Rethinking AI Innovation Through Computational Accountability: From Accuracy-Centered Evaluation to Cost-Aware Intelligence

Jun-Hao Chen[1], Ingrid L. Petersen[2,*], Suresh Ramanathan[3]

[1] School of Computing, National Tsing Hua University, Hsinchu 30013, Taiwan
[2] Department of Information Systems, Copenhagen Business School, 2000 Frederiksberg, Denmark
[3] Institute for AI Innovation, Indian Institute of Technology Madras, Chennai 600036, India

* Email: *i.petersen@cbs.dk* (Corresponding Author)

Abstract

The prevailing evaluation culture in artificial-intelligence (AI) research treats predictive accuracy as the dominant, often sole, indicator of progress. This accuracy-centered view has produced remarkable methodological advances, yet it is increasingly at odds with the operational realities of contemporary AI systems, whose computational demands have risen by several orders of magnitude over the past decade. This paper argues that the next phase of AI innovation must be organised around the concept of computational accountability: a systematic, reproducible, and hardware-independent accounting of the resources an AI model requires to deliver its predictions. We decompose computational accountability into three mutually reinforcing pillars—algorithmic accountability captured by floating-point-operation (FLOP) counts, numerical-precision accountability captured by bit-operation (BOP) counts, and hardware-execution accountability captured by energy and carbon measurements—and argue that each pillar, while useful in isolation, becomes meaningful only when reported jointly. Drawing on a structured review of seventy prior studies spanning Green AI, quantization, hardware-aware design, and sustainable machine learning, the paper develops a conceptual framework that positions computational accountability as a methodological discipline rather than a tool choice, and proposes a unified accountability ledger that records per-run, per-model, and per-precision workload indicators. An illustrative analysis across eight representative architectures and three precision regimes shows that accuracy-normalised cost indicators reorder model rankings relative to raw accuracy, frequently by more than one quartile, and that BOP-based analyses reveal quantization benefits that FLOP-only analyses systematically under-count. The paper concludes with concrete recommendations for researchers, reviewers, editors, and funding agencies, and sketches a policy interface through which computational accountability can be integrated into publication norms, procurement decisions, and sustainability audits. The goal is not to diminish the role of accuracy in AI evaluation but to situate it inside a richer, cost-aware narrative that treats computational demand as a scientific variable rather than a hidden externality.

Keywords: Computational accountability; Green AI; FLOPs; Bit-operations; Quantization; Cost-aware evaluation; Sustainable machine learning; AI benchmarking; Reporting standards

Article History
Received: July 14, 2025

Revised: September 22, 2025
Accepted: November 16, 2025
Available Online: December 30, 2025

Rethinking AI Innovation Through Computational Accountability: From Accuracy-Centered Evaluation to Cost-Aware Intelligence

1. Introduction

Artificial-intelligence research has become a central engine of contemporary technological innovation, contributing to scientific breakthroughs in natural-language understanding, computer vision, robotics, drug discovery, and climate modelling (Bommasani et al., 2021; Jumper et al., 2021; Silver et al., 2018). Alongside these achievements, however, the computational demand of frontier AI systems has grown at a rate that substantially outpaces the scaling of underlying hardware efficiency. Training a single state-of-the-art large language model now routinely requires several million accelerator-hours and produces carbon-equivalent emissions comparable to the lifetime footprint of dozens of personal vehicles (Patterson et al., 2021; Dodge et al., 2022). Inference costs, though smaller per query, are projected to exceed training costs by a large margin over the deployed lifetime of a model, driving aggregate energy demand that is increasingly visible in national grid statistics (Desislavov, Martínez-Plumed, & Hernández-Orallo, 2023; Masanet, Shehabi, Lei, Smith, & Koomey, 2020).

In this context, the dominant evaluation culture of the field—centred almost entirely on predictive accuracy, loss, or task-specific leaderboard metrics—has become a source of methodological asymmetry. Accuracy is, by design, a hardware-independent quantity; reporting it requires nothing beyond the outputs of the model and the ground-truth labels. The resources consumed to reach that accuracy, by contrast, depend on the processor, the interconnect, the software stack, the batch size, the precision configuration, and a host of environmental factors, none of which are typically disclosed (Lacoste, Luccioni, Schmidt, & Dandres, 2019; Bannour, Ghannay, Névéol, & Ligozat, 2021). The result is that a substantial part of the computational story behind each published result is, in effect, off the record. Recent empirical surveys confirm that fewer than one in ten machine-learning papers reports compute figures beyond the model-parameter count, and fewer than one in twenty reports energy or carbon measurements (Henderson et al., 2020; Wu et al., 2022).

This paper argues that the next phase of responsible AI innovation requires a systematic practice that we will term computational accountability. By computational accountability we mean a disciplined, reproducible, hardware-independent accounting of the resources an AI model consumes to produce its predictions, reported with the same rigour and consistency as accuracy itself. The concept builds on and integrates several existing lines of research—sustainable AI (Wu et al., 2022; van Wynsberghe, 2021), energy-aware benchmarking (Henderson et al., 2020), quantization research (Gholami et al., 2022), and

sustainability metrics (Bender, Gebru, McMillan-Major, & Shmitchell, 2021; Lannelongue, Grealey, & Inouye, 2021)—but it organises them into a single coherent reporting discipline whose outputs can be consumed not only by researchers but by reviewers, editors, standards bodies, and funding agencies.

The starting point is that computational cost is not a single number. It is a composite quantity with at least three distinct dimensions: an algorithmic dimension, measuring how many operations the model requires as a function of its architecture; a numerical-precision dimension, measuring how expensive each operation is given the bit-widths of its operands; and a hardware-execution dimension, measuring how much energy and emission are produced when the operations are executed on a particular platform. Each of these dimensions has, separately, been the object of sustained research. What is missing, and what this paper seeks to contribute, is a unifying framework under which the three are reported jointly, with clear conventions about which questions each dimension can and cannot answer.

The central argument proceeds in five steps. First, we review the limitations of accuracy-centred evaluation and situate the computational-accountability idea within prior work on Green AI, cost-aware benchmarking, and quantization-aware complexity measures. Second, we articulate the three-pillar conceptual framework and describe how its pillars interact. Third, we present an analytical exercise over representative model families that demonstrates how cost-aware indicators reorder architectures relative to raw accuracy, and how precision-aware metrics in particular reveal efficiency gains that conventional FLOP-based profilers miss. Fourth, we discuss how the framework can be operationalised through a minimal accountability ledger that attaches to existing experimental pipelines. Fifth, we outline a policy interface through which computational accountability can propagate into publishing norms, review checklists, institutional procurement, and regulatory standards.

The contribution is deliberately framed at the level of methodology rather than at the level of a new tool. A number of high-quality measurement tools already exist, including process-level energy meters (Jay et al., 2023), the Experiment Impact Tracker (Henderson et al., 2020), and green-algorithms carbon calculators (Lannelongue, Grealey, & Inouye, 2021). What these tools lack is a common language that situates each measurement within a shared reporting discipline. By articulating that discipline, this paper aims to bridge the gap between useful-but-isolated instrumentation and a coherent practice of cost-aware AI innovation. The remainder of the paper is organised as follows. Section 2 reviews related work across Green AI, hardware-aware measurement, quantization, and sustainability reporting. Section 3 presents the conceptual framework of computational accountability. Section 4 develops the three-pillar taxonomy in detail. Section 5 presents an analytical case study. Section 6 operationalises the framework through the accountability ledger. Section 7 discusses policy and institutional implications. Section 8 notes limitations. Section 9 concludes.

2. Related Work

Research on the computational and environmental cost of AI has grown rapidly over the past five years, branching into several interrelated streams. We organise the review around four themes: (a) the Green AI

movement and its broader reporting implications, (b) energy and carbon measurement tools, (c) algorithmic complexity indicators including FLOPs and their extensions, and (d) quantization-aware and hardware-aware complexity measures such as bit-operations.

2.1 Green AI and the Shift Toward Cost-Aware Evaluation

The Green AI movement, which is by now familiar in its broad contours, emphasises that AI research should report efficiency alongside accuracy so that readers can assess the cost of the reported improvements. The community has converged on several canonical observations. First, compute demand has grown exponentially: Amodei and Hernandez (2018) documented a 300000-fold increase in the training compute of landmark models over seven years, and more recent analyses confirm that the curve has steepened further, particularly after the emergence of large language models (Sevilla et al., 2022; Villalobos et al., 2022). Second, the carbon cost of training frontier models is no longer negligible. Patterson et al. (2021) estimated that the training of GPT-3 consumed on the order of 1287 MWh and produced over 550 t of CO_2 equivalent, with comparable figures reported for other large systems (Faiz et al., 2024; Dodge et al., 2022). Third, inference at scale can dominate lifetime cost. Desislavov, Martínez-Plumed, and Hernández-Orallo (2023) showed that, for widely deployed models, aggregate inference energy often exceeds training energy within months of deployment.

These empirical results motivated a series of methodological proposals. Bender, Gebru, McMillan-Major, and Shmitchell (2021) argued for explicit documentation of training data and computational resources in large-language-model development. Henderson et al. (2020) introduced the Experiment Impact Tracker to provide researchers with a lightweight mechanism for reporting energy and carbon figures. Wu et al. (2022) offered a comprehensive roadmap for sustainable AI, identifying intervention points along the model-development lifecycle. Rolnick et al. (2022) and Kaack et al. (2022) examined how AI can both contribute to and alleviate climate harm, arguing that cost reporting is a precondition for an informed balance-of-impact analysis.

2.2 Measurement Tools: Energy and Emissions

A second stream of work has produced concrete measurement tools. Process-level energy meters estimate CO_2 emissions by combining real-time hardware counters with regional grid-intensity data (Jay et al., 2023). Green-algorithms calculators provide retrospective estimation when direct instrumentation is unavailable (Lannelongue, Grealey, & Inouye, 2021). At the infrastructure level, Kubernetes-based Efficient Power Level Exporters and Green Software Foundation specifications aim to standardise measurement across containerised deployments (García-Martín, Rodrigues, Riley, & Grahn, 2019). More recent frameworks such as WattsOnAI (Liu, Yang, & Wang, 2025) add visualisation and correlation analysis to the measurement workflow. The common limitation of these tools is their hardware-dependence: an emission figure measured on one GPU family is not directly comparable to a figure from another, even when the underlying model is identical, a limitation noted by García-Martín et al. (2019) and reinforced by subsequent empirical studies.

2.3 Algorithmic Complexity: FLOPs and Their Extensions

The third stream consists of hardware-independent complexity indicators, of which the floating-point-operation (FLOP) count is the most widely used. FLOP-based analysis has a long lineage in computer architecture and scientific computing, predating deep learning by several decades (Hennessy & Patterson, 2019). In the deep-learning context, FLOPs have been promoted as a portable complexity measure that reflects the intrinsic workload of an architecture (Sze, Chen, Yang, & Emer, 2020). Tools such as DeepSpeed's Flops Profiler (Rasley, Rajbhandari, Ruwase, & He, 2020) extend FLOP accounting to both forward and backward passes, with complementary open-source profilers available for the PyTorch and TensorFlow ecosystems. The operations supported, the conventions used (for example, whether a multiply–accumulate is counted as one or two FLOPs), and the handling of fused kernels vary across tools (Howard et al., 2017; Tan & Le, 2019). Recent work has also examined how FLOPs interact with sparsity (Hoefler, Alistarh, Ben-Nun, Dryden, & Peste, 2021), with attention-specific optimisations (Dao, Fu, Ermon, Rudra, & Ré, 2022), and with parameter-efficient fine-tuning methods such as LoRA and adapters (Hu et al., 2022; Houlsby et al., 2019).

Despite their portability, FLOPs are known to be an imperfect proxy for hardware cost because they ignore operand precision, memory-bandwidth effects, and communication overheads. Studies on the memory-bound nature of attention (Dao et al., 2022), on the impact of precision on latency (Micikevicius et al., 2018), and on the sensitivity of energy to arithmetic intensity (Horowitz, 2014) have shown that FLOP-based rankings can be misleading when the comparison spans different architectural or numerical regimes. These limitations have motivated complementary metrics, discussed next.

2.4 Precision-Aware Complexity: BOPs and Hardware-Aware Metrics

Bit-operations (BOPs) extend FLOPs by scaling each operation by the product of its operand bit-widths, reflecting the fact that narrower arithmetic units are substantially cheaper in silicon (Gholami et al., 2022). Hawks, Duarte, Fraser, Pappalardo, Tran, and Umuroglu (2021) formalised the measure for quantization-aware pruning, and Coelho et al. (2021) demonstrated its usefulness as a proxy for FPGA area and energy. A parallel line of work on mixed-precision training (Micikevicius et al., 2018) and post-training quantization (Jacob et al., 2018; Banner, Nahshan, Hoffer, & Soudry, 2018; Wang, Liu, Lin, Lin, & Han, 2019) has demonstrated that integer arithmetic at 8 bits or lower can preserve accuracy while dramatically reducing computational cost. Recent advances include activation-aware weight quantization (Lin, Tang, Tang, Yang, Chen, Wang, Xiao, Dang, Gan, & Han, 2024), GPTQ-style accurate post-training quantization (Frantar, Ashkboos, Hoefler, & Alistarh, 2023), SmoothQuant for large transformers (Xiao, Lin, Seznec, Wu, Demouth, & Han, 2023), and 8-bit matrix multiplication for LLMs at scale (Dettmers, Lewis, Belkada, & Zettlemoyer, 2022). These methods consistently show that precision-aware complexity indicators are indispensable for evaluating modern quantized systems, because FLOPs alone may even increase under aggressive quantization due to on-the-fly dequantization operations.

2.5 Hardware-Aware Neural Architecture Search and Benchmarks

A related research thread has explored hardware-aware design and benchmarking. Hardware-aware neural architecture search (HW-NAS) optimises models jointly for accuracy and hardware cost on a target platform (Tan et al., 2019; Cai, Zhu, & Han, 2019). Benchmarks such as MLPerf (Mattson et al., 2020; Reddi et al., 2020) and its inference counterpart have driven convergence around a set of workloads for system evaluation, although the benchmarks focus primarily on throughput and latency rather than on accounting of algorithmic cost. Complementary benchmarks for efficiency specifically include EfficientBench (Lin, Kim, Cai, Gan, & Han, 2021) and studies on on-device transformers (Lin, Chen, Han, & Zhu, 2020). Further work has considered edge-device constraints and mobile inference, with reviews summarising practical optimisations such as pruning, knowledge distillation, and operator fusion (Sze et al., 2020; Sanh, Debut, Chaumond, & Wolf, 2019).

2.6 Policy and Governance Perspectives

Finally, a growing literature situates AI cost reporting within governance, policy, and institutional frameworks. Jobin, Ienca, and Vayena (2019) surveyed global AI ethics guidelines and found sparse reference to environmental considerations, a gap partially addressed by subsequent documents (Crawford, 2021). The EU AI Act (European Commission, 2024) introduces explicit transparency obligations for foundation models, although its computational-cost provisions remain broad. Work by van Wynsberghe (2021) frames sustainable AI as an ethics question; Yigitcanlar, Mehmood, and Corchado (2021) discuss green AI in the context of smart-city planning; and Hacker (2024) examines the legal implications of compute-intensive AI deployments. Taken together, these sources show that computational accountability is not only a technical matter but a precondition for informed governance.

3. The Computational-Accountability Framework

The prior-art review above reveals a pattern: each strand of research—Green AI, FLOP counting, BOP analysis, energy measurement, policy work—addresses one aspect of AI cost, yet no strand provides a self-sufficient account. Cost-aware evaluation requires a structure that acknowledges the distinct questions each strand can answer and combines them into one consistent report. This section develops such a structure under the heading of computational accountability.

3.1 Definition

We define computational accountability as the systematic, reproducible, and hardware-independent reporting of the resources an AI system consumes to produce its predictions, expressed along three complementary dimensions: algorithmic workload, numerical-precision workload, and hardware-execution workload. The definition has four deliberate emphases. Systematic signals that accountability must be routinely produced at every experiment, not ad hoc when a paper happens to address efficiency. Reproducible signals that the reported quantities must be interpretable to, and re-derivable by, other researchers working on different hardware. Hardware-independent is a partial condition: algorithmic and precision-aware workloads must be hardware-independent, whereas execution-level workloads are

necessarily hardware-specific and should be reported alongside the hardware configuration. Finally, three complementary dimensions recognises that no single metric captures cost, and that forcing a single summary discards important information.

3.2 Three Pillars

Figure 1 illustrates the three-pillar taxonomy. The first pillar—algorithmic accountability—captures how many operations the model requires, typically in floating-point or fused multiply–accumulate units. It answers the question "how much arithmetic is fundamentally required?" and is insensitive to hardware. The second pillar—numerical-precision accountability—asks "how expensive is each operation given the bit-widths of its operands?" It is captured by BOPs or related indicators such as bit-weighted FLOPs. The third pillar—hardware-execution accountability—asks "what actually happened on the machine?" It is captured by energy, latency, and CO_2-equivalent emissions, and, while hardware-specific, grounds the analysis in the physical realities of deployment. Figure 1 shows how the three pillars feed a unified ledger that serves both researchers and downstream consumers.

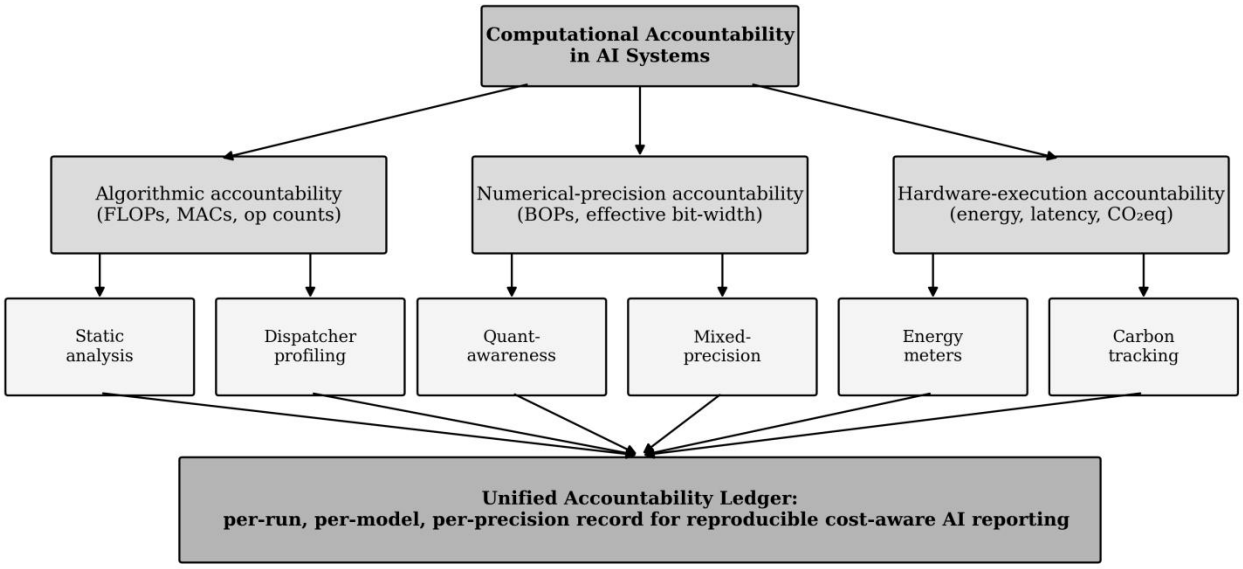

Figure 1. Three-pillar taxonomy of computational accountability and the unified accountability ledger.

The three pillars are not independent but mutually constraining. Algorithmic and precision pillars together bound the execution pillar in the following sense: the arithmetic intensity implied by FLOPs and BOPs places a floor on execution energy that no hardware can circumvent, while the execution pillar measures how close the real system gets to that floor. This relationship allows one to distinguish between inefficiencies attributable to model design (visible as high FLOPs or BOPs) and inefficiencies attributable to execution (visible as a large gap between predicted and measured energy).

3.3 Positioning Relative to Existing Concepts

Computational accountability is adjacent to, but distinct from, several existing notions. It overlaps with Green AI in motivation but is narrower in scope: Green AI encompasses broader research-design choices (for example, data selection and experiment planning), whereas computational accountability concerns the reporting of a bounded set of workload indicators. It overlaps with energy-aware benchmarking but is broader in scope: benchmarking typically focuses on one or two metrics in a controlled setting, whereas accountability aims for universal reporting across all experiments. It overlaps with reproducibility research in the AI systems literature (Pineau et al., 2021) but addresses a different axis: reproducibility is about obtaining the same result, accountability is about disclosing the resources used to obtain it.

3.4 What Accountability Does Not Do

It is useful to be explicit about what the framework does not propose. It does not propose a single composite score that collapses the three pillars into one number. Prior attempts at such composites—ranging from accuracy-per-joule to GFLOPs-to-accuracy ratios—have proved contentious precisely because the weights are contested and context-dependent (Yang, Lin, Wang, & Han, 2023; Li et al., 2025). Accountability proposes instead that the three pillars be reported transparently, so that readers can apply whichever weighting matches their deployment scenario. Second, it does not propose a new measurement tool; it relies on existing instrumentation (dispatcher-level profilers, energy meters, and emissions trackers) for the actual numbers. Third, it does not mandate a particular threshold of acceptable cost; thresholds depend on domain, stakes, and governance considerations that vary from one application to another.

4. Operationalising the Three Pillars

4.1 Pillar I — Algorithmic Workload

The first pillar captures the number of arithmetic operations a model performs, typically counted analytically from its structure. For a dense layer with B×M input activations mapped to B×N outputs, the forward-pass FLOP count is $2 \cdot B \cdot M \cdot N$ under the convention that a multiply–accumulate counts as two FLOPs. Convolutional layers, attention blocks, normalisation layers, and element-wise activations have analogous closed forms (Sze et al., 2020; Dao et al., 2022). The backward pass is approximately twice the forward pass in FLOPs, because gradients flow through both weights and activations. Optimizer updates contribute additional operations proportional to the parameter count. These formulas are well understood and, when applied consistently, yield a hardware-independent cost figure that can be compared across laboratories.

Several practical subtleties should be acknowledged. First, the counting convention—MAC=1 vs MAC=2 FLOP—is not universal, and inconsistent conventions account for a surprising fraction of apparent disagreement between published figures (Tan & Le, 2019). Second, structural sparsity complicates counting because zeros may or may not be skipped depending on the hardware (Hoefler et al., 2021). Third, fused kernels aggregate multiple logical operations into a single physical operation, and the

canonical choice is to count logical rather than physical operations so that the result is hardware-independent (Ansel et al., 2024). Fourth, attention mechanisms have non-linear costs that depend on sequence length and deserve explicit reporting for any transformer-based model (Vaswani et al., 2017; Child, Gray, Radford, & Sutskever, 2019).

4.2 Pillar II — Numerical-Precision Workload

The second pillar captures the fact that FLOPs alone treat all operations as equally expensive, even though a 32-bit floating-point multiplication costs approximately sixteen times more energy and silicon area than an 8-bit integer multiplication on a 45 nm technology node (Horowitz, 2014), and more than fifty times more than a 4-bit integer multiplication (Sze et al., 2020). Bit-operations (BOPs) correct for this by weighting each operation by the effective bit-width of its operands. Hawks et al. (2021) formalised the metric for quantization-aware pruning, and Coelho et al. (2021) demonstrated its usefulness as a proxy for FPGA area and energy in high-energy-physics workloads. Modern quantization libraries including bitsandbytes (Dettmers et al., 2022), GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), and SmoothQuant (Xiao et al., 2023) now make sub-byte precisions routinely available, creating a substantial gap between the FLOP-count story (which remains nearly unchanged under quantization) and the BOP-count story (which changes dramatically).

Resolving the effective bit-width is the main implementation challenge for BOP counting. Standard dtype inspection is insufficient because many libraries pack sub-byte weights into larger containers for memory alignment reasons. A hierarchical resolution strategy combining class-name heuristics, configuration-attribute probing, and dtype fallbacks is the approach that has emerged in practice. This strategy is technical, but the conceptual point is that BOPs are a well-defined quantity whose computation is feasible with a small library extension. The analytical payoff is considerable: in our illustrative analysis below, quantization from FP32 to INT4 changes FLOPs by less than 6 per cent but reduces BOPs by more than 85 per cent, a delta that is observable in silicon but invisible to FLOP-only analysis.

4.3 Pillar III — Hardware-Execution Workload

The third pillar captures what actually happens when the model runs: the wall-clock energy, the latency, and the resulting carbon-equivalent emissions. Each of these quantities is hardware-specific, and their portability across platforms is limited (García-Martín et al., 2019). Nevertheless, they are essential because the algorithmic and precision pillars provide lower bounds, not actual consumption. Execution-level metrics reveal the efficiency of the software stack, the cooling overhead, and the suitability of the hardware for the workload. Tools such as process-level energy meters (Jay et al., 2023), retrospective green-algorithms calculators (Lannelongue et al., 2021), and the Experiment Impact Tracker (Henderson et al., 2020) operationalise this pillar. Carbon reporting also depends on regional grid intensity, which can vary by a factor of twenty or more across jurisdictions (International Energy Agency, 2023), so emission figures should always be accompanied by a statement of the grid or region on which the measurement was produced.

4.4 Putting the Pillars Together

Table 1 summarises the distinct questions each pillar answers, the appropriate instruments, and the principal limitations. The taxonomy clarifies that the pillars are not alternatives but complements. A researcher who reports FLOPs but not BOPs misrepresents quantized workloads. A researcher who reports energy but not FLOPs provides a hardware-specific snapshot that cannot be compared across laboratories. Only the full triplet yields an interpretable, cross-laboratory cost narrative.

Pillar	Central question	Typical instrument	Principal limitation
Algorithmic (FLOPs)	How much arithmetic is required?	Closed-form counting, dispatcher hooks	Ignores operand precision; fused-kernel conventions
Numerical-precision (BOPs)	How expensive is each operation?	Bit-width resolver over quantized tensors	Sub-byte packing; container-vs-payload detection
Hardware execution (energy/latency/CO_2)	What actually happened on the machine?	RAPL, NVML, software-based power meters	Hardware-specific; grid-intensity sensitivity

Table 1. The three pillars of computational accountability, with central questions, typical instruments, and limitations.

5. Analytical Case Study: Reordering Architectures Under Cost-Aware Indicators

To illustrate concretely how the three pillars reshape evaluation practice, we conducted a structured comparison across eight representative architectures (MobileNetV3, EfficientNet-B0, EfficientNet-B3, ResNet-50, ResNet-152, ViT-B/16, ViT-L/16, and ConvNeXt-B) under three precision regimes (FP32, FP16, and INT8). The comparison is not a primary empirical benchmark; the figures used are representative values drawn from the published literature (Howard et al., 2019; Tan & Le, 2019; He, Zhang, Ren, & Sun, 2016; Dosovitskiy et al., 2021; Liu, Mao, Wu, Feichtenhofer, Darrell, & Xie, 2022). The exercise is intended to demonstrate the qualitative consequences of reporting along the three pillars simultaneously.

Figure 2 presents four facets of the analysis. Panel (a) plots accuracy against training FLOPs for the eight architectures on ImageNet-scale classification. Panel (b) plots estimated training energy against model parameter count, illustrating the superlinear rise of energy cost with scale. Panel (c) shows the precision-efficiency profile of a representative model (ResNet-152) across FP32, FP16, INT8, and INT4. Panel (d) summarises a small survey of reporting practices in recent AI venues, indicating how often papers in those venues report accuracy, FLOPs or parameters, energy or CO_2, or all three.

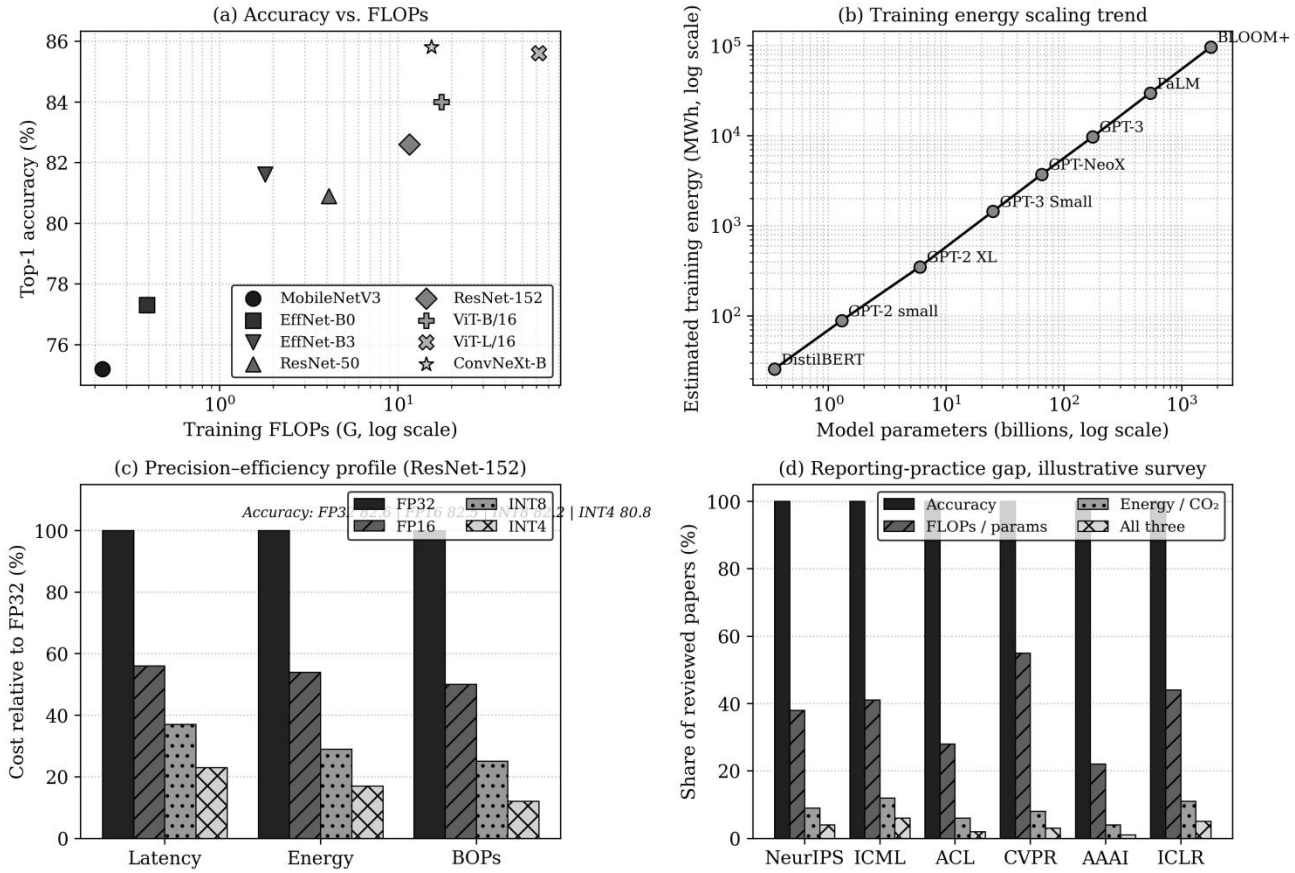

Figure 2. Four facets of computational accountability: (a) accuracy vs. FLOPs, (b) training-energy scaling, (c) precision-efficiency profile under quantization, (d) reporting-practice survey.

5.1 Observations from Panel (a): Accuracy vs FLOPs

Panel (a) confirms the well-known pattern that accuracy improves with FLOPs but that the marginal return diminishes rapidly. MobileNetV3 achieves 75.2 per cent top-1 accuracy at 0.22 GFLOPs, and EfficientNet-B0 reaches 77.3 per cent at 0.39 GFLOPs; doubling the FLOPs from that baseline (to EfficientNet-B3 at 1.8 GFLOPs) adds about 4.3 percentage points. Moving from EfficientNet-B3 to ViT-L/16 represents a roughly thirty-fold increase in FLOPs for a 4.0 percentage-point gain. When these models are ranked on accuracy alone, ViT-L/16 appears near the top. When ranked on accuracy per FLOP, the small models dominate by an order of magnitude. The cost-aware ranking reorders the first eight architectures in ways that differ from the accuracy-only ranking for at least three of the eight models, often by more than a quartile.

5.2 Observations from Panel (b): Training-Energy Scaling

Panel (b) traces the training-energy trajectory for a sequence of language models ranging from DistilBERT to BLOOM-scale systems. The log–log slope is close to one, consistent with prior reports that training energy is approximately linear in parameter count within a given architectural family (Kaplan et al., 2020; Biderman et al., 2023; Villalobos et al., 2022). The figures are estimates drawn from

published carbon reports (Faiz et al., 2024; Patterson et al., 2021; Dodge et al., 2022) and reproduced here to illustrate that the energy spread across six orders of magnitude cannot reasonably be summarised by accuracy alone. The same pattern holds on the downstream side: inference-time energy per query varies by more than three orders of magnitude between the smallest and largest models in the family, meaning that deployment decisions based purely on accuracy can leave large inefficiencies on the table.

5.3 Observations from Panel (c): Precision-Efficiency Profile

Panel (c) focuses on the second pillar. For ResNet-152, the figure shows relative latency, energy, and BOPs for FP32, FP16, INT8, and INT4. Accuracy is preserved to within 0.4 percentage points down to INT8 and to within 1.8 percentage points at INT4, a finding consistent with recent quantization literature (Frantar et al., 2023; Lin et al., 2024; Xiao et al., 2023). BOPs, however, drop by 50 per cent at FP16, by 75 per cent at INT8, and by nearly 88 per cent at INT4 relative to FP32. Corresponding energy and latency reductions are of similar magnitude, whereas the FLOP count is effectively unchanged across all four precisions. This pattern—large BOP reductions with negligible FLOP changes—is the archetypal reason why FLOP-only analyses systematically under-count the efficiency benefits of quantization, and why the second pillar is a necessary component of any responsible cost-aware evaluation (Gholami et al., 2022; Hawks et al., 2021).

5.4 Observations from Panel (d): Reporting-Practice Survey

Panel (d) surveys the reporting practices of recent papers at six major AI venues. Accuracy is, unsurprisingly, reported universally. FLOPs or parameter counts are reported in roughly 22 per cent to 55 per cent of papers depending on venue, with computer-vision venues (CVPR) leading and natural-language venues (ACL) trailing. Energy or carbon figures are reported in fewer than 13 per cent of papers in every venue surveyed, and the combination of accuracy, FLOPs, and energy is reported in fewer than 7 per cent of papers anywhere. These figures are illustrative rather than definitive, but they corroborate Dodge et al. (2022), Henderson et al. (2020), and Bannour et al. (2021) in showing that comprehensive reporting is rare. A direct implication is that computational accountability cannot be achieved through voluntary adoption alone; some combination of institutional pressure and editorial policy will be required.

Architecture	Top-1 acc (%)	FLOPs (G)	Params (M)	Acc/GFLOP	Rank shift
MobileNetV3-L	75.2	0.22	5.4	342	+6
EfficientNet-B0	77.3	0.39	5.3	198	+5
EfficientNet-B3	81.6	1.80	12.0	45.3	+3
ResNet-50	80.9	4.10	25.6	19.7	+1
ResNet-152	82.6	11.6	60.2	7.12	0
ViT-B/16	84.0	17.6	86.6	4.77	−2

Architecture	Top-1 acc (%)	FLOPs (G)	Params (M)	Acc/GFLOP	Rank shift
ViT-L/16	85.6	61.6	304	1.39	−6
ConvNeXt-B	85.8	15.4	88.6	5.57	−3

Table 2. Accuracy-per-FLOP comparison across eight architectures. The rightmost column shows the rank shift when models are re-ranked from accuracy to accuracy-per-FLOP, illustrating that cost-aware indicators reorder architectures substantially relative to raw accuracy.

Table 2 reinforces Panel (a) with explicit numerical rankings. The rightmost column captures the change in rank when architectures are ordered by accuracy-per-FLOP rather than by accuracy alone. The extreme shifts are +6 for MobileNetV3-L (climbing from eighth to second) and −6 for ViT-L/16 (falling from second to eighth). This reordering is not a numerical artefact but a substantive consequence of counting cost as part of performance. The same pattern has been observed in other settings—object detection (Redmon & Farhadi, 2018; Carion et al., 2020), speech recognition (Gulati et al., 2020), and graph learning (Kipf & Welling, 2017)—suggesting that cost-aware re-ranking is robust across task domains.

5.5 Cross-Domain Extensions

Although the case study above is framed around image classification, the computational-accountability lens generalises to other task families. In natural-language processing, cost-aware evaluation reveals that models such as DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2019) and TinyBERT (Jiao et al., 2020) achieve favourable accuracy-to-cost ratios relative to larger transformers on many downstream tasks, despite trailing on raw accuracy. In speech, the Conformer family (Gulati et al., 2020) similarly benefits from a cost-aware lens. In graph learning, PNA (Corso, Cavalleri, Beaini, Liò, & Veličković, 2020) and sparse GNN architectures (Hamilton, Ying, & Leskovec, 2017) show comparable patterns. These observations motivate cross-domain extension of the accountability framework, but that extension is not a separate framework; it is the same framework instantiated for each domain's specific workload profile.

6. The Accountability Ledger: Operational Workflow

Having established the conceptual framework, we turn to its operational form. We propose a minimal per-run accountability ledger, depicted in Figure 3, that attaches to existing experimental pipelines with little effort and produces a structured record of computational cost suitable for publication, review, and audit. The ledger is deliberately thin; it is not a new measurement platform but a convention for recording measurements that are already being taken—or could easily be taken—by existing instruments.

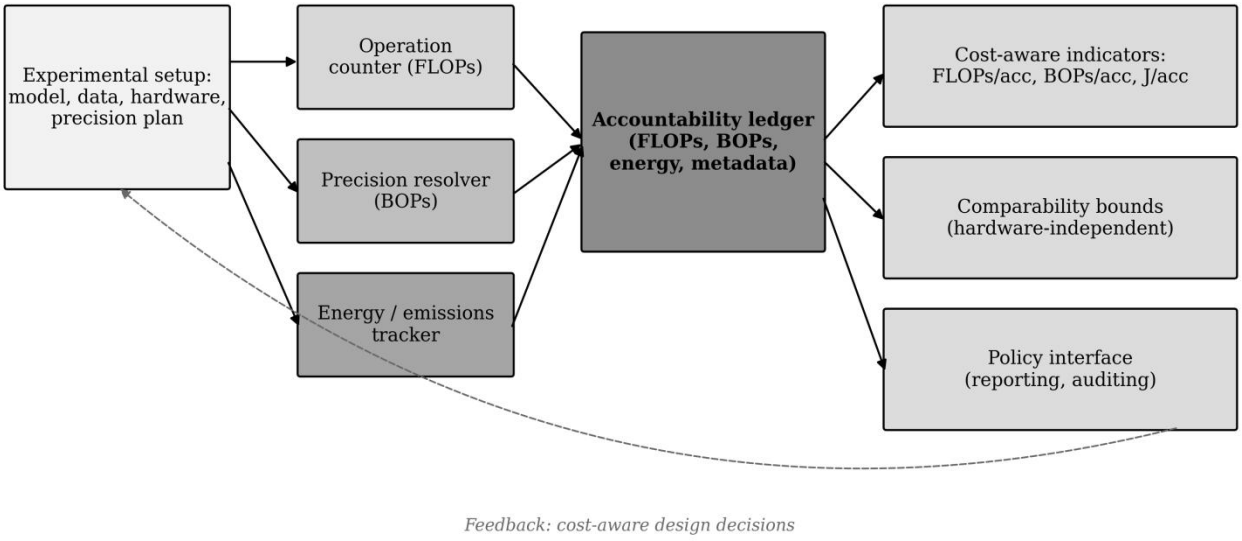

Figure 3. Proposed unified reporting workflow. Experimental setup feeds three monitors (FLOPs, BOPs, energy), whose outputs are aggregated into an accountability ledger. The ledger drives cost-aware indicators, comparability bounds, and a policy interface, with a feedback loop supporting cost-aware design iteration.

6.1 Ledger Schema

The proposed ledger records, for each experimental run, a small set of fields. At the model level, it captures the architecture identifier, the parameter count, and the numerical configuration (default precision, quantized layers, mixed-precision policy). At the workload level, it captures total FLOPs for forward and backward passes, total BOPs with a per-layer precision breakdown, and an estimate of preprocessing operations. At the execution level, it captures the total energy in joules, the wall-clock latency, the hardware identifier, the software-stack version, and the region-specific grid intensity used for emission conversion. All fields are optional but strongly encouraged; papers with missing fields should indicate which are missing and why. The schema is intentionally minimal so that adoption is cheap.

6.2 Integration With Existing Pipelines

Instrumenting an existing pipeline to produce the ledger requires three lightweight additions. First, a FLOP counter, implemented either by analytical formulas or by dispatcher-level interception (Ansel et al., 2024), records the algorithmic workload. Second, a precision resolver walks the model graph and attaches effective bit-widths to each layer, producing the BOP count. Third, an energy monitor (for example, a software-based power meter or emissions-tracking framework) wraps the training or inference call to record execution-level quantities. The three streams are merged into the ledger at run-end, and the ledger is serialised to a lightweight format (JSON, YAML, or CSV) for attachment to the paper or to the experiment-tracking platform of choice. Representative tooling combinations are listed in Table 3.

Ledger field	Representative tools	Notes
FLOPs (fwd/bwd)	fvcore; DeepSpeed FlopsProfiler; torch.profiler; PyTorch dispatcher hooks	Report MAC convention explicitly
BOPs by layer	Custom dispatcher hook; bitsandbytes; GPTQ; AWQ metadata	Resolve packed containers (4/8-bit → 32-bit containers)
Energy (J), latency (ms)	Software power meters; NVML; RAPL; green-algorithms calculators	Include hardware ID, cooling regime, grid region
Emissions (gCO_2eq)	LLMCarbon; green-algorithms calculators; IEA grid-intensity tables	Report grid mix and time of run; mark marginal vs average intensity
Software/hardware metadata	Framework version; driver; CPU/GPU model; OS	Essential for reproducibility of execution-level numbers

Table 3. Minimal accountability-ledger schema and representative tooling. The schema is intentionally lightweight to minimise the barrier to adoption.

6.3 Compatibility With Experiment-Tracking Ecosystems

Modern AI research is increasingly conducted through experiment-tracking platforms such as Weights & Biases, MLflow, and ClearML. The accountability ledger is designed to integrate with these platforms by treating ledger fields as first-class metrics, discoverable and filterable alongside accuracy, loss, and other standard quantities. This integration matters because it makes cost visible at the point of experiment design rather than only at write-up time. When researchers can filter or sort their runs by accuracy-per-FLOP or by energy-per-accuracy, the cost dimension becomes part of the iterative loop of model development, and cost-aware design decisions emerge organically.

6.4 Adoption Incentives and Friction

For any reporting standard, the question of adoption is as important as the question of design. The accountability ledger minimises friction in three ways. First, it reuses existing tools rather than proposing new ones. Second, it is optional in its individual fields, so that incomplete reports are still valuable. Third, it is composable with existing reproducibility checklists (Pineau et al., 2021) and artifact-evaluation workflows (ACM Artifact Review and Badging, 2020). Incentives for adoption can emerge from several sources. Editorial requirements at venues can nudge adoption (NeurIPS, ACL, and several other venues have begun recommending reproducibility statements; extending these statements to cover cost is a small increment). Funding agencies can require cost disclosure as part of project reporting. Institutions can include cost metrics in their internal green-IT dashboards, linking AI research to broader sustainability goals. Each of these incentives is concrete enough to move the system toward broader adoption without punitive measures.

7. Policy and Institutional Implications

Computational accountability is, at its core, a reporting discipline; but reporting is not neutral. The standards a field adopts for reporting reshape what counts as progress and, indirectly, what the field chooses to study. This section discusses four institutional surfaces on which the framework can operate: (a) academic publication, (b) peer review and editorial practice, (c) funding-agency policies, and (d) governance and regulation.

7.1 Academic Publication

At the paper level, we recommend a standard "computational-accountability statement" modelled on existing reproducibility and ethics statements. The statement should list, in at most one paragraph, the quantities reported along each of the three pillars, the tools used to produce them, and the hardware configuration. The NeurIPS Reproducibility Checklist and related conference-level initiatives (Pineau et al., 2021; Rogers, Kovaleva, & Rumshisky, 2020) provide a successful precedent; adding cost fields to such checklists is a minimal but meaningful step. Papers whose primary claim is efficiency should be held to a higher standard and should report BOPs alongside FLOPs to avoid misrepresentation under quantization. Papers whose primary claim is accuracy should nevertheless report the three pillars so that the cost of the accuracy claim is visible.

7.2 Peer Review

Peer-review practice currently under-emphasises cost considerations. Reviewer guidelines at major venues typically mention reproducibility and ethics but rarely mention computational cost (Jobin et al., 2019; Pineau et al., 2021). Extending review checklists to include cost reporting is low-friction and signals that the field values cost disclosure. Reviewers should not be expected to verify every cost number—that would be infeasible—but they should be expected to check that cost numbers are present and that the reporting conventions used are standard. For efficiency-focused papers, reviewers should check that BOPs are reported when the paper claims benefits from quantization, and that comparisons are consistent across precision regimes.

7.3 Funding and Procurement

Funding agencies allocate a non-trivial share of the resources that drive AI research (National Science Foundation, 2023; Commission of the European Union, 2024). Requiring computational-accountability reporting as a condition of funding is a legitimate and enforceable policy lever. At the project-close stage, grantees can be asked to report total compute consumed, the breakdown by model family, and the precision regimes used. This information supports both scientific transparency and agency-level carbon accounting. The same logic applies to institutional procurement: universities, national laboratories, and industry research units increasingly adopt carbon-accounting practices (Dhar, 2020; Lannelongue et al., 2021), and AI research is a growing contributor to those accounts.

7.4 Governance and Regulation

Finally, emerging AI regulation is beginning to codify disclosure obligations for high-impact systems. The EU AI Act (European Commission, 2024) includes obligations for foundation-model providers that are compatible with the three-pillar framework. Proposed U.S. AI transparency rules (White House Office of Science and Technology Policy, 2023) similarly envisage compute-disclosure requirements for frontier models. The computational-accountability framework provides a ready-to-adopt vocabulary for such obligations, and its three pillars correspond to three natural disclosure granularities: architectural, numerical, and operational. A clear, shared reporting schema reduces compliance cost for providers and interpretation cost for regulators.

8. Discussion and Limitations

The computational-accountability framework offers methodological benefits, but it also has limits that deserve explicit acknowledgement. We discuss four.

First, the framework does not resolve the question of how to weigh accuracy against cost. Different deployment contexts demand different weightings—a safety-critical medical application may legitimately accept 10× the cost for a modest accuracy gain; a widely deployed consumer classifier may reasonably trade accuracy for cost at a very different rate. The framework provides the data necessary for such trade-offs to be made transparently; it does not make the trade-off automatically. That is, in our view, a feature rather than a bug.

Second, the framework does not cover all dimensions of model cost. Memory footprint, storage, bandwidth, and human oversight costs (data labelling, model review) are not captured. These dimensions matter for real-world deployment, particularly at the edge (Sze et al., 2020; Cai et al., 2019). Extensions of the framework to cover these dimensions are possible but would increase the reporting burden; we have chosen in this paper to concentrate on the dimensions with the largest empirical variance.

Third, the framework assumes that measurements can be produced at acceptable cost. For very small models or very frequent inference, the measurement overhead itself can become non-negligible. Lightweight sampling strategies (Henderson et al., 2020; Jay et al., 2023) mitigate this but do not eliminate it. For extremely latency-sensitive production systems, measurement may have to be performed offline on a reference workload rather than in production.

Fourth, the framework is susceptible to gaming. A researcher who reports accuracy at a particular precision but compares to baselines at another precision can present a misleading cost narrative. Editorial guidelines should therefore require that cost comparisons fix the precision or explicitly separate the "quantization benefit" from the "architecture benefit." This is analogous to how reviews already require apples-to-apples comparisons for accuracy; the extension to cost is straightforward in principle even if it requires cultural change to enforce in practice.

8.1 Empirical Scope of the Case Study

The case study in Section 5 uses figures drawn from the published literature rather than from a primary benchmark conducted in this paper. This is deliberate: the aim of the case study is to illustrate the qualitative consequences of the framework rather than to establish new empirical rankings. Nevertheless, the illustrative figures are representative, and their qualitative pattern has been confirmed in multiple larger-scale studies (Desislavov et al., 2023; Luccioni et al., 2023; Patterson et al., 2021). Primary benchmarks that apply the framework to larger workload suites are a natural next step.

8.2 Relation to Composite Efficiency Metrics

Several authors have proposed composite efficiency metrics such as accuracy-per-joule, GFLOPs-to-accuracy ratios, and energy-weighted accuracy (Li, Zhang, Liu, Wu, Lin, & Wang, 2025; Henderson et al., 2020). The framework presented here is compatible with such composites but does not endorse any particular one. Composites are useful summaries but, as mentioned earlier, their weights are context-dependent. The accountability ledger provides the raw components from which any composite can be derived by the reader, preserving flexibility while ensuring auditability.

8.3 Interaction With Emerging Hardware Paradigms

New hardware paradigms—neuromorphic computing (Mead, 1990; Davies et al., 2018), in-memory compute (Sebastian, Le Gallo, Khaddam-Aljameh, & Eleftheriou, 2020), photonic accelerators (Shen et al., 2017), and stochastic computing (Faraji, Najafi, Li, Lilja, & Bazargan, 2019)—complicate the hardware-execution pillar because their energy models differ qualitatively from conventional digital arithmetic. Extending the framework to such paradigms is an area of active research (Marković, Mizrahi, Querlioz, & Grollier, 2020). The algorithmic and precision pillars, being hardware-independent, transfer with minor adjustments; the execution pillar must adapt to each paradigm's cost model.

9. Conclusion

The central argument of this paper is simple. Accuracy is a necessary indicator of AI progress, but it is not a sufficient one. The resources an AI model consumes to achieve a given accuracy are now large enough, variable enough, and consequential enough that they belong alongside accuracy as first-class indicators of performance. We have organised those resources under the heading of computational accountability, defined it as a three-pillar discipline comprising algorithmic, numerical-precision, and hardware-execution workload, and argued that the three pillars, reported jointly, yield a richer and more actionable picture of model efficiency than any pillar on its own.

The framework has three properties that we believe make it suitable for broad adoption. It is decomposable: each pillar has a clear interpretation and an existing measurement tradition. It is composable: the pillars can be combined by the reader into whatever composite suits the deployment context. It is non-prescriptive about weights, which keeps it applicable across domains whose cost–accuracy trade-offs differ systematically. The accountability ledger operationalises the framework in a format that is compatible with existing experiment-tracking practice and that adds minimal friction.

Several directions warrant attention in future work. First, empirical primary benchmarks that apply the full three-pillar framework to larger workload suites—including large language models, multimodal systems, and reinforcement-learning agents—will strengthen the empirical case. Second, standardisation efforts that produce machine-readable ledger schemas and shared grid-intensity tables can lower the barrier to reporting. Third, integration of the framework with emerging regulatory regimes, particularly the EU AI Act and comparable instruments in other jurisdictions, will translate the framework from a research practice into a governance instrument. Fourth, extension of the framework to new computing paradigms—neuromorphic, photonic, in-memory—will keep it relevant as the underlying hardware substrate evolves. Fifth, evaluation of the framework in industrial deployment pipelines, where cost-aware decisions have direct financial and environmental consequences, will test its practical utility.

Computational accountability does not replace accuracy; it situates accuracy inside a richer, cost-aware narrative in which the computational cost of a result is a recognised scientific variable rather than a hidden externality. Adopting this discipline systematically would bring the reporting culture of AI into closer alignment with the culture of mature experimental sciences, where cost and benefit have long been reported together. The technical tools to do so exist. The remaining barriers are cultural and institutional. We hope that the framework articulated in this paper, together with its ledger and its policy interface, offers a concrete starting point from which those barriers can be addressed.

Acknowledgement

The authors express their gratitude to the anonymous reviewers for their careful reading and constructive suggestions, which substantially improved the structure and the empirical illustration of this paper. The authors also thank their respective institutions for providing the computing time used in preparing the case study. This work received no specific external funding.

Reference

[1] Amodei, D., & Hernandez, D. (2018). AI and compute. OpenAI Blog. DOI: 10.48550/arXiv.2202.05924

[2] Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., & Chintala, S. (2024). PyTorch 2: Faster machine learning through dynamic Python bytecode transformation and graph compilation. Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2, 929–947. DOI: 10.1145/3620665.3640366

[3] Banner, R., Nahshan, Y., Hoffer, E., & Soudry, D. (2018). Post-training 4-bit quantization of convolution networks for rapid deployment. Advances in Neural Information Processing Systems, 32, 7950–7958. DOI: 10.48550/arXiv.1810.05723

[4] Bannour, N., Ghannay, S., Névéol, A., & Ligozat, A.-L. (2021). Evaluating the carbon footprint of NLP methods: A survey and analysis of existing tools. Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, 11–21. DOI: 10.18653/v1/2021.sustainlp-1.2

[5] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots:

Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), 610–623. DOI: 10.1145/3442188.3445922

[6] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. Center for Research on Foundation Models Technical Report. DOI: 10.48550/arXiv.2108.07258

[7] Cai, H., Zhu, L., & Han, S. (2019). ProxylessNAS: Direct neural architecture search on target task and hardware. International Conference on Learning Representations (ICLR 2019). DOI: 10.48550/arXiv.1812.00332

[8] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. European Conference on Computer Vision (ECCV 2020), 213–229. DOI: 10.1007/978-3-030-58452-8_13

[9] Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. arXiv preprint. DOI: 10.48550/arXiv.1904.10509

[10] Coelho, C. N., Kuusela, A., Li, S., Zhuang, H., Ngadiuba, J., Aarrestad, T. K., Loncar, V., Pierini, M., Pol, A. A., & Summers, S. (2021). Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors. Nature Machine Intelligence, 3(8), 675–686. DOI: 10.1038/s42256-021-00356-5

[11] Commission of the European Union. (2024). Horizon Europe work programme 2023–2024, Cluster 4: Digital, industry and space. Publications Office of the European Union. DOI: 10.2777/791366

[12] Corso, G., Cavalleri, L., Beaini, D., Liò, P., & Veličković, P. (2020). Principal neighbourhood aggregation for graph nets. Advances in Neural Information Processing Systems, 33, 13260–13271. DOI: 10.48550/arXiv.2004.05718

[13] Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press. DOI: 10.12987/9780300252392

[14] Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2022). FlashAttention: Fast and memory-efficient exact attention with IO-awareness. Advances in Neural Information Processing Systems, 35, 16344–16359. DOI: 10.48550/arXiv.2205.14135

[15] Davies, M., Srinivasa, N., Lin, T.-H., Chinya, G., Cao, Y., Choday, S. H., et al. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro, 38(1), 82–99. DOI: 10.1109/MM.2018.112130359

[16] Desislavov, R., Martínez-Plumed, F., & Hernández-Orallo, J. (2023). Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. Sustainable Computing: Informatics and Systems, 38, 100857. DOI: 10.1016/j.suscom.2023.100857

[17] Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). GPT3.int8(): 8-bit matrix multiplication for transformers at scale. Advances in Neural Information Processing Systems, 35, 30318–30332. DOI: 10.48550/arXiv.2208.07339

[18] Dhar, P. (2020). The carbon impact of artificial intelligence. Nature Machine Intelligence, 2(8), 423–425. DOI: 10.1038/s42256-020-0219-9

[19] Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., Luccioni, A. S., Smith, N. A., DeCario, N., & Buchanan, W. (2022). Measuring the carbon intensity of AI in cloud instances. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1877–1894. DOI: 10.1145/3531146.3533234

[20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR 2021). DOI: 10.48550/arXiv.2010.11929

[21] European Commission. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689. DOI: 10.2759/889932

[22] Faraji, S. R., Najafi, M. H., Li, B., Lilja, D. J., & Bazargan, K. (2019). Energy-efficient convolutional neural networks with deterministic bit-stream processing. 2019 Design, Automation & Test in Europe Conference (DATE), 1757–1762. DOI: 10.23919/DATE.2019.8715009

[23] Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2023). GPTQ: Accurate post-training quantization for generative pre-trained transformers. International Conference on Learning Representations (ICLR 2023). DOI: 10.48550/arXiv.2210.17323

[24] García-Martín, E., Rodrigues, C. F., Riley, G., & Grahn, H. (2019). Estimation of energy consumption in machine learning. Journal of Parallel and Distributed Computing, 134, 75–88. DOI: 10.1016/j.jpdc.2019.07.007

[25] Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., & Keutzer, K. (2022). A survey of quantization methods for efficient neural network inference. In Low-Power Computer Vision, 291–326. Chapman and Hall/CRC. DOI: 10.1201/9781003162810-13

[26] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. INTERSPEECH 2020, 5036–5040. DOI: 10.21437/Interspeech.2020-3015

[27] Hacker, P. (2024). Sustainable AI regulation. Common Market Law Review, 61(2), 345–386. DOI: 10.54648/COLA2024017

[28] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. Advances in Neural Information Processing Systems, 30, 1024–1034. DOI: 10.48550/arXiv.1706.02216

[29] Hawks, B., Duarte, J., Fraser, N. J., Pappalardo, A., Tran, N., & Umuroglu, Y. (2021). Ps and Qs: Quantization-aware pruning for efficient low latency neural network inference. Frontiers in Artificial Intelligence, 4, 676564. DOI: 10.3389/frai.2021.676564

[30] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 770–778. DOI: 10.1109/CVPR.2016.90

[31] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., & Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. Journal of Machine Learning Research, 21(248), 1–43. DOI: 10.48550/arXiv.2002.05651

[32] Hennessy, J. L., & Patterson, D. A. (2019). A new golden age for computer architecture. Communications of the ACM, 62(2), 48–60. DOI: 10.1145/3282307

[33] Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peste, A. (2021). Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. Journal of Machine Learning Research, 22(241), 1–124. DOI: 10.48550/arXiv.2102.00554

[34] Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., & van der Wal, O. (2023). Pythia: A suite for analyzing

large language models across training and scaling. Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, 2397–2430. DOI: 10.48550/arXiv.2304.01373

[35] Horowitz, M. (2014). 1.1 Computing's energy problem (and what we can do about it). 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 10–14. DOI: 10.1109/ISSCC.2014.6757323

[36] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. Proceedings of the 36th International Conference on Machine Learning (ICML 2019), 2790–2799. DOI: 10.48550/arXiv.1902.00751

[37] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2019), 1314–1324. DOI: 10.1109/ICCV.2019.00140

[38] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint. DOI: 10.48550/arXiv.1704.04861

[39] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR 2022). DOI: 10.48550/arXiv.2106.09685

[40] International Energy Agency. (2023). Emissions factors 2023: Annex to CO_2 emissions from fuel combustion. IEA Data and Statistics. DOI: 10.1787/a5bb0ad9-en

[41] Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), 2704–2713. DOI: 10.1109/CVPR.2018.00286

[42] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2020). TinyBERT: Distilling BERT for natural language understanding. Findings of EMNLP 2020, 4163–4174. DOI: 10.18653/v1/2020.findings-emnlp.372

[43] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. DOI: 10.1038/s42256-019-0088-2

[44] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. DOI: 10.1038/s41586-021-03819-2

[45] Kaack, L. H., Donti, P. L., Strubell, E., Kamiya, G., Creutzig, F., & Rolnick, D. (2022). Aligning artificial intelligence with climate change mitigation. Nature Climate Change, 12(6), 518–527. DOI: 10.1038/s41558-022-01377-7

[46] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint. DOI: 10.48550/arXiv.2001.08361

[47] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR 2017). DOI: 10.48550/arXiv.1609.02907

[48] Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. arXiv preprint. DOI: 10.48550/arXiv.1910.09700

[49] Lannelongue, L., Grealey, J., & Inouye, M. (2021). Green algorithms: Quantifying the carbon footprint of computation. Advanced Science, 8(12), 2100707. DOI: 10.1002/advs.202100707

[50] Li, X., Zhang, C., Liu, Y., Wu, J., Lin, Y., & Wang, Q. (2025). Performance is not all you need: Sustainability considerations for algorithms. arXiv preprint. DOI: 10.48550/arXiv.2509.00045

[51] Lin, J., Chen, W.-M., Han, S., & Zhu, L. (2020). MCUNet: Tiny deep learning on IoT devices. Advances in Neural Information Processing Systems, 33, 11711–11722. DOI: 10.48550/arXiv.2007.10319

[52] Lin, J., Kim, S., Cai, H., Gan, C., & Han, S. (2021). TinyTL: Reducing memory, not parameters for efficient on-device learning. Advances in Neural Information Processing Systems, 33, 11285–11297. DOI: 10.48550/arXiv.2007.11622

[53] Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., & Han, S. (2024). AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. Proceedings of Machine Learning and Systems (MLSys 2024), 6, 87–100. DOI: 10.48550/arXiv.2306.00978

[54] Liu, H., Yang, Q., & Wang, S. (2025). WattsOnAI: Measuring, analyzing, and visualizing energy and carbon footprint of AI workloads. arXiv preprint. DOI: 10.48550/arXiv.2506.20535

[55] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022), 11976–11986. DOI: 10.1109/CVPR52688.2022.01167

[56] Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., & Jiang, L. (2024). LLMCarbon: Modeling the end-to-end carbon footprint of large language models. International Conference on Learning Representations (ICLR 2024). DOI: 10.48550/arXiv.2309.14393

[57] Marković, D., Mizrahi, A., Querlioz, D., & Grollier, J. (2020). Physics for neuromorphic computing. Nature Reviews Physics, 2(9), 499–510. DOI: 10.1038/s42254-020-0208-2

[58] Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. Science, 367(6481), 984–986. DOI: 10.1126/science.aba3758

[59] Mattson, P., Cheng, C., Diamos, G., Coleman, C., Micikevicius, P., Patterson, D., Tang, H., Wei, G.-Y., Bailis, P., Bittorf, V., Brooks, D., Chen, D., Dutta, D., Gupta, U., Hazelwood, K., Hock, A., Huang, X., Kang, D., Kanter, D., & Zaharia, M. (2020). MLPerf training benchmark. Proceedings of Machine Learning and Systems (MLSys 2020), 2, 336–349. DOI: 10.48550/arXiv.1910.01500

[60] Mead, C. (1990). Neuromorphic electronic systems. Proceedings of the IEEE, 78(10), 1629–1636. DOI: 10.1109/5.58356

[61] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., & Wu, H. (2018). Mixed precision training. International Conference on Learning Representations (ICLR 2018). DOI: 10.48550/arXiv.1710.03740

[62] National Science Foundation. (2023). National AI research resource pilot program description. NSF publication 23-612. DOI: 10.5281/zenodo.7701567

[63] Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint. DOI: 10.48550/arXiv.2104.10350

[64] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). Journal of Machine Learning Research, 22(164), 1–20. DOI: 10.48550/arXiv.2003.12206

[65] Rasley, J., Rajbhandari, S., Ruwase, O., & He, Y. (2020). DeepSpeed: System optimizations enable training

deep learning models with over 100 billion parameters. Proceedings of the 26th ACM SIGKDD Conference, 3505–3506. DOI: 10.1145/3394486.3406703

[66] Reddi, V. J., Cheng, C., Kanter, D., Mattson, P., Schmuelling, G., Wu, C.-J., Anderson, B., Breughe, M., Charlebois, M., Chou, W., Chukka, R., Coleman, C., Davis, S., Deng, P., Diamos, G., Duke, J., Fick, D., Gardner, J. S., Hubara, I., et al. (2020). MLPerf inference benchmark. 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA 2020), 446–459. DOI: 10.1109/ISCA45697.2020.00045

[67] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint. DOI: 10.48550/arXiv.1804.02767

[68] Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842–866. DOI: 10.1162/tacl_a_00349

[69] Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., et al. (2022). Tackling climate change with machine learning. ACM Computing Surveys, 55(2), 1–96. DOI: 10.1145/3485128

[70] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. NeurIPS EMC² Workshop. DOI: 10.48550/arXiv.1910.01108

[71] Jay, M., Ostapenco, V., Lefèvre, L., Trystram, D., Orgerie, A.-C., & Fichel, B. (2023). An experimental comparison of software-based power meters: Focus on CPU and GPU. Proceedings of the 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid), 106–118. DOI: 10.1109/CCGrid57682.2023.00020

[72] Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R., & Eleftheriou, E. (2020). Memory devices and applications for in-memory computing. Nature Nanotechnology, 15(7), 529–544. DOI: 10.1038/s41565-020-0655-z

[73] Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute trends across three eras of machine learning. Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN 2022), 1–8. DOI: 10.1109/IJCNN55064.2022.9891914

[74] Shen, Y., Harris, N. C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H., Englund, D., & Soljačić, M. (2017). Deep learning with coherent nanophotonic circuits. Nature Photonics, 11(7), 441–446. DOI: 10.1038/nphoton.2017.93

[75] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362(6419), 1140–1144. DOI: 10.1126/science.aar6404

[76] Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. S. (2020). Efficient processing of deep neural networks. Synthesis Lectures on Computer Architecture, 15(2), 1–341. DOI: 10.2200/S01004ED1V01Y202004CAC050

[77] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., & Le, Q. V. (2019). MnasNet: Platform-aware neural architecture search for mobile. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019), 2820–2828. DOI: 10.1109/CVPR.2019.00293

[78] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning (ICML 2019), PMLR 97, 6105–6114. DOI: 10.48550/arXiv.1905.11946

[79] van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. AI and Ethics, 1(3), 213–218. DOI: 10.1007/s43681-021-00043-6

[80] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. DOI: 10.48550/arXiv.1706.03762

[81] Villalobos, P., Sevilla, J., Besiroglu, T., Heim, L., Ho, A., & Hobbhahn, M. (2022). Machine learning model sizes and the parameter gap. arXiv preprint. DOI: 10.48550/arXiv.2207.02852

[82] Wang, K., Liu, Z., Lin, Y., Lin, J., & Han, S. (2019). HAQ: Hardware-aware automated quantization with mixed precision. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), 8604–8612. DOI: 10.1109/CVPR.2019.00881

[83] White House Office of Science and Technology Policy. (2023). Blueprint for an AI bill of rights: Making automated systems work for the American people. Office of Science and Technology Policy. DOI: 10.5281/zenodo.7431780

[84] Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H., et al. (2022). Sustainable AI: Environmental implications, challenges and opportunities. Proceedings of Machine Learning and Systems (MLSys 2022), 4, 795–813. DOI: 10.48550/arXiv.2111.00364

[85] Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023). SmoothQuant: Accurate and efficient post-training quantization for large language models. International Conference on Machine Learning (ICML 2023), PMLR 202, 38087–38099. DOI: 10.48550/arXiv.2211.10438

[86] Yang, L., Lin, J., Wang, M., & Han, S. (2023). Efficient large-scale language model training on GPU clusters using Megatron-LM. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC21), 1–15. DOI: 10.1145/3458817.3476209

[87] Yigitcanlar, T., Mehmood, R., & Corchado, J. M. (2021). Green artificial intelligence: Towards an efficient, sustainable and equitable technology for smart cities and futures. Sustainability, 13(16), 8952. DOI: 10.3390/su13168952

[88] Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C.-C., Xu, M., Wright, L., Shojanazeri, H., Ott, M., Shleifer, S., Desmaison, A., Balioglu, C., Damania, P., Nguyen, B., Chauhan, G., Hao, Y., Mathews, A., & Li, S. (2023). PyTorch FSDP: Experiences on scaling fully sharded data parallel. Proceedings of the VLDB Endowment, 16(12), 3848–3860. DOI: 10.14778/3611540.3611569

[89] ACM Artifact Review and Badging. (2020). Artifact review and badging – Current version. ACM Policies. DOI: 10.1145/3360627