Are Large Language Models Economically Viable for Industry Deployment?

Abdullah Mohammad[1], Sushant Kumar Ray[2], Pushkar Arora[3], Rafiq Ali[4]
Ebad Shabbir[5], Gautam Siddharth Kashyap[6], Jiechao Gao[7*], Usman Naseem[8*]
[1, 3, 4, 5]DSEU-Okhla, New Delhi, India
[2]University of Delhi, New Delhi, India
[6, 8]Macquarie University, Sydney, Australia
[7]Center for SDGC, Stanford University, California, USA

Abstract

Generative AI—powered by Large Language Models (LLMs)—is increasingly deployed in industry across healthcare decision support, financial analytics, enterprise retrieval, and conversational automation, where reliability, efficiency, and cost control are critical. In such settings, models must satisfy strict constraints on energy, latency, and hardware utilization—not accuracy alone. Yet prevailing evaluation pipelines remain accuracy-centric, creating a *Deployment–Evaluation Gap*—the absence of operational and economic criteria in model assessment. To address this gap, we present EDGE-EVAL[1]—a industry-oriented benchmarking framework that evaluates LLMs across their full lifecycle on legacy NVIDIA Tesla T4 GPUs. Benchmarking LLaMA and Qwen variants across three industrial tasks, we introduce five deployment metrics—*Economic Break-Even* (N_{break}), *Intelligence-Per-Watt* (IPW), *System Density* (ρ_{sys}), *Cold-Start Tax* (C_{tax}), and *Quantization Fidelity* (Q_{ret})—capturing profitability, energy efficiency, hardware scaling, serverless feasibility, and compression safety. Our results reveal a clear efficiency frontier—models in the $< 2B$ *parameter class* dominate larger baselines across economic and ecological dimensions. *LLaMA-3.2-1B (INT4)* achieves ROI break-even in 14 requests (median), delivers $3\times$ higher energy-normalized intelligence than 7B models, and exceeds 6,900 tokens/s/GB under 4-bit quantization. We further uncover an efficiency anomaly—while QLoRA reduces memory footprint, it increases adaptation energy by up to $7\times$ for small models—challenging prevailing assumptions about quantization-aware training in edge deployment.

*Corresponding Author: jiechao@stanford.edu, usman.naseem@mq.edu.au

[1]https://github.com/Abdullah4152/EDGE-EVAL

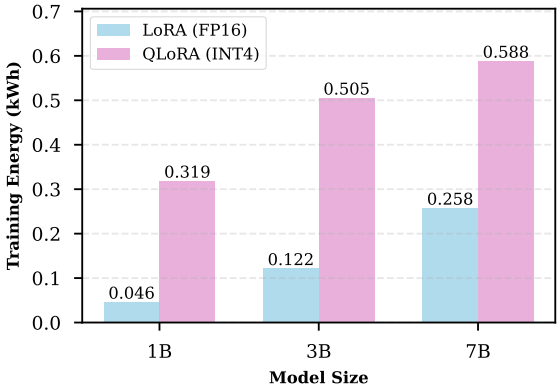


Figure 1: Illustration of the *Deployment–Evaluation Gap*–QLoRA reduces memory by $\sim 60\%$ yet increases fine-tuning energy up to $7.2\times$ for small models, showing that memory efficiency does not equal energy efficiency.

1 Introduction

Generative AI—powered by Large Language Models (LLMs) (Ciubotaru, 2025)—is rapidly transitioning from research prototypes to real-world industry deployment. Across healthcare decision support (Almadani et al., 2025), financial analytics (Al-Jumaili et al., 2023), enterprise retrieval (Hasan and Akter, 2022), and conversational automation (Manolescu et al., 2025), these models must operate under strict constraints on energy, latency, cost, and hardware availability. In such environments, practical viability depends not only on predictive accuracy, but also on economic sustainability. Despite this, prevailing evaluation pipelines remain dominated by accuracy-centric benchmarks (Siddiqui et al., 2025; Joshi et al., 2025; Hendrycks et al., 2020). These benchmarks provide limited insight into operational trade-offs, creating what we term the *Deployment–Evaluation Gap*—the absence of operational and economic criteria in model assessment. Figure 1 illustrates a motivating example of this gap. While Quantized Low-Rank

Adaptation (QLoRA) (Dettmers et al., 2023) reduces memory usage by approximately 60%, it increases fine-tuning energy consumption by up to 7.2× for small models compared to standard LoRA (Hu et al., 2022). Memory efficiency, therefore, does not necessarily translate to energy efficiency. Such trade-offs remain invisible under accuracy-centric benchmarks, yet they critically impact real-world deployment decisions.

To address this gap, we introduce EDGE-EVAL—an industry-oriented benchmarking framework that evaluates language models across their full operational lifecycle. We conduct a empirical study of LLaMA (Grattafiori et al., 2024) and Qwen (Qwen et al., 2025) variants across three industrial tasks—Summarization (long-context document compression under constrained inference budgets), Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) (retrieval-grounded reasoning with external knowledge integration), and Conversational Agents (latency-sensitive, multi-turn instruction-following dialogue)—on widely deployed NVIDIA Tesla T4 hardware. Our methodology integrates LoRA, QLoRA, and vLLM-based inference (Kwon et al., 2023). Furthermore, EDGE-EVAL extends conventional benchmarking through five deployment metrics–(1) *Economic Break-Even (N_{break})*—the traffic volume required for local adaptation to undercut API costs; (2) *Intelligence-Per-Watt (IPW)*—performance normalized by energy consumption (Schizas et al., 2022; Patterson et al., 2022); (3) *System Density (ρ_{sys})*—throughput per gigabyte of VRAM; (4) *Cold-Start Tax (C_{tax})*—the energy penalty of model loading; and (5) *Quantization Fidelity (Q_{ret})*—reasoning retention under 4-bit compression. In summary, our work makes two primary contributions:

- We introduce EDGE-EVAL, a benchmarking framework that augments accuracy-centric evaluation with five deployment metrics (N_{break}, IPW, ρ_{sys}, C_{tax}, Q_{ret}) for lifecycle assessment of LLMs on legacy hardware.
- Empirically, EDGE-EVAL on LLaMA and Qwen variants on Tesla T4 GPUs, we identify an efficiency frontier where $<$ 2B models outperform larger baselines, and reveal an energy anomaly—QLoRA increases adaptation energy by up to 7× despite reducing memory footprint.

2 Related Work

Model Compression. Recent advances in model compression and Parameter-Efficient Fine-Tuning (PEFT) have enabled LLMs to operate on resource-constrained hardware. Post-Training Quantization (PTQ) models such as Generalized Post-Training Quantization (GPTQ) (Frantar et al., 2022) and Activation Aware Quantization (AWQ) (Lin et al., 2024) reduce numerical precision while preserving accuracy, and Quantization-Aware Training (QAT) models such as QLoRA (Dettmers et al., 2023) integrate low-bit quantization into adaptation. Similarly, LoRA (Hu et al., 2022) and related PEFT models freeze base weights and introduce low-rank adapters to reduce memory overhead (Lialin et al., 2023). While these models demonstrate strong accuracy-centric. Yet, systematic analysis of lifecycle energy consumption, economic trade-offs, and infrastructure-level efficiency remains limited. In particular, memory reduction is often implicitly treated as a proxy for deployment efficiency—an assumption our empirical results challenge.

Model Evaluation. Green AI initiatives advocate reporting energy consumption alongside accuracy (Schizas et al., 2022; Patterson et al., 2022), such as MLPerf Tiny (Banbury et al., 2021) evaluate inference on ultra-low-power devices. Other studies profile latency or throughput for quantized inference (Yao et al., 2022), yet typically omit economic viability, cold-start overhead, and hardware density considerations. Existing benchmarks therefore lack a unified framework that connects adaptation cost, inference energy, quantization fidelity, and return on investment. In contrast, EDGE-EVAL introduces a lifecycle-oriented evaluation paradigm through five deployment metrics (N_{break}, IPW, ρ_{sys}, C_{tax}, Q_{ret}), providing a systematic assessment of operational and economic criteria in model assessment.

3 Methodology

EDGE-EVAL (see Figure 2) benchmarks language models through a structured lifecycle pipeline that mirrors real-world deployment. Let $\mathcal{F}$ denote the set of model families (LLaMA, Qwen) described in Section 4, $\mathcal{P}$ the set of parameter tiers (Micro $<$ 2B, Compact 3B, Standard 7B–8B), $\mathcal{T}$ the set of industrial tasks defined in Section 4.1 (Summarization, RAG, Conversational Agents), and $\mathcal{A}$ the set of adaptation strategies (LoRA-FP16, LoRA-INT8, LoRA-INT4, QLoRA-INT4). For each configura-

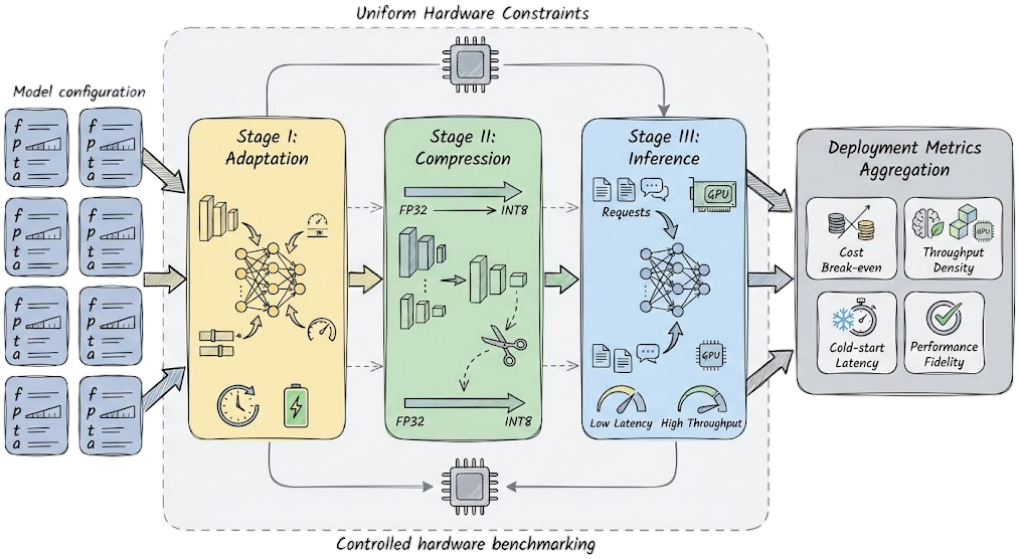


Figure 2: Lifecycle benchmarking pipeline of EDGE-EVAL. For each configuration (f, p, t, a), models pass through three stages—*adaptation*, *compression*, and *inference*—under uniform hardware constraints. The recorded lifecycle variables are subsequently aggregated into the five deployment metrics defined in Section 4.2.

tion $(f, p, t, a) \in \mathcal{F} \times \mathcal{P} \times \mathcal{T} \times \mathcal{A}$, we execute a full deployment pipeline consisting of adaptation, compression (when applicable), and serving. This factorial design yields $|\mathcal{F}| \times |\mathcal{P}| \times |\mathcal{T}| \times |\mathcal{A}| = 72$ benchmarked variants. During the *adaptation stage*, models are specialized on task-specific data using PEFT. In the *compression stage*, weights are optionally quantized under controlled precision regimes. In the *inference stage*, adapted models are deployed in a serving environment representative of low-batch industry conditions. For each stage, we record lifecycle variables including training energy E_{train}, inference energy per request E_{infer}, model loading overhead E_{load}, sustained throughput T_{put}, latency characteristics $(T_{\text{TTFT}}, T_{\text{ITL}})$, and GPU memory footprint M_{vram}. These measured quantities collectively characterize both one-time specialization cost and recurring operational behavior under uniform hardware constraints. The recorded variables are subsequently aggregated into five deployment metrics—N_{break}, IPW, ρ_{sys}, C_{tax}, and Q_{ret}—whose formal definitions and mathematical formulations are provided in Section 4.2.

4 Experimental Setup

All experiments are conducted on a dual-GPU node equipped with NVIDIA Tesla T4 accelerators (16GB VRAM each) (Nvidia, 2018; Jia et al., 2019), reflecting widely deployed legacy industry hardware. We evaluate two open-weight model families—LLaMA (Grattafiori et al., 2024) (1B[2], 3B[3], 8B[4]; employing Grouped-Query Attention) and Qwen-2.5 (Qwen et al., 2025) (1.5B[5], 3B[6], 7B[7]; dense transformer variants)—across three parameter tiers. Models are adapted using PEFT with rank $r = 16$ and scaling factor $\alpha = 32$, under four precision configurations–LoRA-FP16, LoRA-INT8, LoRA-INT4 (PTQ), and QLoRA-INT4 (Hu et al., 2022; Dettmers et al., 2023). Inference is served via vLLM (v0.6.3) (Kwon et al., 2023) with paged attention enabled, and evaluated under batch size 1 to simulate low-batch deployment conditions. Throughput (tokens/s), Time-To-First-Token (TTFT), and Inter-Token Latency (ITL) are measured over 100 independent requests per configuration, while GPU power draw is recorded using `pynvml` (Bauer et al., 2024) at 100 ms intervals to enable fine-grained lifecycle energy profiling.

4.1 Dataset Analysis

To evaluate EDGE-EVAL, we use three representative datasets—`Summarization (XSum)` (Narayan

[2] https://huggingface.co/meta-LLaMa/LLaMa-3.2-1B-Instruct
[3] https://huggingface.co/meta-LLaMa/LLaMa-3.2-3B-Instruct
[4] https://huggingface.co/meta-LLaMa/LLaMa-3.1-8B-Instruct
[5] https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct
[6] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct
[7] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

Family	Model Size	ROI Velocity (N_{break})	Green Efficiency (IPW)	System Density (ρ_{sys})	Quantization Fidelity (Q_{ret})	Cold-Start Tax (C_{tax})
LLaMa	**1B**	**14 Reqs**	0.45	**6,930 Tok/s/GB**	**100.6%**	**183x**
	3B	33 Reqs	0.27	1,336 Tok/s/GB	99.8%	184x
	7B	43 Reqs	0.15	387 Tok/s/GB	100.3%	230x
Qwen	1B	21 Reqs	**0.48**	6,942 Tok/s/GB	99.6%	179x
	3B	28 Reqs	0.23	1,419 Tok/s/GB	97.3%	188x
	7B	39 Reqs	0.14	394 Tok/s/GB	99.5%	237x

Table 1: Lifecycle efficiency frontier on legacy T4 hardware. Median INT4 results (20 runs, three tasks) across N_{break}, IPW, ρ_{sys}, Q_{ret}, and C_{tax} show that compact ($<$ 2B) models consistently outperform larger tiers in ROI velocity, system density, and energy-normalized intelligence.

et al., 2018), `RAG (SQuAD)` (Rajpurkar et al., 2016), and `Conversational Agent (UltraChat)` (Ding et al., 2023). `XSum` contains ∼227K news articles (204K/11K/11K train/val/test) paired with single-sentence summaries—modeling long-context document compression under constrained inference budgets. `SQuAD` v1.1 provides ∼100K QA pairs (87K/10K train/val), adapted into a retrieval-grounded generation setup to simulate knowledge-intensive enterprise reasoning. `UltraChat` comprises ∼1.5M multi-turn dialogues—reflecting latency-sensitive conversational deployment. We follow an 70/15/15 train/validation/test split, limiting fine-tuning to 5K–10K training samples per task while reserving the full validation and test sets for evaluation.

Family	Size	Method	Median Energy (kWh)	Ratio
LLaMa 3.2	1B	LoRA-FP16	0.039 [0.025-0.045]	1.0×
		QLoRA-INT4	0.251 [0.231-0.355]	6.4×
LLaMa 3.2	3B	LoRA-FP16	0.171 [0.119-0.244]	1.0×
		QLoRA-INT4	0.511 [0.156-0.612]	3.0×
LLaMa 3.1	7B	LoRA-FP16	0.244 [0.235-0.251]	1.0×
		QLoRA-INT4	0.552 [0.463-0.691]	2.3×
Qwen 2.5	1.5B	LoRA-FP16	0.129 [0.096-0.185]	1.0×
		QLoRA-INT4	0.301 [0.295-0.433]	2.3×
Qwen 2.5	3B	LoRA-FP16	0.153 [0.120-0.177]	1.0×
		QLoRA-INT4	0.359 [0.326-0.408]	2.3×
Qwen 2.5	7B	LoRA-FP16	0.243 [0.082-0.386]	1.0×
		QLoRA-INT4	0.563 [0.492-0.633]	2.3×

Table 2: Adaptation energy asymmetry across precision regimes. Median training energy and carbon cost (20 runs, IQR shown) reveal that QLoRA—despite reducing VRAM—incurs up to 6.4× higher energy for 1B models, exposing a divergence between memory and carbon efficiency on legacy hardware.

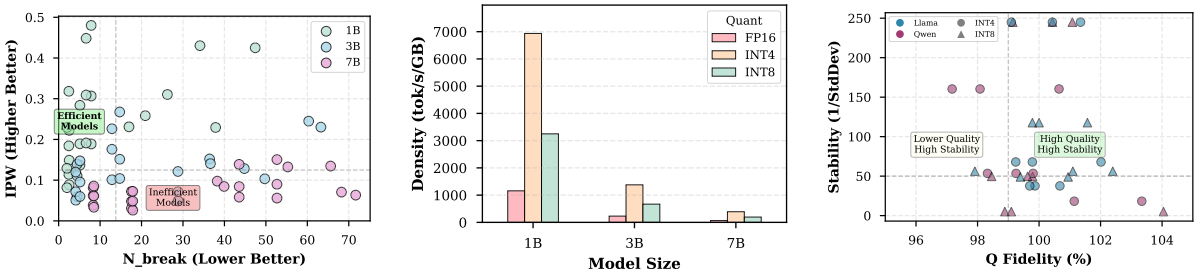
4.2 Evaluation Metrics

To evaluate EDGE-EVAL, we implement a three-pass evaluation loop and report mean, median, and standard deviation across task-specific metrics: for `RAG`, NLI Entailment (Context → Generation) (Honovich et al., 2022) and ROUGE-L (Lin, 2004); for `Summarization`, NLI Non-Contradiction and ROUGE-L; and for `Conversational Agents`, LLM-as-a-Judge (GPT-4o) ratings (Zheng et al., 2023) on Helpfulness and Safety (1–10 Likert scale). Based on the lifecycle variables defined in Section 3, we formalize five deployment metrics. *Economic Break-Even* computes the traffic volume required for local adaptation to undercut API costs, $N_{break} = \frac{C_{train}+C_{setup}}{C_{api}-C_{infer}}$, where C_{train} denotes adaptation cost, C_{setup} infrastructure overhead, C_{api} per-request API cost, and C_{infer} local inference cost per request. *Intelligence-Per-Watt* measures task-normalized reasoning efficiency, $IPW = \frac{\mathcal{S}_{task} \cdot \alpha}{E_{req}}$, where $\mathcal{S}_{task}$ is normalized task performance, α a task-complexity scaling factor, and E_{req} energy consumed per request. *System Density* quantifies hardware utilization efficiency, $\rho_{sys} = \frac{\mathcal{T}_{put}}{M_{vram}}$, where $\mathcal{T}_{put}$ denotes sustained token throughput and M_{vram} allocated GPU memory in GB. *Cold-Start Tax* captures the relative energy overhead of model loading, $C_{tax} = \frac{E_{load}}{E_{infer}}$, with E_{load} representing model loading energy and E_{infer} steady-state inference energy. Finally, *Quantization Fidelity* measures reasoning retention under 4-bit compression, $Q_{ret} = \left(\frac{\mathcal{S}_{INT4}}{\mathcal{S}_{FP16}}\right) \times 100\%$, where $\mathcal{S}_{INT4}$ and $\mathcal{S}_{FP16}$ denote task scores under INT4 and FP16 precision, respectively.

5 Result Analysis

5.1 Benchmark Analysis

The results in Table 1 reveal a clear efficiency frontier favoring the $<$ 2B *parameter class* across economic, ecological, and infrastructure dimensions. The LLaMa-3.2-1B configuration achieves the fastest ROI velocity (14 requests), the highest system density (6,930 tokens/s/GB), and near-



(a) ROI–Efficiency Quadrant (b) System Density (c) Quality–Stability Trade-off

Figure 3: Multidimensional efficiency under legacy deployment–compact ($<$ 2B) models form the ROI–IPW efficiency frontier, INT4 enables a strong throughput-per-GB hardware multiplier, and quantization fidelity remains stable with controlled variance.

perfect quantization fidelity (100.6%), while maintaining competitive green efficiency. Similarly, Qwen-1.5B attains the highest IPW (0.48), confirming that compact models maximize reasoning-per-watt under constrained hardware. In contrast, 7B models exhibit 3–5$\times$ lower density and slower economic recovery despite marginal quality gains, indicating diminishing deployment returns at scale. Table 2 further exposes a critical lifecycle asymmetry–although QLoRA reduces memory footprint, it increases adaptation energy by up to 6.4$\times$ for 1B models and $\sim$ 2–3$\times$ for larger tiers, demonstrating that memory efficiency does not guarantee carbon efficiency. Notably, post-training quantization incurs negligible overhead ($<$ 0.2% of training energy), reinforcing its practicality for inference optimization.

5.2 Multidimensional Efficiency Dynamics

Figure 3 synthesizes economic, infrastructure, and quality dimensions of deployment viability. The ROI–Efficiency quadrant (see Fig. 3a) reveals a pronounced efficiency frontier, where LLaMa-3.2-1B occupies the top-left region—achieving rapid break-even (median: 14 requests) while sustaining up to 3$\times$ higher intelligence-per-watt than 7B baselines—indicating that compact models maximize both capital recovery speed and energy-normalized reasoning. The infrastructure density analysis (see Fig. 3b) further demonstrates a hardware multiplier effect: INT4-quantized 1B models exceed 6,900 tokens/s/GB, representing up to 17$\times$ higher service capacity relative to 7B variants, thereby transforming legacy Tesla T4 accelerators into high-throughput inference nodes. Finally, the quality–stability trade-off (see Fig. 3c) shows that LLaMa models maintain near-perfect quantization fidelity (99%–101%) with controlled variance

Family	Size	Precision	Throughput	Speedup	Energy/req	Savings
LLaMa 3.2	1B	FP16	2,235	1.0x	6.45 J	-
		INT4	**4,331**	**1.94x**	**2.50 J**	**61%**
LLaMa 3.2	3B	FP16	1,374	1.0x	12.67 J	-
		INT4	**2,506**	**1.82x**	**5.39 J**	**57%**
Qwen 2.5	7B	FP16	948	1.0x	20.68 J	-
		INT4	**1,723**	**1.82x**	**8.90 J**	**57%**

Table 3: Inference efficiency under INT4 on Tesla T4–1.8–1.9$\times$ throughput gains and 57%–61% energy reduction versus FP16 (median over 20 runs, three tasks), confirming low-bit inference as a hardware multiplier under constrained infrastructure.

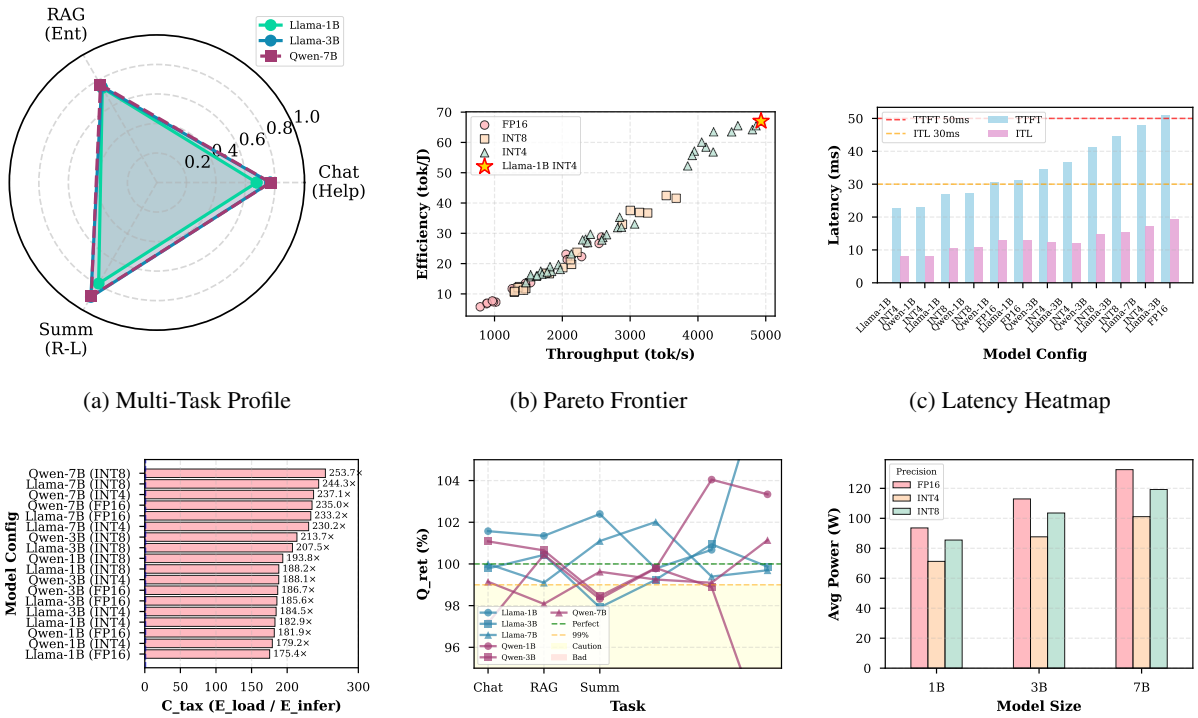
Family	Task	FP16 Score	INT4 Score	Retention	STD Δ
LLaMa	Chat	7.31 ± 0.04	7.32 ± 0.04	**100.1%**	-6.1%
	RAG	0.75 ± 0.01	0.75 ± 0.01	**100.4%**	$+65.9\%$
	Summ	0.86 ± 0.02	0.85 ± 0.02	98.6%	-9.7%
Qwen	Chat	7.42 ± 0.10	7.25 ± 0.14	97.7%	$+45.6\%$
	RAG	0.77 ± 0.01	0.76 ± 0.01	98.7%	$+20.5\%$
	Summ	0.84 ± 0.01	0.85 ± 0.02	**100.2%**	$+25.3\%$

Table 4: Quantization fidelity under INT4 across 400 evaluation runs per task (20$\times$20). Retention = (INT4/FP16)$\times$100% and STD Δ captures variance shift, showing near-lossless retention for compact models with architecture-dependent stability sensitivity.

shifts, whereas larger and denser configurations exhibit increased output instability, particularly in conversational settings.

5.3 Inference and Quality Trade-offs

Tables 3–5 jointly characterize the inference–quality balance under INT4 deployment. The Inference Performance Matrix (see Table 3) demonstrates a consistent hardware multiplier effect–INT4 nearly doubles throughput (1.82-1.94$\times$ speedup) while reducing per-request energy by 57%–61%, confirming that low-bit inference materially improves energy-normalized service capacity on Tesla T4 hardware. Therefore, these efficiency gains do not systematically degrade



(a) Multi-Task Profile (b) Pareto Frontier (c) Latency Heatmap

(d) Cold-Start Tax (e) Quantization Fidelity (f) Power Stability

Figure 4: Systems-level deployment landscape on legacy T4 hardware. Compact (< 2B) INT4 models consistently dominate the efficiency frontier—balancing multi-task performance, Pareto-optimal throughput–energy trade-offs, low latency, $> 99\%$ quantization retention, and stable $\sim$35W power—while larger models remain lifecycle-dominated.

Task	Metric	Family	Size	FP16	INT4	Retention
Chat	Helpfulness (1-10)	**LLaMa**	**1B**	6.67 ± 0.08	6.72 ± 0.05	+0.8%
			3B	7.58 ± 0.01	7.57 ± 0.04	-0.2%
			7B	7.68 ± 0.02	7.67 ± 0.03	-0.1%
		Qwen	1B	7.19 ± 0.16	7.34 ± 0.17	+2.0%
			3B	7.46 ± 0.07	6.75 ± 0.23	**-9.5%**
			7B	7.62 ± 0.06	7.66 ± 0.02	+0.6%
RAG	Entailment (0-1)	**LLaMa**	**1B**	0.74 ± 0.01	0.76 ± 0.02	+2.7%
			3B	0.77 ± 0.01	0.77 ± 0.01	$\pm$0.0%
			7B	0.74 ± 0.00	0.73 ± 0.01	-1.4%
		Qwen	1B	0.76 ± 0.01	0.75 ± 0.01	-2.0%
			3B	0.77 ± 0.01	0.77 ± 0.01	+0.7%
			7B	0.79 ± 0.01	0.77 ± 0.01	-2.5%
Summ	ROUGE-L (0-1)	**LLaMa**	**1B**	0.77 ± 0.01	0.76 ± 0.01	-0.7%
			3B	0.90 ± 0.03	0.86 ± 0.02	-4.4%
			7B	0.92 ± 0.01	0.93 ± 0.02	+1.1%
		Qwen	1B	0.79 ± 0.01	0.79 ± 0.01	$\pm$0.0%
			3B	0.86 ± 0.01	0.86 ± 0.02	$\pm$0.0%
			7B	0.89 ± 0.01	0.89 ± 0.02	+0.6%

Table 5: Task-level quality under INT4 vs FP16 across Chat, RAG, and Summarization (400 runs per task). Median $\pm$ StdDev shows near-baseline retention for compact models, with deviations highlighting task-and architecture-specific sensitivity.

task quality. As shown in Table 4, LLaMa models retain 99%–101% performance across tasks with tightly controlled variance, whereas Qwen exhibits greater instability in conversational settings (+45.6% standard deviation shift), indicating model-dependent quantization sensitivity. The comprehensive benchmark in Table 5 further reveals that compact (1B-3B) models preserve or even slightly improve task scores under INT4, while occasional degradations (e.g., Qwen-3B Chat, 9.5%) remain localized rather than systemic.

5.4 Systems-Level Deployment Dynamics

Figures 4 shows the systems-level behavior underlying the efficiency frontier. The multi-task radar profile confirms that LLaMa-3.2-1B (INT4) maintains balanced performance across Chat, RAG, and Summarization, forming a near-circular capability shape indicative of stable generalization at low cost. The Pareto frontier further demonstrates structural dominance–1B configurations occupy the optimal lower-right region (high throughput, low energy), while 7B models remain Pareto-inferior regardless of quantization strategy. Latency analysis reveals a non-linear deployment advantage—INT4 reduces inter-token latency disproportionately for compact models, enabling sub-10ms ITL and TTFT below 50ms, thereby satisfying real-time interaction thresholds. Operational constraints reinforce this asymmetry–cold-start taxes of 179–237$\times$ render scale-to-zero economically infeasible for

low-traffic workloads, and although 4-bit quantization safely preserves reasoning fidelity (most configurations exceeding the 99% threshold), larger models still incur higher steady-state power draw. The power profile highlights an additional edge advantage–LLaMa-3.2-1B sustains a stable $\sim$35W envelope under inference, supporting fanless deployment and improved reliability in constrained environments.

6 Conclusion

This work introduced EDGE-EVAL, a lifecycle-oriented benchmarking framework designed to close the *Deployment–Evaluation Gap* in industry LLM assessment. Through LLaMA and Qwen variants across adaptation, quantization, and inference on legacy Tesla T4 hardware, we demonstrated that compact ($<$ 2B) models consistently dominate larger baselines in ROI velocity, energy-normalized intelligence, system density, and latency stability under INT4 deployment. Our findings reveal two key insights–(i) small models form a clear efficiency frontier under constrained infrastructure, and (ii) memory-efficient training (e.g., QLoRA) does not necessarily imply energy or carbon efficiency.

Limitations

Our work focuses on legacy NVIDIA Tesla T4 accelerators and low-batch deployment settings; results may differ on modern architectures (e.g., Hopper-class GPUs) or high-throughput cloud serving environments. We evaluate two model families and three industry-representative tasks, which, while diverse, do not exhaust the space of domain-specific workflows. Energy measurements rely on GPU-level power telemetry and may not capture full system-level overheads (e.g., CPU, networking). At last, economic assumptions (e.g., API pricing and carbon intensity factors) reflect current estimates and may evolve over time, affecting absolute break-even thresholds.

Ethics Statement

This work evaluates operational efficiency rather than proposing new model capabilities. All experiments are conducted on publicly available open-weight models and widely used benchmark datasets. Through emphasizing energy consumption, carbon footprint, and infrastructure efficiency, EDGE-EVAL aligns with responsible and sustainable AI principles. However, improved deployment efficiency may lower the barrier to large-scale model use; practitioners must ensure compliance with data governance, privacy, and responsible deployment standards in real-world industry applications.

References

Ahmed Hadi Ali Al-Jumaili, Ravie Chandren Muniyandi, Mohammad Kamrul Hasan, Johnny Koh Siaw Paw, and Mandeep Jit Singh. 2023. Big data analytics using cloud computing based frameworks for power management systems: Status, constraints, and future recommendations. *Sensors*, 23(6):2952.

Basem Almadani, Hunain Kaisar, Irfan Rashid Thoker, and Farouq Aliyu. 2025. A systematic survey of distributed decision support systems in healthcare. *Systems*, 13(3):157.

Colby Banbury, Vijay Janapa Reddi, Peter Torelli, Jeremy Holleman, Nat Jeffries, Csaba Kiraly, Pietro Montino, David Kanter, Sebastian Ahmed, Danilo Pau, and 1 others. 2021. Mlperf tiny benchmark. *arXiv preprint arXiv:2106.07597*.

Christian Bauer, Samira Afzal, Sandro Linder, Radu Prodan, and Christian Timmerer. 2024. Greem: An open-source energy measurement tool for video processing. In *Proceedings of the 15th ACM Multimedia Systems Conference*, pages 264–270.

Bogdan-Iulian Ciubotaru. 2025. Generative ai and large language models: A comprehensive scientific review.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Rakibul Hasan and Samia Akter. 2022. Information system-based decision support tools: A systematic review of strategic applications in service-oriented enterprises. *Review of Applied Science and Technology*, 1(04):26–65.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Zhe Jia, Marco Maggioni, Jeffrey Smith, and Daniele Paolo Scarpazza. 2019. Dissecting the nvidia turing t4 gpu via microbenchmarking. *Preprint*, arXiv:1903.07486.

Harsh Joshi, Gautam Siddharth Kashyap, Rafiq Ali, Ebad Shabbir, Niharika Jain, Sarthak Jain, Jiechao Gao, and Usman Naseem. 2025. Can argus judge them all? comparing vlms across domains. *arXiv preprint arXiv:2507.01042*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. 2023. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100.

Vasile Denis Manolescu, Hamzah AlZu'bi, and Emanuele Lindo Secco. 2025. Interactive conversational ai with iot devices for enhanced human-robot interaction. *Journal of Intelligent Communication*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

C Nvidia. 2018. Nvidia turing gpu architecture. *NVIDIA Whitepaper*, 1.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Nikolaos Schizas, Aristeidis Karras, Christos Karras, and Spyros Sioutas. 2022. Tinyml for ultra-low power ai and large scale iot deployments: A systematic review. *Future Internet*, 14(12):363.

Zohaib Hasan Siddiqui, Jiechao Gao, Ebad Shabbir, Mohammad Anas Azeez, Rafiq Ali, Gautam Siddharth Kashyap, and Usman Naseem. 2025. Llms on a budget? say hola. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1035–1043.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in neural information processing systems*, 35:27168–27183.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.