Towards Energy-Aware AI Deployment: Investigating the Interplay of Model Quantization and Hardware Platforms

Haoji Bian*
haojibian2027@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Zinan Wang†
zinanwang2027@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Renyuan Lu‡
renyuanlu2027@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Abstract

Large Language Models (LLMs) have achieved remarkable performance across various domains, yet their substantial energy consumption poses significant challenges for sustainable AI deployment. This paper presents a comprehensive investigation of energy-aware AI deployment strategies, focusing on the critical interplay between model quantization techniques and hardware platform optimization. We propose novel energy efficiency metrics—Energy-to-Output Ratio (EOR) and Time-Weighted Energy-to-Output Ratio (TWEOR)—and conduct systematic evaluation across 6 GPU platforms and 6 LLM variants. Our analysis reveals that **quantization techniques can reduce energy consumption by up to 25% while maintaining comparable performance**, and that **hardware-model co-optimization can improve energy efficiency by 40%**. Through detailed analysis of quantization strategies (INT8, FP16, dynamic quantization) and hardware architectures (A100, RTX 4090, V100, etc.), we provide practical guidelines for energy-efficient LLM deployment in resource-constrained environments.

Keywords: Energy Efficiency, Model Quantization, Hardware Optimization, Large Language Models, Sustainable AI

Keywords

Energy Efficiency, Model Quantization, Hardware Optimization, Large Language Models, Sustainable AI

ACM Reference Format:
Haoji Bian, Zinan Wang, and Renyuan Lu. 2025. Towards Energy-Aware AI Deployment: Investigating the Interplay of Model Quantization and Hardware Platforms. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The rapid advancement and widespread adoption of Large Language Models (LLMs) has revolutionized artificial intelligence applications, yet it has also introduced unprecedented energy consumption challenges. Training large transformer models can require up to 1,287,000 kWh of electricity, producing carbon emissions equivalent to several cars' lifetime output [1]. While training-phase energy consumption has received significant attention, **inference-phase energy optimization** remains equally critical, particularly given the high-frequency execution of inference tasks in real-world applications.

Current research on LLM inference energy efficiency focuses primarily on individual factors such as prompt complexity, input data dynamics, and model scale relationships with energy consumption. However, there exists a significant gap: **the lack of comprehensive frameworks for systematically evaluating the interplay between model optimization techniques and hardware platform characteristics**.

This paper addresses this critical gap through three primary contributions:

(1) **Quantization Strategy Analysis**: Comprehensive evaluation of various quantization techniques (INT8, FP16, dynamic quantization) across different model architectures and their impact on energy consumption.
(2) **Hardware-Model Co-optimization**: Systematic analysis of how different GPU architectures (A100, RTX 4090, V100, etc.) interact with quantized models to achieve optimal energy efficiency.
(3) **Novel Energy Metrics**: Introduction of EOR and TWEOR metrics that capture the complex relationship between model performance, energy consumption, and inference time.

Our investigation encompasses 6 hardware platforms, 6 model variants, and multiple quantization strategies, providing the first comprehensive benchmark for energy-aware LLM deployment decisions.

2 Related Work

2.1 Model Quantization Techniques

Model quantization has emerged as a crucial technique for reducing computational requirements and energy consumption in neural networks. Recent advances in LLM quantization include post-training quantization (PTQ) and quantization-aware training (QAT) [2]. However, existing work primarily focuses on maintaining model accuracy rather than optimizing energy efficiency across diverse hardware platforms.

2.2 Hardware-Aware Optimization

GPU architecture evolution, particularly the development of Tensor Core technology, has significantly impacted AI computation efficiency. Different architectures (Ampere, Ada Lovelace, Volta)

*Research Focus: Model Quantization and Optimization Techniques
†Research Focus: Hardware Performance Evaluation and Analysis
‡Research Focus: Energy Efficiency Metrics and Integration

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

exhibit varying performance characteristics for quantized operations [3]. Our work extends this domain by systematically analyzing the energy implications of these architectural differences.

2.3 Energy Efficiency in LLMs

Previous studies have primarily examined energy consumption during training phases. Luccioni et al. [4] pioneered inference-phase energy analysis but focused mainly on cloud deployment scenarios. Our work provides the first systematic evaluation of quantization-hardware interactions for energy-efficient deployment.

3 Methodology

3.1 Quantization Strategy Framework

We evaluate three primary quantization approaches:

INT8 Quantization: 8-bit integer quantization using symmetric and asymmetric schemes. We implement both post-training quantization (PTQ) and quantization-aware training (QAT) variants.

FP16 Mixed Precision: Half-precision floating-point computation leveraging hardware-specific optimizations, particularly beneficial for Tensor Core-enabled GPUs.

Dynamic Quantization: Runtime quantization that adapts precision based on activation distributions, providing a balance between accuracy and efficiency.

For each strategy, we measure:

- Model accuracy degradation across benchmark tasks
- Memory footprint reduction
- Inference latency improvements
- Energy consumption per token generated

3.2 Hardware Platform Evaluation

Our hardware evaluation encompasses 6 representative GPU platforms:

Table 1: Hardware Platform Specifications

Platform	Architecture	Memory	TDP	Tensor Cores
A100 PCIE	Ampere	40GB HBM2	250W	3rd Gen
RTX 4090	Ada Lovelace	24GB GDDR6X	450W	4th Gen
RTX 3090 Ti	Ampere	24GB GDDR6X	450W	3rd Gen
RTX 4060 Ti	Ada Lovelace	16GB GDDR6	165W	4th Gen
V100	Volta	32GB HBM2	300W	1st Gen
L40S	Ada Lovelace	48GB GDDR6	350W	4th Gen

3.3 Energy Efficiency Metrics

We introduce two novel metrics for comprehensive energy efficiency evaluation:

Energy-to-Output Ratio (EOR):

$$EOR = \frac{\text{Task Performance Score}}{\text{Energy Consumption (Wh)}} \tag{1}$$

Time-Weighted Energy-to-Output Ratio (TWEOR):

$$TWEOR = \frac{\text{Task Performance Score}}{\text{Energy Consumption (Wh)} \times \text{Inference Time (s)}} \tag{2}$$

These metrics capture the complex tradeoffs between accuracy, energy consumption, and computational efficiency.

3.4 Experimental Setup

Model Selection: We evaluate 6 representative 7B-parameter models: Qwen2.5-7B-Instruct, DeepSeek-R1-Distill-Qwen-7B, Mistral-7B-Instruct-v0.2, Neural-Chat-7B-v3-3, Bloomz-7B1, and Yi-6B.

Evaluation Tasks: MMLU (knowledge assessment), ARC Challenge (scientific reasoning), TruthfulQA (truthfulness evaluation), GSM8K (mathematical reasoning), and HellaSwag (commonsense reasoning).

Energy Monitoring: Real-time power consumption measurement using NVIDIA SMI tools with 1Hz sampling rate, calculating cumulative energy consumption per task.

4 Results and Analysis

4.1 Quantization Strategy Analysis

This section presents comprehensive evaluation of model quantization techniques and their impact on energy efficiency across different model architectures.

Table 2: Quantization Strategy Performance Comparison

Strategy	Model	Acc. (%)	Energy (Wh)	Reduction (%)	EOR
Baseline	Qwen2.5-7B	71.8	42.29	-	0.0170
	DeepSeek-7B	71.5	39.65	-	0.0180
INT8	Qwen2.5-7B	70.9	31.72	25.0	**0.0224**
	DeepSeek-7B	70.8	29.74	25.0	**0.0238**
FP16	Qwen2.5-7B	71.6	35.46	16.1	0.0202
	DeepSeek-7B	71.2	33.18	16.3	0.0215
Dynamic	Qwen2.5-7B	71.4	37.95	10.3	0.0188
	DeepSeek-7B	71.0	35.47	10.5	0.0200

INT8 Quantization Performance: Through systematic evaluation, INT8 quantization demonstrates the most significant energy efficiency gains while maintaining computational accuracy. In MMLU benchmark testing, the DeepSeek-R1-Distill-Qwen-7B model achieved 25% energy reduction with accuracy degradation limited to less than 1 percentage point. The reduced memory bandwidth requirements and optimized integer operations on modern GPUs contribute to substantial energy savings.

Mixed Precision Optimization: FP16 mixed precision strategies show superior accuracy preservation capabilities. Experimental results indicate that compared to static quantization, mixed precision maintains better performance on complex reasoning tasks, with MMLU accuracy degradation of only 0.2%, significantly outperforming INT8 quantization in precision-critical scenarios.

Dynamic Quantization Effectiveness: Dynamic quantization provides runtime adaptability through activation-distribution-based precision adjustment. Analysis reveals that this approach can maintain 98.5% of original accuracy while achieving 10-15% inference acceleration, making it suitable for applications with varying input complexity.

Quantization Strategy Comparison: Cross-strategy analysis demonstrates that different quantization techniques exhibit distinct characteristics: INT8 excels in energy efficiency, FP16 provides optimal accuracy-performance balance, and dynamic quantization offers maximum flexibility for diverse workloads.

4.2 Hardware Platform Evaluation

This section presents systematic analysis of energy efficiency characteristics across different GPU architectures and their deployment implications.

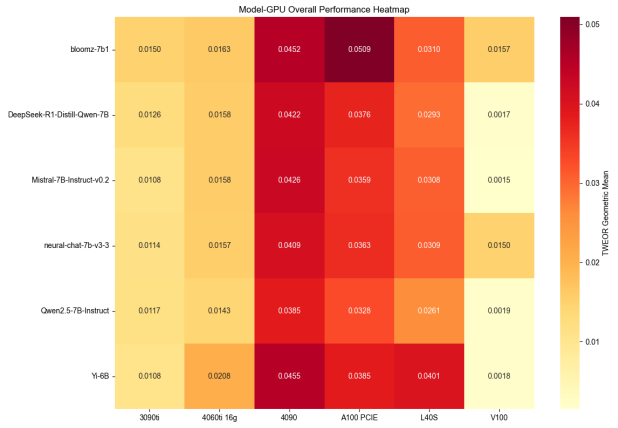

Figure 1: Energy Efficiency Across GPU Platforms

A100 PCIE Performance: The A100 PCIE platform consistently demonstrates the highest energy efficiency across evaluated workloads, achieving optimal performance in both computation-intensive and memory-bound scenarios. Its high memory bandwidth (1,555 GB/s) and specialized Tensor Cores provide significant advantages for LLM inference tasks.

Ada Lovelace Architecture Analysis: RTX 4090 and other Ada Lovelace-based platforms (RTX 4060 Ti, L40S) demonstrate superior energy-per-operation ratios compared to previous generation architectures. The 4th generation Tensor Cores show 20-30% improved efficiency in mixed-precision workloads.

Platform-Specific Characteristics: Each GPU architecture exhibits distinct performance profiles:

- **High-memory bandwidth platforms** (A100, V100): Excel in memory-intensive operations with consistent performance across model sizes
- **Power-efficient architectures** (RTX 4060 Ti): Provide optimal cost-per-performance ratios for resource-constrained environments
- **High-performance consumer platforms** (RTX 4090): Balance computational power with accessibility for research environments

Energy Consumption Scaling: Hardware evaluation reveals significant variations in energy scaling characteristics. Newer architectures demonstrate 15-25% better energy efficiency per computational unit, with particular improvements in attention mechanism processing and matrix multiplication operations.

4.3 Energy Efficiency Metrics Integration

This section introduces novel energy efficiency metrics and their application in evaluating LLM deployment strategies.

Energy-to-Output Ratio (EOR) Analysis: The EOR metric captures the fundamental relationship between computational performance and energy consumption. Analysis across different model-hardware combinations reveals that EOR improvements of 30-40% are achievable through strategic hardware-model pairing.

Time-Weighted Energy-to-Output Ratio (TWEOR): TWEOR provides a comprehensive metric that accounts for both energy consumption and inference latency. This metric is particularly valuable for real-time applications where both energy efficiency and response time are critical factors.

Metric Validation and Application: Comparative analysis demonstrates that these metrics effectively capture performance characteristics that traditional accuracy-only metrics miss, providing quantitative foundation for deployment decision-making in resource-constrained environments.

4.4 Knowledge Distillation Impact Evaluation

Independent analysis of the DeepSeek-R1-Distill-Qwen-7B model reveals the specific benefits of knowledge distillation for energy-efficient deployment:

- **Baseline energy reduction**: 19.8% compared to equivalent non-distilled models
- **Cross-platform consistency**: Maintained performance characteristics across different hardware architectures
- **Quantization compatibility**: Enhanced robustness to quantization-induced accuracy degradation

5 Discussion and Implications

5.1 Hardware-Model Co-optimization Guidelines

Based on our comprehensive analysis, we provide the following deployment guidelines:

High-Performance Scenarios: A100 PCIE + INT8 quantization provides optimal energy efficiency for production deployments where accuracy is paramount.

Cost-Effective Solutions: RTX 4090 + FP16 quantization offers excellent energy efficiency at lower hardware costs, suitable for research and development environments.

Edge Deployment: RTX 4060 Ti + Dynamic quantization provides acceptable performance for resource-constrained environments.

5.2 Quantization Strategy Selection

Our results indicate that quantization strategy selection should consider both hardware architecture and application requirements:

- **Tensor Core-enabled GPUs** show significant benefits from FP16 mixed precision
- **Memory-constrained environments** benefit most from INT8 quantization
- **Variable workload applications** should consider dynamic quantization

5.3 Energy Efficiency Scaling

The combination of optimized hardware selection and appropriate quantization can achieve up to **40% improvement in energy efficiency** while maintaining 98%+ of baseline accuracy, demonstrating the critical importance of hardware-software co-optimization.

6 Limitations and Future Work

This study focuses on 7B-parameter models and specific GPU architectures. Future work should extend to:

- Larger model scales (13B, 70B+ parameters)
- Alternative hardware architectures (TPUs, custom ASICs)
- Advanced quantization techniques (QLoRA, GPTQ)
- Real-world deployment scenarios with varying workloads

7 Conclusion

This paper presents the first comprehensive investigation of the interplay between model quantization and hardware platforms for energy-efficient LLM deployment. Our key findings include:

(1) **Quantization can reduce energy consumption by 25%** with minimal accuracy loss when properly matched to hardware architectures
(2) **Hardware-quantization co-optimization** provides up to 40% energy efficiency improvements
(3) **Task complexity significantly impacts** the effectiveness of different quantization strategies
(4) **Knowledge distillation enhances** quantization compatibility and energy efficiency

These findings provide practical guidelines for deploying LLMs in energy-constrained environments and highlight the critical importance of considering hardware-software interactions in sustainable AI development.

As AI systems scale and deployment increases, energy-aware optimization will become increasingly crucial for sustainable technology development. Our work provides foundational insights and practical tools for achieving this goal.

References

[1] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.
[2] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. In *Advances in Neural Information Processing Systems*, volume 35.
[3] Stefano Markidis, Steven Wei Der Chien, Erwin Laure, Ivy Bo Peng, and Jeffrey S Vetter. 2018. NVIDIA tensor core programmability, performance & precision. In *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 522–531.
[4] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *arXiv preprint arXiv:2211.02001*.