ICPRAM 2026

15th International Conference on Pattern Recognition Applications and Methods

PROCEEDINGS

Marbella, Spain

2 - 4 March, 2026

EDITORS

Modesto Castrillon-Santana

Daniel Riccio

Ana Fred

Maria De Marsico

https://icpram.scitevents.org

SPONSORED BY

INSTICC

PAPERS AVAILABLE AT

SCITEPRESS DIGITAL LIBRARY

ICPRAM 2026

Proceedings of the
15th International Conference on
Pattern Recognition Applications and Methods

Marbella - Spain

March 2 - 4, 2026

Sponsored by
INSTICC - Institute for Systems and Technologies of Information, Control and Communication

IEEE Technically co-sponsored by
CIS - IEEE Computational Intelligence Society

ACM In Cooperation
SIGAI - ACM Special Interest Group on Artificial Intelligence

Endorsed by
IAPR - International Association for Pattern Recognition

In Cooperation with
AAAI - Association for the Advancement of Artificial Intelligence
AEPIA - Spanish Association of Artificial Intelligence
INNS - International Neural Network Society
European Association for Signal Processing
EAB - European Association for Biometrics

Copyright © 2026 by SCITEPRESS – Science and Technology Publications, Lda.

Edited by Modesto Castrillon-Santana, Daniel Riccio, Ana Fred and Maria De Marsico

Printed in Portugal
ISSN: 2184-4313
ISBN: 978-989-758-797-9
DOI: 10.5220/0000217000004067
Depósito Legal: 557933/25

https://icpram.scitevents.org
icpram.secretariat@insticc.org

The Hidden Price Tag behind LLMs and Their Environmental Cost

Nermeen Abou Baker[a] and Uwe Handmann[b]
Computer Science Institute, Ruhr West University of Applied Science, Luetzowstr. 5, Bottrop, Germany

Keywords: Large Language Models, Sustainable AI, Energy Consumption, Carbon Emissions, Water Usage, Environmental Sustainability.

Abstract: Artificial intelligence systems consume substantial environmental resources, which are often hidden due to inconsistent measurement methodologies and limited transparency. This study analyses the energy consumption, carbon emissions, and water usage of thirteen major language models to estimate the environmental costs of AI development and deployment. The analysis indicates that training energy consumption varies by 41,000× across models, which far exceeds parameter scaling ratios. Geographic training location creates an additional 6× difference in emissions for similar computational requirements. However, critical transparency gaps exist: seven of the thirteen major models lack verified environmental impact data, which prevents accurate cross-model comparisons. True environmental costs are underestimated by traditional assessments, as shown by a comprehensive measurement methodology. Moreover, infrastructure imposes substantial additional costs. Analysis of water consumption demonstrates a significant shift in impact from training to inference, with significant geographic variations in efficiency, and inference consumption is now exceeding training requirements. These findings suggest that strategic choices about measurement methodology, training location, and deployment can produce significant environmental improvements. This study recommends the standardisation of environmental reporting to enable the accurate assessment and systematic optimisation of AI systems. Accurate assessments are essential for the sustainable development of AI as the technology scales globally.

1 INTRODUCTION

1.1 The Rise of Large Language Models

The recent success of deep learning has been driven mainly by large language models (LLMs), which have grown dramatically in size, increasing from millions to around a trillion parameters. Yet scaling LLMs presents significant environmental challenges. Training and inference consume substantial amounts of electricity, which contributes to CO_2 emissions. The true environmental cost, nevertheless, remains poorly understood due to inconsistent measurement methodologies and limited corporate transparency.

1.2 The Transparency and Measurement Problem

The fundamental challenge is that detailed information about energy consumption, emissions, and resource usage is often unavailable due to proprietary restrictions, inconsistent reporting practices between different model versions, or the absence of systematic measurement. Google's comprehensive production analysis demonstrates the importance of choosing the right measurement approaches. Different methodological approaches yield significantly divergent results across studies (Elsworth et al., 2025).

The environmental impact of LLMs spans their entire development and deployment lifecycle. As Figure 1 shows, focusing only on training energy significantly underestimates the total environmental footprint.

Several critical metrics present significant measurement challenges:

- Number of parameters: While some parameter counts are widely known, others vary across versions or remain hidden.
- Training energy (kWh): The total electricity consumed when training a model, including GPU/TPU computation, CPU processing, memory usage, and cooling systems. It depends on the efficiency of the hardware, the duration of the training, and the data centre infrastructure, and it represents the raw energy demand, regardless of the source.
- CO_2 emissions, measured in tCO_2e (tonnes

[a] https://orcid.org/0000-0002-9683-5920
[b] https://orcid.org/0000-0003-1230-9446

Abou Baker, N. and Handmann, U.
The Hidden Price Tag behind LLMs and Their Environmental Cost.
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 15th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2026)*, pages 250-261
ISBN: 978-989-758-797-9; ISSN: 2184-4313
Proceedings Copyright © 2026 by SCITEPRESS – Science and Technology Publications, Lda.

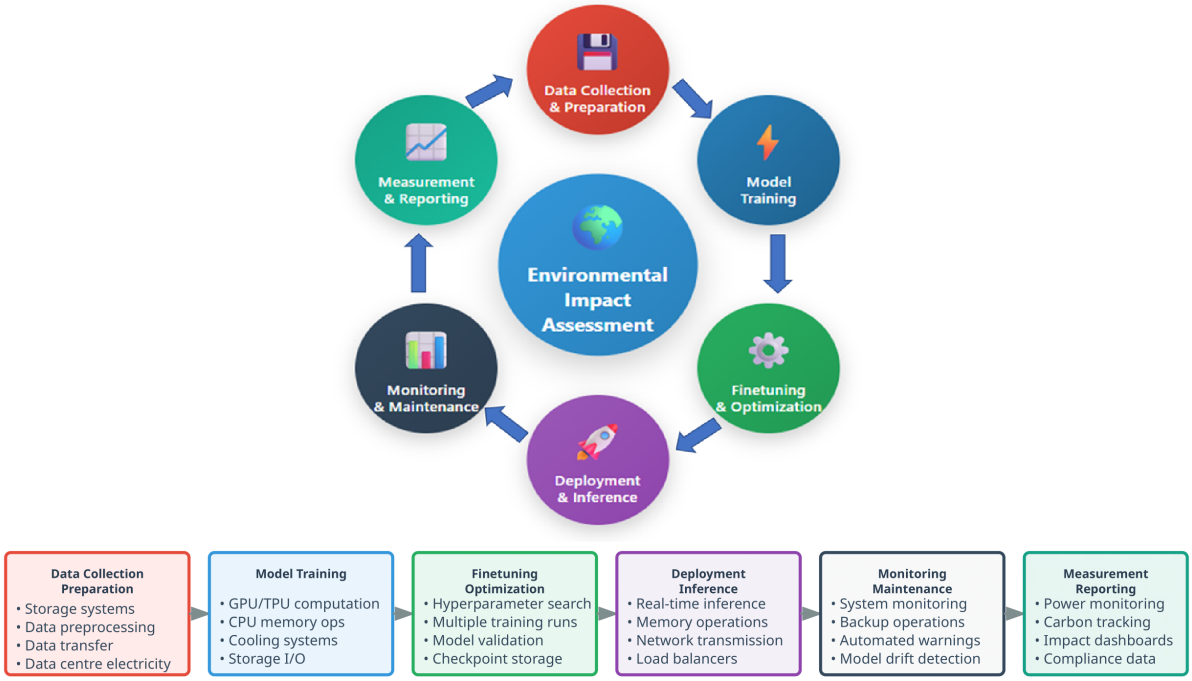

Figure 1: ML lifecycle stages and environmental impact.

of carbon-dioxide equivalent): Carbon-dioxide-equivalent emissions resulting from electricity used during training. Emission levels vary by location, even with identical energy consumption, because electricity sources differ widely (coal, natural gas, nuclear, renewables).

- Water consumption: Data centre cooling requirements create substantial freshwater demands.

1.3 Problem Statement

These transparency gaps create three major problems:

1. Incomparable environmental assessments, which prevent meaningful comparisons of models or tracking of progress.
2. Hidden environmental costs that remain invisible to users, researchers, and policymakers.
3. A lack of responsibility mechanisms for the environmental consequences of organisations.

1.4 Research Questions

This paper addresses the following research questions:
RQ1: What is the current state of environmental impact reporting for major LLMs, and where do significant transparency gaps exist?
RQ2: How do methodological differences in environmental assessment (e.g. boundaries, metrics, and treatment of indirect impacts) affect comparability and accuracy of reported impacts?
RQ3: What technical and operational factors cause actual environmental differences between LLM systems? These factors include hardware efficiency, architecture, training duration, and geographic location.
RQ4: What standardised methodologies and reporting frameworks are needed to enable transparent and comparable environmental assessment of AI systems?

1.5 Research Contributions

This paper provides a detailed analysis of the energy consumption, carbon emissions, and water usage of major large language models, namely: BERT (Base) (Devlin et al., 2019), BLOOM (Scao et al., 2023), LLaMA 2 (Base, 13B, 65B, 70B) (Touvron et al., 2023), OPT (Zhang et al., 2022), PaLM Large (Chowdhery et al., 2022), Gopher (Rae et al., 2022), DeepSeek-V3 (DeepSeek-AI et al., 2025), GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), and Mistral Large 2 (Mistral AI, 2024). It illustrates the actual environmental impact of both developing and deploying these models. It systematically identifies the methodological differences that cause significant variations in reported impacts, and emphasises the urgent need for consistent measurement practices. Based on this analysis, the study proposes a standardised framework to enable transparency and meaningful cross-model comparisons. Finally, the paper pro-

vides actionable recommendations to guide research, industry, and policy towards more sustainable pathways for the future of AI.

2 BACKGROUND AND RELATED WORK

2.1 Environmental Assessment Studies and Methodologies

Carbon accounting and reporting in AI is becoming increasingly recognised, although it is not yet standard practice. Conferences such as NeurIPS encourage submissions to include emissions estimates (NeurIPS, 2021), and recent studies have documented the substantial environmental costs involved. Training large language models such as T5 can emit hundreds of tons of CO_2, with scaled experiments for models like T0 producing around 8.9 tCO_2e (Sanh et al., 2022). The carbon footprint depends on factors such as model size, training duration, hardware type, server placement, and the carbon intensity of the electricity grid, which varies dramatically by region based on the energy mix used for electricity generation (Islam et al., 2023).

The studies listed in Table 1 address various aspects of these measurement challenges, yet each has important limitations. Strubell et al.'s foundational work established baseline energy estimates for transformer models, focusing on a limited number of architectures (Strubell et al., 2019). Li et al. advanced the understanding of water consumption specifically for GPT-3, demonstrating that inference costs can exceed training costs (Li et al., 2025). Elsworth et al.'s production-level analysis of Google Gemini represents the most comprehensive methodology to date. It considers infrastructure overhead and idle capacity that were overlooked in previous studies (Elsworth et al., 2025). However, no study has yet compared the environmental impact of multiple models from different organisations using consistent evaluation criteria. This is an important gap to fill, as it prevents identifying best practices, assessing industry-wide progress, and establishing accountability mechanisms. This work addresses this gap by analysing 13 major LLMs across organisations using standardised metrics where data reliability permits.

A growing number of tools have been developed to measure, monitor, and reduce emissions from machine learning (ML) activities. Real-time tracking tools such as Experiment-Impact-Tracker (Henderson et al., 2020), NeuralPower (Cai et al., 2017), WattsOn (Mittal et al., 2012), and Joulemeter (Kansal et al., 2009) monitor power usage and identify energy-intensive operations. Post-hoc estimation tools such as Carbontracker (Anthony et al., 2020), ML CO_2 Impact (Lacoste et al., 2019), eco2AI (Budennyy et al., 2022), and CodeCarbon (Courty et al., 2024) provide carbon and energy usage estimates based on training metadata. These tools enable practitioners to generate reports, visualise consumption, and optimise workloads, encouraging transparency and responsibility throughout the model lifecycle.

Although transparent reporting practices are lacking in LLM development, they exist in some computer vision domains. DINOv3 demonstrates transparent reporting by clearly documenting its methodology (PUE of 1.1 and carbon intensity of 0.385 kg CO_2e/kWh), energy consumption of 47 MWh, and emissions of 18 tCO_2e (Siméoni et al., 2025). Transfer learning approaches in computer vision have also been shown to facilitate the development of efficient models (Abou Baker et al., 2022). This level of methodological transparency enables meaningful comparisons across studies.

2.2 Mitigation Strategies and Sustainability Challenges

Mitigation strategies focus on technical and operational improvements. Model distillation, low-precision computation, and hardware-aware optimisation can all help to reduce energy consumption (Eilam et al., 2023). Modern hardware, such as NVIDIA H100 GPUs, delivers better performance per watt than older generations. Data centre efficiency has a strong impact on environmental footprint: efficient facilities achieve a PUE of 1.12 (12% overhead), whereas inefficient facilities can exceed 2.0, doubling energy use. Operators can reduce emissions by scheduling workloads during periods of high renewable energy generation and selecting regions with lower-carbon electricity (Mehta, 2024).

The rapid growth of AI workloads emphasizes the urgency of addressing these issues. Global AI energy consumption is projected to reach 85–134 TWh per year by 2027 (de Vries, 2023), and water use for cooling data centres presents significant environmental concerns. Sustainable AI requires a comprehensive approach that includes accurate measurement tools, improvements in hardware and algorithm efficiency, and adoption of low-carbon energy sources.

Table 1 summarises previous studies on the environmental impact of LLMs and shows how this work differs in scope, metrics, and focus on transparency.

Table 1: Comparative overview of prior studies on the environmental impact of LLMs vs this work.

Study	Models Analysed	Metrics	Data Source	Transparency Focus
(Strubell et al., 2019)	Limited models	Energy, CO_2	Estimates	Limited
(Li et al., 2025)	GPT-3 focus	Water consumption	Third-party analysis	Moderate
(Elsworth et al., 2025)	Google Gemini	Energy, CO_2, Water	Production data	High
This study	13 models	Energy, CO_2, Water	Mixed sources	Primary focus

3 LLM DEVELOPMENT LIFECYCLE AND ENVIRONMENTAL IMPACT

The environmental footprint of AI systems extends far beyond the computational demands of model training, which is a common misconception that systematically underestimates the true environmental impact. Figure 1 shows the whole ML development lifecycle. It demonstrates how environmental costs accumulate across six interdependent stages. Each stage has distinct resource demands that compound throughout a model's operational lifetime.

Traditional environmental assessments have a key methodological weakness: they focus on the energy consumption of training, while ignoring the wider infrastructure needed to support production AI systems. This narrow definition of the boundary fails to capture several substantial cost categories. Pre-training phases, such as data acquisition, storage, and preprocessing, require dedicated infrastructure that is independent of model computation. Post-training activities, such as continuous inference serving, model monitoring, and periodic retraining, often exceed one-time training costs when spread over deployment lifespans. Furthermore, the measurement and reporting infrastructure itself represents considerable overhead, though it is essential for enabling optimisation.

A comprehensive lifecycle analysis can fundamentally change estimates of environmental impact. Google's production-level methodology shows that including idle capacity, supporting infrastructure, and operational overheads increases narrow, training-focused assessments by 2.2–4.5× (Elsworth et al., 2025). This divergence suggests that optimisation strategies targeting individual lifecycle stages, while potentially valuable, cannot achieve systemic sustainability without coordinated improvements throughout the entire development and deployment pipeline. The implication is clear: to accurately govern the environmental impact of AI systems, measurement frameworks must capture whole-lifecycle resource consumption, rather than prioritising computationally intensive stages at the expense of ongoing operational demands.

3.1 Data Reliability Classification

Given the transparency challenges identified in AI environmental reporting. This work establishes a three-tier reliability classification for the environmental data analysed in this study:

- *Tier 1 (official verified data):* Direct reports from model developers published in peer-reviewed papers or official technical reports. Examples include BLOOM documentation from BigScience (Scao et al., 2023) and Google's Gemini production measurements (Elsworth et al., 2025).
- *Tier 2 (third-Party verified estimates):* Independent research using disclosed parameters with documented methodologies. Examples include BERT energy estimates from Strubell et al. (Strubell et al., 2019).
- *Tier 3 (industry estimates):* Industry analyses and extrapolations with limited methodology disclosure. Examples include various GPT-4 energy estimates from industry reports.

This classification enables transparent discussion of data limitations while providing meaningful analysis of available information. The reliability tier is indicated where relevant to help in interpreting the confidence level of different data points.

3.2 Analysis Methodology

This study uses a systematic three-stage process to synthesise environmental data from official technical reports, peer-reviewed papers and verified third-party analyses for 13 major LLMs.

1. **Data Collection:** We gathered environmental metrics from diverse sources spanning official disclosures, academic publications, and industry analyses.
2. **Reliability Classification:** Each data point was assigned to one of three tiers as defined in Section 3.1: Tier 1 (official, verified data) includes BLOOM and LLaMA 2 variants; Tier 2 (third-party, verified estimates) includes BERT and OPT; Tier 3 (industry estimates) includes GPT-4 and Mistral Large 2. This classification directly

affects the reliability of cross-model comparisons, with Tier 1 providing the greatest confidence for identifying optimisation strategies.

3. **Comparative Analysis:** We normalised metrics across models where data reliability permitted, focusing on energy consumption (MWh), emissions (tCO_2e), and water usage (kL) to reveal significant differences in environmental efficiency.
4. **Data Limitations:** Seven of the 13 models lack verified environmental impact data, and water consumption data is only available for 2 models. These transparency gaps prevent comprehensive cross-model comparisons in some dimensions.

Figures 2 and 3 show the training energy consumption and CO_2 emissions for major LLMs. The bubbles show the parameter counts, and the emissions are presented as common activity equivalents to emphasise the importance of training methodology and infrastructure choices. Section 4 investigates the environmental impact of thirteen major language models, aiming to identify the main factors causing variations in energy consumption, carbon emissions, and water usage.

4 DISCUSSION OF AI EFFICIENCY AND SUSTAINABILITY

4.1 Energy Differences between Models

Analysis of the training energy consumption and CO_2 emissions across major language models, illustrated in Figure 2, shows that the relationship between the number of parameters in a model and its environmental impact is not straightforward. Training energy consumption ranges from 0.001507 MWh for BERT (Strubell et al., 2019) to 62,300 MWh for GPT-4 (Ludvigson, 2023), which represents a difference of over **41,000×**, far exceeding parameter scaling ratios. Similarly, Figure 3 shows that CO_2 emissions range from 0.65 tons for BERT to 20,400 tons for Mistral Large 2 (Digit, 2024), which is a variation of over 31,000× that cannot be explained by computational requirements alone.

Recent comparative analyses show that emissions from AI queries are increasing. Table 2 summarises the query-level performance comparison (Claude AI, 2025). It shows that there is a clear efficiency gradient. While Grok is the most efficient system, it is approximately 25× more efficient than the least efficient GPT-4 implementation. Additionally, GPT-4o combines very low CO_2 emissions per query with detailed reporting of energy and water consumption, reflecting environmental and operational efficiency.

Table 2: Query-level performance comparison.

Model	**CO_2 (g)**	**Notes**
Grok	0.17	Most efficient
Gemini	1.6	High efficient
Claude	3.2	Moderate
GPT-4	4.32	Lower efficient
GPT-4o	0.03	Best: 0.24 Wh, 0.26 mL

4.2 How Companies Train Models Differently

Organisational training practices have a significant impact on environmental efficiency. Models developed by the same organisation tend to demonstrate consistent scaling patterns, which suggests that they share infrastructure and methodological approaches. For example, the LLaMA series demonstrates predictable energy scaling with parameter increases, with LLaMA 2 13B consuming approximately double the energy of LLaMA 2 Base, which is consistent with the increase in parameters (Touvron et al., 2023). Cross-organisational comparisons, by contrast, reveal substantial variations. BLOOM achieved a comparable number of parameters to GPT-3 while using 66% less training energy, which indicates that there are substantial differences in training infrastructure, algorithmic choices, and measurement methodologies between organisations.

4.3 Geographic Impact on Environmental Footprint

Although GPT-3 and BLOOM have nearly identical sizes, the geographic training location appears to dominate environmental impact. As shown in Figures 2 and 3, BLOOM's energy-to-emissions ratio is 0.069 tCO_2/MWh, compared to GPT-3's 0.429 tCO_2/MWh, which is a **6×** difference in carbon intensity. According to (Claude AI, 2025), identical AI queries produce 77% higher emissions when processed in West Virginia's coal-heavy electricity grid than in California's renewable energy mix. Regional grid carbon intensity can create variations in the environmental impact of up to 10× for identical computational workloads. Strategic data centre placement thus enables substantial carbon footprint reduction through geographic optimisation.

A consistent pattern of carbon intensity was observed among the LLaMA 2 variants analysed, which

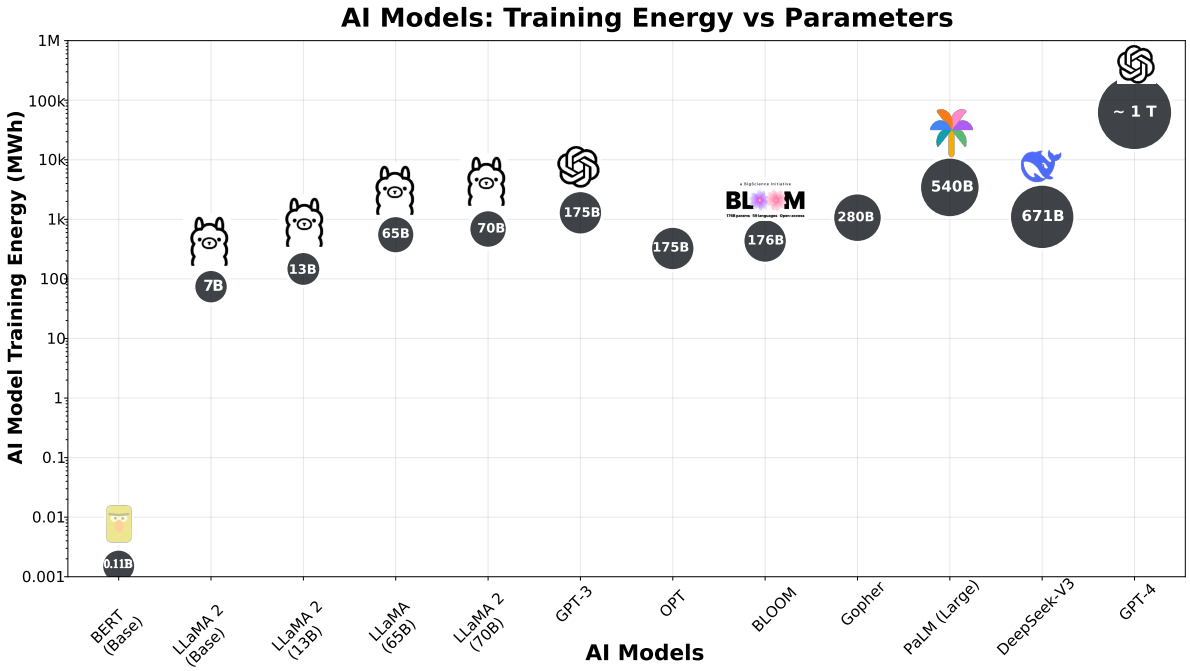

Figure 2: Training energy consumption for the major LLM models (with their parameters in billions).

were trained in similar geographic regions. In contrast, OPT's emissions vary depending on its training infrastructure, preventing a direct comparison. DeepSeek-V3, Gopher, and PaLM do not disclose their training locations either, making it impossible to assess the geographic contribution to their carbon intensity. This lack of transparency could hide potential efficiency variations of 5–10× based only on grid composition.

Beyond geographic carbon intensity differences, combining location decisions with optimisation strategies most effectively reduces environmental impact. Table 3 quantifies these combined effects and shows how the interaction between geographic and optimisation factors creates significant variations in efficiency. The impact factors are expressed as multiplication ratios, except for grid emissions (annual percentage reduction) and model selection (percentage emission reduction).

4.4 Infrastructure Overhead in Production Systems

Recent first-party data from Google's production infrastructure offers the most detailed measurement of AI inference energy consumption to date. Analysis of Gemini Apps shows that the median text prompt uses 0.24 Wh of energy and 0.26 ml of water when measured using a comprehensive methodology that includes all infrastructure components (Elsworth et al., 2025). This methodology includes active AI accelerator power (58% of the total energy used), CPU and DRAM requirements (25%), idle machine capacity (10%), and data centre overhead (8%).

Table 3: Geographic and optimisation impact.

Factor	Range	Examples
Geographic variations		
Regional carbon intensity	6×	BLOOM vs GPT-3
US regional grids	1.8×	West Virginia vs California (77% higher)
Overall geographic impact	5–10×	Location-dependent
Optimisation potential		
Production optimisation	33×	Google Gemini (12 months) (Elsworth et al., 2025)
Combined optimisation	44×	Google Gemini emissions
Hardware advancement	30×	Custom accelerators
Clean energy transition	30%/yr	Grid decarbonisation
Model selection	96%	Efficient vs inefficient models

Comprehensive measurement quantifies the substantial cost impact of infrastructure. Data centre infrastructure multiplies AI's environmental impact by 2.2–4.5×, with average facilities operating at 1.56 PUE. This means that 56% of additional energy goes towards cooling, power distribution, and facility operations. Advanced cooling systems for AI workloads consume an additional 40–100% of energy compared to processing alone. Embodied carbon from hardware manufacturing contributes an additional 20–30% of annual operational emissions (Claude AI, 2025). These differences in measurement approaches explain the substantial variations in

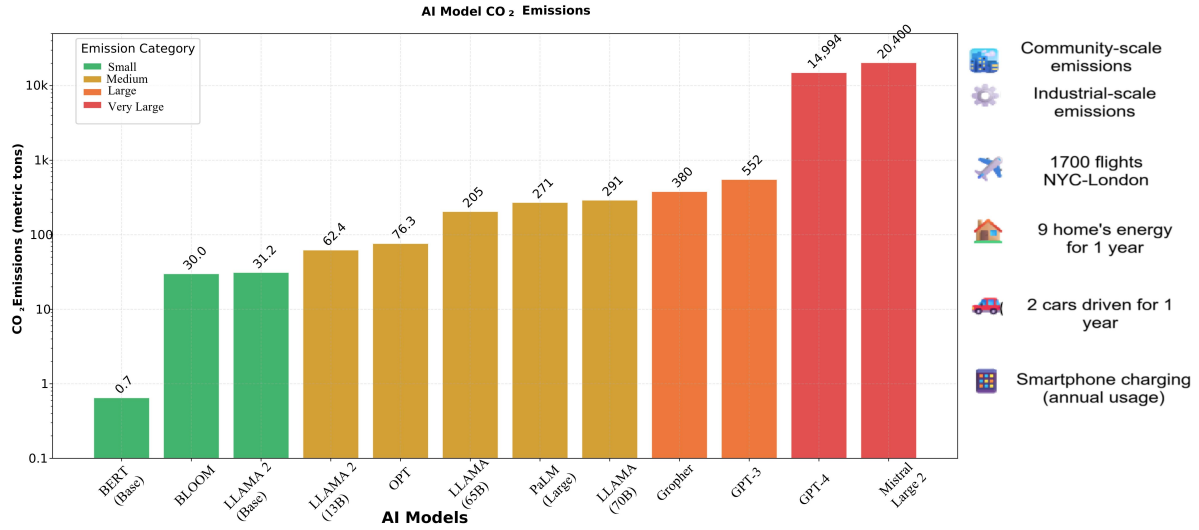

Figure 3: Training CO_2 emissions for major LLM models.

reported environmental impacts across studies and demonstrate why narrow measurement approaches systematically underestimate the true environmental costs.

4.5 Water Consumption Analysis

Water consumption represents a substantial yet frequently overlooked environmental impact that varies widely by geographic location and infrastructure efficiency. Infrastructure efficiency also differs between providers and regions, which results in large variations in water usage for the same models and workloads. Water consumption represents the most severe data gap in this analysis. Of 13 models, only GPT-3 and GPT-4 have associated water data, and these figures derive from third-party estimates rather than official disclosure. The remaining 11 models, including BLOOM, all LLaMA 2 variants, OPT, PaLM, Gopher, DeepSeek-V3, and Mistral Large 2, provide no water consumption information. This 85% data gap prevents comparative analysis of water efficiency across models or organisations. Without systematic water reporting, the AI industry cannot evaluate the impact of regional water shortages, optimise the efficiency of cooling systems, or make informed decisions about the location of data centres in regions where water is scarce. This lack of transparency is concerning, given the demonstrated shift from training-dominated to inference-dominated water consumption shown in available GPT data.

Table 4 illustrates these variations and indicates that inference currently consumes more water than training. As Table 4 indicates, the 'hidden' water footprint consists of freshwater that evaporates and is permanently lost from local ecosystems rather than being recycled. Differences in infrastructure efficiency mean that Microsoft Azure's Water Usage Effectiveness (WUE) is among the best. Regional variations, nonetheless, substantially affect the environmental impact of identical computational workloads.

Water scarcity is an important factor in achieving sustainable AI. Training and inference in areas experiencing water shortage have a greater local impact and must be considered alongside carbon emissions. Shifting from one-time training to continuous inference increases cumulative water usage, and focusing only on training underestimates the total AI water footprint.

Water consumption data is also unavailable for the following models beyond GPT: BLOOM, LLaMA 2 variants, OPT, PaLM, Gopher, DeepSeek-V3, and Mistral Large 2. This data gap prevents systematic comparisons of water efficiency, which results in the most severe lack of transparency in environmental reporting. Without comprehensive water disclosure, the AI industry cannot accurately evaluate the impact of regional water shortages or optimise the placement of data centres in areas where water is scarce.

4.6 Optimisation Evidence Is Possible

Efficient AI methods substantially reduce environmental impact across training and deployment of large language models (Abou Baker et al., 2024). Transferability assessment techniques can reduce the computational cost of model selection by enabling models to be ranked without exhaustive finetuning (Abou Baker and Handmann, 2024). Table 5 provides an overview of recent LLMs, and the efficiency and

Table 4: Water consumption analysis (2 of 13 models).

Metric	Value	Note
Consumption Volumes		
GPT-3 training	0.7 ML	1.4 Olympic pools (Li et al., 2025)
GPT-4o inference (annual)	1.4 ML	2.8 pools (Li et al., 2025)
Infrastructure Efficiency (WUE)		
Microsoft Azure	1.15 L/kWh	Industry leading
China avg data centers	6.0 L/kWh	5.2× less efficient (Jegham et al., 2025)

finetuning techniques they employ.

Designs that are parameter-efficient, such as ALBERT's cross-layer sharing (Lan et al., 2020) or DeepSeek-V3's sparse Mixture-of-Experts (MoE), minimise the number of active parameters per forward pass. This reduces computation and energy use while maintaining model quality. Knowledge distillation, employed in Gemini Flash and DeepSeek-R1, shrinks large models into smaller variants that deliver comparable performance at a lower cost to train and infer. Quantisation of LLaMA 2 inference variants further reduces memory and power requirements on GPUs and edge devices. Efficient attention mechanisms, including Multi-Head Latent Attention and Grouped-Query Attention (GQA), enable models to handle long contexts without significantly increasing computational demands. Finally, finetuning approaches such as Low-Rank Adaptation (LoRA) and Reinforcement Learning with Human Feedback (RLHF) provide scalable methods of aligning large base models with downstream tasks while keeping resource usage under control.

Comprehensive optimisation across the AI serving stack resulted in a reduction of 33× in per-prompt energy consumption and 44× in per-prompt emissions (Elsworth et al., 2025). These improvements were achieved through coordinated enhancements to model architectures, MoE designs, and advanced inference techniques, including speculative decoding. Other contributing factors were custom hardware offering 30× better energy efficiency, optimised data centre operations with a PUE of 1.09 and clean energy procurement strategies that reduced grid emissions intensity by 30% year on year.

Analyses of efficiency variations identify critical thresholds for sustainable AI deployment. Model selection has the greatest impact: efficient models can reduce emissions by up to 96% compared to inefficient alternatives (Claude AI, 2025). An environmental threshold is reached when AI query emissions, including infrastructure costs, exceed 2–3 g CO_2 e. Beyond this point, even traditional approaches may become competitive (Claude AI, 2025). Hardware upgrades through newer GPUs can improve efficiency by 2–3×, while variations in infrastructure efficiency, as measured by PUE, can change environmental impact by 15–150% (Claude AI, 2025). Overall, decisions regarding deployment, such as model selection, hardware configuration and system location, can have a significant impact on environmental performance, even when functionality remains constant.

These efficiency improvements focus on operational optimisation. However, the environmental costs of the hardware lifecycle, including manufacturing emissions, resource extraction and disposal, are not captured in operational metrics. Hardware manufacturing accounts for 20–30% of annual operational emissions in the form of embodied carbon. Therefore, achieving true sustainability requires balancing the need for hardware upgrades with the option of maximising equipment utilisation. While frequent hardware upgrades may offset operational efficiency gains, this is a trade-off that requires further investigation.

5 ENVIRONMENTAL TRANSPARENCY PATTERNS AND THEIR IMPLICATIONS

5.1 Measurement and Methodological Inconsistencies

Different organisations use varying methods to measure environmental impact, creating incomparable results. Google's comprehensive approach shows that narrow measurement approaches capture less than 60% of actual energy consumption, leading to a 2.4× underestimate of true environmental impact (Elsworth et al., 2025).

The absence of standardised frameworks creates several inconsistencies.

- Boundary variations: Some studies include comprehensive development costs and multiple training runs; others report only the final training run.
- Infrastructure accounting: Treatment varies from complete exclusion to full PUE accounting.
- Emissions calculations: Methods vary between location-based and market-based accounting approaches.

Table 5: Comparison of the major LLMs and their efficiency/finetuning approaches.

Model	Publisher	Published Year	Efficiency & Finetuning
GPT-3	OpenAI	2020	None, extensive model/data parallel training (Brown et al., 2020)
GPT-4	OpenAI	2023	Not publicly reported; RLHF, massive parallelism (OpenAI, 2023)
BERT (Base)	Google	2019	None
ALBERT (Base)	Google	2019	Parameter sharing
PaLM (Large)	Google	2022	None, Prompting only (Chowdhery et al., 2022)
Gemini	Google	–	Distillation (Flash), Pathways, Chat: supervised + RLHF (IBM, 2024)
RoBERTa (Base)	Facebook AI	2019	None (Liu et al., 2019)
OPT	Meta AI	2022	None, Minimal/derivatives use LoRA
LLaMA (65B)	Meta AI	–	None, Minimal/derivatives use LoRA
LLaMA 2 (Base, 13B, 70B)	Meta AI	2023	Quantised inference variants, RLHF (Chat)
BLOOM	BigScience	2022	None, Base + instruction variants (Scao et al., 2023)
Gopher	DeepMind	2021	None, Base + RLHF Chat version (Rae et al., 2022)
Mistral Large 2	Mistral AI	2024	None, Instruction + likely RLHF
Qwen	Alibaba Cloud	2023	None, Chat: supervised + RLHF (Bai et al., 2023)
DeepSeek-V3	DeepSeek	2024	Knowledge distillation, PEFT (LoRA), Supervised + RLHF (DeepSeek-AI et al., 2025)

- Water usage: Inconsistent inclusion of water consumption for cooling versus power generation.

5.2 Variability in AI Sustainability Reporting

AI organisations vary greatly in the way they report environmental data. This creates transparency gaps that prevent cross-model comparisons. The variation in disclosure practices is substantial: BigScience provides comprehensive data for BLOOM, documenting 433,196 kWh of training energy and 25 tCO_2e emissions (Scao et al., 2023), while OpenAI provides minimal disclosure for GPT-4 despite its widespread deployment and substantial computational requirements. Similarly, DeepSeek-V3 developers reported using 2.788 million H800 GPU-hours for training their 671-billion-parameter model (DeepSeek-AI et al., 2025), but provided no energy consumption or emissions figures.

Regional regulatory frameworks influence disclosure patterns. European organisations like BigScience often face stronger environmental disclosure requirements and have demonstrated more comprehensive reporting practices. Companies make strategic decisions about sharing environmental data, with organisations achieving favorable environmental performance often sharing detailed information, while those with less favorable metrics may choose limited disclosure.

5.3 Need for Standardised Environmental Governance

Transparency gaps prevent an accurate assessment of the total environmental impact of the AI industry and limit the development of evidence-based policies. Without consistent environmental data, markets cannot reward environmental improvements effectively, and researchers cannot identify the most sustainable development approaches.

Current voluntary disclosure approaches are insufficient. The systematic nature of reporting gaps suggests that standardised environmental reporting frameworks are necessary to ensure comparable environmental impact information across organisations. Such frameworks would allow researchers to identify the most effective practices, enable markets to reward environmental efficiency, and provide policymakers with the accurate data needed to develop sustainable policies.

The substantial variations in environmental im-

pact documented above raise critical questions about the consistency and transparency of measurements. Systematic gaps in environmental reporting prevent accurate comparisons and informed decision-making.

6 ANSWERING THE RESEARCH QUESTIONS

RQ1: Transparency Status: Seven of the thirteen major AI models examined lack verified environmental impact data. This includes widely used models from Google (BERT, RoBERTa, ALBERT and PaLM) and Chinese companies (DeepSeek and Qwen). The level of transparency ranges from BigScience's comprehensive BLOOM documentation to OpenAI's minimal GPT-4 disclosure. These transparency gaps make it difficult to accurately evaluate the total environmental impact of the AI industry, which limits the development of evidence-based policies.

RQ2: Methodological Impact: Methodological differences cause large variations in reported environmental impact. Organisations range from using narrow approaches that consider only direct emissions to comprehensive assessments that include infrastructure costs, which can multiply the impact by 2.2× to 4.5×. Studies also differ in terms of whether they include development costs, the methods they use to calculate emissions and how they account for geographical factors. These differences can create a variation of 5–10× for identical workloads. Indirect impacts, such as cooling water and carbon emissions from hardware (20–30% of emissions), as well as lifecycle stages beyond training, are often not considered. This explains why identical models can show dramatically different environmental footprints across studies.

RQ3: Key Factors Influencing Environmental Footprint: Geographic location has the greatest impact on the environment, creating differences in carbon intensity of up to 6× between regions and 77% higher emissions for identical workloads depending on grid composition. Infrastructure efficiency is the second most important factor, with comprehensive measurement revealing a 2.4× effect from data centre costs that narrow approaches ignore. The greatest improvement opportunity lies in optimisation potential, with systematic approaches demonstrating a 33× reduction in energy consumption and model selection enabling reductions in emissions of up to 96%.

RQ4: Required Standardisation Framework: Transparent and comparable environmental assessment of AI systems requires standardised methodologies that define broad measurement boundaries. This should cover infrastructure costs, idle capacity, and supporting systems, not just accelerators. Reporting should include mandatory disclosures such as the location of training, energy sources, grid carbon intensity, and full PUE. Organisations must report metrics consistently using standardised approaches, with clear documentation of methods for energy use (kWh), emissions (tCO_2e), and water consumption. Assessments must cover the entire AI lifecycle, including both training and inference, and include real-time tracking within development workflows instead of relying on forecasts.

These answers confirm that achieving environmental sustainability in AI requires overcoming technical optimisation and transparency challenges.

7 CONCLUSION

This study identifies significant differences in the environmental impact of major LLMs. Training energy consumption can vary by up to 41,000×, and carbon emissions for identical computational workloads can differ by up to 6× according to geographic location. These variations far exceed what can be explained by computational requirements only, demonstrating significant gaps in measurement methodologies and organisational transparency practices.

Seven of thirteen major AI models lack verified environmental impact data. Such data gaps prevent accurate cross-model comparisons and systematic optimisation efforts. Methodological inconsistencies create additional challenges, with narrow measurement approaches underestimating true environmental costs by excluding infrastructure costs, which multiply the impact by between 2.2× and 4.5×. Geographic factors have been identified as the dominant influence on environmental footprint, while systematic optimisation strategies have demonstrated the potential to reduce energy consumption in production systems by up to 33×.

An analysis of water consumption indicates a significant shift towards inference-dominated impact, with the annual cost of inference now exceeding the one-time cost of training for major deployed models. Regional variations in infrastructure efficiency create differences in water usage effectiveness of up to 5×, which emphasises the importance of strategic deployment decisions. These findings confirm the urgent need for standardised environmental reporting frameworks that allow transparent and comparable assessments of AI systems. Current voluntary disclosure approaches are insufficient for accurately evaluating the industry-wide impact or developing poli-

cies. Standardised methodologies must include comprehensive measurement boundaries, mandatory geographic and infrastructure disclosures, and lifecycle assessment approaches that capture the training and inference phases.

The transition from unstructured to systematic environmental governance is essential for the sustainable development of AI. As AI systems continue to scale up globally, implementing consistent measurement and reporting practices will enable markets to reward environmental efficiency, help researchers to identify optimisation strategies, and assist policymakers in developing evidence-based sustainability regulations. Only through such systematic approaches can the AI industry ensure that technological advancement aligns with environmental sustainability goals.

Future research should address the following limitations: Firstly, it should establish industry-wide, mandatory environmental disclosure requirements through regulatory frameworks or industry partnerships. Secondly, it should focus on developing comprehensive lifecycle assessment methodologies that consider the impacts of hardware manufacturing and disposal alongside operational metrics. Thirdly, environmental tracking should be integrated into ML development workflows in real time rather than being estimated post hoc. Finally, it is necessary to investigate optimisation strategies that balance hardware efficiency improvements against carbon costs.

ACKNOWLEDGEMENTS

This work has been funded by the Ministry of Economy, Innovation, Digitisation, and Energy of the State of North Rhine-Westphalia, Germany, and the European Union within the projects Digital.Zirkulär.Ruhr and Transferhub.

REFERENCES

Abou Baker, N. and Handmann, U. (2024). One size does not fit all: Benchmarking modelselection scores for image classification. *Scientific Reports (Sci Rep)*, 14(30239):1–26.

Abou Baker, N., Rohrschneider, D., and Handmann, U. (2022). Battery detection of xray images using transfer learning. In *The 30th European Symposium on Artificial Neural Networks (ESANN 2022)*, pages 241–246, Bruges, Belgium.

Abou Baker, N., Rohrschneider, D., and Handmann, U. (2024). Parameter-efficient fine-tuning of large pretrained models for instance segmentation tasks. *Machine Learning and Knowledge Extraction*, 6(4):2783–2807.

Anthony, L. F. W., Kanding, B., and Selvan, R. (2020). Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems. arXiv:2007.03051.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., and et al., F. H. (2023). Qwen technical report.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., and et al., A. A. (2020). Language models are few-shot learners.

Budennyy, S. A., Lazarev, V. D., Zakharenko, N. N., Korovin, A. N., Plosskaya, O. A., Dimitrov, D. V., Akhripkin, V. S., Pavlov, I. V., Oseledets, I. V., Barsola, I. S., Egorov, I. V., Kosterina, A. A., and Zhukov, L. E. (2022). eco2ai: Carbon emissions tracking of machine learning models as the first step towards sustainable ai. *Doklady Mathematics*, 106(1):S118–S128.

Cai, E., Juan, D.-C., Stamoulis, D., and Marculescu, D. (2017). Neuralpower: Predict and deploy energy-efficient convolutional neural networks.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., and et al., S. G. (2022). Palm: Scaling language modeling with pathways.

Claude AI (2025). AI query vs sustainable paper: the surprising environmental comparison. Accessed: September 2025.

Courty, B., Schmidt, V., Luccioni, S., Goyal-Kamal, MarionCoutarel, Feld, B., Lecourt, J., LiamConnell, Saboni, A., and et al., I. (2024). mlco2/codecarbon: v2.4.1.

de Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, 7(10):2191–2194.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., and et al., C. Z. (2025). Deepseek-v3 technical report.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Digit (2024). Mistrial ai study highlights the environmental impact of llms. *Digit*. Accessed: September 2024.

Eilam, T., Bello-Maldonado, P., Bhattacharjee, B., Costa, C., Lee, E. K., and Tantawi, A. (2023). Towards a methodology and framework for ai sustainability metrics. In *Proceedings of the 2nd Workshop on Sustainable Computer Systems*, HotCarbon '23, New York, NY, USA. Association for Computing Machinery.

Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S., Townsend, B., Ranganathan, P., Dean, J., Vahdat, A., Gomes, B., and Manyika, J. (2025). Measuring the environmental impact of delivering ai at google scale.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the systematic re-

porting of the energy and carbon footprints of machine learning.

IBM (2024). Google gemini: advanced multimodal ai (ibm think article). https://www.ibm.com/think/topics/google-gemini#:~:text=In\%20addition\%20to\%20a\%20transformer,4. Accessed: 2025-09-26.

Islam, M. S., Zisad, S. N., Kor, A.-L., and Hasan, M. H. (2023). Sustainability of machine learning models: An energy consumption centric evaluation. In *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6.

Jegham, N., Abdelatti, M., Elmoubarki, L., and Hendawi, A. (2025). How hungry is AI? benchmarking energy, water, and carbon footprint of llm inference.

Kansal, A., Zhao, F., Liu, J., Kothari, N., and Bhattacharya, A. (2009). Joulemeter: Virtual machine power measurement and management. Technical Report MSR-TR-2009-103, Microsoft Research. MSR Tech Report.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations.

Li, P., Yang, J., Islam, M. A., and Ren, S. (2025). Making ai less "thirsty": Uncovering and addressing the secret water footprint of ai models.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Ludvigson, K. (2023). The carbon footprint of gpt-4. *Towards Data Science*. Accessed: September 2025.

Mehta, S. (2024). How much energy do llms consume? unveiling the power behind ai. Accessed: 2025-09-24.

Mistral AI (2024). Mistral large 2 (v2.0). https://mistral.ai/news/mistral-large-2407. Accessed: 2025-09-30.

Mittal, R., Kansal, A., and Chandra, R. (2012). Empowering developers to estimate app energy consumption. In *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, Mobicom '12, page 317–328, New York, NY, USA. Association for Computing Machinery.

NeurIPS (2021). Neurips 2021 paper checklist. https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist. Accessed: 2025-09-24.

OpenAI (2023). Gpt-4 technical report. https://openai.com/index/gpt-4-research/. Accessed: 2025-09-26.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., and et al., S. Y. (2022). Scaling language models: Methods, analysis and insights from training gopher.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., and et al., A. R. (2022). Multitask prompted training enables zero-shot task generalization.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., and et al., A. M. R. (2023). Bloom: A 176b-parameter open-access multilingual language model.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., and et al., M. R. (2025). Dinov3.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., and et al., S. B. (2023). Llama 2: Open foundation and fine-tuned chat models.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022). Opt: Open pre-trained transformer language models.

Proceedings of ICPRAM 2026

15th International Conference on Pattern Recognition Applications and Methods

https://icpram.scitevents.org

EVENT MANAGEMENT SYSTEM:

LOGISTICS:

TECHNICALLY CO-SPONSORED BY:

ENDORSED BY:

IN COOPERATION WITH:

PROCEEDINGS WILL BE SUBMITTED FOR EVALUATION FOR INDEXING BY:

Copyright © 2026 by SCITEPRESS - Science and Technology Publications, Lda. All Rights Reserved

ISSN: 2184-4313
ISBN: 978-989-758-797-9