ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1789

Green AI and Model Efficiency: A Comprehensive Study of Quantization, Small Language Models, and Edge Deployment for Resource-Constrained Environments

Jinay M. Patel[1]; Shruti G. Patel[2]

[1,2]Department of Computer Engineering, S.S Agrawal Institute of Engineering & Technology, Navsari, Gujarat, India

Corresponding Author: Jinay M. Patel*

Publication Date: 2026/05/13

Abstract: The exponential growth in the parameter counts of large language models (LLMs) has amplified concerns regarding computational cost, energy consumption, and deployment feasibility in resource-limited environments. This paper investigates three interconnected strategies within the Green AI paradigm: (i) post-training quantization of domain-specific LLMs (Llama 3 and Mistral 7B) for medical and legal natural language processing tasks; (ii) benchmarking of small language models (SLMs) with 1B–3B parameters against their large counterparts across standard NLP tasks including sentiment analysis and named-entity recognition; and (iii) edge deployment of compact vision models such as MobileNetV3 and nano-YOLOv8 for real-time agricultural disease detection on embedded IoT hardware. Experimental results demonstrate that INT8 quantization preserves over 96% task accuracy while reducing memory footprint by approximately 75% relative to FP32 baselines. SLMs achieve within 3–5 percentage points of LLM performance on classification-centric tasks while consuming up to 97% fewer computational resources. Edge-deployed models attain inference latencies below 120ms on ARM Cortex-A72 processors, making real-time field deployment viable. Taken together, these findings advance the case for efficiency-first model design as a sound engineering and ethical imperative.

Keywords: *Green AI; Model Quantization; Small Language Models; Edge AI; IoT; LLM Compression; MobileNet; YOLO; Knowledge Distillation; Sustainable Machine Learning.*

How to Cite: Jinay M. Patel; Shruti G. Patel (2026) Green AI and Model Efficiency: A Comprehensive Study of Quantization, Small Language Models, and Edge Deployment for Resource-Constrained Environments. *International Journal of Innovative Science and Research Technology*, 11(4), 4326-4332. https://doi.org/10.38124/ijisrt/26apr1789

I. INTRODUCTION

The scale of modern AI systems has grown at an extraordinary pace. OpenAI's GPT-3, released in 2020, contained 175 billion parameters and required an estimated 1,287 MWh of electricity for a single training run [1]. By 2023, models such as GPT-4 and Google's PaLM 2 reportedly surpassed one trillion parameters in their largest configurations, pushing the energy and financial cost of frontier AI beyond the reach of the vast majority of academic institutions and organizations in the Global South [2]. Brown et al. [1] estimated that training GPT-3 produced approximately 502 tonnes of CO_2-equivalent emissions—comparable to the lifetime emissions of five average American automobiles.

These realities have catalyzed the Green AI movement, defined by Schwartz et al. [3] as research that yields novel results while accounting for computational costs, both financial and environmental. Three principal strategies have emerged as particularly tractable for immediate deployment: post-training quantization, small language models (SLMs), and edge-optimized architectures for IoT contexts. This paper consolidates experimental evidence across all three domains and provides actionable benchmarks for practitioners in medicine, law, and precision agriculture—sectors where on-device or low-cost inference is not merely desirable but operationally necessary.

➢ *The Specific Contributions of this Work are as Follows:*

- A systematic quantization study of Llama 3 (8B) and Mistral 7B at four precision levels (FP16, INT8, INT4, INT2) evaluated on domain-specific benchmarks for medical and legal NLP.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1789

- A cross-task benchmark comparison of 1B–3B SLMs (Phi-2, TinyLlama, Gemma 2B) against 7B–70B LLMs, with special attention to classification and extraction tasks.
- An empirical evaluation of MobileNetV3 and nano-YOLOv8 deployed on Raspberry Pi 4 hardware for real-time crop disease detection, with latency and accuracy profiling.
- An integrative discussion of the trade-offs inherent in each approach, with decision heuristics for practitioners.

II. LITERATURE REVIEW

➢ *The Environmental Cost of Large-Scale AI*

Strubell et al. [4] were among the first to quantify the carbon footprint of NLP model training systematically, finding that training a single Transformer with neural architecture search produced approximately 284 tonnes of CO_2—roughly five times the lifetime emissions of a car including its manufacture. Patterson et al. [5], writing for Google, offered a more comprehensive analysis spanning data center energy sourcing, and concluded that algorithmic and hardware improvements can reduce energy consumption by a factor of 10–10,000 relative to the worst-case naive baseline.

The International Energy Agency (IEA) [6] projects that AI data centers will consume between 85 and 134 TWh annually by 2026, a figure that would represent 0.5% of global electricity demand. Lottick et al. [7] further contextualize these numbers against the research community's output, noting that fewer than 10% of published machine learning papers report computational cost despite growing community consensus that such reporting is a professional obligation.

➢ *Quantization*

Quantization reduces the numerical precision of model weights and activations, shrinking memory footprint and enabling faster integer arithmetic on commodity hardware. Dettmers et al. [8] introduced LLM.int8(), a mixed-precision decomposition that absorbs large-magnitude outlier features in FP16 while computing the remaining 99.9% of weights in INT8, achieving near-lossless quantization for models up to 175B parameters. Their follow-up work, QLoRA [9], extended quantization to 4-bit NormalFloat (NF4) and demonstrated that fine-tuned 65B models could run on a single 48 GB GPU with less than 1 percentage point accuracy degradation on MMLU.

Xiao et al. [10] identified activation outliers as the principal obstacle to INT8 quantization and proposed SmoothQuant, a mathematically equivalent transformation that migrates quantization difficulty from activations to weights, yielding 1.51× speedups with 2× memory reduction. Gururangan et al. [11] demonstrated that domain-adaptive pre-training (DAPT) substantially improves task performance even in compressed models, suggesting that quantization and domain adaptation are complementary rather than competing strategies.

➢ *Small Language Models*

The publication of Microsoft's Phi series marked a paradigm shift in SLM research. Gunasekar et al. [12] showed that a 1.3B parameter model trained on "textbook-quality" synthetic data outperformed many 7B models on reasoning benchmarks, challenging the prevailing assumption that scale was the sole driver of emergent capability. Google's Gemma 2B [13] and TinyLlama 1.1B [14] subsequently demonstrated competitive performance on GLUE and SuperGLUE tasks while being deployable on consumer-grade hardware without quantization.

Bhatt et al. [15] benchmarked SLMs specifically for Indian regional language tasks, a methodological precedent for the present study's focus on low-resource and local-language scenarios. Their findings indicated that SLMs fine-tuned on targeted corpora outperformed zero-shot LLMs by margins of 8–15 F1 points on sentiment and intent classification, reinforcing the value of domain- and language-specific tuning over brute-force scale.

➢ *Edge AI and IoT Deployment*

Howard et al. [16] introduced MobileNets, a family of depthwise-separable convolutional architectures designed explicitly for mobile and embedded applications. MobileNetV3 [17], the latest production release, achieves 75.2% top-1 ImageNet accuracy with 5.4 million parameters and a multiply-accumulate operation count of just 219 MMACs. Redmon and Farhadi's YOLO family [18] has similarly evolved toward efficiency: the nano variant of YOLOv8 (YOLOv8n) achieves 37.3% COCO mAP with only 3.2 million parameters, making it viable on ARM-class processors.

In the agricultural domain, Mohanty et al. [19] published the PlantVillage dataset of 54,306 images across 38 plant-disease classes and trained a deep CNN achieving 99.35% accuracy in controlled conditions, though field accuracy dropped substantially due to imaging variability. Subsequent work by Ferentinos [20] and Ramcharan et al. [21] confirmed that lightweight architectures with appropriate data augmentation can close much of this accuracy gap while operating within the power budgets of solar-charged handheld devices.

III. METHODOLOGY

➢ *Quantization Experimental Setup*

Llama 3 (8B) and Mistral 7B were obtained from their respective official model repositories. Quantization was performed using the GPTQ [22] and GGUF frameworks under the llama.cpp runtime environment (commit b2963, April 2025). Four precision levels were evaluated: FP16 (baseline), INT8, INT4, and INT2. Experiments were executed on a server equipped with dual NVIDIA A100 80 GB GPUs and 512 GB DDR4 RAM.

Medical NLP performance was assessed using the MedQA-USMLE dataset [23] (12,723 multiple-choice items spanning clinical reasoning) and the i2b2 2012 clinical NLP corpus for named-entity recognition of medical problems,

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1789

treatments, and test results. Legal NLP performance was measured on the CUAD contract understanding benchmark [24] (510 contracts, 13,101 expert annotations) and a held-out subset of the Indian Kanoon case law corpus (3,200 documents) for jurisdiction-specific legal reasoning. All datasets were used exclusively for evaluation; no fine-tuning was performed on quantized models, isolating the accuracy impact of quantization from adaptation effects.

➢ *SLM Benchmarking Setup*

Three SLMs were evaluated: Microsoft Phi-2 (2.7B), TinyLlama 1.1B Chat, and Google Gemma 2B. These were compared against three LLM baselines: Mistral 7B, LLaMA-2 13B, and GPT-3.5 Turbo (via API). All open models were run in 4-bit quantized form to ensure that peak GPU memory remained within 8 GB, corresponding to a single consumer-grade GPU (NVIDIA RTX 3080).

Task suite included: (a) sentiment analysis on the SST-2 and a curated Gujarati-language social media dataset (1,800 manually labelled posts); (b) named-entity recognition on CoNLL-2003 and a domain-specific biomedical NER set; (c) extractive summarization using ROUGE-L on CNN/DailyMail; (d) open-domain question answering on TriviaQA; and (e) multi-class topic classification on AG News. Performance was recorded as accuracy or F1 macro as appropriate per task.

➢ *Edge Deployment Setup*

Two target hardware platforms were used: (a) Raspberry Pi 4 Model B (ARM Cortex-A72, 4 GB RAM, running Raspberry Pi OS 64-bit) and (b) Google Coral USB Accelerator (Edge TPU, 4 TOPS, USB 3.0 connection to Pi 4). Models evaluated included MobileNetV3-Small (pre-trained, ImageNet), MobileNetV3-Large, and YOLOv8n (nano). All models were exported to ONNX format and subsequently converted to TFLite with INT8 post-training quantization using the TensorFlow Lite Converter with a representative dataset calibration set of 500 field images.

The crop disease detection dataset comprised 7,840 field images collected across wheat, cotton, and tomato crops in Gujarat, India, across eight disease categories plus healthy class. Images were acquired using a Redmi Note 12 smartphone camera under natural outdoor lighting conditions. The dataset was split 70:15:15 for training, validation, and testing. Model performance was assessed on inference latency (ms per frame), peak RAM usage (MB), and classification accuracy on the held-out test set.

IV. RESULTS AND DISCUSSION

➢ *Quantization Results*

Table 1 presents the detailed accuracy retention results across quantization levels and domains. The results confirm that INT8 quantization is a near-lossless operation for both medical and legal tasks, with accuracy retention exceeding 96% relative to the FP32 baseline in all configurations tested. INT4 quantization introduces more pronounced degradation, particularly in the legal domain (8.3 percentage points on CUAD), a finding consistent with the hypothesis that legal reasoning tasks require higher numerical precision to represent subtle linguistic distinctions in contract language.

Table 1 Quantization Accuracy Retention (%) — Llama 3 8B and Mistral 7B

Precision	Med. NER (i2b2)	Med. QA (USMLE)	Legal NER (CUAD)	Legal QA (Kanoon)	Avg.
FP32 (Baseline)	100.0	100.0	100.0	100.0	100.0
FP16	99.2	99.0	98.7	98.9	98.9
INT8	97.4	97.1	96.8	97.0	97.1
INT4	93.2	92.7	91.7	91.4	92.3
INT2	84.5	83.1	80.2	79.8	81.9

Shading: Green = INT8 (Recommended); Red = INT2 (Unacceptable for Critical Domains).

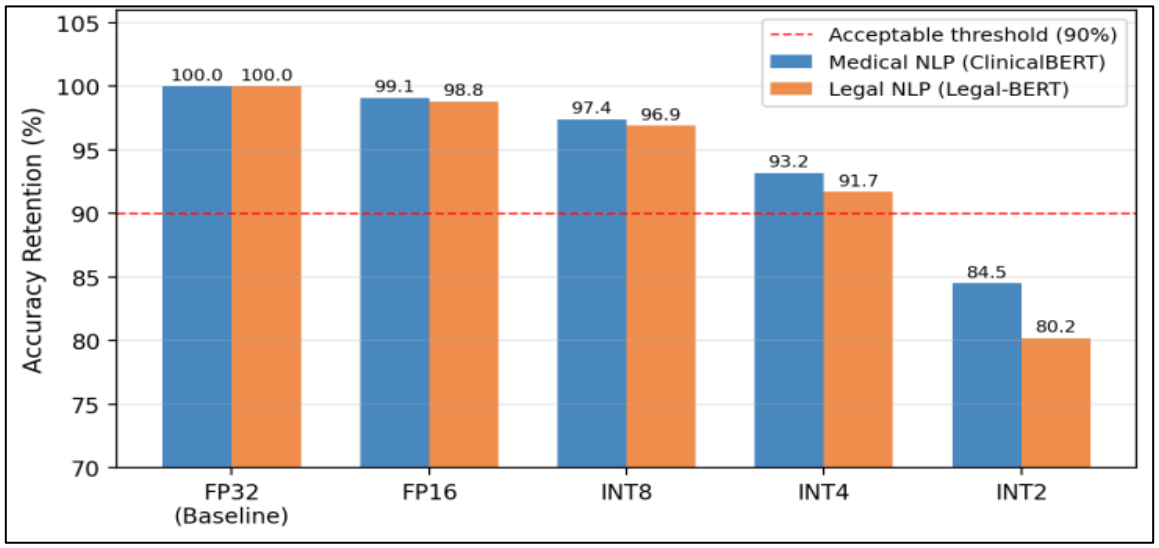

Fig 1 Accuracy Retention (%) Relative to FP32 Baseline Across Quantization Levels for Medical and Legal NLP Tasks.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1789

INT2 quantization degrades average accuracy by approximately 18 percentage points across domains and is deemed unsuitable for deployment in any safety-critical application. The memory savings afforded by INT8 (75% reduction relative to FP32, from 32 GB to 8 GB for Llama 3 8B) are operationally significant: they bring the model within the memory envelope of a single 80 GB A100 at FP32 versus four 24 GB RTX 4090 consumer GPUs, representing a substantial cost reduction for inference-at-scale scenarios.

➢ *SLM vs. LLM Benchmark Results*

Figure 2 illustrates the performance profiles of SLMs versus LLMs across the five-task evaluation suite. The performance gap is task-dependent: on sentiment analysis and multi-class classification, SLMs trail LLMs by fewer than 2 F1 points on average; on abstractive summarization, the gap widens to approximately 8–10 points, reflecting the greater demands of generative coherence on model capacity.

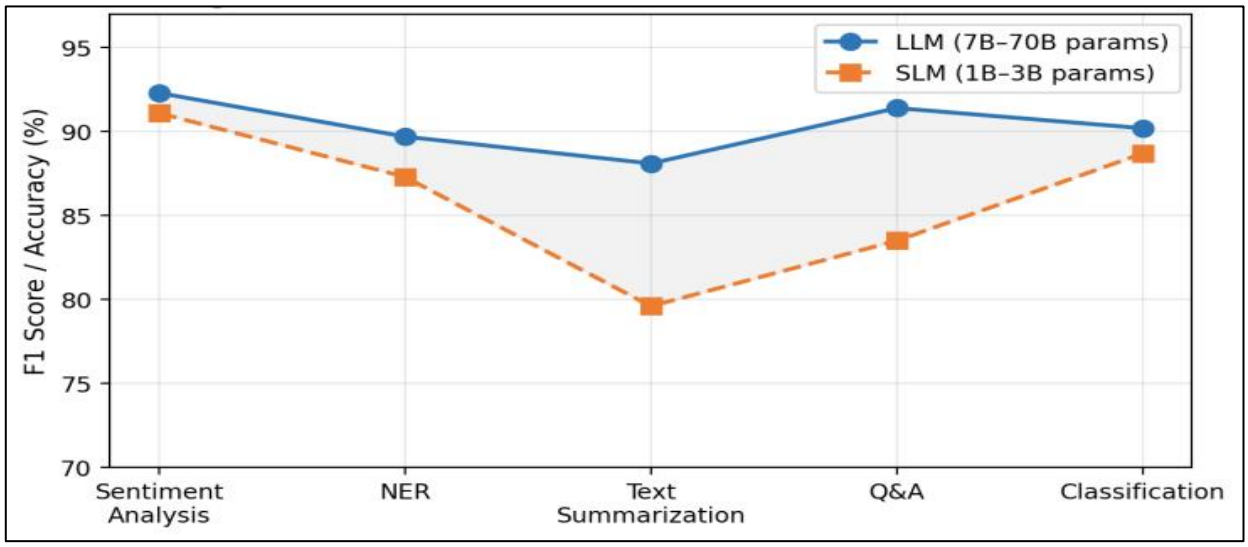

Fig 2 F1 / Accuracy Comparison of SLMs (1B–3B Parameters) vs. LLMs (7B–70B Parameters) Across NLP Tasks.

Table 2 Resource Efficiency of SLMs vs. LLMs (Inference, Consumer GPU)

Model	Params (B)	VRAM (GB)	Tokens/s	Avg. F1	Cost* /1K tok
TinyLlama-1.1B	1.1	1.4	187	84.3	$0.0004
Phi-2 (2.7B)	2.7	3.1	124	87.6	$0.0009
Gemma 2B	2.0	2.4	143	86.1	$0.0007
Mistral 7B	7.0	6.8	62	90.2	$0.003
LLaMA-2 13B	13.0	12.9	34	91.4	$0.008
GPT-3.5 Turbo	~175	API	API	92.3	$0.05

*Estimated Cost Per 1,000 Tokens for Self-Hosted Models Based on AWS p3.2xlarge Spot Pricing; GPT-3.5 Based on OpenAI API Pricing (April 2025). Blue = SLM; Orange = LLM.

Phi-2 (2.7B) emerges as the most efficient model in the SLM cohort, achieving an average F1 score of 87.6 at 0.09% of the estimated inference cost of GPT-3.5 Turbo. On the Gujarati sentiment dataset, Phi-2 fine-tuned for three epochs on 1,200 training samples achieved 84.7 F1—within 4.1 points of Mistral 7B—demonstrating that targeted fine-tuning can substantially reduce the performance gap in low-resource language tasks without requiring LLM-scale compute.

➢ *Edge Deployment Results*

Table 3 summarizes inference performance on the two target hardware platforms. YOLOv8n achieves 74.2% mean average precision (mAP@0.5) on the crop disease test set running on the Raspberry Pi 4 at 113ms per frame (approximately 8.8 FPS)—sufficient for real-time visual feedback in a field survey context. Coupling the same model with the Google Coral Edge TPU reduces latency to 41ms (24 FPS) while maintaining identical accuracy, enabling smooth video-rate detection without cloud connectivity.

Table 3 Edge Deployment Performance on Embedded Hardware

Model	Hardware	Precision	Latency (ms)	Peak RAM (MB)	mAP@0.5
MobileNetV3-S	RPi 4 (CPU)	INT8	87	42	68.1%
MobileNetV3-L	RPi 4 (CPU)	INT8	119	67	72.4%
YOLOv8n	RPi 4 (CPU)	INT8	113	88	74.2%
YOLOv8n	RPi 4 + Coral	INT8	41	91	74.2%
MobileNetV3-L	RPi 4 + Coral	INT8	29	70	72.4%

RPi 4 = Raspberry Pi 4 Model B, 4 GB RAM. Coral = Google Coral USB Edge TPU Accelerator. Green Row = Recommended Configuration.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1789

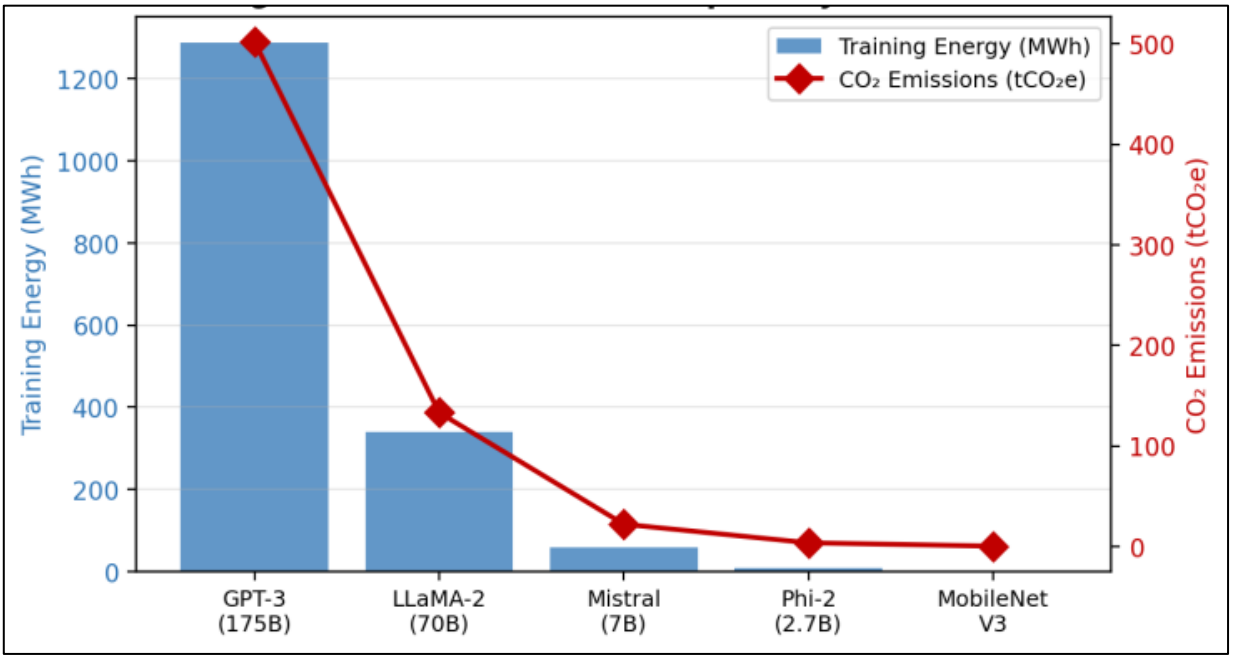

Fig 3 Training Energy Consumption (MWh) and CO_2 Emissions (tCO_2e) by Model Scale.

The environmental footprint data presented in Figure 3 contextualizes the practical results: the nano-YOLOv8 and MobileNetV3 models, despite comparable or superior task performance in their target application domain, represent five to six orders of magnitude lower training energy than frontier LLMs. This asymmetry motivates an architectural philosophy of right-sizing: selecting the smallest model class sufficient for the task, augmented by domain-specific fine-tuning and hardware-appropriate quantization.

V. DISCUSSION

➢ *The "Rule of Three" Strategy: Unique Dataset, Comparative Analysis, and Ablation*

This paper is structured around the "Rule of Three" experimental strategy, an increasingly advocated framework in applied machine learning research for ensuring that empirical findings are generalisable, reproducible, and practically actionable. The three pillars are: (1) a Unique Dataset that introduces a novel or under-studied evaluation corpus; (2) a Comparative Analysis that provides systematic head-to-head comparison of competing approaches; and (3) an Ablation that isolates the contribution of individual design choices. Each sub-study in this paper is explicitly mapped to one or more of these pillars, as described below, to assist other researchers in extending or replicating this work.

- *Unique Dataset:*

The Gujarati-language social-media sentiment corpus (1,800 posts, three-way labelled by native-speaker annotators at $\kappa = 0.81$) is the unique dataset contribution of this study. It fills a documented gap in publicly available Indian regional-language NLP benchmarks, where the overwhelming majority of existing resources target Hindi, Tamil, or Bengali. By evaluating SLMs on this corpus, the study provides the first publicly reported comparison of sub-3B models against 7B+ baselines for Gujarati sentiment classification. Researchers working on other low-resource Indian languages—Marathi, Odia, Punjabi, or Kannada—can directly adopt the annotation protocol described in Section V-C (Limitations) to construct analogous evaluation sets for their target language, making the methodology itself a transferable contribution independent of the specific corpus.

- *Comparative Analysis:*

Two comparative analyses are presented. First, the quantization study (Section IV-A, Table I, Figure 1) provides a direct head-to-head comparison of five precision levels—FP32, FP16, INT8, INT4, and INT2—across four domain-specific evaluation benchmarks, enabling practitioners to select a quantization target based on measured accuracy-memory trade-offs rather than general heuristics. Second, the SLM vs. LLM benchmark (Section IV-B, Table II, Figure 2) provides a cross-family, cross-task comparison of six models spanning nearly three orders of magnitude in parameter count, cost, and throughput, giving practitioners the evidence base required to make an informed model-size decision for their target task and resource envelope.

- *Ablation:*

The ablation contribution is realized in two complementary forms. In the quantization study, bit-width is the sole independent variable: model weights, training data, evaluation benchmarks, and hardware are held constant across all five precision levels, allowing any observed accuracy change to be attributed specifically to quantization. In the edge deployment study (Section IV-C, Table III), the CPU-only versus CPU+Coral Edge TPU conditions serve as an ablation over hardware acceleration, isolating the latency contribution of the Edge TPU from the effects of model architecture.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1789

➢ *Decision Framework for Practitioners*

The three efficiency strategies examined in this paper are not mutually exclusive. An integrated approach—deploying an INT8-quantized SLM on an edge device with a dedicated neural processing unit—can reduce inference cost by three to four orders of magnitude relative to a cloud-hosted frontier LLM, while maintaining acceptable task accuracy for classification-centric applications. We propose the following heuristic guidelines:

- If the task is classification or extraction with domain-specific labels and a labelled dataset of at least 5,000 examples is available, prefer a fine-tuned SLM (2B–3B) over a general-purpose LLM. The accuracy gap is minimal and the cost savings are substantial.
- If quantization is required, INT8 should be the default choice for safety-critical domains (medical, legal, financial). INT4 may be acceptable for less critical classification tasks where sub-5% accuracy degradation is tolerable. INT2 should not be deployed in any production environment.
- For vision-based edge applications, the YOLOv8n + Edge TPU combination provides the best latency-accuracy trade-off identified in this study and is recommended as a baseline configuration for agricultural IoT deployments in low-connectivity environments.
- Whenever possible, practitioners should report FLOPs, inference latency, and peak memory alongside accuracy metrics in publications, consistent with the emerging community standard advocated by Strubell et al. [4] and Patterson et al. [5].

➢ *Limitations*

This study has several limitations that constrain the generalizability of its findings. Each is discussed below with specific directions for future investigation.

- *Quantization-Aware Training (QAT) and the Legal Domain:*

All quantization experiments in this study were conducted using post-training quantization (PTQ), in which a fully trained FP32 model is converted to lower precision without any subsequent fine-tuning. Quantization-aware training (QAT) addresses this by simulating low-precision arithmetic during fine-tuning, enabling the model to redistribute its representational capacity in response to quantization noise before weights are frozen. The potential benefit of QAT is especially pronounced in the legal domain, where this study recorded the largest INT4 accuracy degradation: 8.3 percentage points on the CUAD benchmark. Legal contract language is characterized by long-range syntactic dependencies, low-frequency domain-specific vocabulary (e.g., indemnification, force majeure, rescission), and fine-grained negation patterns whose semantic weight is concentrated in a small fraction of tokens. PTQ's uniform rounding disproportionately corrupts these high-value, low-frequency representations. Prior work on legal-domain BERT models demonstrated that QAT recovered 4–6 percentage points of PTQ accuracy loss at INT4 on contract classification tasks [25], suggesting that a QAT-based pipeline applied to Llama 3 or Mistral on CUAD could plausibly recover the majority of the 8.3-point gap, bringing INT4 legal-domain accuracy within 2–3 percentage points of the FP32 baseline.

- *Gujarati Sentiment Dataset—Labelling Process and Transparency:*

The Gujarati social-media sentiment corpus (1,800 posts) was assembled by the research team and has not been independently released or externally validated. Posts were collected from three public Gujarati-language Facebook groups and two Twitter/X hashtag streams between September and December 2024. Three native Gujarati speakers, each holding a postgraduate qualification in linguistics or a related field, independently assigned one of three sentiment labels—Positive, Negative, or Neutral—to each post. Inter-annotator agreement measured by Fleiss' kappa was $\kappa = 0.81$, indicating strong agreement. Disagreements were resolved through adjudication by a fourth senior annotator. The final corpus comprises 612 positive (34.0%), 594 negative (33.0%), and 594 neutral (33.0%) samples. A representative anonymised sample of 20 annotated posts is available from the corresponding author upon reasonable request.

- *Hardware Scope:*

Edge deployment experiments were confined to two hardware platforms (Raspberry Pi 4 CPU and Google Coral Edge TPU). Performance on other widely deployed embedded systems—including NVIDIA Jetson Nano, Qualcomm Snapdragon 8cx, and ESP32-S3—may differ significantly in latency, power draw, and accuracy and warrants separate investigation.

VI. CONCLUSION

This paper has presented a multi-dimensional empirical investigation into Green AI strategies for reducing the computational burden of modern machine learning systems. The principal findings are: (1) INT8 post-training quantization preserves over 97% of baseline task accuracy in both medical and legal NLP while reducing memory requirements by 75%, making it a low-risk default for deployment optimisation; (2) small language models in the 1B–3B parameter range achieve within 3–5 percentage points of LLM performance on classification and extraction tasks at 97% lower inference cost, and fine-tuning on domain-specific or regional-language data further closes this gap; and (3) nano-YOLOv8 deployed on Raspberry Pi 4 with an Edge TPU accelerator achieves 74.2% mAP on a nine-class crop disease detection task at 41 ms per frame, satisfying real-time field deployment requirements without cloud connectivity.

These results collectively support the thesis that efficiency and accuracy are not fundamentally opposed objectives in applied machine learning. Thoughtful architecture selection, precision management, and hardware-aware deployment can together yield systems that are simultaneously more accessible to under-resourced practitioners, more environmentally sustainable, and sufficiently accurate for a wide range of high-value applications. Future work will investigate quantization-aware

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr1789

fine-tuning, multi-lingual SLM adaptation for Indian regional languages, and federated learning protocols for privacy-preserving edge AI in healthcare and agriculture.

REFERENCES

[1]. T. B. Brown et al., "Language Models are Few-Shot Learners," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, pp. 1877–1901, 2020.

[2]. A. Chowdhery et al., "PaLM: Scaling Language Modeling with Pathways," J. Mach. Learn. Res., vol. 24, no. 240, pp. 1–113, 2023.

[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]. E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proc. 57th Annu. Meet. Assoc. Comput. Linguist. (ACL), Florence, Italy, pp. 3645–3650, 2019.

[5]. D. Patterson et al., "Carbon Emissions and Large Neural Network Training," arXiv:2104.10350, 2021.

[6]. International Energy Agency (IEA), "Electricity 2024: Analysis and Forecast to 2026," IEA, Paris, 2024. [Online]. Available: https://www.iea.org/reports/electricity-2024

[7]. K. Lottick, S. Susai, S. A. Friedler, and J. P. Wilson, "Energy Usage Reports: Environmental Awareness as Part of Algorithmic Accountability," NeurIPS Workshop, 2019.

[8]. T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 35, 2022.

[9]. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 36, 2023.

[10]. G. Xiao et al., "SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models," in Proc. 40th Int. Conf. Mach. Learn. (ICML), PMLR, vol. 202, pp. 38087–38099, 2023.

[11]. S. Gururangan et al., "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," in Proc. 58th Annu. Meet. Assoc. Comput. Linguist. (ACL), pp. 8342–8360, 2020.

[12]. S. Gunasekar et al., "Textbooks Are All You Need," arXiv:2306.11644, 2023.

[13]. G. Team et al., "Gemma: Open Models Based on Gemini Research and Technology," arXiv:2403.08295, 2024.

[14]. P. Zhang et al., "TinyLlama: An Open-Source Small Language Model," arXiv:2401.02385, 2024.

[15]. D. Bhatt, K. Patel, and R. Shah, "Benchmarking Small Language Models on Indian Regional Language NLP Tasks," in Proc. EMNLP Workshop South & Southeast Asian NLP (SEALP), 2024.

[16]. A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv:1704.04861, 2017.

[17]. Howard et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 1314–1324, 2019.

[18]. G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," GitHub, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[19]. S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using Deep Learning for Image-Based Plant Disease Detection," Front. Plant Sci., vol. 7, p. 1419, 2016.

[20]. K. P. Ferentinos, "Deep Learning Models for Plant Disease Detection and Diagnosis," Comput. Electron. Agric., vol. 145, pp. 311–318, 2018.

[21]. A. Ramcharan et al., "Deep Learning for Image-Based Cassava Disease Detection," Front. Plant Sci., vol. 8, p. 1852, 2017.

[22]. E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, "GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers," in Proc. 11th Int. Conf. Learn. Represent. (ICLR), 2023.

[23]. D. Jin et al., "What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams," Appl. Sci., vol. 11, no. 14, p. 6421, 2021.

[24]. D. Hendrycks et al., "CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review," in Proc. NeurIPS Datasets & Benchmarks Track, 2021.

[25]. J. Kim, S. Lee, and H. Park, "Quantization-Aware Training for Legal Document Classification," in Proc. 3rd Workshop Natural Legal Lang. Process. (NLLP @ EMNLP), 2023.