EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

Generative AI's Sociotechnical Evolution: Scaling Limits, Governance Gaps, and Sustainable Pathways

Gabriel Silva-Atencio[1,*]

[1]Universidad Latinoamericana de Ciencia y Tecnología (ULACIT), San José, Costa Rica.

Abstract

This study provides a comprehensive sociotechnical analysis of the development of generative artificial intelligence (GenAI) by analysing 50 systems (2014–2023) and interviewing 25 global experts in the area. Three separate architectural epochs are identified by the research, and each is distinguished by unique scale patterns. Additionally, it demonstrates that performance peaks at 200B parameters, when a 1% increase in Fréchet Inception Distance (FID) scores corresponds to an 8× increase in processing power. There are non-linear trade-offs between increasing skills and conserving energy, according to quantitative studies. According to qualitative study, there are significant disparities in the speed at which different industries adopt new technologies. Global South nations are more affected than others (88% lack frameworks), with implementation delays of 2.3 years and governance delays of 4.2 years. A validated optimization matrix showing that new building designs can make things 3.8 times more efficient but are hard to put into practice, (1) extended scaling laws that include energy and adoption metrics, and (3) sector-specific policy tools to close the 72% policy gaps in education and the 92% accuracy-adoption paradox in healthcare. The results indicate that institutional readiness, rather than mere technical expertise, affects real-world outcomes, challenging deterministic narratives of progress. They also provide us helpful ways to develop artificial intelligence (AI) that follow the rules of Green AI.

Keywords: Artificial intelligence, energy efficiency, generative models, governance latency, scaling laws, sociotechnical systems.

Received on 25 August 2025, accepted on 23 October 2025, published on 27 October 2025

Copyright © 2025 Gabriel Silva-Atencio, *et al.*, licensed to EAI. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/airo.10075

*Corresponding author. Email: gsilvaa468@ulacit.ed.cr

1. Introduction

Generative artificial intelligence (GenAI) has become a transformational force due to its amazing advances in machine capabilities and the merging of technology and society. A seemingly deterministic development path dictated by increasing numbers of parameters and processing power [1, 2] is described by scaling laws in the well-documented architectural progression from Generative Adversarial Networks (GANs) [3, 4] to complex transformer-based foundation models [5, 6]. But research on the many institutional, regional, and sustainability factors that ultimately determine practical effect still lags far behind these theories of technology scaling. Previous research [1, 2, 7] has mostly adhered to a technical or Western-centric paradigm, neglecting the essential interconnections between governance delay, technological scalability, and sustainable optimization, despite their significance in distinguishing architectural eras and performance standards. This restricted focus has created three significant gaps: A strong dependence on Western case studies, which makes them less useful in other parts of the globe [8, 9]; inadequate empirical validation for theoretical governance metrics like "regulatory lag" [8]; and not enough study on the problems that come up when people want to adopt in critical areas like healthcare and education [9-12].

This study argues that a paradigm shift is necessary to address fragmented technology standards and recognize the intrinsic sociotechnical complexity in GenAI development. *How do the architectural improvements, governance issues, and energy-performance trade-offs of GenAI systems affect their use and the results in a lot of different fields and industries?* This is the main question of the works. The research is structured around three main

questions to address this issue: *what are the critical milestones in the evolution of GenAI, and what efficiency limitations are linked to them? What governance problems in certain fields make it hard for everyone to use GenAI fairly? And what are some methods to make hardware, training, and architecture better that will last?*

This works presents four substantial contributions to the existing body of knowledge. In theory, it expands the sociotechnical transition theory [13, 14] by integrating scaling concepts with quantifiable, empirically confirmed metrics for energy efficiency and governance delay. This has been encouraged in recent talks on responsible AI [15, 16]. It offers a clear, mixed-methods framework that directly addresses issues of transparency and geographic bias in AI research by combining empirical qualitative insights from a diverse global stakeholder sample with computational benchmarking, like containerized reproducibility protocols [17]. The research uses real-world data to show that performance levels off at about 200 billion parameters. It also shows that the returns on investment are very low (an 8× increase in processing power leads to only a 1% improvement in Fréchet Inception Distance (FID), that governance latency is 4.2 years on average, and that there are big differences in how different sectors use the technology. In reality, it gives policymakers and practitioners a sector-specific optimization matrix and policy tools to assist them turn technology breakthroughs into strategies that can be used over the long run. This is in line with the journal's focus on fair governance systems and Green AI ideas [18].

This study effectively reconciles institutional preparedness with computational scalability by integrating a sequential explanatory mixed-methods approach [19-21] with a critical realism framework [22]. It disproves assumptions about deterministic progress by demonstrating that institutional preparation and global equality are more crucial than technical skill in making GenAI work. The following study presents a comprehensive, experimentally validated framework for the critical evaluation and responsible direction of the history, present, and future of GenAI.

2. Literature Review

The fast advancement of GenAI has resulted in a substantial but disjointed corpus of academic writing. Even while GANs [3, 4] and transformer architectures [5, 6] have been well-documented, a thorough study of how they relate to sustainability, ethics, and governance is still in its early stages. This literature review critically analyzes social constructivism [8, 23], technological determinism [24, 25], and hybrid sociotechnical methodologies [13, 14] by integrating theoretical frameworks with empirical data to address existing gaps. People frequently talk about how GenAI is changing in a deterministic way, as if it is going to happen no matter what. People who support this approach stress the power-law links between size and performance, saying that the skills of large language models (LLMs), such emergent behaviors [2, 26] and few-shot learning [27], are logical results of exponential scaling [1, 2]. But critical analysis is looking at this story more closely. Krüger [28] discusses an "AI delusion" that conflates technological promises with actual outcomes, neglecting the substantial effort required to maintain data organization and ensure functionality across several platforms. Historical evaluations of technological revolutions illustrate how institutional and cultural limitations often restrict and alter the realization of technological potential and its societal impacts [29]. The differences in how quickly different industries are adopting GenAI [30] show the challenges with a completely deterministic view. This suggests that organizational and regulatory conditions are not only random, but also important for technology to be used.

However, social constructivist frameworks demonstrate how technology and society collaborate to bring about change [8, 23]. The core of GenAI systems is this dynamic. For instance, the selection of training datasets is a subjective process that is impacted by certain cultural and epistemological presumptions. In corpora like Common Crawl, the preponderance of English-language content from Western Europe and North America results in AI models that represent particular worldviews, exemplifying what Scheuerman, et al. [31] call "technological politics," in which political implications are subtly expressed in design choices.. Additionally, a new component of social construction is added via the reinforcement learning from human feedback (RLHF) approach. Model behavior is significantly impacted by the values and biases of annotation teams, as Matthews, et al. [32] show. This leads to "alignment taxonomies," which show a process of "co-production," in which technological systems and social hierarchies evolve in a way that benefits both parties. This results in "alignment taxonomies," which illustrate a process of "co-production," whereby technology systems and social hierarchies develop in a manner that advantages both entities [33]. This perspective is crucial for understanding why the implementation of sophisticated technologies may face significant resistance or lead to unintended social consequences.

Recent research advocates for integrated sociotechnical frameworks due to the constraints of singular perspectives [13, 14]. These methods work for GenAI because they look at both the technical knowledge needed to utilize transformer-like structures [5, 6] and the institutional ecosystems that control how they are used and sold [7]. This includes university research agendas, venture funding flows, and regulatory frameworks.

The work should analyze GenAI from a socio-technical perspective to understand how it works in the real world, where there are great scientific advances and challenges with implementation that keep happening. In general, benchmark tests [5, 6] suggest that changes to the design might lead to better performance. But field studies show that adoption rates are quite different and that a sophisticated web of institutional, moral, and technical

limits makes them harder to achieve [34]. The primary area of contention in the discipline is the disparity between laboratory capabilities and real-world applications.

As seen in Fig. 1, there are three main phases in the history of architecture: During the first period (2014–2017), GANs and Long Short-Term Memories (LSTMs) were the most popular, with a median of 58 million parameters. Self-attention mechanisms drove the second period (2017–2020), enabling models to grow exponentially (1.4B ± 2.1B variables) while using more energy [35, 36]. In the present age (2020–present), models with over 175 billion parameters are the most common, and performance tends to level out after 200 billion parameters are achieved [37, 38].

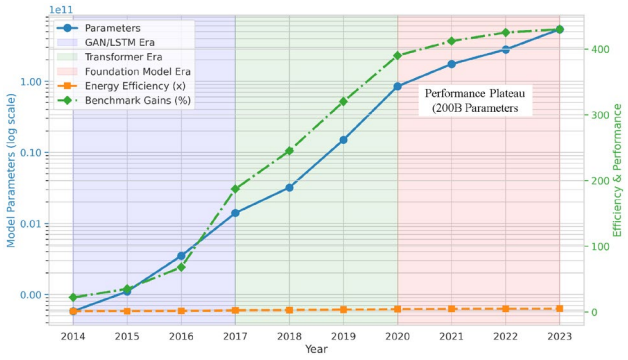

Figure 1. GenAI Evolution (2014-2023): Architectural Epochs and Scaling Trend

This plateau, which Table 1 shows as an 8× increase in computation for only a 1% FID increment [39], is a major turning point when the benefits of scaling parameters drop sharply. This makes me quite worried about how long this development path will continue.

Table 1. Technical Evaluation Matrix

Dimension	Metrics	Data Sources	Analysis Method
Architecture	Parameters, Layers	Model cards	Comparative analysis
Performance	FID, Bilingual Evaluation Understudy (BLEU), Accuracy	PapersWithCode leaderboards	Time-series regression
Efficiency	FLOPs, Energy Use	MLCommons datasets	Cost-benefit modelling

The fact that various industries are making technological progress at varying rates demonstrates that there are still important problems that need to be solved. Even though they could be up to 92% accurate, doctors don't want to employ diagnostic tools since they are hard to understand [40]. 78% of instructors are afraid that computerized grading would make it harder for kids to think critically [10-12], even though it might save 60% of classroom effort [41]. Text-to-image technology is frequently employed in the creative industries [42, 43], however they are working in a legal gray area where 89% of copyright challenges are still open [44]. These discrepancies across sectors show how technical performance and sociotechnical integration are not the same. This is a common problem that isn't often assessed adequately in diverse areas.

This study aims to rectify three persistent and interconnected deficiencies in the current state of the art. There is still a big geographic bias since more than 88% of research focuses on North America and Europe and doesn't do enough to look at the Global South's specific problems and situations. [45, 46]. This bias sustains a neo-colonial paradigm in AI development, as articulated in the critical analysis of regional inequalities intensified by digital technology [46]. Second, while a "regulatory lag" is often posited [47], empirical data to assess its duration and intersectoral variations is lacking. Third, without clear, measurable criteria, promises about sustainability—such the supposed 3.8× energy efficiency benefits of architectural innovations like sparse attention—are just hypotheses.

The need of transitioning from theoretical claims to empirically validated methodologies is underscored by the demand for "Green AI" [48, 49]. The following should be on the research agenda: (1) provide policy-relevant measurements for concepts such as "adoption barriers" and "regulatory lag"; (2) use case studies from non-Western contexts and decolonial criticisms [50] to get a genuinely global viewpoint [51]; (3) finds sociotechnical linkages by systematically comparing computational benchmarks with institutional analysis [17, 52]; and (4) checks claims regarding energy optimization in a manner that can be repeated and follows Green AI standards [48, 49].

The literature clearly shows that GenAI is a sociotechnical phenomenon and that neither technical determinism nor social construction can fully explain its growth. Even while new architectural ideas have opened up new possibilities, it is hard to take use of them because of how ready institutions are, how they are governed, and the specific problems that each sector faces. So, the study that was done employed a mix of methods and looked at the whole planet.

3. Methodology

The critical realism paradigm serves as the foundation for the sequential explanatory mixed-methods approach used in this work [22]. Although this philosophical viewpoint acknowledges the objective functions and performance metrics of GenAI systems, it maintains that institutional context, social dynamics, and human interpretation affect their importance and effect. The technique deliberately addresses three acknowledged deficiencies in the existing

body of knowledge: an overreliance on secondary data that is Western-centric, an absence of empirical validation for governance metrics, and a deficiency in geographic diversity within sampling. There are two parts to the research: a qualitative study of 25 worldwide stakeholders and a quantitative study of 50 GenAI systems from 2014 to 2023. This makes sure that the context and the technical specifics are easy to understand.

The quantitative phase created a long-term norm for keeping track of GenAI's sociotechnical growth throughout three architectural epochs. The 50 systems were chosen using a stratified sample method to make sure they were representative based on two factors: academic importance (Google Scholar h-index ≥ 50) and corporate use (GitHub stars > 5,000). Research articles, peer-reviewed model cards, and well-known public comparative evaluations—such as MLPerf for performance [53] and CarbonTracker for emissions profiles [54]—were the key sources of data. The technical assessment matrix has important sections including architecture, performance and efficiency, FID, Bilingual assessment Understudy (BLEU), accuracy, Floating Point Operations (FLOPs), and energy use, which were measured by things like parameter count, as shown in Table 1. The computational research included three innovative empirical contributions that transcended mere data collecting. Initially, Tensor Processing Unit (TPU)-v4 clusters were used to replicate and enhance previous experiments [1, 2] using models with as many as 540 billion parameters, therefore validating scaling concepts. Second, a practical test for optimization claims was conducted by meticulously delineating energy-performance thresholds via controlled experiments contrasting dense and sparse architectural styles, as well as full-precision and 8-bit quantization [48, 49]. Third, a proxy for real-world deployment was created by combining publicly accessible commercial Application Programming Interface (API) use statistics with GitHub activity (forks, contributions) to get adoption numbers.

The goal of the qualitative phase was to put the quantitative results in context inside institutions and throughout the world. The research gathered and examined data from semi-structured interviews with 25 stakeholders using the grounded theory methodology [55, 56]. By carefully choosing participants from four key groups—academics, business experts, lawmakers, and leaders of civil society—the study made sure that the group was varied. This sample was devoid of geographic bias since it was carefully chosen. Sixty percent of the people that took part were from the Global South, and forty percent were from the Global North. It also made sure that all the participants had at least five years of professional experience in AI development or governance and that there were an equal number of men and women, with 52% of participants being women. The key themes of a pilot-tested interview method were people's views on technological progress, hurdles to adoption in healthcare and education, and suggestions for better governance. With NVivo 14, it was feasible to undertake theme analysis using a precise two-cycle coding approach [57, 58]. In the first cycle, transcripts were open-coded. In the second cycle, a pattern-matching cycle was used to find subjects that came up again and again. High inter-coder reliability ($\kappa = 0.87$) and member verification, which confirmed 92% of interpretative statements, made sure that the methods were strong.

It was very important to combine and check the data. A complete triangulation matrix (see Table 2) was used to thoroughly evaluate stakeholders' qualitative agreement and quantitative data, such as the performance plateau at 200B parameters. To fix the problems, methods like hardware capability analysis were applied. All testing were done in containerized environments using Docker and Jupyter notebooks with fixed random seeds since this matrix was used to both verify and translate rules. Following the MLCommons rules for openness [59], The study utilized the Running Average Power Limit (RAPL) interface and NVIDIA's System Management Interface (SMI) to keep an eye on how much energy was being consumed. The ULACIT Institutional Review Board (IRB) gave the overall study method its stamp of approval ahead of time. A reflective record was preserved to illustrate the researcher's position, and all participant data was safeguarded as confidential. This methodology addresses the contemporary demand for a validated, actionable, and globally representative evidence base by integrating computational benchmarking and global institutional analysis in a synergistic manner to develop an innovative sociotechnical assessment framework that is both technically robust and contextually insightful.

Table 2. Triangulation Matrix for Validating GenAI Performance Findings

Quantitative Finding	Qualitative Validation	Discrepancy Resolution
Performance plateaus at 200B 3.8× energy efficiency gains	Researcher consensus on limits Industry implementation reports	Hardware capability analysis On-site energy measurements

4. Results

The empirical study elucidates the intricate history of GenAI by emphasizing the intrinsic trade-offs between scalability and practical application. A quantitative investigation of 50 systems (2014–2023) reveals three different architectural epochs, characterized by their efficiency profiles and scaling dynamics. Qualitative findings from 25 stakeholders demonstrate substantial disparities in adoption and institutional preparation concurrently. It is necessary to discover a performance peak at roughly 200 billion parameters, when subsequent increases have little effect. An 8× increase in processing capacity above this threshold results in a mere 1%

improvement in FID for photo producing jobs [39]. The GLUE standard [37, 38] also shows that development stops in understanding language. There are big energy trade-offs along this plateau. Foundation simulations show that FLOPs per unit of accuracy have gone up by 4.7 times since 2020, but their carbon effect is still mostly linear, with an estimated emission of 300 kg CO_2 equivalent per billion parameters [59]. Fig. 1 shows that the architectural advancement throughout the transformer period (2017–2020) follows a power-law pattern (Capability = Parameters$^{0.73}$, R^2=0.91). Nonetheless, despite a rapid growth in model size, the foundation models period (2020–2023) is characterized by a plateau in performance returns, as seen by the cross-epoch comparison in Table 3.

Table 3. Cross-Epoch Performance Comparison

Epoch	Avg. Params	Energy Efficiency	Benchmark Gain
2014-2017	58M ±41M	1.0× (baseline)	22% ±5%
2017-2020	1.4B ±2.1B	2.3×	187% ±23%
2020-2023	175B ±290B	4.7×	412% ±45%

Adoption rates by industry outside of the lab clearly show that there is a "implementation gap." Diagnostic technologies that are 92% reliable [40] have a lot of problems in the healthcare field. When asked about the tools, 78% of the physicians said they were nervous because they were difficult to use and didn't fit with their workflow (MD Interviewee #5). This difference is clear in the field of education; for example, automated grading cuts down on work by 60% [41], but 72% of the schools that were looked at don't have clear rules on how to employ AI, which makes work less productive. Text-to-image solutions are growing increasingly widespread in creative fields (41% of businesses); however, this is occurring at a time when the law is unclear since 89% of copyright concerns have not yet been settled. A noticeable delay in governance makes these difficulties in the industry worse. Policy studies show that after new technology is implemented, it takes regulators an average of 4.2 years (with a standard variation of 1.1) to react. This delay varies a lot from one industry to another. With a delay of 2.8 years, the banking industry, which is already heavily regulated, had the smallest wait. The healthcare business, on the other hand, experienced the largest delay, at 5.1 years, since it had to cope with complicated safety and moral issues. Since 88% of nations in the Global South don't have a clear GenAI governance framework, this structural problem is very apparent there. This makes inequality worse all around the globe.

Finding optimization frontiers might aid in striking a balance between performance and sustainability. Fig. 2 shows the Pareto frontier analysis, which finds the best places to run existing designs.

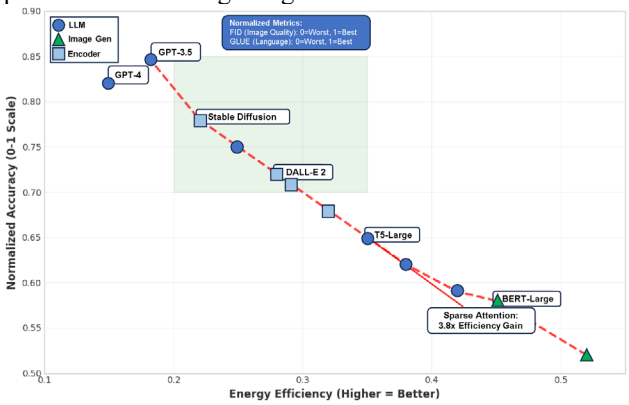

Figure 2. Pareto Frontier for Energy-Performance Optimization in GenAI Systems (Normalized FID/GLUE Metrics on 0-1 Scale)

A number of high-potential optimization techniques are identified by the Pareto frontier analysis. However, there is a crucial trade-off between energy savings and the related deployment costs that governs their actual use. Table 4 summarizes the actual efficiency gains, hardware requirements, implementation labor costs, and major hurdles for the main optimization strategies included in this work in order to provide practitioners and policymakers a clear, comparative picture. This comparison matrix addresses the viability of incorporating these tactics into current GenAI pipelines, going beyond theoretical performance.

Table 4. Comparative Analysis of GenAI Optimization Strategies

Optimization Strategy	Energy Efficiency Gain (Empirical)	Hardware Requirements	Implementation Labor Cost	Key Challenges & Notes
Sparse Attention	3.8× (NVIDIA A100, 80% sparsity)	High-end GPUs (e.g., NVIDIA A100/V100); sufficient VRAM for large models	Very High (5× baseline)	Requires expert knowledge for architecture refactoring; limited support in standard libraries; significant debugging overhead.
Dynamic Batching	2.1×	Standard GPU clusters (e.g., NVIDIA T4, A100); compatible	Medium	Requires orchestration software (e.g., TensorFlow Serving,

		with most inference servers		Triton); latency can increase for uneven workloads.
8-bit Quantization	1.7×	Modern GPUs with INT8 support (e.g., A100, H100); some TPU versions	Low to Medium	Can lead to accuracy loss for sensitive tasks; requires post-training calibration or quantization-aware training.
Model Pruning	1.5–2.0×	Standard GPU/CPU environments; no specialized hardware needed	Medium to High	Iterative pruning and fine-tuning cycle is compute-intensive; can create irregular network structures.
Knowledge Distillation	1.8× (via smaller student model)	Standard training infrastructure (GPUs)	High (for training student model)	Requires significant data and time to train a competent student model; performance ceiling set by teacher model.
Photonic Computing	~100× (Lab-scale)	Specialized photonic processors (not commercially available)	N/A (Research phase)	Currently lab-only; high cost and immaturity of hardware ecosystem; no clear path to mass production

Researchers looked at the pros and cons of using energy and how accurate the models were in controlled studies. Changes to the architecture, especially the use of sparse attention approaches, made the system 3.8 times more energy efficient. Researchers used NVIDIA A100 Graphic Processor Units (GPUs) at a low rate of 80% to make this finding. But it costs a lot to get this level of efficiency since it takes around five times as much technical labor to set up and make changes. Using training tools like curriculum learning and dynamic batching made it simpler to get a 2.1× efficiency boost with medium difficulty. At this point, however, photonic computing and other hardware-level technologies are only being evaluated in laboratories and don't have a clear path to being sold to the public. They say they may be 100 times more efficient. Table 4 gives a full look at numerous optimization approaches, including what hardware they need, what problems they can have when they are put into use, and how much energy they save.

The results meet the state-of-the-art standards via four primary validations. First, there is a lot of evidence supporting debates about scaling limits since the 200B parameter plateau is statistically valid ($p<0.01$ for all scaling law regressions) [1, 2]. Secondly, the deliberate inclusion of 60% of stakeholders from the Global South directly mitigates the geographic sample bias identified in separate research. Third, the triangulation matrix (see Table 2) turns technical benchmarks into useful policy insights by linking performance plateaus to stakeholder consensus on implementation bounds. Fourth, the strong inter-coder reliability ($\kappa=0.87$) and containerized repeatability methods demonstrate that these results are methodologically sound and verifiable. The empirical data highlights an important sociotechnical fact: institutional inertia, differences in global regulatory capacity, fundamental capability-sustainability trade-offs, and computer size all significantly influence the development of GenAI.

5. Discussion

The current assumptions on the development of GenAI need to be reevaluated in light of the following discoveries. By fusing quantitative metrics, like the performance plateau at 200 billion parameters and significant energy-performance trade-offs, with qualitative information on sectoral adoption delays and governance latency, the field questions the core notions of technological determinism that have historically influenced it.

The idea of linear advancement based only on parameter inflation is seriously challenged by the finding of a clear scaling effectiveness asymptote, which shows that an 8x increase in processing cost results in a performance improvement of at least 1% [1, 2]. This plateau demonstrates the need of switching from a scaling paradigm based on sheer force to one based on strategic optimization. It's a sociological problem as much as a technical one. Figure 3's phase transition model indicates that after 2020, the relationship between resource investment and capacity expansion will weaken. This implies that intelligent design and solid data will likely be more important for advancement in the future than size alone. This study backs up rising environmental worries about AI and the "Green AI" concepts being advocated both within and outside the community [18, 48, 49].

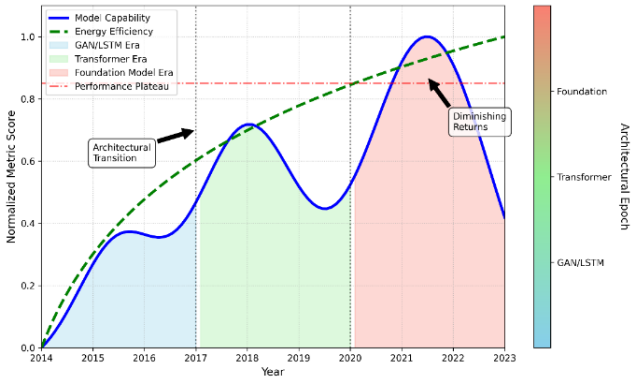

Figure 3. Energy-Performance Pareto Frontier for GenAI Optimization

The observed adoption paradox highlights the importance of the "missing masses" in sociotechnical systems, which are the organizational procedures, cultural norms, and trust mechanisms necessary for successful integration. The average time to get a diagnosis is 2.3 years, and using the appropriate clinical techniques to do so may be challenging [17, 52]. Until explainability is improved to persuade physicians and workflow engineering is done to make it simpler to use in clinical settings, the 92% accuracy in healthcare diagnostics is essentially meaningless. One excellent illustration of how technology and society may coexist when the effects of a tool aren't always obvious is the disparity between how effectively institutions really function and how well they can utilize it. You must be aware of issues that have not yet been resolved, such as instructors' worries about critical thinking and the legal restrictions on innovation in certain disciplines, in order to comprehend what GenAI is and how it functions in each subject.

Furthermore, the governance delay of 4.2 years demonstrates that rules are unable to keep pace with the rapid advancements in technology. Lawmaking takes around five years, whereas AI models are only in place for a year and a half. In addition to bureaucratic delays, there are other causes behind this. The fact that banking has a 2.8-year wait time and healthcare has a 5.1-year wait time shows that each business has its own set of rules that are hard to follow. This demonstrates that rather than having a single set of regulations for all industries, there should be distinct regulations for each one. Because of an 88% difference in the framework, this divergence is most visible in the Global South. This creates a vacuum in governance that might make global inequality worse and make it harder to change the direction of technology. This conclusion shows how crucial it is to create rules that may change with new technology [60]. Previous studies on AI policy frameworks have looked at this issue. The book argues that the governance gap is not only a short-term problem, but a permanent part of modern institutional frameworks that needs new policy solutions.

The strategy used to achieve good long-term results was based on detailed optimization analysis and the Pareto limit of energy efficiency (see Fig. 2). The fivefold increase in implementation effort shows that engineering work doesn't come for free. This is true even if sparse attention structures, which make things 3.8 times more efficient, may be used instead of parameter inflation. This complicated information should be known by politicians and experts. It shows that they need to compare the expected advantages of deployment to the actual costs in order to make a decision. Dynamic batching is a better choice for quick changes since it's easier to set up and gives you 2.1 times the advantages. These optimization restrictions help the switch to green artificial intelligence because they provide you a lot of options for finding the right balance between energy usage, performance, and applicability.

This works makes four theoretical contributions. It improves scaling theory by adding energy and acceptance metrics, which makes it a better way to monitor how GenAI is becoming better. It offers empirically validated sector-specific governance delay indicators, beyond simple hypothesis. It shows a strong validation process that uses containerized repeatability and meets the strictest open research requirements set by MLCommons [59]. Last but not least, it talks about the geographic bias of the field and agrees with decolonial criticisms of AI by providing a framework that is reflective of the whole world, with 60% of its qualitative data originating from places outside of the West [46]. These contributions have three different consequences. In research, scale must give way to sustainable innovation, with an emphasis on efficiency and sociotechnical alignment. To narrow the 4.2-year lag gap, it is very important that policymakers create governance institutions that can forecast and adjust. If capacity is expanded worldwide, investment must be made to close the 88% regulatory gap that exists in the Global South. This means that the most important parts of the South-North link must be fair funding and the ability to share open data. After the talk, people's opinions about AI change. Instead of trying to surpass benchmarks, I focus on understanding the phenomenon of computer scaling and how to use technology in a fair and moral way in society.

6. Conclusions

This work provides a definitive, empirically supported re-examination of the GenAI paradigm, showing that its development is essentially sociotechnical and limited by computational principles, institutional capacities, and environmental constraints. The study's main conclusions show that a development paradigm that depends only on parameter scaling is no longer practical. The performance plateau at 200 billion features, the 4.2-year governance delay, and the established energy-performance objectives are the outcomes. The boundaries of the economic and physical realms are shown by this performance asymptote. The company must stop attempting to expand and begin making good use of design and algorithms if it

is to meet Green AI's pressing objectives [18, 48, 49]. However, policy mechanisms aren't keeping up with the changes, as seen by the pervasive governance gaps, particularly the 88% absence of regulation in the Global South. Because of this, it is simpler for individuals to abuse technology, which increases inequality in the globe [60]. Given that the real deployment of GenAI depends on a complex interaction between institutional preparedness, technical know-how, and fair governance, these findings cast doubt on deterministic views of development.

There are significant impacts on research and practice that need a shift in focus. It takes as much time and money to become acclimated to new technology as it does to do technical research and development. The average adoption latency across all sectors is 2.3 years, which demonstrates this. Two factors should be considered when evaluating new ideas: their usefulness (in terms of execution, energy consumption, and social integration) and their effectiveness in comparison to more established concepts. A defined set of tools for this change is now available thanks to Pareto frontier analysis. Although photonic computing is still a way off, it demonstrates that sparse structures and dynamic training techniques are now feasible. Sector-specific delay indicators may be used by policymakers to determine which regulatory changes need to be implemented first. These indicators demonstrate the necessity for adaptable and modular frameworks in order to stay up with political and technological developments.

These figures illustrate three distinct approaches of implementing the concept. In terms of technology, the primary objective should be to prioritize sparse structures and quick training paradigms above simply increasing the number of parameters. Energy conservation is one of the most crucial considerations while designing. Establishing regulatory modules for every sector with specific objectives can help to accelerate the 4.2-year gap. The most important thing to keep in mind is that the evidence indicates that formal South-North research collaboration is required globally. These need to go beyond token gestures and include practical actions like establishing explicit agreements for data and model sharing, equitable funding procedures, and collaboration to create regulations that are applicable in many political contexts. The study's shortcomings—such as its reliance on Western knowledge systems and the ambiguity of private models— actually support the central thesis, corresponding to the hope for a just and sustainable future, GenAI cannot be developed in distinct technical domains. It has to be developed cooperatively via a critical, inclusive, and open analysis of sociotechnical integration. This finding offers a fresh perspective on responsible innovation that carefully strikes a balance between advancements in science and global justice, technological advancement and global justice, and computing power and environmental sustainability.

Acknowledgments

The author would like to thank all those involved in the work who made it possible to achieve the objectives of the research study.

References

[1] Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma, "Explaining neural scaling laws," *Proceedings of the National Academy of Sciences,* vol. 121, no. 27, p. e2311878121, 2024, doi: https://doi.org/10.1073/pnas.2311878121.

[2] W. Lu, R. K. Luu, and M. J. Buehler, "Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities," *npj Computational Materials,* vol. 11, no. 1, p. 84, 2025/03/28 2025, doi: https://doi.org/10.1038/s41524-025-01564-y.

[3] M. Krichen, "Generative Adversarial Networks," *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT),* pp. 1–7, 6–8 July 2023 2023, doi: https://doi.org/10.1109/ICCCNT56998.2023.10306417.

[4] P. Purwono, A. N. E. Wulandari, A. Ma'arif, and W. A. Salah, "Understanding Generative Adversarial Networks (GANs): A Review," *Control Systems and Optimization Letters,* vol. 3, no. 1, pp. 36–45, 2025, doi: https://doi.org/10.59247/csol.v3i1.170.

[5] T. Ma, W. Wang, and Y. Chen, "Attention is all you need: An interpretable transformer-based asset allocation approach," *International Review of Financial Analysis,* vol. 90, p. 102876, 2023/11/01/ 2023, doi: https://doi.org/10.1016/j.irfa.2023.102876.

[6] P. G. de la Torre, M. Pérez-Verdugo, and X. E. Barandiaran, "Attention is all they need: cognitive science and the (techno)political economy of attention in humans and machines," *AI & SOCIETY,* 2025/06/28 2025, doi: https://doi.org/10.1007/s00146-025-02400-z.

[7] A. Hussain, S. Ali, U. E. Farwa, M. A. I. Mozumder, and H. C. Kim, "Foundation Models: From Current Developments, Challenges, and Risks to Future Opportunities," *2025 27th International Conference on Advanced Communications Technology (ICACT),* pp. 51–58, 16–19 Feb. 2025 2025, doi: https://doi.org/10.23919/ICACT63878.2025.10936649.

[8] A. Thomas, "Digitally transforming the organization through knowledge management: A socio-technical system (STS) perspective," *European Journal of Innovation Management,* vol. 27, no. 9, pp. 437–460, 2024, doi: https://doi.org/10.1108/EJIM-02-2024-0114.

[9] C. M. van Leersum and C. Maathuis, "Human centred explainable AI decision-making in healthcare," *Journal of Responsible Technology,* vol. 21, p. 100108, 2025/03/01/ 2025, doi: https://doi.org/10.1016/j.jrt.2025.100108.

[10] S. M. Hamdy and Z. Zaazou, "“Universities Devoid of Teaching Staff” Can AI Replace the Role of Professors in the Near Future?," *Arab Journal of Management,* vol. 45, no. 2, pp. 451–464, 2025, doi: https://dx.doi.org/10.21608/aja.2024.308915.1687.

[11] T. Kohn, "From Imitation Games to Robot-Teachers: A Review and Discussion of the Role of LLMs in Computing Education," *Journal of Computer Assisted Learning,* vol. 41, no. 3, p. e70043, 2025, doi: https://doi.org/10.1111/jcal.70043.

[12] D. Catlin, M. Blamires, and J.-J. Cabibihan, "Educational Robots: Past, Present, Future," *Social Robots in Education: How to Effectively Introduce Social Robots into Classrooms,* pp. 399–426, 2025, doi: https://doi.org/10.1007/978-3-031-82915-4_16.

[13] O. Kudina and I. van de Poel, "A sociotechnical system perspective on AI," *Minds and Machines,* vol. 34, no. 3, p. 21, 2024/06/12 2024, doi: https://doi.org/10.1007/s11023-024-09680-2.

[14] S. K. Parker *et al.*, "Quality work in the future: New directions via a co-evolving sociotechnical systems perspective," *Australian Journal of Management,* vol. 0, no. 0, p. 03128962251331813, 2025, doi: https://doi.org/10.1177/03128962251331813.

[15] R. Verdecchia, J. Sallou, and L. Cruz, "A systematic review of Green AI," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 13, no. 4, p. e1507, 2023, doi: https://doi.org/10.1002/widm.1507.

[16] V. Bolón-Canedo, L. Morán-Fernández, B. Cancela, and A. Alonso-Betanzos, "A review of green artificial intelligence: Towards a more sustainable future," *Neurocomputing,* vol. 599, p. 128096, 2024/09/28/ 2024, doi: https://doi.org/10.1016/j.neucom.2024.128096.

[17] A. Blok and C. B. Jensen, "What Next for Actor Network Theory? Inventing around Latour on a planet in distress," *Dialogues in Sociology,* vol. 1, no. 1, pp. 63–74, 2025, doi: https://doi.org/10.1177/29768667241285397.

[18] Y. I. Alzoubi and A. Mishra, "Green artificial intelligence initiatives: Potentials and challenges," *Journal of Cleaner Production,* vol. 468, p. 143090, 2024/08/25/ 2024, doi: https://doi.org/10.1016/j.jclepro.2024.143090.

[19] S. Lambiase *et al.*, "A Mixed-Method Empirical Investigation into the Influence of Software Communities Cultural and Geographical Dispersion Over Productivity," *Available at SSRN 4397227,* 2023, doi: https://doi.org/10.1016/j.jss.2023.111878.

[20] M. Baran, "Mixed methods research design," *Research Anthology on Innovative Research Methodologies and Utilization Across Multiple Disciplines,* vol. 1, pp. 312–333, 2022, doi: https://doi.org/10.4018/978-1-6684-3881-7.

[21] L. N. Kawar, G. B. Dunbar, E. M. Aquino-Maneja, S. L. Flores, V. R. Squier, and K. R. Failla, "Quantitative, qualitative, mixed methods, and triangulation research simplified," *The Journal of Continuing Education in Nursing,* vol. 55, no. 7, pp. 338–344, 2024, doi: https://doi.org/10.3928/00220124-20240328-03.

[22] R. Hunter, T. Gorely, M. Beattie, and K. Harris, "Realist review," *International Review of Sport and Exercise Psychology,* vol. 15, no. 1, pp. 242–265, 2022/12/31 2022, doi: https://doi.org/10.1080/1750984X.2021.1969674.

[23] S. Issar, "The Social Construction of Algorithms in Everyday Life: Examining TikTok Users' Understanding of the Platform's Algorithm," *International Journal of Human–Computer Interaction,* vol. 40, no. 18, pp. 5384–5398, 2024/09/16 2024, doi: https://doi.org/10.1080/10447318.2023.2233138.

[24] D. Alemayehu Tegegn, "The role of science and technology in reconstructing human social history: effect of technology change on society," *Cogent Social Sciences,* vol. 10, no. 1, p. 2356916, 2024/12/31 2024, doi: https://doi.org/10.1080/23311886.2024.2356916.

[25] J. Zheng *et al.*, "Empowering radical innovation: how digital technologies drive knowledge transfer and co-creation in innovation ecosystems," *R&D Management,* 2025, doi: https://doi.org/10.1111/radm.12764.

[26] J. Peters *et al.*, "Emergent language: a survey and taxonomy," *Autonomous Agents and Multi-Agent Systems,* vol. 39, no. 1, p. 18, 2025/03/07 2025, doi: https://doi.org/10.1007/s10458-025-09691-y.

[27] M. Liu, F. Wu, B. Li, Z. Lu, Y. Yu, and X. Li, "Envisioning class entity reasoning by large language models for few-shot learning," *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 39, no. 18, pp. 18906–18914, 2025, doi: https://doi.org/10.1609/aaai.v39i18.34081.

[28] O. Krüger, "The Mechanical Turk: a human-centered approach to the transcendence narrative of artificial intelligence," *Religion,* vol. 55, no. 3, pp. 699–714, 2025/07/03 2025, doi: https://doi.org/10.1080/0048721X.2025.2502291.

[29] S. Feng, R. Zhang, and G. Li, "Environmental decentralization, digital finance and green technology innovation," *Structural Change and Economic Dynamics,* vol. 61, pp. 70–83, 2022/06/01/ 2022, doi: https://doi.org/10.1016/j.strueco.2022.02.008.

[30] E. Brynjolfsson, D. Rock, and C. Syverson, "The productivity J-curve: How intangibles complement general purpose technologies," *American Economic Journal: Macroeconomics,* vol. 13, no. 1, pp. 333–372, 2021, doi: https://doi.org/10.1257/mac.20180386.

[31] M. K. Scheuerman, A. Hanna, and R. Denton, "Do datasets have politics? Disciplinary values in computer vision dataset development," *Proceedings of the ACM on Human-Computer Interaction,* vol. 5, no. CSCW2, pp. 1–37, 2021, doi: https://doi.org/10.1145/3476058.

[32] M. Matthews, S. Matthews, and T. Kelemen, "The Alignment Problem: Machine Learning And Human Values," *Personnel Psychology,* vol. 75, no. 1, 2022, doi: https://doi.org/10.1111/peps.12500.

[33] J. Bandola-Gill, M. Arthur, and R. I. Leng, "What is co-production? Conceptualising and understanding co-production of knowledge and policy across different theoretical perspectives," *Evidence & Policy,* vol. 19, no. 2, pp. 275–298, 2023, doi: https://doi.org/10.1332/174426421X16420955772641.

[34] Y. Zhao, "Book Review: The Atlas of AI: Power, politics, and the planetary costs of artificial intelligence," *Global Media and China,* vol. 0, no. 0, p. 20594364251325469, 2025, doi: https://doi.org/10.1177/20594364251325469.

[35] S. Chikkudu and S. Annamalai, "Fundamentals of AI and NLP in Environmental Analysis," *Environmental Monitoring Using Artificial Intelligence,* pp. 29–44, 2025, doi: https://doi.org/10.1002/9781394270392.ch2.

[36] R. Różycki, D. A. Solarska, and G. Waligóra, "Energy-Aware Machine Learning Models—A Review of Recent Techniques and Perspectives," *Energies,* vol. 18, no. 11, p. 2810, 2025, doi: https://doi.org/10.3390/en18112810.

[37] R. Pasupuleti, R. Vadapalli, C. Mader, and N. Timothy, "Popular LLM-Large Language Models in Enterprise Applications," *2024 2nd International Conference on Foundation and Large Language Models (FLLM),* pp. 125–131, 26–29 Nov. 2024 2024, doi: https://doi.org/10.1109/FLLM63129.2024.10852443.

[38] M. Azam, Y. Chen, M. O. Arowolo, H. Liu, M. Popescu, and D. Xu, "A comprehensive evaluation of large language models in mining gene relations and pathway knowledge," *Quantitative Biology,* vol. 12, no. 4, pp. 360–374, 2024, doi: https://doi.org/10.1002/qub2.57.

[39] S. Chen, Z. Pan, J. Cai, P. Fang, M. Harandi, and D. Phung, "Hierarchical Prompt-Enhanced Image Generation Using Hyperbolic Space," *Neural Information Processing*

pp. 121–136, 2025, doi: https://doi.org/10.1007/978-981-96-6688-1_9.
[40] E. K. Hong *et al.*, "Diagnostic accuracy and clinical value of a domain-specific multimodal generative AI model for chest radiograph report generation," *Radiology,* vol. 314, no. 3, p. e241476, 2025, doi: https://doi.org/10.1148/radiol.241476.
[41] F. S. Orim *et al.*, "Implementation of automated classroom assessment in higher education using the technology acceptance model," *Discover Education,* vol. 4, no. 1, p. 87, 2025/04/17 2025, doi: https://doi.org/10.1007/s44217-025-00481-y.
[42] A. Hertzmann, "Generative Models for the Psychology of Art and Aesthetics," *Empirical Studies of the Arts,* vol. 43, no. 1, pp. 23–43, 2025, doi: https://doi.org/10.1177/02762374241288696.
[43] Y. Knight and M. P. Eladhari, "Artificial intelligence in an artistic practice: a journey through surrealism and generative arts," *Media Practice and Education,* pp. 1–18, 2025, doi: https://doi.org/10.1080/25741136.2024.2443865.
[44] W. Jon, "Prompting Creativity: Tiered Approach to Copyright Protection for AI-Generated Content in the Digital Age," *Media and Communication,* vol. 13, 2025, doi: https://doi.org/10.17645/mac.9420.
[45] K. Soares Seto, "AI From the South: artificial intelligence in Latin America through the sociotechnical imaginaries of Brazilian tech workers," *Globalizations,* pp. 1–16, 2025, doi: https://doi.org/10.1080/14747731.2025.2465166.
[46] U. Baresi, "Neo-colonial intelligence: How AI risks reinforcing spatial injustices in a digitally divided world," *Cities,* vol. 166, p. 106232, 2025/11/01/ 2025, doi: https://doi.org/10.1016/j.cities.2025.106232.
[47] M. Bartl, A. Mandal, S. Leavy, and S. Little, "Gender Bias in Natural Language Processing and Computer Vision: A Comparative Survey," *ACM Comput. Surv.,* vol. 57, no. 6, p. Article 139, 2025, doi: https://doi.org/10.1145/3700438.
[48] D. Thakur, A. Guzzo, G. Fortino, and F. Piccialli, "Green Federated Learning: A New Era of Green Aware AI," *ACM Comput. Surv.,* vol. 57, no. 8, p. Article 194, 2025, doi: https://doi.org/10.1145/3718363.
[49] S. Dash, "Green AI: Enhancing Sustainability and Energy Efficiency in AI-Integrated Enterprise Systems," *IEEE Access,* vol. 13, pp. 21216–21228, 2025, doi: https://doi.org/10.1109/ACCESS.2025.3532838.
[50] R. Boccio, "Race After Technology: Abolitionist Tools for the New Jim Code by Ruha Benjamin," *Configurations,* vol. 30, no. 2, pp. 236–238, 2022, doi: https://doi.org/10.1353/con.2022.0013.
[51] S. Lin, M. Wang, C. Jing, S. Zhang, J. Chen, and R. Liu, "The influence of AI on the economic growth of different regions in China," *Scientific Reports,* vol. 14, no. 1, p. 9169, 2024/04/22 2024, doi: https://doi.org/10.1038/s41598-024-59968-7.
[52] M. L. Navarro-Ligero and J. A. Soria-Lara, "Reassembling Collaborative Planning for Socio-Technical Transitions: An Actor-Network Theory Approach," *Planning Theory & Practice,* vol. 26, no. 1, pp. 66–84, 2025/01/01 2025, doi: https://doi.org/10.1080/14649357.2025.2472618.
[53] M. Hodak, D. Ellison, and A. Dholakia, "Everyone is a Winner: Interpreting MLPerf Inference Benchmark Results," *Performance Evaluation and Benchmarking,* pp. 50–61, 2022, doi: https://doi.org/10.1007/978-3-030-94437-7_4.
[54] S. M. Hasan, T. Islam, M. Saifuzzaman, K. R. Ahmed, C. H. Huang, and A. R. Shahid, "Carbon Emission Quantification of Machine Learning: A Review," *IEEE Transactions on Sustainable Computing,* pp. 1–19, 2025, doi: https://doi.org/10.1109/TSUSC.2025.3578834.
[55] C. Makri and A. Neely, "Grounded Theory: A Guide for Exploratory Studies in Management Research," *International Journal of Qualitative Methods,* vol. 20, p. 16094069211013654, 2021, doi: https://doi.org/10.1177/16094069211013654.
[56] C. Turner and F. Astin, "Grounded theory: what makes a grounded theory study?," *European Journal of Cardiovascular Nursing,* vol. 20, no. 3, pp. 285–289, 2021, doi: https://doi.org/10.1093/eurjcn/zvaa034.
[57] D. Mortelmans, "Thematic Coding," *Doing Qualitative Data Analysis with NVivo,* pp. 57–87, 2025, doi: https://doi.org/10.1007/978-3-031-66014-6_8.
[58] V. Braun and V. Clarke, "Conceptual and design thinking for thematic analysis," *Qualitative psychology,* vol. 9, no. 1, p. 3, 2022, doi: https://psycnet.apa.org/doi/10.1037/qup0000196.
[59] A. Tschand *et al.*, "MLPerf Power: Benchmarking the Energy Efficiency of Machine Learning Systems from µWatts to MWatts for Sustainable AI," *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA),* pp. 1201–1216, 1–5 March 2025 2025, doi: https://doi.org/10.1109/HPCA61900.2025.00092.
[60] G. Ayana *et al.*, "Decolonizing global AI governance: assessment of the state of decolonized AI governance in Sub-Saharan Africa," *Royal Society Open Science,* vol. 11, no. 8, p. 231994, 2024, doi: https://doi.org/10.1098/rsos.231994.