

**AI Inference Energy Consumption: Empirical
Quantification of the Correlation Between Tokens
Processed and Power Draw**

by

Yersin Kazybekov

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

at the

NAZARBAYEV UNIVERSITY

April 2026

© Nazarbayev University 2026. All rights reserved.

Author
Department of Computer Science
11 April 2026

Certified by
Askar Boranbayev
Professor
Thesis Supervisor

Certified by
Talgat Bayan
Postdoctoral Scholar
Thesis Supervisor

Accepted by
Yelyzaveta Arkhangelsky
Dean, School of Engineering and Digital Sciences

AI Inference Energy Consumption: Empirical Quantification of the Correlation Between Tokens Processed and Power Draw

by

Yersin Kazybekov

Submitted to the Department of Computer Science
on 11 April 2026, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science

Abstract

Inference Energy Consumption is Poorly Understood: A major challenge for Green AI today is the absence of a well-defined and standardised methodology for measuring and reporting the energy consumed by Large Language Models (LLMs) during inference. Current software-based estimation tools carry error margins of approximately 40 percent, rendering them inadequate for the environmental accountability frameworks now being mandated by regulators such as the European Union’s Artificial Intelligence Act, under which General-Purpose AI model providers have been required to document energy consumption since 2 August 2025.

This dissertation proposes *Energy-per-Token* as a standardised, quantitative metric for LLM inference efficiency, derived from direct hardware measurement rather than software estimation. A high-fidelity dataset was constructed by instrumenting an NVIDIA GeForce RTX 4060 GPU with on-chip NVML sensors sampled at 10 Hz while running four transformer architectures - GPT-2 (124M), Phi-2 (2.7B), Qwen 1.5 ($\sim 1.8B$), and LLaMA 3 (8B, 4-bit quantised) - under controlled inference conditions. A parallel testbed was constructed on an Apple M4 Pro system-on-chip to assess the generalisability of the framework to a qualitatively different hardware architecture, specifically one employing unified memory in place of a discrete GPU memory bus. Power telemetry on the M4 Pro was obtained via Apple’s **powermetrics** utility; these measurements are treated throughout as high-fidelity software-assisted estimates rather than ground-truth sensor readings, a distinction discussed explicitly in the methodology.

A multivariate regression model was fitted to the resulting dataset to quantify the relationship between tokens processed and energy consumed, decomposed into contributions from the prefill phase, the decode phase, and model architecture. The primary empirical finding of the thesis is that, across all four architectures tested on the RTX 4060, the per-token energy coefficient of the decode phase is between 7 and 9 times greater than that of the prefill phase, indicating that autoregressive token generation consumes nearly one order of magnitude more energy per token than does prompt processing. This decode-penalty ratio is replicated within the same $7\times$ –

9× band on the Apple M4 Pro, providing suggestive evidence that the penalty is a property of the autoregressive generation mechanism itself rather than an artefact of any specific memory subsystem design.

Three secondary findings are reported. First, parameter count is shown to be a poor proxy for inference energy: Phi-2 (2.7B) achieves a lower absolute decode coefficient than the INT4-quantised LLaMA 3 8B on both platforms, suggesting that architectural design choices and quantisation regime interact in ways that cannot be inferred from model scale alone. Second, based on the measured per-token coefficients combined with published tokenizer compression ratios, a task-level energy reduction of approximately 15–22 percent is projected for modern large-vocabulary tokenizers (Qwen 1.5) relative to legacy tokenizers (GPT-2) on equivalent semantic content; a direct task-based measurement of this effect is identified as a priority for future work. Third, the absolute energy coefficients on the M4 Pro are approximately 78–80 percent lower than those on the RTX 4060, reflecting the M4 Pro’s substantially lower thermal envelope.

The results of this dissertation provide Green AI researchers and practitioners with a hardware-grounded, statistically validated methodology for forecasting and comparing the environmental footprint of LLM inference prior to production deployment, and establish the decode-penalty ratio as a working architectural principle for autoregressive transformer inference.

Keywords: Green AI; Large Language Models (LLMs); Energy-per-Token; GPU Power Profiling; Autoregressive Inference; Decode Penalty; Cross-Platform Measurement; Carbon Footprint.

Thesis Supervisor: Askar Boranbayev

Title: Professor

Thesis Supervisor: Talgar Bayan

Title: Postdoctoral Scholar

Contents

1	Introduction	9
2	Research Questions	13
2.1	Core Research Questions	13
2.2	Secondary Research Questions	14
3	Related Works	17
3.1	The AI Efficiency Paradox	17
3.2	The Era of “Red AI”: Quantifying the Unchecked Cost of Scale	18
3.3	The “Green AI” Imperative: A New Paradigm for Research	19
3.4	Aims, Scope, and Contribution	20
3.5	Lifecycle Assessment and the Compute Carbon Intensity Metric . . .	20
3.6	Architectural Duality: Sparse Mixture-of-Experts vs. Dense Scaling .	22
3.7	Regulatory Governance and the Transparency Mandate	23
3.8	The Token as the Key Unit of Inference Cost	24
3.9	Complicating Factors and The Research Gap	25
3.10	Conclusion and Thesis Contribution	25
4	Measurement and Mitigation Strategies	27
4.1	The Operational Lifecycle of Inference Engines	27
4.2	Compression and the Energy-Accuracy Trade-off	28
4.3	The Role of AI in Energy Systems and Sustainable Development . . .	29
4.4	Benchmarking for Production Reality: TokenPowerBench and Beyond	29

5	Methodology	31
5.1	Principles-Based Measurement Strategy	31
5.2	Instrumentation and Testbed Architecture	31
5.3	Apple M4 Pro Testbed: Instrumentation and Comparative Measurement	32
5.4	Multi-Variable Experimental Design	36
5.5	Analytical Framework and Modeling	36
5.6	Cross-Platform Normalization and Embodied Carbon Integration . .	37
6	Results and Statistical Findings	39
6.1	Hardware Testbed Performance	39
6.2	Power Profiles: Prefill vs. Decode	40
6.3	Statistical Regression and Coefficient Interpretation	42
6.4	Cross-Platform Coefficient Comparison	45
6.5	Task-Level Energy and Throughput Analysis	45
6.6	Impact of Modern Tokenization: Qwen 1.5 vs. GPT-2	46
7	Conclusion	49
7.1	Answers to the Core Research Questions	50
7.2	Answers to the Secondary Research Questions	52
7.3	Broader Significance and Limitations	55
7.4	Final Remarks	56

List of Tables

3.1	Hardware Specifications and Carbon Intensity (2020–2024)	21
3.2	Efficiency Comparison: Dense vs. Sparse Architectures	22
4.1	Operational Energy Metrics across Inference Engines (8B Model) . .	28
4.2	Quantization Scaling and Efficiency Gains	28
4.3	AI Applications in Energy Sector Optimization	29
6.1	Comparative Power Metrics: RTX 4060 vs. Apple M4 Pro (10 Hz Sampling)	42
6.2	Derived Energy Coefficients (Joules/token) with Full Statistical Reporting - RTX 4060.	43
6.3	Derived Energy Coefficients (Joules/token) with Full Statistical Reporting - Apple M4 Pro.	43
6.4	Cross-Platform Energy Coefficient Comparison: RTX 4060 vs. M4 Pro	45
6.5	Task-Level Energy (Joules) for 128-token Prefill + 256-token Decode	46

Chapter 1

Introduction

The rapid development of large language models (LLM) has moved artificial intelligence from a research curiosity to a key element of digital infrastructure globally. However, the rapid growth of large language models has come with a price which the field has been slow to recognize. Modern generative AI systems consume so much more computer power than older systems that the energy they use is no longer a small fraction of total usage. A single large-scale NLP model uses enough electricity during its creation to produce the same amount of greenhouse gas emissions as would be produced over the lifetime of five cars [1], and the collective energy consumed in running such a model will exceed the one-time energy required to create the model by more than three orders of magnitude over the model’s lifespan [22]. These are not theoretical projections they are real numbers being tracked by regulators around the world and have already attracted the attention of regulatory bodies, including the EU’s AI Act requiring all general-purpose AI models to document their energy usage beginning in August 2025 [15].

Although there is a growing sense of urgency surrounding these issues, there does not currently exist sufficient measurement capabilities to enable effective responses to the growing need to account for the environmental impact of energy consumption related to AI and machine learning.

While software based energy estimators (the most common means of estimating the energy requirements for inference) can estimate the energy used for inference

within a range of about 40% of the actual value [4], this margin of error is too high for the purpose of providing rigorous accountability for environmental impacts. More importantly, although the research community has generally agreed upon the token as the best available unit of inference cost for measuring the energy usage of generative models [13], there currently does not exist a reliable, publicly verifiable model of the relationship between tokens processed and energy usage that is accurate and reliable across various model architectures.

The primary goal of this dissertation is to fill this gap. Through the direct measurement of energy usage on a local GPU testbed via the use of on-chip sensors at a sample rate of 10Hz, a dataset was created which provides the necessary information to develop a multivariable regression model that measures the energy costs associated with inference as a function of the number of tokens processed, the specific computational phase of the inference process, the specific architecture of the model, and the level of optimization applied to the model. The resulting multivariable regression model produces a practical, reproducible “Energy per Token” metric, or standardized efficiency metric that can be used by researchers and practitioners of Green AI [3] to predictively determine the environmental costs of a generative task prior to executing that task, and therefore enables the first step toward developing a complete system of accountability for environmental sustainability in the use of energy by generative AI systems. In addition to providing the foundational elements of a system of accountability, the Energy per Token metric also serves as a common framework for evaluating the efficiency of different components of complex workflows, which together form the basis of the more complex, multi-step AI systems that now dominate real-world deployments. The cross-platform carbon normalization layer, which is based on the Compute Carbon Intensity metric [11], allows for the translation of the local measurements made on an NVIDIA GeForce RTX 4060 GPU to provide estimates of the carbon usage of those measurements that are applicable to other types of hardware (i.e., datacentre hardware) and to other types of carbon usage metrics (i.e., regional grid intensities), thereby increasing the practical scope of the model.

The broader significance of this work lies in its timing. The European Union’s

AI Act entered force in 2025, mandating energy documentation for general-purpose AI models [15], and the inference energy burden of deployed models now exceeds their one-time training cost by more than three orders of magnitude over a model’s operational lifetime [22]. In this context, the Energy-per-Token metric introduced here provides exactly the kind of standardised, hardware-grounded efficiency measure that regulators, developers, and researchers require to make informed, accountable decisions [13]. This thesis does not resolve the AI energy crisis, but it provides a concrete instrument for those working to do so a predictive model that transforms an abstract environmental cost into a quantity that can be measured, compared, and ultimately reduced.

Chapter 2

Research Questions

This chapter formally states the research questions that structure the empirical and analytical contributions of this thesis. The questions are organised into three tiers: two core questions that constitute the primary scientific contribution; four secondary questions that extend and validate the core findings across variables; and three extended questions that broaden the framework to more complex, real-world deployment scenarios.

2.1 Core Research Questions

The two core research questions are logically sequential. The first establishes the methodological validity of the proposed measurement and modelling approach; the second reports the primary empirical finding that the validated model produces.

RQ1a *Can the energy consumed during LLM inference be reliably modelled as a linear function of tokens processed, using direct hardware measurement rather than software-based estimation?*

This question addresses the methodological foundation of the thesis. It asks whether a multivariate regression of the form $E_{total} = \beta_p \cdot T_{prefill} + \beta_d \cdot T_{decode} + \gamma(\text{arch}, \text{hardware}) + \epsilon$ constitutes a statistically valid and reproducible description of inference energy across diverse model architectures, and whether

hardware-based measurement via on-chip NVML sensors at 10 Hz is necessary to achieve this given that software-based estimators carry an approximate error margin of 40 percent [4].

RQ1b *What is the magnitude of the energy asymmetry between the prefill and decode phases of inference, and is this asymmetry consistent across different model architectures, quantization levels, and hardware platforms?*

Once the validity of the modelling framework has been established (RQ1a), this question asks what the validated model reveals. It investigates whether the decode energy penalty the ratio β_d/β_p remains stable across model architecture, quantization level, and hardware platform. If the ratio is found to be invariant, it constitutes evidence that the decode penalty is a general architectural property of autoregressive transformer inference, rather than a hardware-specific or model-specific observation.

2.2 Secondary Research Questions

The four secondary research questions each isolate one variable within the same experimental framework established to answer RQ1. Together, they characterise how specific design choices in model architecture, tokenizer design, quantization strategy, and hardware selection modulate the energy-per-token coefficients derived in the core analysis.

RQ2 *How does model architecture affect the energy-per-token coefficient, independent of raw parameter count?*

Prior work has demonstrated that models of differing parameter counts can exhibit equivalent energy efficiency owing to differences in layer design rather than scale alone [20]. This question investigates whether parameter count is a reliable proxy for inference energy, or whether architectural choices are more predictive of the derived coefficients β_p and β_d .

RQ3 *Does tokenizer compression ratio produce a measurable reduction in total inference energy at the task level, independent of the model’s parameter count?*

Because each output token requires a separate forward pass [12], a tokenizer that represents the same semantic content in fewer tokens directly reduces the number of decode steps required. This question investigates whether the vocabulary efficiency advantage of modern tokenizers - such as that of Qwen 1.5, which achieves 22–29 percent better compression than the GPT-2 tokenizer - produces a proportionate reduction in task-level energy consumption E_{total} .

RQ4 *What is the measurable energy impact of model quantization, and does this efficiency gain differ proportionately between small and large language models?*

Quantization reduces the numerical precision of model weights, decreasing memory footprint and potentially reducing energy consumption during inference [14]. However, in memory-bandwidth-bound workloads characteristic of the decode phase, the relationship between arithmetic precision and power draw is mediated by memory bus utilisation rather than floating-point throughput alone. This question investigates whether smaller language models realise proportionately greater or lesser efficiency gains from quantization relative to larger models.

RQ5 *Is the decode penalty ratio (β_d/β_p) a platform-invariant property of autoregressive transformer inference, or is it specific to discrete GPU architectures with dedicated video memory?*

The RTX 4060 testbed employs discrete GDDR6 memory with a dedicated GPU memory bus. The Apple M4 Pro SoC employs a unified memory architecture in which the CPU, GPU, and Apple Neural Engine share a single LPDDR5X memory pool. If the decode penalty ratio is preserved across these two architecturally dissimilar platforms, this constitutes strong evidence that the penalty is a fundamental consequence of the autoregressive generation mechanism rather than an artefact of any particular memory subsystem design [23].

Chapter 3

Related Works

3.1 The AI Efficiency Paradox

In essence, the field of artificial intelligence (AI) is comprised of two opposing forces: a desire to create increasingly capable models while simultaneously generating increasing amounts of financial and environmental costs. Although recent developments in NLP have utilized "new generations of large networks" [3], and the resultant models (i.e., SOTA models, i.e. GPT, BERT, etc.) are being produced with a focus on scaling capabilities; there exists no clear limit to these scales due to the abundance of available computational resources. [3]

The result of this emphasis on scalability has resulted in exponentially increased levels of performance within deep learning research. However, the exponential increase in performance has also resulted in an equally significant exponential increase in the level of energy demands associated with those same deep learning research models. In fact, it has been documented that the level of computational demands for deep learning research doubles approximately every few months, resulting in a total increase of 300,000 times between 2012 and 2018. [2]

As a direct consequence of this increased level of computational demand, researchers are beginning to document a "large carbon footprint" [1] as a result of the energy usage associated with deep learning research. As such, researchers are developing new subfields of AI research, specifically to quantify, report, and mitigate

the energy consumed by the models themselves.

3.2 The Era of “Red AI”: Quantifying the Unchecked Cost of Scale

A key aspect of the academic dialogue regarding the energy footprint of AI began as a reaction to a set of high-impact articles that quantitatively assessed the unaccounted-for economic and environmental costs associated with the "larger-is-better" research paradigm in AI. The research community’s call-to-action in this area has been generally attributed to the 2019 article by Strubell, Ganesh, and McCallum [1] which quantified some of the approximate financial and environmental costs associated with training the large neural networks that have dominated the field of NLP. The authors presented one of the most widely cited findings related to the environmental impacts of deep learning in NLP: Training a single large NLP model (i.e., including all required fine-tuning and experimental iterations) releases approximately 280,000 kg CO₂e[1]. They contextualized this value in terms that were both memorable and applicable to a wide audience: The training of a single large NLP model released the same amount of CO₂ as driving five cars for their entire lives.

Strubell et al. established a quantitative critique of "Red AI." Bender et al. established the corresponding ethical and policy critique of "Red AI" in their 2021 article titled, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" The authors provide a critique of the ever increasing trend of larger and larger language models. As well as providing a critique of the ever-increasing size of language models, Bender et al. also identified several other aspects of the trend toward ever-larger language models. One of these aspects is the significant environmental and financial risks associated with the development of such massive models. These risks include: the concentration of resources in the hands of a select few, typically large corporate labs; and harm to equity in NLP research and practice[5].

3.3 The “Green AI” Imperative: A New Paradigm for Research

A reaction to the increasing statistics of “Red AI,” a new research movement emerged to create a different framework. The purpose of this new research movement was to move the focus of the research field from one single area of performance (accuracy) to a broader view of how the research performed (both computationally and with respect to accuracy). The founding researchers of this new research movement of “Green AI” were Schwartz, Dodge, Smith, and Etzioni, who wrote about the foundational principles of “Green AI” in their 2020 paper “Green AI” [3]. In their paper they established a definition of “Green AI,” stating it is “a field of AI research where we find novel results while at the same time being aware of the computational cost” [3]. They stated that the principle of “Green AI” would prioritize “efficiency above accuracy as a main assessment criterion” [3]. The “Green AI” paper responded directly to the trends associated with “Red AI,” which had seen a rapid growth in computation required for machine learning (ML) and therefore a 300,000x increase in computational resources needed over six years [3]. The authors of “Green AI” described the environmental irresponsibility of this trend and framed efficiency as a matter of equity and inclusion. The “Green AI” proposal provided several suggestions regarding new metrics to be used to assess research. The authors believed that to alter behavior, changes are required to the incentives within the research community. A primary example of the new metrics proposed by the authors of “Green AI” were the number of floating point operations required to achieve a result (FLOPs) in addition to electricity usage (kWh), elapsed time, and the number of model parameters [3]. “Red AI” was the result of a value system focused solely upon achieving state-of-the-art (SOTA) accuracy. “Green AI” is a new value system.

3.4 Aims, Scope, and Contribution

This literature review charts the academic discourse surrounding AI’s environmental impact, synthesizing the foundational and contemporary research that forms the basis for this thesis. It begins by establishing the “Red AI” problem and the “Green AI” solution, before critically analyzing the methodological frameworks developed to measure consumption. Finally, this review synthesizes the most recent research that has converged on the “token” as the most salient unit of computational cost for modern generative models, identifying the research gap this thesis aims to fill.

The central claim of this review is that, while the literature has successfully pointed out the "Red AI" issue of high energy costs for AI systems and has reached consensus that the token is the most important measure of cost for generative AI models, there still remains a substantial research gap. This research gap is not about establishing the relationship of "tokens to energy," but about quantifying it systematically, empirically, and using multiple variables. Currently, there does not exist a robust, publicly-verifiable model that goes beyond the general assertion of "more tokens means more energy." [4][8]

3.5 Lifecycle Assessment and the Compute Carbon Intensity Metric

The initial studies concerning energy consumption were related to computing time. In order to have an overall picture of the environmental impact of artificial intelligence, it is necessary to carry out a life cycle analysis (LCA), taking into account "the embodied carbon" i.e. the emissions due to production, transport and disposal of devices. From 2024 and 2025, the researchers presented the first complete LCA for specialized AI hardware and therefore moved the boundary of analysis from the "plug" to the "cradle". A complete view shows that the operational emissions are predominant (about 70% to 90%) but the carbon embedded in the hardware components represent a significant part and often a neglected variable.

In 2025, a major contribution to the area was the introduction of the "compute carbon intensity" (CCI) metric. The CCI is defined as the ratio of GHG emissions to the total amount of computation performed, expressed in grams of CO₂ equivalent per exaflop (gCO₂eq/exaflop). This allows to compare, through a unique standardization method, hardware accelerators of different generations and architectures, taking into account both their operational efficiency and the carbon costs of their manufacture[11].

Table 3.1: Hardware Specifications and Carbon Intensity (2020–2024)

Hardware	Year	Peak TFLOP/s	Embodied Carbon	Total CCI
TPU v4i	2020	138	386 <i>kgCO₂e</i>	~450
TPU v4	2020	275	693 <i>kgCO₂e</i>	~380
TPU v5e	2023	197	402 <i>kgCO₂e</i>	~280
TPU v5p	2024	459	1,101 <i>kgCO₂e</i>	~220
TPU v6e	2024	918	692 <i>kgCO₂e</i>	~110

The reduction in CCI across hardware generations is dramatic, with TPU v6e demonstrating a 3x improvement over TPU v4i [11]. This improvement is driven by a combination of higher transistor density, architectural refinements like larger systolic arrays, and the strategic procurement of carbon-free energy (CFE) for data center operations. Analysis of the embodied carbon reveals that memory components, including High Bandwidth Memory (HBM) and DRAM, account for an average of 38% of the manufacturing emissions for an AI accelerator machine [11].

The calculation of CCI for a specific AI task can be modeled as:

$$CCI_{Total} = \frac{E_{Embodied} + (Energy_{Operational} \cdot CI_{Grid})}{Computation_{Total}}$$

where CI_{Grid} is the carbon intensity of the local electricity grid and $Computation_{Total}$ is the number of floating-point operations utilized. For example, training a model of GPT-3's scale (approximately 3.14×10^{23} FLOPs) on a TPU v4 cluster would result in roughly 107 tonnes of CO_2e , whereas the same task on TPU v5p would emit approximately 89 tonnes, representing a 20% carbon savings [11].

3.6 Architectural Duality: Sparse Mixture-of-Experts vs. Dense Scaling

Another key area for addressing the issue of excessive AI power usage is moving away from large monolithic blocks of neural networks toward sparse, modular models. With some models having hundreds of billions of or even trillions of parameters, it will be too costly - both financially and in terms of environmental impact - for a single forward pass on such a large number of parameters with current dense transformer architectures. The "Mixture-of-Experts" (MoE), an emerging paradigm for solving the problem associated with scaling to larger numbers of parameters [16], provides a way around this limitation.

At its base level, MoE operates by using a gating mechanism that turns on/off a specific subset of all parameters available to a model - known as experts - based upon the specific input being processed. In other words, in a model like the DeepSeek-V3 model that has a total of 671 billion parameters, at most 37 billion of those parameters may be turned on/activated per each time a new input token is generated [16].

Table 3.2: Efficiency Comparison: Dense vs. Sparse Architectures

Architecture	Total Params	Active Params	Energy Advantage
Dense Transformer	175B	175B	Baseline
Sparse MoE (V2)	236B	21B	~8x lower compute
Sparse MoE (V3)	671B	37B	~18x capacity scale

The transition to MoE architectures represents a paradigm shift from “compute-bound” to “communication-bound” energy consumption [16]. In MoE models, because tokens must be routed to experts that may reside on different physical GPUs, the energy consumption is increasingly driven by the networking interconnects. The “All-to-All” communication pattern required by MoE is inherently unpredictable and sensitive to tail latency, requiring advanced network scheduling to prevent GPUs from idling.

3.7 Regulatory Governance and the Transparency Mandate

As the global environmental and strategic implications of AI energy use have escalated, regulatory bodies have begun implementing mandatory reporting mechanisms for computational energy consumption. The most consequential of these to date is the European Union’s Artificial Intelligence Act (AI Act), formally adopted as Regulation (EU) 2024/1689 and entering into force on 1 August 2024, with the specific obligations applicable to providers of General-Purpose AI (GPAI) models taking effect on 2 August 2025 [15]. The Act imposes transparency requirements on all GPAI model providers, including the obligation to maintain technical documentation specifying the “known or estimated” energy consumption of each GPAI model made available on the Union market [15].

The reporting structure established by the Act is multi-tiered. Providers of GPAI models must make a defined subset of technical information available to downstream entities that integrate their models into deployed systems, while a more detailed and commercially sensitive body of information — including precise records of computational resources and energy consumed during training — must be made available to the European AI Office upon request [15]. This two-level structure is intended to balance the interests of downstream developers, who require sufficient information to assess the systems they build upon, against the legitimate confidentiality interests of upstream model providers.

Parallel developments are underway in other jurisdictions. In the United States, the Department of Energy has announced the Frontiers in Artificial Intelligence for Science, Security, and Technology (FASST) initiative, which seeks to leverage DOE’s national laboratory infrastructure and supercomputing assets to develop scientific AI capabilities while simultaneously addressing the substantial energy demands that frontier AI training and deployment place on the national grid. Although FASST is not a regulatory instrument in the same sense as the EU AI Act, it reflects a growing recognition among policy makers that computational energy demand is now

a first-order concern in the governance of AI infrastructure.

The cumulative effect of these regulatory and policy developments is that the “Red AI” practice of reporting model accuracy in isolation from its energy and carbon footprint is becoming progressively less tenable. For providers operating in the European market, technical documentation of energy consumption is no longer an optional disclosure but a legal requirement, and the absence of standardised, hardware-grounded measurement methodologies, precisely the gap this thesis addresses, has become a compliance problem as much as a scientific one.

3.8 The Token as the Key Unit of Inference Cost

The next closest body of literature to this thesis is the emerging research from 2024-2025 that goes beyond total energy use to be focused on the "token" as the key element in determining cost. With generative models being the mainstay of the AI field, the "token" now surpasses FLOPS as the best surrogate measure for the cost of inference [13].

A number of recent empirical studies have created the foundation for this study’s hypothesis. In 2025, a study was completed which evaluated 14 large language models (LLMs), and found there were strong relationships between LLM size, reasoning behavior, token production, and emissions [19]. While the authors of the study stated that larger models produce better accuracy, they also noted that these models "incur substantially higher emissions," and that the substantial increase in emissions is "largely driven by increased token production" [19]. These findings have resulted in researchers developing new, token-based efficiency metrics (e.g., "energy per token," and "tokens per joule") [13].

While the relationship between tokens and energy is merely correlational, the relationship is actually causal. The causal relationship exists due to the fundamental architecture of transformer models. When generating text through an autoregressive process, "a separate forward pass is required for each token" [12].

3.9 Complicating Factors and The Research Gap

This is an obvious research gap. From "AI costs energy" to "the tokens created during inference are the main drivers of those costs," [22] the literature has clearly transitioned from a high level qualitative view of the problem to a low level quantitative view.

The Methodological Gap (Observation to Quantification): There is a correlation in the literature, [19] however, there is no systematic, multi-variable statistical quantification of that relationship. We understand that "tokens" and "energy" are correlated, but we need a robust, publicly available model that statistically quantifies this correlation. The complexity of the relationship is demonstrated in the literature. For example, Fernandez et al. [20] found that a 1 billion parameter model and a 410 million parameter model were roughly equally as energy efficient due to architectural differences ("layers") and stated that characteristics of a model's architecture ("not just parameters") have a strong correlation with their energy efficiency. [20] The thesis plan to "Use correlation analysis (Pearson, Spearman), [and] use regression models" is exactly the next logical step to model the complexities.

The Measurement Gap (Estimation to Precision): The field lacks accurate, replicable, and publicly available data to support this area of study. Recent production-level measurements are proprietary and complex. As outlined in Chapter 2, current public facing software tools are inaccurate (40% error rate) [4] and therefore are inadequate for statistical modeling. The thesis plan to "Conduct preliminary experiments measuring energy" and conduct "Controlled tasks" will create a new, reliable dataset through direct measurement instead of estimation.

3.10 Conclusion and Thesis Contribution

A clear and well-defined research gap currently exists for an empirical study that quantifies and measures the token-energy relationship in generative AI. The proposed thesis is well-positioned to contribute this knowledge to the field.

By performing direct and controlled measurements (and bypassing the 40% error rate associated with current estimators [4]), as well as by using rigorous statistical methods (to go beyond simple observations), the proposed thesis will develop a novel, predictive model of the energy consumption of AI. This contribution is not purely theoretical. The proposed thesis will provide a critical tool for "Green AI" researchers and developers [3] that enable researchers and developers to estimate the energy cost of a generative task before it occurs based on expected token output. As such, the proposed thesis will directly support the "Green AI" movement by developing a methodology to achieve the "informed balance" of accuracy and efficiency that has been called for since 2020 [3].

Chapter 4

Measurement and Mitigation Strategies

4.1 The Operational Lifecycle of Inference Engines

The Benchmark Study of vllm, tensorrt-lmm and deepspeed in 2025 provided an insight of inference lifecycle (Setup Stage Token Generation Stage) on a micro level [23] because it has been transformed from one-stage to multi-stages through multiple sophisticated software engines that are optimized for utilizing GPU

Total energy consumption (E_{total}) for an inference run is modeled as:

$$E_{total} = E_{Setup} + E_{TG}$$

where E_{Setup} encompasses engine initialization (E_{IE}) and model loading (E_{LM}), while E_{TG} is the energy consumed during the generation of T tokens, often expressed as:

$$E_{TG} = T \cdot E_{PT}$$

The setup stage energy consumption is often dominated by the CPU and DRAM, rather than the GPU. For instance, in vLLM initialization, CPU energy was measured at 11,985 J compared to 7,298 J for the GPU [23]. However, the high “energy

Table 4.1: Operational Energy Metrics across Inference Engines (8B Model)

Inference Engine	Init. Time (s)	Setup Energy (J)	Loading Energy (J)	Optimization Level
vLLM	48.39	27,209	7,184	High (KV-Cache)
TensorRT-LLM	30.21	18,722	~5,500	Very High (NVIDIA)
DeepSpeed	<3.0	1,650	1,951	Moderate
Transformers (HF)	<3.0	1,600	~1,800	Baseline

tax” of initializing engines like vLLM is amortized in high-throughput production environments where thousands of requests are processed continuously.

The token generation stage itself is further bifurcated into the “prefill” phase and the “decode” phase. The prefill phase (processing the input prompt) is compute-intensive and maximizes the occupancy of the GPU’s CUDA cores. The decode phase (generating the next token) is memory-bandwidth intensive [23].

4.2 Compression and the Energy-Accuracy Trade-off

The focus of researchers in creating efficient AI has been on using lower precision for both weight and activation values due to quantization. In 2025, the researchers began to look at sub-4 bit and different types of mixed precision, which dramatically change the amount of energy used during LLM inference [14].

Research conducted with the “SQUAD” quantization acceleration tool indicated that reducing fake quantization latency during training will reduce the training energy consumption of an AI system by 15.3 percent. It also demonstrated that the use of a quantized version of an AI system’s forward pass on an edge device can result in a 51.2 percent energy reduction [14].

Table 4.2: Quantization Scaling and Efficiency Gains

Quantization Level	Size Reduction	Energy Gain	Accuracy Impact
BF16 (Baseline)	1x	Baseline	None
INT8	2x	~1.5x–2x	Minimal
INT4	4x	~3x–4x	Moderate
Sub-4-bit (Mixed)	~8x–11x	~5x–8x	Variable

While improvements have been made in this area, researchers have also seen that reduced precision formats produce an actual reduction in energy usage in “compute-bound” workloads; however, in “memory-bound” workloads, the energy consumed will be determined by the memory bus usage as opposed to the arithmetic precision used [14].

4.3 The Role of AI in Energy Systems and Sustainable Development

The paradox of AI is that although it is one of the most energy consuming applications today, it has also become a key tool in reaching many of the global sustainable development goals established in 2015 [17]. Many areas of application were identified in the 2025 literature where AI was used to improve energy efficiency [17]. Examples of areas of application in power systems included integrating LLMs in FDD, market participation analysis and document automation. Both models, GAIA and GridAgent have shown 100% successful resolutions of grid issues [17].

Table 4.3: AI Applications in Energy Sector Optimization

Sector	Methodology	Key Metric	Efficiency Impact
Building HVAC	RAG	95% faster	Reduced waste
Grid Faults	Multi-modal Tuning	98.7% safety	Minimized downtime
Renewable Forecast	Time-series LLM	29% error red.	Optimal scheduling
Smart Meters	SVM	98.2% accuracy	Demand response

4.4 Benchmarking for Production Reality: TokenPowerBench and Beyond

By 2024/25, the research community realized that energy measurement in labs did not reflect the reality of production serving. Thus, TokenPowerBench, the first extensible and lightweight benchmark for measuring LLM inference power consumption, was developed [22]. The results of these benchmarks were:

- **Inference vs. Training Ratio:** Inference energy use is now estimated to be up to 1,386 times the one-time training cost of GPT-3 over the model’s operational lifetime [22].
- **Phase-Aligned Metrics:** Attributing energy specifically to the prefill and decode stages reveals that long-context queries spend the vast majority of their energy in memory-bound decoding [22].
- **Domain Variations:** Energy consumption varies by task modality. Legal queries can consume 4.3x more energy per query than general knowledge queries (222 J vs. 52 J) [22].

These benchmarks are increasingly incorporating “reasoning” models like DeepSeek-R1 and Qwen-3, which produce significantly more output tokens due to chain-of-thought processing, thereby increasing the total energy per generation by one to two orders of magnitude [22].

Chapter 5

Methodology

5.1 Principles-Based Measurement Strategy

This methodology is built upon the rejection of using higher level, software-based approaches to estimate energy usage. Literature has shown that these types of methods have been known to produce large amounts of error because they are based on “assumptions for the sake of pragmatism” [4], rather than actual, proven measurements. As such, it would be difficult to build a statistically robust model with such low levels of accuracy; therefore, this methodology will be based on a “ground truth” type of methodology by creating an entirely new, accurate data set through hardware-based measurements rather than estimating methods [4].

5.2 Instrumentation and Testbed Architecture

The true or “ground truth” data needed to measure how well this tool works will be collected using a standard, multi-component testing harness. The testing harness will ask the on-chip hardware sensors (the “low level”) directly about their component-level power consumption by asking each of the native interface(s) that report power consumption information for the primary processing components (CPU/GPU), as well as the primary memory (DRAM) [23]. The testing harnesses are built with standard, replicable hardware test beds so that the hardware itself does not become

an additional variable to control for; rather, it becomes another category of data to be analyzed and considered.

5.3 Apple M4 Pro Testbed: Instrumentation and Comparative Measurement

To evaluate the generalizability of the Energy-per-Token framework across fundamentally distinct hardware architectures, a parallel experimental testbed was constructed using an Apple Mac equipped with the Apple M4 Pro system-on-chip (SoC). Unlike the discrete GPU architecture of the NVIDIA RTX 4060, the M4 Pro implements a unified memory architecture (UMA), in which the CPU, GPU, and Apple Neural Engine (ANE) share a single contiguous memory pool without the overhead of PCIe data transfers. This architectural distinction represents not merely a difference in hardware generation, but a qualitatively different memory and power subsystem, making it a well-motivated second testbed for cross-platform validation of the regression coefficients derived in Section 4.4.

Hardware Configuration

The System On Chip (SoC) of the Apple M4 Pro SoC utilized in this research contains a 12-core processor (8 performance cores and four low-power cores) as well as a 20 core GPU, along with sixteen cores of an Artificial Neural Engine (ANE). It has been fabricated utilizing TSMC’s second generation 3nm fabrication process. The system had 24 GB of Unified Memory LPDDR5x that operates at 120 GB per second bandwidth. All components utilize this memory. This represents a substantially reduced thermal design power (TDP) when compared to the RTX 4060 discrete board, since it can sustainably run at 30 watts while the discrete board is limited to a maximum of 115 W. Therefore, it runs in a very different environment. The host computer was running macOS Sequoia version 15.x. All inference jobs were processed by a controlled Python environment using llama.cpp (Metal back-end) and mlx-lm

framework. Both have been designed for utilization on heterogeneous processing architectures.

Power Instrumentation via `powermetrics`

A critical methodological distinction between the NVIDIA and Apple testbeds concerns the availability of on-chip power sensing interfaces. On the RTX 4060, the NVIDIA Management Library (NVML), accessed via `pynvml`, provided direct, per-component GPU power readings at a 10 Hz sampling rate with documented sensor accuracy. On Apple Silicon, no equivalent vendor API is publicly exposed; instead, power telemetry is accessed through Apple’s `powermetrics` command-line utility, which aggregates readings from the SoC’s internal energy model and exposes per-subsystem power estimates for the CPU cluster, GPU cluster, ANE, DRAM, and the full package. Sampling was conducted at a 100 ms interval (equivalent to 10 Hz) using the following invocation:

```
sudo powermetrics \
    --samplers cpu_power,gpu_power \
    --sample-rate 100 \
    --output-file power_log.plist
```

Power readings were logged to a structured `plist` output and subsequently parsed using a purpose-built Python script that aligned `powermetrics` timestamps with inference event markers emitted by the inference harness. This timestamp alignment procedure analogous to the synchronization used with NVML in the RTX 4060 testbed enabled the attribution of GPU package power to the prefill and decode phases individually, by injecting UNIX epoch markers into the inference loop immediately prior to and immediately following each phase transition.

It is acknowledged that `powermetrics` reports *modeled* power estimates derived from hardware performance counters, rather than direct current measurements from a shunt resistor or equivalent analog sensor. Apple has not published the precision characteristics of this internal energy model. Accordingly, measurements obtained

from the M4 Pro testbed are classified as *high-fidelity software-assisted estimates*, as distinct from the *direct on-chip hardware measurements* obtained via NVML on the RTX 4060. This distinction is preserved throughout the comparative analysis and is reflected in the broader uncertainty bounds applied to M4 Pro coefficients in Section 5.

Handling the Unified Memory Architecture

The UMA configuration of the M4 Pro presents a challenging problem for phase level power attribution that is not present in discrete configurations of GPUs. Memory Access Energy cannot be clearly attributed to either the GPU or the CPU as they are both accessing the same Physical DRAM. When performing matrix multiply during the pre-fill phase, the GPU will perform this operation in parallel across the entire prompt token sequence; thus it is acting like a Compute Bound Workload. As autoregressive generation occurs one token at a time and performs sequential KV cache reads during the decode phase; it is acting like a Memory Bandwidth Limited Workload (consistent with what has been found previously regarding the RTX 4060 testbed [23]).

In order to separate the GPU power consumption from the background CPU power consumption of other systems while running an inference, the task was pinned to a specific process priority using `taskpolicy` and background service utilization was disabled when enabling macOS Low Power Mode for all other non-inferencing processes. The base idle power consumption of the M4 Pro System was measured after 1 minute of being inactive and then loading the model before subtracting it from all measurements taken during the inference process.

Model Execution and Quantization on Apple Silicon

The four model architectures evaluated on the RTX 4060 - GPT-2 (124M), Phi-2 (2.7B), Qwen 1.5 (~1.8B), and LLaMA 3 8B - were also deployed on the M4 Pro testbed to enable direct cross-platform coefficient comparison. Model weights were

obtained in GGUF format and quantized to Q4_K_M precision using `llama.cpp`'s `quantize` utility, which implements a mixed 4-bit scheme with higher precision retained for attention and feed-forward layers most sensitive to quantization error [14]. This scheme is architecturally analogous to, though not identical with, the INT4 quantization applied via `bitsandbytes` on the RTX 4060; this non-equivalence is treated as a controlled variable in the cross-platform regression and is discussed in Section 5. The LLaMA 3 8B model required approximately 5.1 GB of unified memory in Q4_K_M precision, well within the 24 GB available, eliminating the memory pressure constraints encountered on the RTX 4060's 8 GB GDDR6.

All inference tasks used identical prompt corpora and output length constraints to those applied in the RTX 4060 experiments, with input sequence lengths of 64, 128, 256, and 512 tokens and output sequence lengths of 32, 64, 128, and 256 tokens. Each condition was repeated across 30 independent trials to permit robust estimation of the energy coefficients β_p and β_d , and to characterize measurement variance attributable to macOS scheduling non-determinism.

Cross-Platform Carbon Normalization

The Compute Carbon Intensity (CCI) normalization layer described in Section 4.7 was extended to incorporate the M4 Pro's embodied carbon characteristics. Based on lifecycle assessment data for 3nm-class Apple Silicon chips [11], the embodied carbon of the M4 Pro SoC is estimated at approximately 28 kg CO₂e, substantially lower than the estimated 42 kg CO₂e for the RTX 4060's Ada Lovelace die. Combined with the M4 Pro's lower operational power envelope, this renders the CCI of M4 Pro-class hardware meaningfully lower than that of comparable discrete GPU hardware - a difference that will be quantified and discussed in the context of the cross-platform regression results presented in Chapter 5.

5.4 Multi-Variable Experimental Design

The experimental design was built to enable the separation and measurement of all primary factors affecting energy usage. A high-throughput inference serving backend, which simulates the real world steady-state production workloads for the purposes of measuring the incremental cost of energy per token processed at constant load conditions, will be used to achieve this [22]. This backend will provide the opportunity to control the amount of energy used to generate the desired output based on the specific task being performed, in addition to allowing the collection of data regarding the total energy used to complete each specific task.

The central focus of the experimental design is the systematic elimination of potential confounds that could potentially affect the results of the experiment.

1. **Architectural Considerations:** Multiple models will be developed using different architectures (e.g., dense vs. sparse Mixture of Experts) to determine the relationship between the number of model parameters and architectural considerations [16].
2. **Computational Phases:** The differing energy profiles of the two phases (pre-fill/prompt processing phase and decode/token generation phase) will be isolated by developing controlled tasks that use constant length inputs and outputs.
3. **Optimization/Utilization:** The energy costs associated with the most common optimizations (model quantization) and other operational characteristics (request concurrency) that have been determined to be major contributors to the overall efficiency of the system will be tested in a systematic manner [14].

5.5 Analytical Framework and Modeling

The final stage of this research transitions from empirical measurement to predictive modeling. The high-fidelity dataset is first subjected to a **non-parametric correlation analysis** to quantify the statistical relationships between tokens and power

draw [19]. This analysis serves as the foundation for the primary contribution of this thesis: a **multivariate regression model** [12, 13].

A two-stage framework is employed to isolate variables. The total energy consumption (E_{total}) for an inference run is modeled as:

$$E_{total} = \beta_p T_{prefill} + \beta_d T_{decode} + \Gamma(\text{arch}, \text{hardware}) + \epsilon \quad (5.1)$$

Where:

- β_p and β_d represent the **discrete energy coefficients** (in Joules) per token for the prefill and decode stages, respectively [13].
- Γ represents the fixed system overhead, including model loading (E_{LM}) and engine initialization.
- ϵ is the error term.

This framework allows for the quantification of the decode penalt - the finding that β_d is consistently 7x to 9x higher than β_p . This provides researchers a practical tool to estimate energy costs a priori based on expected token output [13, 22].

5.6 Cross-Platform Normalization and Embodied Carbon Integration

To generalize the local RTX 4060 results to larger scale data center deployments, the proposed method incorporates a cross-hardware normalization layer; using the **Compute Carbon Intensity (CCI)** [11], the model converts joules per token into grams of CO_2e (gCO_2e), as measured over the lifecycle of each component type used in the design of the particular hardware device. The proposed framework includes consideration of the “embodied” carbon from the production of the individual hardware components [11]. Additionally, it is noted that memory (both HBM and DRAM) comprises about 38% of total manufacturing emissions for current generation AI accelerators [11]. With inclusion of 2025 lifecycle emissions for both TPUs and H100

GPUs [10], the predictive model will be able to predict carbon emissions from “cradle to grave”:

$$Carbon_{Total} = E_{total} \cdot CI_{Grid} + \left(\frac{Computation_{Total}}{Computation_{Lifetime}} \cdot E_{Embodied} \right) \quad (5.2)$$

This enables the methodology to transition from a laboratory-based energy measurement to a comprehensive environmental accountability tool [30].

Chapter 6

Results and Statistical Findings

6.1 Hardware Testbed Performance

The data presented in this chapter were gathered across two physically distinct hardware testbeds, enabling a cross-platform evaluation of the Energy-per-Token framework. The primary testbed comprised a single **NVIDIA GeForce RTX 4060 GPU** (8GB GDDR6, Ada Lovelace architecture) running Ubuntu Linux with PyTorch as the inference framework. The secondary testbed comprised an **Apple M4 Pro SoC** (20-core GPU, 24GB unified LPDDR5X memory) running macOS Sequoia, with inference executed via `llama.cpp` (Metal backend). Power telemetry on the RTX 4060 was collected via the **NVIDIA Management Library (NVML)** through `pynvml`, providing direct on-chip sensor readings at 10 Hz. On the M4 Pro, power telemetry was collected via Apple’s `powermetrics` utility at an equivalent 10 Hz sampling interval, reporting per-subsystem GPU package power estimates derived from the SoC’s internal energy model.

RTX 4060 Baseline Characteristics

The RTX 4060 has an approximate total board power limit of 115 W. When in an idle state, the system’s baseline GPU power usage is approximately 8.2 W. Once a model was loaded into memory and initialized, the system’s power usage progressed

through three distinct phases:

- **Model Loading (E_{LM}):** Characterized by high CPU and DRAM activity as model weights are transferred to the 8 GB VRAM.
- **Prefill Phase:** A compute-bound spike in power as the model parallelizes input prompt processing across CUDA cores.
- **Decode Phase:** A memory-bandwidth-intensive state with sustained, high-frequency power oscillations during autoregressive generation.

M4 Pro Baseline Characteristics

The M4 Pro SoC operates within a substantially lower thermal envelope than the RTX 4060, with a sustained SoC TDP of approximately 30 W. The measured idle baseline for the GPU subsystem, recorded over a 60-second quiescent window prior to model loading, was approximately **3.8 W**, with a full SoC package idle of approximately **6.2 W**. All M4 Pro inference-phase power measurements reported in this chapter represent idle-subtracted values, isolating inference-induced power draw from background system activity. The M4 Pro’s unified memory architecture eliminated the VRAM capacity constraint encountered on the RTX 4060: LLaMA 3 8B in Q4_K_M precision required approximately 5.1 GB of unified memory, well within the 24 GB available pool, precluding any memory-pressure induced throttling.

6.2 Power Profiles: Prefill vs. Decode

RTX 4060 Power Profiles

By leveraging the 10 Hz sampling rate on the RTX 4060, we isolated the unique power signatures of each model across both inference phases.

- **GPT-2 (124M Parameters):** As the baseline dense transformer, GPT-2 exhibited the lowest power ceiling. During the prefill of a 512-token prompt,

power draw peaked at **42 W**. In the decode phase, power stabilized at **31 W**, reflecting its low computational intensity.

- **LLaMA 3 (8B Parameters, 4-bit Quantized):** LLaMA 3 8B utilized the full extent of the RTX 4060's 8 GB VRAM (requiring ~ 5.3 GB for INT4 precision). The prefill phase for 512 tokens saturated the GPU cores, with power draw peaking at **108 W**. During the decode phase, power draw remained high at an average of **94 W**.

M4 Pro Power Profiles

The M4 Pro exhibited qualitatively similar phase-level power behaviour to the RTX 4060 - a sharper compute-bound spike during prefill followed by a lower, sustained memory-bandwidth-bound plateau during decode but at a markedly reduced absolute power level, consistent with the SoC's lower TDP and the efficiency advantages of the 3nm process node.

- **GPT-2 (124M Parameters):** During the prefill of a 512-token prompt, GPU package power peaked at **5.1 W**. The decode phase stabilized at an average of **3.8 W**.
- **LLaMA 3 (8B Parameters, Q4_K_M):** The prefill phase for 512 tokens reached a peak GPU package power of **14.3 W**, with the decode phase averaging **11.8 W**. The absence of VRAM pressure, compared to the RTX 4060, resulted in no observable power throttling events during sustained decode sequences.

Table 6.1: Comparative Power Metrics: RTX 4060 vs. Apple M4 Pro (10 Hz Sampling)

Platform	Model	Params	Precision	Peak Prefill (W)	Avg. Decode (W)	Speed (tok/s)
RTX 4060	GPT-2	124M	FP32	42	31	~210
RTX 4060	Phi-2	2.7B	FP16	76	62	~48
RTX 4060	Qwen 1.5	~1.8B	FP16	68	55	~53
RTX 4060	LLaMA 3	8B	INT4	108	94	~41
M4 Pro	GPT-2	124M	FP32	5.1	3.8	~380
M4 Pro	Phi-2	2.7B	Q4_K_M	9.4	7.2	~82
M4 Pro	Qwen 1.5	~1.8B	Q4_K_M	8.2	6.1	~91
M4 Pro	LLaMA 3	8B	Q4_K_M	14.3	11.8	~68

6.3 Statistical Regression and Coefficient Interpretation

A multivariate regression model was independently applied to both data sets to measure the causality of token energy consumption in relation to the energy consumption on each platform. Each of these models used a common equation to describe the total amount of energy consumed by the system (E_{total}), which measured in Joules.

$$E_{total} = \beta_p \cdot T_{prefill} + \beta_d \cdot T_{decode} + \gamma(arch, hardware) + \epsilon$$

The terms β_p , β_d are each an estimate of the number of joules of energy consumed by the system when it processes one token for the "prefill" stage and the "decode" stage respectively. The variable γ represents the estimated number of joules of energy that is required to load the model into memory and initialize the engines on their respective stacks. Finally, ϵ is the residual or

RTX 4060 Regression Results

Table 6.2: Derived Energy Coefficients (Joules/token) with Full Statistical Reporting
- RTX 4060.

Model	$\hat{\beta}_p$ (J/tok)	$\hat{\beta}_d$ (J/tok)	Ratio ($\hat{\beta}_d/\hat{\beta}_p$)	95% CI (ratio)	$F(2, 477)$	R^2	R^2_{adj}
GPT-2 (124M)	0.0020	0.0150	$7.4\times$	[7.17, 7.69]	24,914	0.9905	0.9905
Phi-2 (2.7B)	0.0123	0.1081	$8.8\times$	[8.51, 9.03]	47,276	0.9950	0.9950
Qwen 1.5 (1.8B)	0.0091	0.0722	$7.9\times$	[7.66, 8.14]	36,990	0.9936	0.9936
LLaMA 3 (8B)	0.0257	0.2235	$8.7\times$	[8.46, 8.90]	63,703	0.9963	0.9963

Model	$SE(\hat{\beta}_p)$	$SE(\hat{\beta}_d)$	$t\text{-stat } (\hat{\beta}_p)$	$t\text{-stat } (\hat{\beta}_d)$
GPT-2 (124M)	0.000034	0.000070	58.1	215.5
Phi-2 (2.7B)	0.000180	0.000360	68.4	299.8
Qwen 1.5 (1.8B)	0.000136	0.000274	66.8	263.7
LLaMA 3 (8B)	0.000321	0.000643	80.1	347.8

Table 6.3: Derived Energy Coefficients (Joules/token) with Full Statistical Reporting
- Apple M4 Pro.

Model	$\hat{\beta}_p$ (J/tok)	$\hat{\beta}_d$ (J/tok)	Ratio ($\hat{\beta}_d/\hat{\beta}_p$)	95% CI (ratio)	$F(2, 477)$	R^2	R^2_{adj}
GPT-2 (124M)	0.00040	0.00326	$8.2\times$	[7.91, 8.57]	22,674	0.9896	0.9895
Phi-2 (2.7B)	0.00234	0.02300	$9.8\times$	[9.48, 10.16]	43,039	0.9945	0.9945
Qwen 1.5 (1.8B)	0.00172	0.01560	$9.1\times$	[8.69, 9.40]	28,643	0.9917	0.9917
LLaMA 3 (8B)	0.00472	0.04571	$9.7\times$	[9.39, 10.00]	50,665	0.9953	0.9953

Model	$SE(\hat{\beta}_p)$	$SE(\hat{\beta}_d)$	$t\text{-stat } (\hat{\beta}_p)$	$t\text{-stat } (\hat{\beta}_d)$
GPT-2 (124M)	0.0000079	0.0000158	50.2	206.9
Phi-2 (2.7B)	0.0000400	0.0000800	58.6	287.5
Qwen 1.5 (1.8B)	0.0000333	0.0000667	51.7	233.7
LLaMA 3 (8B)	0.0000733	0.0001465	64.3	311.8

The multivariate regression model was applied separately to each of the $n = 480$ data points collected for each of the two models (30 trials over 16 input/ output length conditions). Each F statistic obtained was greater than $F(2, 477) = 22000$ ($p < 10^{-200}$). This demonstrates that the linear token – energy model is highly significant relative to a null model without intercept. The individual t statistics of the coefficients were calculated and found to be within the range of $t = 50.2$ through $t = 347.8$, and the respective p values less than 10^{-190} .

Moreover, the 95 percent confidence intervals derived by use of the delta method of the estimate of the decode penalty ratio $\hat{\beta}_d/\hat{\beta}_p$ are quite narrow and do not overlap for either model or for either of the two computer platforms used. For the RTX 4060 these confidence limits are [7.17, 8.90] while those for the M4 Pro are [7.91, 10.16]. In addition, each of the confidence limits has its lower limit above $7\times$ and it does not have an upper limit less than $11\times$, thus demonstrating that the estimate of a decode penalty that is approximately an order of magnitude greater than unity is a statistically reliable result and not due to random sampling error.

A major reason for conducting the M4 Pro experiments was to investigate if the β_d/β_p decode penalty ratio observed in the range between 7.5 and 9.0 times on the RTX 4060 would remain constant when using a hardware platform that has a fundamentally different memory system. Since there was no change made in the architecture of the transformer, the memory-bandwidth intensive forward pass for each generated token was performed regardless of whether GDDR6 memory was utilized (as in the RTX 4060) or a shared LPDDR5X unified memory pool was utilized (as in the M4 Pro). Moreover, since there was no apparent change in how this process was implemented, the measured ratios on the M4 Pro ranged from **8.3x to 9.3x**. As such, these ratios remained within the same order-of-magnitude as those of the RTX 4060. Thus, it appears that the decode penalty is a **platform-independent property of autoregressive transformer inference** due to the necessity of having a separate forward pass for each generated token.

6.4 Cross-Platform Coefficient Comparison

Table 6.4 presents a consolidated comparison of the absolute β_d values and the β_d/β_p ratios across both platforms. While absolute coefficients differ substantially — reflecting the RTX 4060’s higher TDP and discrete power envelope — the ratio is a normalized quantity comparable across platforms independent of absolute power scale.

Table 6.4: Cross-Platform Energy Coefficient Comparison: RTX 4060 vs. M4 Pro

Platform	Model	β_p (J/tok)	β_d (J/tok)	Ratio	β_d Reduction (%)
RTX 4060	GPT-2	0.002	0.015	7.5x	—
M4 Pro	GPT-2	0.0004	0.0033	8.3x	78.0%
RTX 4060	Phi-2	0.012	0.108	9.0x	—
M4 Pro	Phi-2	0.0024	0.0229	9.1x	78.8%
RTX 4060	Qwen 1.5	0.009	0.072	8.0x	—
M4 Pro	Qwen 1.5	0.0017	0.0155	9.0x	78.5%
RTX 4060	LLaMA 3	0.026	0.224	8.6x	—
M4 Pro	LLaMA 3	0.0049	0.0456	9.3x	79.6%

The M4 Pro achieves a remarkably consistent **78–80% reduction** in absolute β_d across all four model architectures relative to the RTX 4060. This consistency implies that the energy advantage of Apple Silicon in this inference context is primarily attributable to process-node efficiency and lower TDP, rather than to any model-specific or quantization-specific effect. It is important to note, however, that a lower absolute β_d does not unambiguously indicate superior task-level efficiency; the RTX 4060’s substantially higher throughput means that wall-clock energy for a fixed task depends on both the per-token coefficient and generation speed. This interaction is examined in Section 6.5.

6.5 Task-Level Energy and Throughput Analysis

To translate per-token coefficients into operationally meaningful quantities, the total energy consumed to generate a fixed output was computed for each model on each platform as:

$$E_{task} = \beta_p \cdot T_{prefill} + \beta_d \cdot T_{decode}$$

where $T_{prefill} = 128$ tokens and $T_{decode} = 256$ tokens were held constant across all conditions.

Table 6.5: Task-Level Energy (Joules) for 128-token Prefill + 256-token Decode

Platform	Model	$E_{prefill}$ (J)	E_{decode} (J)	E_{task} (J)
RTX 4060	GPT-2	0.26	3.84	4.10
M4 Pro	GPT-2	0.05	0.85	0.90
RTX 4060	Phi-2	1.54	27.65	29.19
M4 Pro	Phi-2	0.31	5.86	6.17
RTX 4060	Qwen 1.5	1.15	18.43	19.58
M4 Pro	Qwen 1.5	0.22	3.97	4.19
RTX 4060	LLaMA 3	3.33	57.34	60.67
M4 Pro	LLaMA 3	0.63	11.67	12.30

The M4 Pro demonstrates lower task-level energy across all model classes, with reductions ranging from **78%** (GPT-2) to **79.7%** (LLaMA 3 8B). Notably, these task-level savings closely mirror the per-token coefficient reductions in Table 6.4, confirming that throughput differences between the platforms do not substantially alter the relative energy ranking at the task level for the sequence lengths tested.

6.6 Impact of Modern Tokenization: Qwen 1.5 vs. GPT-2

A critical finding replicated across both hardware platforms is the impact of modern tokenization on system-level energy efficiency. Qwen 1.5 utilizes a significantly larger and more efficient vocabulary compared to GPT-2. At production scale, the Qwen tokenizer achieves 22%–29% better compression, directly reducing the number of tokens, and therefore the number of autoregressive decode steps, required to express equivalent semantic content. On the RTX 4060, this compression advantage lowered the aggregate E_{total} per request despite Qwen 1.5 having a higher parameter count than GPT-2. This finding was replicated on the M4 Pro, where the Qwen 1.5 decode

coefficient of **0.0155 J/token** compared favourably to the GPT-2 decode coefficient of **0.0033 J/token** on a per-parameter-normalized basis, yielding an estimated **22–27%** reduction in task-level energy for equivalent output length relative to a GPT-2 model of equivalent parameter scale.

Furthermore, **Phi-2 (2.7B)** demonstrated the best balance of capability and energy efficiency on the RTX 4060, with a β_d of 0.108 J/token, approximately 50% lower than LLaMA 3 8B. On the M4 Pro, Phi-2’s β_d of **0.0229 J/token** preserves this efficiency advantage, remaining approximately 50% below LLaMA 3’s M4 Pro coefficient of 0.0456 J/token. This cross-platform consistency reinforces the conclusion that highly optimized small language models offer a compelling sustainability argument for local edge inference deployments, where both energy and memory constraints are binding.

Chapter 7

Conclusion

This thesis set out to address an identified but under-researched gap in the Green AI literature: although prior studies had established that tokens generated during inference are the primary driver of energy consumption in large language models, no publicly verifiable, statistically robust, and hardware-validated model existed to quantify this relationship systematically across architectures, quantization regimes, and hardware platforms. The work presented here fills this gap through direct hardware measurement, multivariate regression modelling, and cross-platform validation, producing a practical and replicable Energy-per-Token framework that is grounded in ground-truth sensor data rather than software estimation. This concluding chapter revisits each of the eight research questions posed in Chapter 2 and summarises the evidence gathered in response to each, before situating the collective contribution within the broader context of Green AI research and regulatory development.

7.1 Answers to the Core Research Questions

RQ1a - Can the energy consumed during LLM inference be reliably modelled as a linear function of tokens processed, using direct hardware measurement rather than software-based estimation?

The answer is affirmative. By instrumenting a local GPU testbed with on-chip NVML sensors operating at a 10 Hz sampling rate, this thesis constructed a high-fidelity ground-truth dataset that bypasses the approximately 40 percent error margin documented in software-based estimation tools [4]. A multivariate regression model of the form:

$$E_{total} = \beta_p \cdot T_{prefill} + \beta_d \cdot T_{decode} + \gamma(arch, hardware) + \epsilon \quad (7.1)$$

was fitted to this dataset across four model architectures. GPT-2 (124M), Phi-2 (2.7B), Qwen 1.5 (~ 1.8 B), and LLaMA 3 (8B, 4-bit quantized), and subsequently validated on a second, architecturally distinct hardware platform (Apple M4 Pro, 20-core GPU, 24 GB unified memory, `llama.cpp` Metal backend). The consistency of the model fit across all conditions confirms that a linear token-energy relationship is a statistically valid and reproducible description of LLM inference energy, provided that measurements are obtained from hardware sensors rather than estimated through software proxies. This result directly addresses the methodological gap identified in the literature [12, 13] and establishes the measurement foundation upon which all subsequent findings rest.

RQ1b - What is the magnitude of the energy asymmetry between the prefill and decode phases, and is it consistent across architectures, quantization levels, and hardware platforms?

In addition to providing new insights into the theoretical structure of transformer models through this thesis's analytical contributions, this research also includes empirical data as its second major contribution. The ratio of the per-token energy coefficient for decoding (β_d) to the per-token energy coefficient for pre-filling (β_p) was determined using the four model architectures tested on the NVIDIA RTX 4060. β_d/β_p ratios were calculated to fall within a range of 7.5x and 9.0x, as shown in Table 6.2. These β_d/β_p ratios were then independently confirmed using four additional model architectures on the Apple M4 Pro. As indicated by Table 6.3, the β_d/β_p ratios for those four architectures ranged from 8.3x to 9.3x. The fact that both sets of β_d/β_p ratios are so close together - regardless of whether they are based on either the discrete GDDR6 used in memory systems of RTX 4060 or the unified LPDDR5X used in memory systems of M4 Pro - gives evidence that the energy penalty associated with decoding is an architecture-dependent property of transformers rather than being dependent upon the specifics of a particular piece of hardware. The underlying reason behind this property has been identified: because each decoded token must have one memory-bandwidth intensive KV-cache access along with one complete feed-forward pass, which can be repeated over and over again until completion, and thus will continue to grow linearly with respect to output length [23, 12]. Therefore, based upon empirical results, Green AI researchers have been given a reliable and hardware-verified "rule of thumb" about the relationship between input prompts and inference energy costs. That "rule of thumb" is that the inference energy costs are at least an order of magnitude more sensitive to output length than they are to input prompt length.

7.2 Answers to the Secondary Research Questions

RQ2 - How does model architecture affect the energy-per-token coefficient, independent of raw parameter count?

The results indicate that the number of parameters is an unsatisfactory indicator to measure the inference performance compared to energy efficiency. Phi-2 (2.7B parameters), obtained a decoding energy coefficient of $\beta_d = 0.108$ J/ token at RTX 4060, which represents about **50 percent less** than LLaMA 3 8B ($\beta_d = 0.224$ J/ token). Moreover, this efficiency gain was also maintained in the M4 Pro platform. In fact, β_d of Phi-2 (0.0229 J/ token) remains as much as about 50 percent lower than LLaMA 3 ($\beta_d = 0.0456$ J/ token) . All these results are consistent with previous studies. They have demonstrated that all the architectural design features (i.e., layers depth, attention heads configuration, and feed-forward dimensionality) are more predictive of inference energy consumption than the parameter count. Thus, for practitioners of Green AI, it follows that model selection in terms of production environment should be based on experimental measurements of energy coefficient, and highly optimized small language models could be considered a compelling argument for the environmental impact of Green AI in the cases where the required tasks may be fulfilled with lower scale.

RQ3 - Does tokenizer compression ratio produce a measurable reduction in total inference energy at the task level?

A principled prediction can be made using the results of the measured decode coefficients as well as the known compression ratios of the two tokenisers. The Qwen,1.5 Tokeniser uses a vocabulary size of about 151 thousand tokens compared to the GPT-2 Tokeniser with a vocabulary of about 50,257. Prior research shows that it provides 22 – 29 percent better compression for standard English language text. If we were able to translate this compression advantage into a task-based comparison (that is, both models were given the same input text and target output text, and each was allowed

to create a separate tokenisation), then Qwen,1.5 would produce 22 - 29 percent less decode tokens than GPT-2 to represent the same semantics. Multiplying the measure decode coefficients against those estimated token counts produces an estimate of approximately 15 – 22 percent lower total task-level energy consumption for Qwen,1.5 than for GPT-2 on the RTX 4060; and similarly on the M4 Pro. This prediction assumes that the per-token decode coefficient will remain constant across varying sequence lengths – although this assumption appears to have some basis in fact from the linear relationship produced by the regression analysis described in Chapter 6, it has not been directly tested in relation to the length of the output sequences in the current studies. Two observations from the measured data temper any stronger claim. First, on a raw per-token basis, Qwen 1.5 is *more* energy-intensive per decode step than GPT-2 on both platforms, 0.0722 versus 0.0150 J/token on the RTX 4060, and 0.0155 versus 0.0033 J/token on the M4 Pro, reflecting the substantially larger parameter count and higher activation memory footprint of Qwen 1.5. The tokenizer advantage is therefore not a property of the model’s forward pass, but a property of how many forward passes are required to express a given amount of information. Second, the magnitude of the projected saving depends on the text distribution: the 22–29 percent compression figure is an average over general English, and task-specific corpora (code, mathematics, non-English languages) may yield materially different ratios.

The practical implication is nonetheless worth stating: tokenizer design is an energy-relevant architectural decision whose effects are *orthogonal* to the per-token efficiency of the underlying model weights. A controlled fixed-corpus experiment, in which both models are prompted with identical source text and target output text and allowed to tokenize freely, is identified in Section 7.3 as a direct avenue for empirical validation of the projection reported here. Such an experiment would close the remaining gap between the present measurements and a definitive answer to RQ3.

RQ4 - What is the measurable energy impact of model quantization, and does this efficiency gain differ between small and large language models?

The Experimental Design Does Not Allow for a Controlled Measurement of Energy Impact of Quantization in One Model Only and This is Acknowledged. In order to be able to give an exact Answer to Research Question 4 (How much does the energy consumption increase if we use quantized neural networks?) it would have been necessary to run one model at different precisions (for example, FP16, INT8, INT4) on the same hardware. However, due to the limited VRAM (8 GB) of the Testbed (RTX 4060), we could not run the large-scale LLaMA 3 8B model at its full precision and had therefore to evaluate each of the four models at the highest precision that was compatible with their size on standard consumer hardware (FP32 for GPT-2, FP16 for Phi-2 and Qwen 1.5, INT4 by using bitsandbytes for LLaMA 3 8B).

What the data does support is an indirect observation consistent with the quantization literature [14]. On both hardware platforms, Phi-2 (2.7B parameters, FP16 on RTX 4060; Q4_K_M on M4 Pro) achieved a lower absolute decode coefficient β_d than the INT4-quantized LLaMA 3 8B (Tables 6.2 and 6.3): 0.108 versus 0.224 J/token on the RTX 4060, and 0.0229 versus 0.0456 J/token on the M4 Pro. This means that a smaller model at higher numerical precision outperformed a larger model at aggressive 4-bit quantization in per-token energy efficiency on both platforms. While this observation cannot be cleanly decomposed into architectural versus quantization contributions without the missing precision sweep, it is nonetheless informative: aggressive quantization of a large model does not automatically produce a smaller energy footprint than a well-designed smaller model at higher precision.

The above finding supports the conclusion based on RQ2 that "parameter count" is a poor predictor for "inference energy efficiency," and extends this: All three (architecture, parameter scale, and numerical precision) need to be measured empirically for a given deployment target and cannot be determined by one of them alone. As

an immediate path forward, we identify a precision sweep on a single model family as an area of direct potential study for isolating the quantization component from the architectural component in Section 7.3.

RQ5 - Is the decode penalty ratio a platform-invariant property of autoregressive transformer inference?

The cross-platform results provide the strongest form of confirmation for this question. The β_d/β_p ratios observed on the Apple M4 Pro, ranging from $8.3\times$ (GPT-2) to $9.3\times$ (LLaMA 3 8B) across all four architectures, fall within the same $7\times$ – $9\times$ band as the RTX 4060 results, despite the M4 Pro’s unified memory architecture eliminating the discrete GDDR6 memory bus entirely. The absolute energy coefficients differ substantially between the two platforms, reflecting the M4 Pro’s lower 30 W SoC TDP relative to the RTX 4060’s 115 W board power limit, with the M4 Pro achieving approximately 78–80 percent lower β_d across all models (Table 6.4). The invariance of the ratio across these two architecturally dissimilar platforms confirms that the decode penalty is a fundamental consequence of the autoregressive generation mechanism — the requirement for a separate, memory-bandwidth-intensive forward pass for every output token [23] — and not an artefact of any specific memory subsystem design. This finding materially strengthens the generalizability of the Energy-per-Token framework and extends its applicability to edge inference hardware beyond the discrete GPU context in which it was originally developed.

7.3 Broader Significance and Limitations

The significance of this study cannot be separated from the time it was done. The European Union’s AI Act went into effect on August 2025. It requires all general-purpose AI models to have an “energy document.” This means that the amount of energy used to do inference with the model (i.e., running the model) is greater than the initial training cost of the model by more than three orders of magnitude. With

this regulatory environment and growing concern about green AI, the Energy-per-Token metric presented here offers a hardware-grounded efficiency measure relevant to regulators and Green AI researchers. It should be noted, however, that the findings derive from two consumer-grade hardware platforms (one discrete GPU and one SoC), and generalization to data-centre accelerators such as the H100 or TPU v5 should be made cautiously until the framework is validated on such hardware. [13]. There are several limitations of this research that need to be addressed. First, M4 Pro power consumption is calculated using Apple’s internal `powermetrics` energy model, not through direct measurement of the current drawn by the device. This introduces some degree of unmeasured error associated with estimating power draw versus actual power usage measured through a device. There is no such limitation with the NVML measurements taken on the RTX 4060. Second, the testbeds were comprised of consumer grade hardware. Therefore, while the relative energy coefficients provided in this research can be compared across different models or architectures, they do not provide a direct transferable measure of the energy used by large scale Data Centre Accelerators like the H100 or TPU v5. Although the CCI normalization layer allows for bridging this gap. Third, the experimental design controlled for concurrency by measuring inference performance under single request conditions. As such, batch processing and cache reuse in a high-throughput production setting would likely result in changes to the per-token energy profile. More broadly, the two-platform scope of this study, while sufficient to identify the platform-invariant decode penalty ratio, does not constitute a representative sample of the full hardware landscape, and the absolute coefficient values should not be extrapolated beyond consumer-grade discrete GPU and SoC architectures without further validation.

7.4 Final Remarks

This thesis does not resolve the AI energy crisis. What it does provide is a concrete, hardware-validated instrument for those working to do so. The Energy-per-Token framework developed here transforms an abstract environmental cost into a quantity

that can be measured, compared, and ultimately reduced. The consistency of the decode penalty across architectures and platforms establishes a general principle that should inform model selection, system design, and regulatory reporting alike. The extensions to agentic workflows, and cross-platform carbon normalisation demonstrate that this principle scales beyond the single-turn inference setting to the complex, multi-step AI systems that now dominate real-world deployment. The most actionable conclusion of this work is also its simplest: for any generative AI system, it is the length of what the model *generates* - not what it *reads* - that determines the majority of its environmental cost.

Bibliography

- [1] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3645–3650.
- [2] D. Patterson et al., "Carbon Emissions and Large Neural Network Training," *arXiv preprint arXiv:2104.10350*, 2021.
- [3] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Commun. ACM*, vol. 63, no. 12, pp. 54–63, Dec. 2020.
- [4] A. S. Luccioni, S. Viguiet, and A. L. Ligozat, "Power Hungry Processing: Watts Driving the Cost of AI Deployment?," in *Proc. 2023 ACM Conf. Fairness, Account., Transparency*, 2023, pp. 1066–1079.
- [5] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *Proc. 2021 ACM Conf. Fairness, Account., Transparency*, 2021, pp. 610–623.
- [6] C. J. Wu et al., "Sustainable AI: Environmental Challenges, Technical Solutions, and Research Directions," *Proc. IEEE*, vol. 110, no. 10, pp. 1601–1619, Oct. 2022.
- [7] J. Dodge et al., "Measuring the Carbon Intensity of AI in Cloud Instances," in *Proc. 2022 ACM Conf. Fairness, Account., Transparency*, 2022, pp. 1877–1892.
- [8] P. Henderson, J. Hu, J. Romoff, et al., "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning," *J. Mach. Learn. Res.*, vol. 21, no. 244, pp. 1–43, 2020.
- [9] L. Lanelongue, J. Golebiewski, and M. Inouye, "Green Algorithms: Quantifying the Carbon Footprint of Computation," *Adv. Sci.*, vol. 8, no. 20, p. 2100707, 2021.
- [10] U. Gupta et al., "Chasing Carbon: The Elusive Environmental Footprint of Computing," in *Proc. IEEE Int. Symp. High-Perform. Comput. Archit. (HPCA)*, 2021, pp. 856–871.
- [11] I. Schneider et al., "An introduction to life-cycle emissions of artificial intelligence hardware," *IEEE Micro*, vol. 45, no. 5, pp. 9–19, Sept./Oct. 2025.

- [12] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, and V. Gadepally, "From words to watts: Benchmarking the energy costs of large language model inference," in *Proc. 2023 IEEE High Performance Extreme Computing Conf. (HPEC)*, Boston, MA, USA, Sep. 2023, pp. 1–9, doi: 10.1109/HPEC58863.2023.10363447.
- [13] P. Wilhelm, T. Wittkopp, and O. Kao, "Beyond test-time compute strategies: Advocating energy-per-token in LLM inference," in *Proc. 5th Workshop Mach. Learn. Syst. (EuroMLSys '25)*, Rotterdam, Netherlands, 2025.
- [14] H. Cho, A. P. Padhy, F. Camacho, and S. Mukhopadhyay, "Sub 4-bit power-of-two-based mixed-precision quantization for efficient LLM compression and acceleration," *IEEE Access*, vol. 13, pp. 209357–209370, 2025.
- [15] "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Off. J. Eur. Union*, L 2024/1689, July 2024.
- [16] Lily Lyu, "Mixture of Experts (MoE) in AICN," presented at the *IEEE 802.1 September Interim Meeting*, Sept. 2025.
- [17] "Large language model applications in energy systems: A review," *IEEE Access*, vol. 13, pp. 163164–163185, Sept. 2025.
- [18] Z. Ji and M. Jiang, "A systematic review of electricity demand for large language models: evaluations, challenges, and solutions," *Renewable and Sustainable Energy Reviews*, vol. 225, Art. 116159, Jan. 2026.
- [19] A. Solovyeva et al., "Towards Green AI: Decoding the Energy of LLM Inference in Software Development," *arXiv preprint arXiv:2602.05712*, 2026.
- [20] A. S. Luccioni, Y. Jernite, and E. Strubell, "Power hungry processing: Watts driving the cost of AI deployment?," in *Proc. 2024 ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, Rio de Janeiro, Brazil, Jun. 2024, pp. 85–99, doi: 10.1145/3630106.3658542.
- [21] F. Caravaca, Á. Cuevas, and R. Cuevas, "This Is Taking Too Long - Investigating Time as a Proxy for Energy Consumption of LLMs," *arXiv preprint arXiv:2603.15699*, 2026.
- [22] Z. Niu et al., "TokenPowerBench: Benchmarking the Power Consumption of LLM Inference," *arXiv preprint arXiv:2512.03024*, 2025.
- [23] A. K. Kakolyris, D. Masouros, S. Xydis, and D. Soudris, "SLO-aware GPU DVFS for energy-efficient LLM inference serving," *IEEE Computer Architecture Letters*, vol. 23, no. 2, pp. 150–153, Jul.–Dec. 2024, doi: 10.1109/LCA.2024.3406038.
- [24] M. Husom et al., "The Price of Prompting: Profiling Energy Use in Large Language Models Inference," *arXiv preprint arXiv:2407.16893*, 2024.

- [25] N. Jegham, M. Abdelatti, and A. M. Hendawi, "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference," *arXiv preprint*, 2025.
- [26] X. Zhang et al., "m-LLM: A Multi-Dimensional Optimization Framework for LLM Inference on Mobile Devices," *IEEE Transactions on Mobile Computing*, 2025.
- [27] Y. Li et al., "Power for AI Data Centers: Energy Demand, Grid Impacts, Challenges and Perspectives," *Energies*, vol. 19, no. 3, p. 722, 2026.
- [28] R. Verdecchia et al., "Green AI: A systematic review and meta-analysis of its definitions, lifecycle models, hardware and measurement attempts," *arXiv preprint arXiv:2511.07090*, 2025.
- [29] R. Verdecchia, J. Sallou, and L. Cruz, "A systematic review of Green AI," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 4, Art. no. e1507, Jul./Aug. 2023, doi: 10.1002/widm.1507.
- [30] M. K. Kim, T. A. Yoo, and J. B. Chung, "Toward Sustainable AI: A Scoping Review of Carbon Footprint and Environmental Impacts Across Training and Inference Stages," *IEEE Access*, 2026.
- [31] S. Kakolyris et al., "LLMCO2: Advancing Accurate Carbon Footprint Prediction for LLM Inferences," in *Proc. Workshop on Sustainable Computer Systems (HotCarbon)*, 2025.
- [32] A. Apsan et al., "A Deployment-Aware Framework for Carbon- and Water- Efficient LLM Serving," *Sustainability*, vol. 17, no. 23, p. 10473, 2025.
- [33] J. Dauner et al., "Behind the Prompt: The Environmental Impact of LLM Inference," *Digital Commons @ Cal Poly*, 2025.