org/10.1016/j.rcim.2025.102982

27. R. Zhang, J. He, X. Luo, D. Niyato, J. Kang, Z. Xiong, et al., Toward democratized generative AI in next-generation mobile edge networks, *IEEE Network*, **39** (2025), 251–260. https://doi.org/10.1109/MNET.2025.3541078

28. G. Kim, S. Hwang, B. Jang, Efficient compressing and tuning methods for large language models: a systematic literature review, *ACM Comput. Surv.*, **57** (2025), 1–39. https://doi.org/10.1145/3728636

29. A. Sharshar, L. Khan, W. Ullah, M. Guizani, Vision-language models for edge networks: a comprehensive survey, *IEEE Internet Things J.*, **12** (2025), 32701–32724. https://doi.org/10.1109/JIOT.2025.3579032

30. Y. Guo, Z. Hao, J. Shao, J. Zhou, X. Liu, X. Tong, et al., PT-BitNet: scaling up the 1-bit large language model with post-training quantization, *Neural Networks*, **191** (2025), 107855. https://doi.org/10.1016/j.neunet.2025.107855

31. J. Fernandez, C. Na, V. Tiwari, Y. Bisk, S. Luccioni, E. Strubell, Energy considerations of large language model inference and efficiency optimizations, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025, 32556–32569. https://doi.org/10.18653/v1/2025.acl-long.1563

32. R. Aralimatti, S. Shakhadri, K. Kruthika, K. Angadi, Fine-tuning small language models for domain-specific AI: an edge AI perspective, In: *Intelligent systems and applications*, Cham: Springer, 2025, 503–520. https://doi.org/10.1007/978-3-031-99965-9_31

33. F. Jeanquartier, C. Jean-Quartier, P. Rieder, V. Misirlić, C. Pasero, R. Hohensinner, et al., Assessing the carbon footprint of language models: towards sustainability in AI, *Resour. Conserv. Recy.*, **226** (2026), 108670. https://doi.org/10.1016/j.resconrec.2025.108670


AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)