Advancing Green and Fair AI: A Research Perspective on Environmental and Social Sustainability

Loris Belcastro[1], Riccardo Cantini[1,*], Fabrizio Marozzo[1], Alessio Orsino[1,*], Domenico Talia[1] and Paolo Trunfio[1]

[1] *University of Calabria, Rende, Italy*

Abstract

Artificial Intelligence (AI) systems are increasingly embedded in everyday technologies, posing the need to align their development with principles of environmental and social sustainability. This paper presents the research activities of the ScaLab team at UNICAL, centered on sustainable AI, offering a unified perspective that examines both its *environmental* impact and *societal* implications. On the environmental front, the research investigates interpretable energy estimation for edge AI, efficient knowledge distillation, and lightweight test-time adaptation, all designed to optimize resource utilization in constrained settings. From a social viewpoint, it tackles fairness and inclusivity of Large Language Models by using adversarial analysis to reveal and address hidden biases and discriminatory behavior. Collectively, these efforts aim to foster the development of AI systems that are both resource-efficient and ethically responsible, promoting a more sustainable and equitable digital future.

Keywords

Sustainable AI, Fair AI, Ethical AI, Edge AI, Resource-Efficient AI, Knowledge Distillation, Test-Time Adaptation, Bias Elicitation, LLMs

1. Introduction

Artificial Intelligence (AI) has witnessed rapid adoption across domains such as healthcare, finance, and education, offering unprecedented benefits in automation and decision-making. However, this progress has also raised urgent concerns about the sustainability of AI systems, particularly with the rise of Large Language Models (LLMs) and edge AI. Among the many challenges of sustainable AI, this work focuses specifically on two pressing and interconnected issues: *environmental frugality*, as training and deploying large models require substantial energy and hardware resources—especially challenging in resource-constrained edge settings—and *social fairness*, as LLMs often inherit and amplify societal biases present in their training data, potentially leading to discriminatory behavior across sensitive demographic categories. Addressing these dual concerns demands a unified approach that balances efficiency, interpretability, and ethical responsibility.

This paper presents a comprehensive research perspective aimed at advancing sustainable AI along these two interdependent dimensions, highlighting our recent contributions to both environmental and social sustainability in AI. On the environmental front, our work focuses on *interpretable energy estimation* for edge AI applications [1], XAI-driven knowledge distillation [2], and edge-based test-time adaptation [3]. On the social sustainability front, we introduce *adversarial bias elicitation* benchmarks to uncover hidden biases in LLMs through jailbreak prompting [4, 5].

Together, these contributions support the development of AI systems that are efficient, fair, and transparent, fostering a more sustainable and socially responsible AI paradigm. The remainder of the paper is structured as follows. Sections 2 and 3 discuss our research efforts to enhance environmental and social sustainability in AI systems. Section 4 outlines our vision for an ethical-by-design framework for sustainable AI. Finally, Section 5 concludes the paper.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

[*] Corresponding author.

lbelcastro@dimes.unical.it (L. Belcastro); rcantini@dimes.unical.it (R. Cantini); fmarozzo@dimes.unical.it (F. Marozzo); aorsino@dimes.unical.it (A. Orsino); talia@dimes.unical.it (D. Talia); trunfio@dimes.unical.it (P. Trunfio)
0000-0001-6324-8108 (L. Belcastro); 0000-0003-3053-6132 (R. Cantini); 0000-0001-7887-1314 (F. Marozzo); 00009-0004-6216-8885 (A. Orsino); 0000-0003-1910-9236 (D. Talia); 0000-0002-5076-6544 (P. Trunfio)

2. Environmental Sustainability in AI Systems

2.1. Interpretable Energy Estimation for Edge AI Applications

Deploying AI models on edge devices, such as smartphones, IoT sensors, and embedded systems, is increasingly common due to the need for personalized, real-time, and privacy-preserving inference [6]. However, these devices are typically constrained in energy, memory, and compute power [7], with energy consumption being a critical factor for usability and sustainability. However, traditional energy profiling often requires runtime instrumentation and yields coarse-grained, opaque estimates.

To address this, we proposed a novel methodology for fine-grained, *interpretable energy estimation* that models the energy footprint of edge AI workloads using learnable proxies, each reflecting a specific contributor such as computation, data access, or communication [1]. For a given machine learning algorithm a and dataset d, total energy consumption is decomposed as:

$$E(a,d) = E_{\text{comp}}(a,d) + E_{\text{data}}(a,d) + E_{\text{comm}}(a,d)$$

Each energy component is modeled using an interpretable proxy (e.g., execution time, cache misses, network usage), which is predicted by a regression model $\mathcal{R}_x(f)$ based on a feature vector f. This feature vector describes the application and includes attributes such as dataset size, number of samples, dimensionality, and algorithm type. The regressors are trained on data collected through application monitoring during the execution of various ML algorithms across diverse datasets. The proxy values predicted by the regressors are linearly combined and converted into energy estimates using empirically calibrated scaling factors β_x. Specifically, the total energy estimate is formulated as follows:

$$E(f) = \beta_{\text{comp}} \cdot \mathcal{R}_{\text{comp}}(f) + \beta_{\text{data}} \cdot \mathcal{R}_{\text{data}}(f) + \beta_{\text{comm}} \cdot \mathcal{R}_{\text{comm}}(f)$$

To ensure interpretability, we integrated Explainable AI (XAI) techniques [8], using attribution methods to reveal how input features influence each energy component. For example, high E_{comp} may result from large data dimensionality, while high E_{data} can indicate poor data locality or frequent memory access. This enables accurate, component-level energy estimates with actionable insights into their causes. In addition, the framework includes a classifier that filters algorithm-dataset pairs violating resource constraints (e.g., time or memory), explaining violations via the same attribution methods.

In preliminary experiments on a Raspberry Pi 4, the framework showed high predictive accuracy for proxy variables and total energy consumption, while providing interpretable insights to identify inefficiencies and guide energy-aware scheduling, algorithm design, and selection. Overall, this methodology promotes sustainable AI by equipping developers with practical tools to analyze and optimize the energy footprint of edge AI applications through fine-grained, interpretable estimates.

2.2. XAI-Driven Knowledge Distillation of Large Language Models

While LLMs have recently gained significant traction for their impressive language understanding and generation capabilities, deploying them on resource-constrained devices remains impractical due to their substantial computational and memory demands [9]. To mitigate this challenge, *Knowledge Distillation* (KD) [10, 11] has emerged as an effective approach, transferring knowledge from a large *teacher* model to a smaller, more efficient *student* one with minimal performance degradation. Traditional KD methods train the student to replicate the teacher's soft output distributions but often overlook the teacher's underlying reasoning. As a result, the student may imitate predictions without internalizing the rationale, potentially limiting generalization.

To bridge this gap, we introduced *DiXtill* [2], a novel *XAI-driven knowledge distillation* methodology that exploits local explanations of the teacher's predictions to complement traditional prediction-based supervision. DiXtill leverages Integrated Gradients (IG) to extract word-level attribution scores from the teacher model for each input instance, capturing the underlying rationale behind its predictions and transferring it to the student model. The student is a lightweight, self-explainable network that generates its own explanations via masked attention. The key innovation lies in incorporating explanation alignment into the distillation objective, encouraging the student to replicate not only the teacher's predictions but also its explanation patterns. The distillation loss in DiXtill combines three components:

$$\mathcal{L} = (1 - \alpha) \cdot \mathcal{L}_{\text{CE}} + \alpha \cdot (\mathcal{L}_{\text{KD}}^{(\tau)} + \mathcal{L}_{\text{XAI}})$$

Here, (i) $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss with ground-truth (hard) labels, (ii) $\mathcal{L}_{\text{KD}}$ is the distillation loss based on KL-divergence between teacher and student soft outputs, scaled with temperature τ, and (iii) $\mathcal{L}_{\text{XAI}}$ represents the explanation alignment term, computed as the cosine similarity loss between teacher and student attribution vectors. By jointly optimizing for both accuracy and explanation fidelity, DiXtill produces a self-explainable student model whose predictions and rationales closely align with those of the teacher, thereby enhancing both classification performance and interpretability.

Extensive experiments on a financial sentiment classification task—using a BERT-like model [12] as the teacher and a lightweight Bi-LSTM network with masked attention as the student—show that DiXtill achieves superior classification accuracy compared to traditional knowledge distillation. It also demonstrates higher alignment between teacher and student explanations, along with significant model compression and inference speedup relative to the teacher. Moreover, DiXtill outperforms alternative compression methods such as post-training quantization (PTQ) and structured pruning, which often preserve the original model's black-box nature. In contrast, it produces a compact, fast, and inherently interpretable student model without sacrificing classification performance. Overall, our XAI-driven distillation framework advances sustainable and responsible AI by showing that explainability can guide knowledge transfer into efficient, interpretable models suited for deployment in resource-constrained environments requiring both efficiency and transparency.

2.3. Efficient Test-Time Adaptation on Ultra-Low-Resource Devices

Deep neural networks deployed in real-world applications often face degraded performance due to distribution shifts, which can arise from sensor noise, environmental changes, or domain variations. Test-time adaptation (TTA) has emerged as an effective approach to address such shifts by adapting a pre-trained model to incoming unlabeled samples through entropy minimization [13]. However, most existing TTA methods rely on backpropagation through the full model or updates to normalization layers, both of which are memory-intensive and computationally demanding. These limitations make them unsuitable for deployment on memory-constrained devices such as microcontroller units (MCUs).

To address these challenges, in collaboration with the research group led by Prof. Mascolo at the University of Cambridge, we proposed TinyTTA [3], a resource-efficient framework that enables TTA on devices with strict hardware constraints. TinyTTA enhances memory and energy efficiency during TTA, improving the adaptability of pre-trained models to diverse distribution shifts and facilitating deployment on edge devices. To reduce memory usage, it introduces a self-ensembling strategy that updates only lightweight submodules—groups of consecutive layers from the pre-trained network—while keeping the main layers frozen. This avoids storing full activations for backpropagation, reducing memory overhead. TinyTTA also employs early exits, allowing samples to exit from intermediate submodules instead of passing through the full network, improving latency and energy efficiency. To further minimize memory use and support MCUs, TinyTTA replaces standard normalization layers, which require large batches and are unsupported on many MCUs, with weight standardization [14], enabling batch-agnostic normalization. Finally, to support on-device adaptation, TinyTTA introduces the TinyTTA Engine, a lightweight runtime built on TensorFlow Lite Micro [15], adding backward support for key layers and enabling training on devices with extremely low memory.

The effectiveness of TinyTTA was assessed using four distinct corrupted datasets and four different model architectures on two types of edge devices, a Raspberry Pi Zero 2W and an STM32H747 MCU with extremely low memory. Experimental results demonstrate that TinyTTA improves adaptation accuracy over state-of-the-art TTA baselines when using small batch sizes, drastically reduces memory usage, and achieves lower latency and energy consumption—making it suitable for real-time adaptation on resource-constrained and battery-limited devices. Moreover, it is the only framework able to run TTA on the MCU STM32H747 with a 512 KB memory constraint while maintaining high performance.

3. Social Sustainability in AI Systems

3.1. Adversarial Bias Elicitation in Large Language Models

While LLMs have demonstrated remarkable performance in natural language understanding and generation [16], they remain susceptible to social and representational biases embedded in their training data. These biases often manifest subtly through stereotypical assumptions about gender, ethnicity, religion, sexual orientation, disability, or socioeconomic status, and can seriously impact sensitive domains such as law and healthcare [17, 18]. Traditional fairness assessments often fail to detect such issues, especially when LLMs are aligned to avoid producing overtly harmful content.

To address this gap, we introduce a robust adversarial benchmarking framework designed to evaluate *bias resilience* and *robustness to elicitation* in popular LLMs [4]. Our approach systematically probes models with carefully crafted prompts to reveal hidden biases that standard evaluations may overlook. The proposed methodology follows a two-step pipeline: a standard safety evaluation followed by adversarial analysis using jailbreak prompts. First, each LLM is evaluated using a curated set of sentence completion prompts spanning seven bias categories: age, gender, ethnicity, sexual orientation, disability, religion, and socioeconomic status. For each prompt, the model chooses between two completions, one stereotypical, the other counter-stereotypical. Model behavior is analyzed using metrics capturing key safety aspects, including the tendency to refuse to engage with the prompt (*refusal rate*), to soften biased content (*debiasing rate*), the frequency of stereotyped vs. counter-stereotyped choices, polarization tendency (*fairness*), and an aggregate *safety score* summarizing robustness across bias categories.

Categories exceeding a predefined safety threshold are deemed initially "*safe*" and selected for the subsequent adversarial evaluation phase. Here, we apply jailbreak techniques [19] to test model robustness under adversarial manipulation. These include *role-playing* attacks, where the model adopts fictional personas to justify bias; *machine translation* attacks, which mask prompts using low-resource languages [20]; *obfuscation* (altered spellings); *prompt injection*, embedding harmful instructions within benign inputs; and *reward incentive* attacks, framing biased completions as reward-driven. These adversarial prompts aim to bypass alignment filters and elicit biased outputs even from models deemed safe, enabling to quantify safety degradation as the percentage drop in category-level safety scores.

The analysis—conducted on various widely-used language models at different scales, ranging from high-quality Small Language Models (SLMs) to large-scale LLMs—revealed that no model was fully immune to all forms of adversarial prompting. Even models with strong safety filters can be manipulated to produce biased content under specific jailbreak conditions. Moreover, categories like gender and age were especially vulnerable. Interestingly, model size did not always correlate with robustness, since some SLMs outperformed larger ones on certain bias dimensions.

These findings underscore the limitations of alignment-based filtering and emphasize the need for adversarial red-teaming as a standard in AI safety assessment to better inform bias mitigation strategies.

3.2. Corpus for Linguistic Evaluation of Adversarial Robustness against Bias

To support a thorough assessment of the ethical behavior of language models, we introduced and publicly released *CLEAR-Bias* (*Corpus for Linguistic Evaluation of Adversarial Robustness against Bias*) [5], a systematically curated benchmark dataset designed to assess bias vulnerabilities in LLMs under both standard and adversarial conditions.

CLEAR-Bias comprises 4,400 bias-probing prompts across two task formats: *Choose the Option* (CTO), which assesses model tendency toward biased choices, and *Sentence Completion* (SC), which evaluates the potential for biased generations. The dataset spans ten bias categories, including seven isolated dimensions (age, disability, ethnicity, gender, religion, sexual orientation, and socioeconomic status) and three intersectional categories, designed to capture overlapping stereotypes (ethnicity-socioeconomic status, gender-sexual orientation, and gender-ethnicity), providing a fine-grained taxonomy of bias types across 37 demographic subgroups. Each category contains 20 carefully crafted *base prompts* (10 per task type), totaling 200 base prompts. To simulate diverse adversarial scenarios, each base prompt is systematically expanded using seven jailbreak techniques: *machine translation*, *obfuscation*, *prefix* *injection*, *prompt injection*, *refusal suppression*, *reward incentive*, and *role-playing*. Each technique is implemented in three variants, resulting in 4,200 *adversarial prompts*, which offers a comprehensive testbed for evaluating bias robustness under adversarial manipulation.

4. Toward an Ethical-by-Design Framework for Sustainable AI

In AI, *Ethics by Design* (EbD) refers to the integration of ethical principles–such as fairness, transparency, and accountability–throughout the AI development lifecycle. While EbD traditionally emphasizes values like privacy, fairness, transparency, and accountability, it should also encompass principles of environmental and social sustainability. A unified EbD framework should include strategies for measuring, monitoring, and improving the environmental impact of AI systems, with particular focus on energy consumption and resource efficiency. It must also promote social fairness by addressing structural inequalities and discrimination based on individual or group characteristics. Ethical impact assessments should be embedded throughout the AI lifecycle, alongside robust methods for bias detection, elicitation, and mitigation. Operational transparency, trustworthiness, and explainability are especially critical in high-stakes or sensitive applications and must be prioritized. Furthermore, interdisciplinary collaboration is essential to align AI with human values, legal standards, and societal expectations. The framework should also establish clear accountability and governance structures, including well-defined responsibilities, escalation procedures, and mechanisms to address ethical concerns as they arise.

Technical innovations are also key to achieving this vision. The paradigm shift toward edge-based and decentralized AI—enabling local inference and reducing reliance on cloud infrastructure—enhances data privacy and operational reliability while supporting more sustainable models. These models can maintain their capabilities while reducing resource consumption, thereby improving their environmental footprint. In fact, lower computational demands and increased energy efficiency align with broader environmental goals by reducing the carbon footprint of AI systems, thus contributing to both environmental sustainability and the democratization of AI. Advancing fields such as knowledge distillation and edge-based test-time adaptation is therefore critical. Moreover, accurately modeling and predicting the energy consumption of such applications is essential to informing algorithm design and deployment strategies. Finally, inclusivity must be a foundational principle. Ethical design requires the intentional integration of diverse perspectives across the AI pipeline to prevent systemic bias and promote equitable outcomes. This includes refining datasets, applying debiasing strategies, and adopting inclusive evaluation criteria. Inclusivity should not be treated as an afterthought, but as a core value in AI system design, incorporating diverse cultural contexts and demographic identities at every stage—from training data to evaluation benchmarks.

Together, these efforts lay the foundation for a sustainable AI paradigm that balances deployment constraints with performance, fairness, transparency, and inclusivity.

5. Conclusions

This paper highlighted the pressing need for sustainable AI development by addressing both environmental and societal challenges. Through the research initiatives of the ScaLab team at UNICAL, it demonstrated the potential for creating AI systems that are not only resource-efficient but also socially responsible. By combining interpretable energy estimation, cross-architecture knowledge distillation, efficient test-time adaptation on ultra-low-resource devices, and rigorous evaluations of adversarial robustness in language models, these approaches collectively build toward a unified, ethical-by-design framework aimed at fostering a more sustainable and equitable future for AI.

Acknowledgments

This research was supported by "PNRR MUR project PE0000013-FAIR" - CUP H23C22000860006.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] R. Cantini, A. Orsino, D. Talia, P. Trunfio, Towards interpretable energy estimation for edge ai applications, in: 3rd International Workshop on Intelligent and Adaptive Edge-Cloud Operations and Services, 39th IEEE International Parallel and Distributed Processing Symposium, 2025.

[2] R. Cantini, A. Orsino, D. Talia, Xai-driven knowledge distillation of large language models for efficient deployment on low-resource devices, Journal of Big Data 11 (2024).

[3] H. Jia, Y. Kwon, A. Orsino, T. Dang, et al., Tinytta: Efficient test-time adaptation via early-exit ensembles on edge devices, Advances in Neural Information Processing Systems 37 (2024).

[4] R. Cantini, G. Cosenza, A. Orsino, D. Talia, Are large language models really bias-free? jailbreak prompts for assessing adversarial robustness to bias elicitation, in: International Conference on Discovery Science, 2024.

[5] R. Cantini, A. Orsino, M. Ruggiero, D. Talia, Benchmarking adversarial robustness to bias elicitation in large language models: Scalable automated assessment with llm-as-a-judge, arXiv preprint arXiv:2504.07887 (2025).

[6] H. Hua, Y. Li, T. Wang, et al., Edge computing with artificial intelligence: A machine learning perspective, ACM Computing Surveys 55 (2023).

[7] A. R. Khouas, M. R. Bouadjenek, H. Hacid, et al., Training machine learning models at the edge: A survey, arXiv preprint arXiv:2403.02619 (2024).

[8] R. Dwivedi, D. Dave, H. Naik, et al., Explainable ai (xai): Core ideas, techniques, and solutions, ACM Computing Surveys 55 (2023).

[9] E. Frantar, D. Alistarh, Sparsegpt: Massive language models can be accurately pruned in one-shot, in: International Conference on Machine Learning, 2023.

[10] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[11] J. Gou, B. Yu, S. J. Maybank, D. Tao, Knowledge distillation: A survey, International Journal of Computer Vision 129 (2021).

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[13] D. Wang, E. Shelhamer, S. Liu, B. A. Olshausen, T. Darrell, Tent: Fully test-time adaptation by entropy minimization, in: 9th International Conference on Learning Representations, 2021.

[14] S. Qiao, H. Wang, C. Liu, W. Shen, et al., Micro-batch training with batch-channel normalization and weight standardization, arXiv preprint arXiv:1903.10520 (2019).

[15] R. David, J. Duke, A. Jain, V. Janapa Reddi, et al., Tensorflow lite micro: Embedded machine learning for tinyml systems, Proceedings of Machine Learning and Systems 3 (2021).

[16] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, et al., A survey on evaluation of large language models, ACM Trans. Intell. Syst. Technol. (2024).

[17] R. Navigli, S. Conia, B. Ross, Biases in large language models: origins, inventory, and discussion, ACM J. Data Inf. Qual. 15 (2023).

[18] D. Talia, P. Trunfio, R. Cantini, A. Orsino, The bias problem in healthcare ai, in: M. Cannataro (Ed.), The Evolution of Artificial Intelligence in Healthcare, Springer, 2026. To Appear.

[19] S. Yi, Y. Liu, Z. Sun, T. Cong, et al., Jailbreak attacks and defenses against large language models: A survey, arXiv preprint arXiv:2407.04295 (2024).

[20] S. Ranathunga, E. A. Lee, M. P. Skenduli, R. Shekhar, et al., Neural machine translation for low-resource languages: A survey, ACM Computing Survey (2023).