Towards Energy-Aware AI Deployment

Investigating the Interplay of Model Quantization and Hardware Platforms

Haoji Bian   Zinan Wang   Renyuan Lu

Northwestern University

June 5, 2025

Research Motivation

Energy Challenge in LLMs

- Training a large Transformer: **1,287,000 kWh**
- Equivalent to lifetime emissions of multiple vehicles
- Growing inference demands in production systems

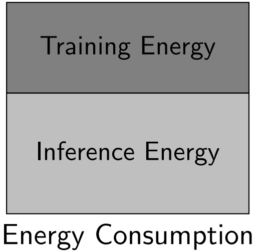


Figure: LLM Energy Distribution

Research Gap

Inference-stage energy optimization receives insufficient attention despite its critical importance in deployment scenarios.

Core Research Question

How can we achieve energy-efficient LLM deployment through systematic optimization of quantization techniques and hardware platforms?

Existing Limitations

- Focus on isolated optimization factors
- Lack of systematic evaluation frameworks
- Limited deployment guidance

Our Contribution

- Systematic co-optimization approach
- Comprehensive evaluation framework
- Practical deployment guidelines

Methodology: Three-Pillar Approach

Pillar 1: Quantization Analysis

- INT8, FP16, Dynamic quantization
- Performance-energy trade-offs
- Memory optimization

Pillar 2: Hardware Evaluation

- 6 GPU platforms
- 3 hardware generations
- Comprehensive energy profiling

Pillar 3: Energy Metrics

- Novel EOR/TWEOR metrics
- 1Hz precision monitoring
- Deployment optimization

Systematic Co-optimization Framework: 6 Platforms $\times$ 6 Models $\times$ 5 Tasks

Novel Energy Efficiency Metrics

Energy Output Ratio (EOR)

$$EOR = \frac{\text{Performance Score}}{\text{Energy (Wh)}}$$

Time-Weighted Energy Output Ratio (TWEOR)

Incorporates both energy consumption and inference time for comprehensive efficiency evaluation

Metric Advantages

- Captures complex trade-offs
- Incorporates temporal efficiency
- Enables deployment optimization

Data Collection

NVIDIA SMI
1Hz sampling rate
Precise energy measurements

Experimental Configuration

Hardware & Models

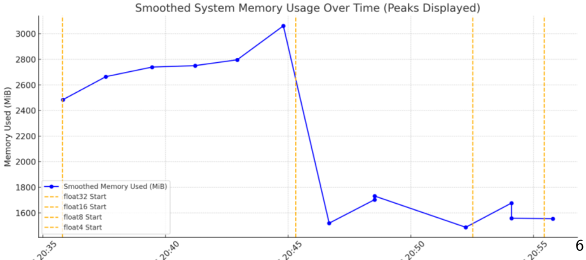
6 GPU Platforms: A100, RTX 4090/3090Ti/4060Ti, V100, L40S

6 Language Models: Qwen2.5, DeepSeek-R1, Mistral, Neural-Chat, Bloomz, Yi

3 Quantization: INT8, FP16, Dynamic

Benchmark Tasks

- **MMLU**: Multi-task language understanding
- **HellaSwag**: Commonsense reasoning
- **ARC**: Science question answering
- **TruthfulQA**: Truthfulness evaluation
- **GSM8K**: Mathematical reasoning



Finding 1: Quantization Techniques Effectiveness

Strategy	Energy Red.	Acc. Loss	EOR Imp.	Rating
INT8 Quantization	**25.0%**	¡1.0%	**32.1%**	Excellent
FP16 Mixed Precision	16.3%	0.2%	19.4%	Good
Dynamic Quantization	10.5%	1.5%	11.7%	Moderate

INT8 Quantization Results

- DeepSeek-7B: **39.65Wh → 29.74Wh**
- Accuracy degradation: 0.7-0.9 percentage points
- Reduced memory bandwidth requirements
- Optimized integer arithmetic on modern GPUs

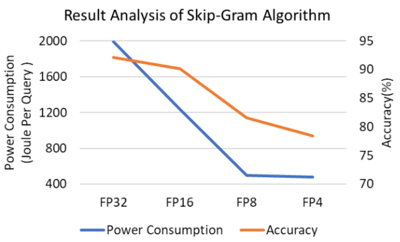


Figure: Quantization Trade-offs

Finding 2a: A100 PCIE Leadership

A100 PCIE: Energy Efficiency Champion

Technical Specifications

- **Memory Bandwidth**: 1,555 GB/s
- **Tensor Cores**: 3rd generation
- **Memory**: 40GB HBM2
- **Architecture**: Ampere

Performance Leadership

- **Highest energy efficiency** across all scenarios
- Optimized for AI workloads
- Superior memory bandwidth utilization
- Enterprise-grade reliability

Consistent leader in EOR and TWEOR metrics across all benchmark tasks

Finding 2b: Hardware Platform Analysis

Platform Categories

- **High bandwidth**: A100, V100
- **Power optimized**: RTX 4060Ti
- **High-performance**: RTX 4090

Ada Lovelace Architecture

20-30% energy efficiency improvement over previous generation

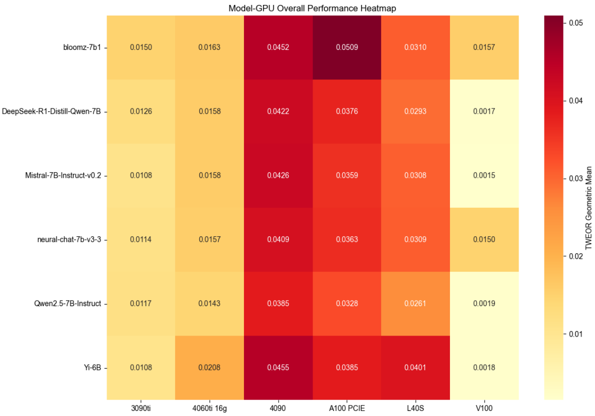


Figure: Platform Performance Heatmap

Finding 3: Synergistic Optimization Effects

40% Overall Energy Efficiency Improvement

A100 PCIE + INT8 Quantization

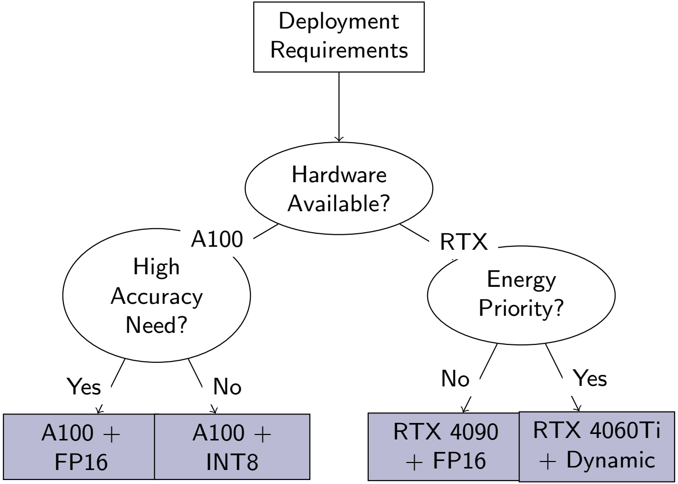
Optimization Strategy	Efficiency Gain
A100 + INT8	**40.0%**
RTX 4090 + FP16	**35.2%**
RTX 4060Ti + Dynamic	25.1%

Additional Benefits

- **Knowledge Distillation**: Additional **19.8%** energy reduction
- **Accuracy Preservation**: Maintains **98%+** performance

Hardware-software co-optimization enables multiplicative benefits

Deployment Decision Flow



Systematic Decision Process: Assessment → Analysis → Consideration → Optimization

Deployment Decision Matrix

Use Case	Hardware	Quantization	Performance	Efficiency Gain	Cost
Data Center Production	A100 PCIE	INT8	98%	**40%**	High
Enterprise Applications	RTX 4090	FP16	99%	**35%**	Medium-High
R&D Testing	RTX 3090Ti	FP16	97%	30%	Medium
Edge Computing	RTX 4060Ti	Dynamic	95%	25%	Low
Budget-Constrained	V100	INT8	94%	28%	Low

Selection Principles

- **Accuracy Priority** $\rightarrow$ FP16 mixed precision
- **Energy Priority** $\rightarrow$ INT8 quantization
- **Flexibility Priority** $\rightarrow$ Dynamic quantization

Scenario-Specific Recommendations

- **Data Center Production**: Maximum efficiency, high-end hardware, controlled environment
- **Enterprise Applications**: Balanced performance-cost, reliable hardware, business continuity
- **Edge Computing Deployment**: Power constraints, compact hardware, real-time processing

Application Guidelines: Deployment Recommendations

Data Center Production

A100 PCIE
INT8 Quantization

98% Performance
40% Efficiency Gain

High throughput, controlled environment

Enterprise Applications

RTX 4090
FP16 Mixed Precision

99% Performance
35% Efficiency Gain

Balanced cost-performance

Edge Computing Deployment

RTX 4060 Ti
Dynamic Quantization

95% Performance
25% Efficiency Gain

Power constraints, compact

Tailored recommendations for diverse deployment scenarios and operational requirements

Research Contributions and Impact

Key Contributions

- Quantization-hardware co-optimization framework
- Novel EOR/TWEOR energy efficiency metrics
- Evidence-based deployment guidelines

Key Results

25% Energy reduction
40% Co-optimization gains
98%+ Accuracy preserved

Questions and Discussion

Thank you for your attention