Hardware-Validated Energy-Aware Reliability Evaluation of Compressed Neural Networks

A Deployment-Aware Framework Bridging Efficient and Trustworthy Machine Learning

Master's Research Proposal — Two-Year Programme

Shohinur Pervez Shohan
B.Sc. in Computer Science & Engineering, Rangamati Science and Technology University, Bangladesh
Prospective Applicant, 2027 Intake

Abstract

Neural network compression — post-training quantization, structured and unstructured pruning, and knowledge distillation — is the standard route to deploying deep models under computational and energy constraints. Evaluation practice, however, rests on two assumptions that have not been tested together: that software-reported energy figures reflect what compressed models actually consume on hardware, and that a compression method which preserves aggregate accuracy also preserves the reliability properties deployment depends on.

Neither assumption is safe. Software energy estimators have been shown to err by up to 40% against physical measurement on general workloads, and have not been validated on compressed models, where quantization changes the executed instruction mix and pruning alters memory-access patterns. Separately, the reliability literature reports contradictory findings: some studies report that pruning improves calibration and corruption robustness, others that compressed networks are less robust than their originals. These conflicting results were obtained under different protocols, on different architectures, with different metrics, and none of them measured energy.

This proposal sets out a controlled evaluation protocol that measures reliability and hardware-grounded energy from the same inference passes, across multiple compression families, and compares compression configurations at matched measured energy rather than at matched FLOPs or compression ratio. Phase 1 (months 1–12) is executable on consumer hardware already available and forms the committed thesis deliverable. Phase 2 (months 13–24) extends the validated protocol to physical power instrumentation, heterogeneous accelerators, and larger model families — each requiring resources and methodological supervision the applicant does not independently possess.

1. Introduction

Deep neural networks have become both more capable and more expensive. Deployment on edge devices, mobile platforms, and energy-constrained infrastructure therefore depends on compression: pruning removes parameters, quantization reduces numerical precision, and knowledge distillation transfers behaviour from a large teacher to a compact student.

Evaluation of these methods has matured along two separate tracks. One community measures efficiency — FLOPs, parameter count, latency, and increasingly, energy on real hardware. The other measures

reliability — calibration, robustness under distribution shift, and class-level or subgroup degradation. Both communities have produced substantial results. Neither has produced results that answer the question a practitioner actually faces at deployment time.

The practical question is this: *when two compression configurations achieve comparable accuracy, which one is the better deployment choice once energy consumption and predictive reliability are both accounted for?* Answering it requires measuring both properties on the same models, under the same protocol, with an energy figure that has been validated rather than assumed. That is what this proposal sets out to do.

2. Research Motivation

2.1 The measurement problem

Compression is routinely justified by reductions in FLOPs or parameter count. Direct hardware measurement has repeatedly shown these proxies to be unreliable: energy depends on memory traffic, kernel availability, cache behaviour, and batch size in ways that operation counts do not capture. This proposal takes that finding as an established premise rather than a contribution.

The consequence, however, has not been followed through. If FLOPs cannot be trusted as the resource variable, then every comparison normalised on FLOPs is compromised — including the comparisons that produce the field's method rankings. And if measured energy is to replace FLOPs as the normalising variable, the measurement itself must be trustworthy.

It currently is not, in a specific and testable way. Software estimators such as CodeCarbon and pyJoules have been validated against physical meters on general AI workloads, with documented errors up to 40%. They have not been validated on compressed models. Post-training quantization alters the ratio of integer to floating-point instructions and dispatches different computational kernels; unstructured pruning introduces irregular memory access. Estimators calibrated on dense, full-precision workloads have a clear theoretical reason to fail in this regime, and that failure has not been measured.

2.2 The reliability problem

A compressed model that retains aggregate accuracy is conventionally treated as successful. A body of evidence shows this is insufficient. Compression has been shown to cause selective forgetting concentrated in a small subset of classes, to produce disparate accuracy impacts across demographic groups, and to alter predictive behaviour in ways that aggregate fairness metrics do not surface.

What the literature does not provide is a consistent account of the direction or magnitude of these effects. Post-hoc pruning has been reported to improve uncertainty calibration and natural-corruption robustness in perception CNNs. Other work reports that compressed networks are less robust to distribution shift than their uncompressed originals, with knowledge distillation and post-training quantization outperforming pruning. Still other work finds that quantization, pruning, and weight clustering can preserve or improve robustness under natural corruption.

These findings are not reconcilable from the published record, because they were obtained on different architectures, at different compression intensities, using different metrics and validation designs. This inconsistency is not a weakness in the proposal; it is its opportunity. **Reconciling contradictory reliability findings under a single controlled protocol — with energy measured alongside — is a more defensible contribution than claiming to discover that compression harms reliability.**

3. Related Work

This section states what prior work established and what it leaves open. Where a gap is described as "not identified in our search" or "remains underexplored," that characterisation reflects systematic searches across Consensus, Scite, Scholar Gateway, Elicit, ResearchRabbit, Google Scholar, and direct publisher retrieval, covering 2019–2026. The phrase "no paper exists" is deliberately avoided: absence of evidence in a search is not proof of absence in the literature.

3.1 Neural network compression evaluation

Compression evaluation is mature and extensively benchmarked. Standard practice compares compressed models against full-precision baselines on top-1 accuracy, parameter count, FLOPs, and increasingly latency. Recent work has explored joint pruning–quantization optimisation via reinforcement learning (Balaskas et al., 2023), achieving 39% average energy reduction for 1.7% average accuracy loss on embedded accelerators. Surveys have proposed evaluation frameworks incorporating latency–accuracy trade-offs and multi-objective Pareto optimisation, and standardised metrics such as Compression and Hardware Agnostic Theoretical Speed have been introduced.

What this solved. Compression can dramatically reduce model size and computational cost. Frameworks exist for comparing methods on accuracy–efficiency trade-offs, and multi-objective formulations over accuracy, FLOPs, and latency are increasingly standard.

What it leaves open. Evaluation remains accuracy-centric. Where efficiency is measured, FLOPs or parameter count typically serve as proxies rather than direct hardware measurements. Reliability properties are rarely part of the comparison protocol. Whether the method ranked best on accuracy retains that ranking under other deployment-relevant objectives has not been systematically tested.

How this thesis differs. This work does not propose a new compression algorithm. It treats the evaluation protocol itself as the object of study, adding reliability and hardware-measured energy to the comparison axes and testing whether method rankings are stable across them.

3.2 Reliability consequences of compression

Hooker et al. (2019) showed that pruning disproportionately degrades a small subset of classes — "selective forgetting" — while aggregate accuracy moves very little; at 50% sparsity, 170 ImageNet classes were significantly affected, rising to 372 at 70%. A companion study (Hooker, Moorosi, Clark, Bengio & Denton, 2020) characterised bias in compressed models through compression-identified exemplars and error concentration. Paganini (2020) examined class, cohort, and individual-level degradation as capacity falls. Tran, Fioretto & Kim (2022) provided theoretical and empirical evidence that pruning creates

disparate group impact via gradient-norm mechanisms. Blakeney et al. (2021) evaluated and partially mitigated bias in pruned networks using knowledge distillation. None of these measures calibration or energy.

On calibration specifically, Mitra, Schwalbe & Klein (2024) is the closest existing work to this thesis. It evaluates unstructured, filter, and channel pruning of perception CNNs on expected calibration error over clean and corrupted data together with natural-corruption robustness, and reports that post-hoc pruning can improve calibration, accuracy, and corruption robustness simultaneously. Mantowsky et al. (2023) examined structured pruning with ECE and reliability diagrams alongside feature-saliency stability. Yuan et al. (2023) benchmarked post-training quantization reliability with attention to calibration-set distribution, calibration-metric choice, distribution shift, and worst-case subgroup performance. Ma, Chen & Yao (2025) combined structured pruning with calibrated uncertainty estimation for mobile deployment. None of these reports hardware-measured energy.

On robustness, Diffenderfer et al. (2021) found that compression can improve out-of-distribution robustness. Shen et al. (2024) compared filter pruning, knowledge distillation, and post-training quantization under domain shift and adversarial attack, finding compressed networks less robust overall with quantization strongest under domain shift. da Silva et al. (2025) evaluated pruning, quantization, clustering, and quantization-aware training on CIFAR-10-C and CIFAR-100-C, reporting that compression can preserve or improve robustness. On compressed large language models, Hong et al. (2024) found that task performance can be retained while trust-related properties degrade, and UniComp (von Rad, Cao & Geiger, 2026; EMNLP 2026) evaluated seven compression techniques across more than forty datasets along performance, reliability, and efficiency — where "reliability" denotes safety, fairness, and privacy benchmarks rather than probability calibration, and efficiency denotes hardware-aware latency analysis rather than measured energy.

Huber, Göhner & Trapp (2025) analysed neural network inference on embedded systems, measuring response time across 100,000 invocations together with Expected Calibration Error as a function of network structure, and evaluated the effect of TensorFlow Lite conversion on both performance and calibration. They conclude that both performance and calibration must be considered when deploying models on embedded systems, and report evidence consistent with pruning improving model calibration. This is the closest existing work to this thesis identified across nine independent literature searches. It measures response time and calibration together, but not hardware-measured energy (no RAPL, NVML, or watt-meter instrumentation), does not sweep multiple compression families, and does not validate software energy estimators. The domain is embedded predictive-maintenance signal processing rather than vision classification.

What this solved. Compression alters model behaviour beyond aggregate accuracy. Selective forgetting, group-level disparate impact, and changes in calibration and robustness are real, replicated, and increasingly well characterised along individual axes.

What it leaves open. The direction of the calibration and robustness effect is not settled by isolated case comparison, though the balance of identified evidence leans toward improvement: Mitra et al. (2024), Huber et al. (2025), Diffenderfer et al. (2021), and da Silva et al. (2025) each report that pruning or

compression can improve calibration, corruption robustness, or out-of-distribution robustness, while Shen et al. (2024) report reduced robustness under compression for domain shift and adversarial settings. A citation-network analysis found no work directly contesting the calibration-improvement finding. What no study provides is the energy cost of any reported reliability gain: none of Mitra et al., Huber et al., Diffenderfer et al., da Silva et al., or Shen et al. measures hardware energy alongside their reliability metric, and per-class disaggregation of calibration — the calibration analogue of Hooker et al.'s per-class accuracy analysis — remains underexplored for vision compression.

How this thesis differs. Reliability metrics and energy are collected from the same inference passes, across multiple compression families, at multiple intensities, under one frozen protocol — the conditions required to reconcile the contradictory findings above rather than add another data point to them.

3.3 Energy measurement and hardware-aware evaluation

Tripp et al. (2024) introduced the BUTTER-E dataset: 63,527 experimental runs with node-level watt-meter measurements across fully connected networks varying in size, shape, and depth. The dataset reveals a hardware- and cache-mediated non-linear relationship between network design and energy, and finds FLOPs and parameter count are not the best energy predictors. BUTTER-E evaluates architecture design space, not post-training compression of a fixed architecture, and reports no reliability axis.

Pachón, Pedraza & Ballesteros (2025) evaluated 180 pruned models across three architectures, five pruning distributions, three methods, and four batch sizes, confirming non-linearity between FLOPs savings and energy savings and demonstrating strong batch-size dependence. Deutel, Woller & Mutschler (2022) measured physical power and energy per inference for pruned and quantized models on ARM Cortex-M microcontrollers. Balaskas et al. (2023) optimised pruning and mixed-precision quantization jointly for measured energy on embedded accelerators. Iyer et al. (2026) profiled pruning and quantization with a software estimator on CPU. Each of these reports accuracy as the sole quality metric.

What this solved. FLOPs are an unreliable energy proxy. Direct hardware measurement reveals non-linear, architecture- and batch-size-dependent energy behaviour, and hardware-aware compression optimisation can yield genuine measured savings.

What it leaves open. Every study in this category reports only accuracy alongside energy. Calibration, robustness, and class-level impact are absent from all hardware-energy compression benchmarks identified in our search. We therefore have no evidence on whether the configuration that minimises energy also preserves calibration or distributes errors evenly across classes.

How this thesis differs. Reliability metrics are added to the same protocol that measures energy, and measured energy — not FLOPs — becomes the variable on which configurations are matched for comparison.

3.4 Software energy estimator validation

Fischer (2025) compared software estimators against external physical meters across hundreds of AI experiments and found estimation errors up to 40%. Aquino-Brítez et al. (2025) developed an energy consumption index comparing CodeCarbon and CarbonTracker against sensor-based hardware

measurement across architectures and GPUs. Rodriguez et al. (2024) conducted a systematic literature review with experiments comparing energy-evaluation tools and methods. Kocher et al. (2025) reported that NVIDIA SMI can correlate poorly with an external power meter under some sampling conditions and proposed calibration improving over CodeCarbon's default assumptions.

de Paula et al. (2025) compared pruning, quantization, and distillation for carbon efficiency using CodeCarbon — the closest prior work to the efficiency half of this proposal — but reports no hardware ground truth, no reliability axis, and operates in a different domain. Rojahn & Grum (2025), in a 103-study systematic review of Green AI, conclude explicitly that the field lacks reproducible, calibrated measurement across hardware tiers and call for a hybrid estimator-plus-metering architecture.

What this solved. Software energy estimators have documented inaccuracies against physical ground truth, and the community increasingly recognises that reported energy figures may be unreliable.

What it leaves open. All validation studies identified in our search evaluate estimators on full-precision models with standard instruction mixes. Whether estimator accuracy holds — or degrades systematically — once quantization changes the instruction mix and pruning changes memory access has not been tested. Rojahn & Grum (2025) name this measurement-integration gap at review level, providing external corroboration independent of this proposal.

How this thesis differs. Work Package 1 runs software estimators concurrently with hardware counters on the same inference passes across a sweep of compression states, and tests whether estimator error changes systematically with compression intensity using an ordered trend test rather than agreement at a single operating point.

4. Research Gap

The preceding sections describe a structural disconnection rather than a missing paper. The hardware-energy literature and the reliability literature have each advanced substantially since 2019 without intersecting. Studies that measure energy on compressed models report only accuracy. Studies that measure calibration, robustness, or class-level impact on compressed models report no energy. A practitioner selecting a compression method therefore holds two disconnected evidence bases with no way to determine whether their rankings agree.

Component combination	Representative prior work	Status
Compression + hardware energy	Pachón et al. 2025; Deutel et al. 2022; Balaskas et al. 2023; Tripp et al. 2024	Well covered — accuracy only
Compression + calibration	Mitra et al. 2024; Mantowsky et al. 2023; Yuan et al. 2023	Covered — no energy
Compression + robustness	Shen et al. 2024; Diffenderfer et al. 2021; da Silva et al. 2025	Covered — no energy, contradictory

Component combination	Representative prior work	Status
Compression + class-level impact	Hooker et al. 2019, 2020; Tran et al. 2022; Paganini 2020	Mature — no energy
Energy + estimator validation	Fischer 2025; Aquino-Brítez et al. 2025; Rodriguez et al. 2024	Covered — no compression
Compression + energy + reliability	—	Not identified in our search
Estimator validation on compressed models	—	Not identified in our search
Matched measured-energy comparison	—	Not identified in our search

Three specific gaps follow from this table.

Gap 1. Software energy estimators have not been validated against hardware counters for compressed models, where quantization changes the instruction mix on which those estimators were calibrated. Fischer (2025) and Aquino-Brítez et al. (2025) validate estimators without compression; de Paula et al. (2025) apply an estimator to compressed models without hardware validation. The compressed-model validation regime remains untested.

Gap 2. Reported reliability effects of compression are contradictory across studies conducted under differing protocols, and probability calibration has not been disaggregated per class in the manner Hooker et al. disaggregated accuracy. A controlled multi-family sweep with a frozen protocol is required to determine under which conditions each reported direction holds.

Gap 3. Compression methods have not been compared at matched measured-energy budgets. Existing comparisons normalise on FLOPs, compression ratio, or sparsity level — proxies whose validity is itself in question. Targeted searches for "iso-energy budget compression comparison" and for rank-correlation analysis between energy and reliability rankings returned no on-topic results across every database searched.

5. Research Questions and Hypotheses

RQ1 — Instrument validity. Do software energy estimators (CodeCarbon, pyJoules) agree with hardware counters (RAPL, NVML) for compressed models, and does agreement degrade systematically as precision falls from FP32 through INT4 and as sparsity increases?

RQ2 — Reliability under compression. Prior work reports that post-hoc pruning and other compression methods can improve calibration and corruption robustness. Under which compression families, intensities, and architectures does that improvement hold — and does it survive when energy is measured alongside, or is any reliability gain purchased at an energy cost that FLOPs-based or latency-based reporting conceals?

RQ3 — Ranking stability under matched energy. When compression configurations are compared within matched measured-energy bands rather than at matched FLOPs, does the ranking under accuracy agree with the rankings under calibration, robustness, and class-level error concentration?

RQ4 (Phase 2) — Generality. Do the Phase 1 conclusions hold against physical-meter ground truth, across heterogeneous accelerators, and in a second model family?

Three primary hypotheses are pre-specified before data collection. All other comparisons are reported as exploratory.

- **H1.** Software-estimator error against hardware counters increases monotonically with compression intensity.
- **H2.** Worst-class calibration error exceeds aggregate calibration error by a margin that grows with compression intensity.
- **H3.** Within matched measured-energy bands, configuration rankings under accuracy and under reliability metrics disagree (Kendall's tau significantly below 1).

6. Phase 1 — Committed Thesis Deliverable (Months 1–12)

Phase 1 requires no laboratory access, cluster time, or funding. It runs on hardware already available: an Intel Core i5-13450HX with RAPL and perf-events, an NVIDIA RTX 3050 6GB with NVML, 24GB DDR5 memory, and an ARM platform with powermetrics.

WP1 — Validation of energy estimators on compressed models

Every compression configuration is executed with CodeCarbon and pyJoules running concurrently with RAPL, NVML, and powermetrics readings taken from the same inference passes, so that comparison is paired and within-run rather than across separate executions.

Agreement at each precision level is quantified using Bland–Altman analysis. Because RQ1 is directional, estimator error is additionally modelled with a linear mixed-effects model using compression intensity as an ordered factor and architecture as a random effect, with a Jonckheere–Terpstra trend test as a distribution-free confirmation. Bland–Altman alone cannot establish whether agreement degrades; the trend test can.

Measurement-variance sub-study. Coefficient of variation is reported per counter across repeated passes; batch-size sensitivity is characterised at 1, 16, 64, and 128; and a minimum detectable energy difference is established for this hardware. No energy-based ranking claim is made anywhere in this thesis for a difference below that bound. A preliminary estimate of this bound is being obtained before submission (see §11).

Output: a standalone methods paper. Target venues — TMLR, Sustainable Computing: Informatics and Systems, or an MLSys / SustainML workshop.

WP2 — Reliability under compression, disaggregated

Mitra et al. (2024) evaluate three pruning variants on aggregate ECE over clean and corrupted data and report improvement. This work extends that result in three directions: from pruning alone to three compression families; from aggregate to per-class calibration; and from a reliability-only protocol to one in which energy is measured alongside.

Metrics computed per configuration: per-class Brier score (primary disaggregated metric), adaptive equal-mass ECE per class (secondary), worst-class calibration and its gap from aggregate, per-class reliability diagrams, confidence-distribution shift relative to the FP32 reference, and all of the above recomputed under corruption.

Estimator choice, stated deliberately. Equal-width ECE is unstable when conditioned on class: with 1,000 CIFAR-10 test images per class across 15 bins, tail bins fall to single-digit counts, and apparent per-class differences may reflect binning variance rather than genuine miscalibration. Brier score requires no binning and is therefore the primary per-class metric; adaptive ECE with a reported bin-count sensitivity analysis is secondary. Aggregate ECE is retained for comparability with Mitra et al.

Calibration-set sensitivity control. Post-training quantization depends on the calibration set used to fit scales, and Yuan et al. (2023) established that calibration-set distribution affects PTQ reliability. Every PTQ result in WP1–WP3 therefore inherits a dependency that must be controlled rather than assumed away. On one architecture, the full metric suite is repeated under two contrasting calibration sets — random and class-balanced — and the resulting variance is compared against between-method variance. This is a validity requirement for the main protocol; it is not claimed as an independent contribution, given Yuan et al.'s prior coverage of the PTQ case.

WP3 — Energy-banded ranking comparison

Compression configurations are not directly comparable at nominal settings: INT4 and 70% sparsity are not the same budget. Configurations are therefore assigned to measured-energy bands, with band width set to twice the minimum detectable energy difference established in WP1, so that configurations within a band are not distinguishable by energy. Rankings are compared within bands.

Interpolation between compression states is explicitly rejected. Quantization intensity is discrete; the sparsity–energy relation is non-monotonic on hardware without sparsity-aware kernels; and interpolation would require evaluating models that do not exist. Where two families never co-occupy a band, that is reported as a finding — non-overlapping operating ranges — rather than bridged by extrapolation.

Ranking granularity. Rankings are computed over configurations (approximately thirteen per architecture), not over the four method families. Kendall's tau on four items has negligible resolution; on thirteen it is informative. Method-level ordering is reported as a derived summary, never as the object of the statistic.

Statistics: Kendall's tau between axis-specific configuration rankings within each band; pairwise inversion counts; Pareto-frontier membership and dominated points; and bootstrap $P(A > B)$ across seeds rather than point-estimate leaderboards.

7. Experimental Protocol

Architectures. ResNet-18 (residual) and MobileNetV3-Small (depthwise-separable). Two structurally contrasting families are used rather than three; architecture-contingency is testable with two, and the reduction protects the Phase 1 critical path.

Compression family	Intensity levels swept
Post-training quantization	FP32, FP16, INT8, INT4
Structured pruning	30% / 50% / 70% sparsity
Unstructured pruning	30% / 50% / 70% sparsity
Knowledge distillation	three student width multipliers

Approximately thirteen configurations per architecture. Structured and unstructured pruning are kept as separate arms rather than pooled, because that split is precisely where existing calibration and robustness findings diverge.

Datasets. CIFAR-10; CIFAR-10-C at severity 3 for the main sweep, with a severity 1–5 ablation on one architecture.

Axis	Metrics
Accuracy	Top-1; per-class recall
Calibration	Per-class Brier (primary); adaptive ECE (secondary); worst-class calibration; aggregate ECE; NLL; reliability diagrams
Robustness	CIFAR-10-C mCE; per-corruption breakdown; change in calibration under shift
Class-level error concentration	Per-class recall shift vs FP32; worst-class accuracy
Energy	RAPL, NVML, powermetrics; CodeCarbon and pyJoules in parallel as instruments under test

Terminology. The fourth axis is class-level error concentration, not fairness. Phase 1 uses synthetic nuisance transformations, which are not protected attributes, and no fairness claim is made. The term "fairness" appears only in Phase 2, where real demographic data is available.

Hardware scope condition, stated in advance. The RTX 3050 lacks practical sparsity-acceleration support for unstructured sparsity. Unstructured pruning will therefore show poor energy returns on this platform. This is a property of the deployment target and a finding about consumer-hardware deployment, not a measurement artefact to be explained after the fact. Server-class sparsity support is a Phase 2 question.

Controls. Batch size fixed and separately swept; warm-up passes discarded; clocks and power states controlled where the platform permits; background load logged; five seeds per configuration; bootstrap confidence intervals throughout.

Reproducibility. The protocol is frozen before data collection. Raw per-example outputs are released, not only summary statistics. Code is public and version-controlled, with experiment configurations tracked.

Explicitly excluded. Explanation and attribution stability is not included. No accepted stability metric exists, attributions shift for reasons unrelated to reliability, and its inclusion would weaken an otherwise tight axis set.

8. Statistical Analysis Plan

Three primary hypotheses (H1–H3, §5) are pre-specified. Pre-specifying a small number of primary tests is preferred over family-wise correction across the full comparison space. Exploratory comparisons are reported with Benjamini–Hochberg false-discovery control and are labelled exploratory in every table and figure.

Effect-size floor. No energy-based claim is made below the minimum detectable energy difference established in WP1. No calibration claim is made below the bin-count sensitivity band established in WP2. A ranking reversal is reported as meaningful only where it exceeds measurement and seed variance.

9. Phase 2 — Conditional on Supervision and Resources (Months 13–24)

Phase 1 establishes and validates the protocol. Phase 2 extends it into regimes the applicant cannot reach independently. Each work package names the specific resource that is missing.

WP4 — Modality generalisation: small language models

Requires: compute beyond a 6GB ceiling; domain supervision. The strongest test of whether this is an evaluation framework rather than a vision artefact. The identical protocol and axis set are applied to a small language model in the 0.5B class under the same compression families. UniComp (2026) occupies adjacent territory for LLM compression evaluation; the differentiators are probability calibration and hardware-measured energy, neither of which that work reports. Placing this in year two rather than as a Phase 1 pilot gives the programme intellectual escalation without diluting the committed deliverable.

WP5 — Physical power ground truth

Requires: bench power meter, current probes, oscilloscope. Phase 1 validates software estimators against hardware counters. RAPL and NVML are themselves modelled estimates with documented error. Establishing the floor beneath WP1 requires external physical instrumentation — equipment that cannot be purchased, calibrated, or safely operated independently.

WP6 — Hardware heterogeneity

Requires: edge accelerators, mobile SoCs, ARM single-board computers, server GPUs with sparsity support. Two consumer platforms cannot support a cross-platform generality claim. Differing memory

hierarchies and kernel support — particularly sparsity acceleration, absent from the Phase 1 platform — are exactly where ranking disagreement should be strongest.

WP7 — Scale to modern architectures

Requires: cluster or multi-GPU access. A 6GB ceiling caps the study near 200M parameters. Extension to Vision Transformers, where deployment stakes actually sit, is not possible on current hardware.

WP8 — Real subgroups

Requires: dataset access, ethics approval, institutional affiliation. Phase 1's per-class analysis is legitimate but is not a fairness result. A fairness claim requires datasets with real demographic attributes, carrying access and ethics requirements an unaffiliated individual cannot meet.

WP9 — Methodological formalisation

Requires: collaboration with specialists. Formalising compression selection as a constrained multi-objective decision problem, and extending to distribution-free coverage guarantees via conformal prediction, would benefit from collaboration with researchers specialising in uncertainty quantification and multi-objective optimisation. This is work that is stronger done jointly, and is a principal reason for seeking a research group rather than continuing independently.

10. Timeline

Months	Activity	Dependency
1–2	Measurement harness built and stabilised	—
2–4	WP1 variance sub-study; minimum detectable energy difference established	Stable harness
4–6	WP1 estimator validation; trend analysis; first paper drafted	MDED bound
6–9	WP2 reliability sweep, two architectures	WP1 bounds
8–10	WP2 calibration-set sensitivity control	WP2 pipeline
9–12	WP3 energy-band ranking analysis	WP1–WP2
12	Phase 1 complete: journal submission, workshop paper, public benchmark	—
13–15	Buffer quarter: onboarding, procurement, ethics applications	Placement
15–19	WP4 SLM extension; WP5 physical ground truth	Compute, instrumentation
18–22	WP6 heterogeneity; WP7 scale; WP8 real subgroups	Lab, approvals
22–24	Thesis writing; second and third publications	—

Harness construction and variance characterisation are deliberately serialised — variance cannot be characterised on an unstable harness. The months 13–15 buffer reflects that laboratory access, procurement, and ethics review do not become available instantaneously on arrival.

Slippage plan. If WP1 overruns, the calibration-set control is reduced from two variants to a single documented choice, and the CIFAR-10-C severity ablation is dropped. WP3 is not cut — it is the contribution.

11. Expected Contributions

Committed (Phase 1, thesis deliverable)

- Validation of software energy estimators on compressed models, with an ordered trend test across compression intensity — a regime not identified as tested in any prior work found across systematic searches.
- A measurement-variance characterisation and minimum detectable energy difference for consumer-hardware compression benchmarking, usable as a reference by subsequent work.
- A controlled multi-family reliability sweep capable of reconciling the contradictory directions reported by Mitra et al. (2024), Shen et al. (2024), and da Silva et al. (2025), with per-class disaggregation using binning-robust estimators.
- Energy-banded ranking comparison, replacing the proxy whose validity is in question with the quantity it was meant to approximate, under pre-specified hypotheses and effect-size floors.
- An open, reproducible benchmark with frozen protocol and released per-example outputs.

Aspirational (Phase 2, resource-dependent)

- Modality generalisation to small language models.
- Physical-meter validation of hardware counters.
- Cross-accelerator generality including sparsity-capable hardware.
- Fairness analysis on real demographic data.

Target venues. TMLR or Sustainable Computing: Informatics and Systems for the measurement work; NeurIPS Datasets & Benchmarks or MLSys for the full framework; SustainML and EfficientML workshops for early-stage results; FAccT for a Phase 2 fairness extension.

12. Feasibility and Applicant Preparation

Hardware in hand. Intel Core i5-13450HX (RAPL, perf-events), NVIDIA RTX 3050 6GB (NVML), 24GB DDR5, ARM platform (powermetrics). Sufficient for all of Phase 1; insufficient for Phase 2, by design.

Preliminary measurement pilot. A short pilot is being conducted before submission: ResNet-18 at FP32 on the RTX 3050, 100 forward passes, RAPL and NVML logged, coefficient of variation computed, and the

expected FP16-to-INT8 energy gap checked against twice the measurement standard deviation. This establishes, before the programme begins, whether the effects of interest exceed the instrument's noise floor. The result will be reported in the final submitted version.

Technical background. Python, PyTorch, scikit-learn, pandas, NumPy, OpenCV; statistical analysis and data visualisation; applied machine-learning and computer-vision project work; Linux development environment.

Methodological background. The applicant's undergraduate thesis reconstructed three decades of land-cover change from satellite imagery using multiple classifiers under spatially-blocked rather than random validation. That single design choice reversed the headline conclusion, while the choice of classifier moved the result by roughly five percentage points. Area-adjusted accuracy with confidence intervals and a threshold sensitivity analysis were reported, and the full pipeline released publicly. The work is under peer review at ICCIT 2026.

The lesson — that evaluation design, not algorithm choice, determined the finding — is the direct origin of this proposal. The subject matter has changed; the disposition has not.

Stated limitations. No accepted publications to date. Mathematical background is sufficient to apply and extend established statistical methods with reference to the literature; the formalisation in WP9 is proposed as collaborative work for that reason.

13. Risks and Mitigation

Risk	Mitigation
Ranking disagreement is not observed	A well-powered null result is publishable because the effect-size floor is established first. The protocol contribution stands regardless of the result's sign.
Measurement noise exceeds the effect	WP1 bounds the minimum detectable difference before any ranking claim. The preliminary pilot tests this before the programme starts.
Per-class calibration differences reflect estimator variance	Brier as primary per-class metric (no binning); adaptive ECE secondary; bin-count sensitivity analysis reported.
Prior work is closer than assessed	Positioning states specific extension axes rather than claiming an empty field. Mitra et al. (2024) is open access (arXiv:2405.20876) and is read in full before submission.
Concurrent work occupies the space	WP1 is deliberately first and fastest; an arXiv preprint on completion establishes priority.
Scope overrun	Two architectures only; grid reduced to ~13 configurations; LLM extension deferred to Phase 2; slippage plan specified in §10.

14. Why Supervision Is Sought

Phase 1 is fully specified and requires nothing that is not already in hand. This is not a proposal to begin work; it is a proposal to extend work that is already designed and about to start.

What Phase 1 cannot reach is enumerated precisely: physical power instrumentation that cannot be purchased or safely operated independently; accelerator diversity, including sparsity-capable hardware, that cannot be acquired; compute beyond a 6GB ceiling; real subgroup data requiring institutional and ethics access; and methodological formalisation that is stronger done collaboratively.

The request is therefore specific: supervision and laboratory access to extend a protocol that will already have been built and validated, into the regime where its conclusions become general.

15. References

Aquino-Brítez, A., García-Sánchez, P., & Ortíz, A. (2025). Towards an energy consumption index for deep learning models: a comparative analysis of architectures, GPUs, and measurement tools. Sensors, 25(3), 846. DOI: 10.3390/s25030846

Balaskas, K., Karatzas, A., Sad, C., Siozios, K., & Anagnostopoulos, I. (2023). Hardware-aware DNN compression via diverse pruning and mixed-precision quantization. IEEE Transactions on Emerging Topics in Computing. DOI: 10.1109/TETC.2023.3346944

Blakeney, C., Huish, N., Yan, Y., & Zong, Z. (2021). Simon says: evaluating and mitigating bias in pruned neural networks with knowledge distillation. arXiv:2106.07849

da Silva, I. W. D., Pereira, E., Barboza, E. de A., dos S. Neto, B. F., & Ribeiro, M. de M. (2025). Evaluating the impact of compression techniques on the robustness of CNNs under natural corruptions. ICMLA 2025. DOI: 10.1109/ICMLA66185.2025.00055

de Paula, D., Soni, A., Upadhyay, R., & Lagos, F. (2025). Comparative analysis of model compression techniques for achieving carbon efficient AI. Scientific Reports. DOI: 10.1038/s41598-025-07821-w

Deutel, M., Woller, P., & Mutschler, C. (2022). Deployment of energy-efficient deep learning models on Cortex-M based microcontrollers using deep compression. arXiv:2205.10369

Diffenderfer, J., Bartoldson, B., Chaganti, S., et al. (2021). A winning hand: compressing deep networks can improve out-of-distribution robustness. arXiv:2106.09129

Fischer, R. (2025). Ground-truthing AI energy consumption: validating CodeCarbon against external measurements. it – Information Technology. DOI: 10.1515/itit-2025-0031; arXiv:2509.22092

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. ICML.

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. ICLR. arXiv:1903.12261

Hong, J., Duan, J., Zhang, C., et al. (2024). Decoding compressed trust: scrutinizing the trustworthiness of efficient LLMs under compression. ICML, PMLR 235. arXiv:2403.15447

Hooker, S., Courville, A., Clark, G., Dauphin, Y., & Frome, A. (2019). What do compressed deep neural networks forget? arXiv:1911.05248

Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). Characterising bias in compressed models. arXiv:2010.03058

Huber, P., Göhner, U., & Trapp, M. (2025). Comprehensive analysis of neural network inference on embedded systems: response time, calibration, and model optimisation. Sensors, 25(15), 4769. DOI: 10.3390/s25154769

Iyer, A., Akshatha, K., & Suma, S. (2026). Green AI: a framework for energy-, latency-, memory- and carbon-efficient deep learning optimisation. IEEE Access. DOI: 10.1109/ACCESS.2026.3714889

Kamal, M., & Talbert, D. (2025). Downsized and compromised? Assessing the faithfulness of model compression. arXiv:2510.06125

Ma, R., Chen, S., & Yao, S. (2025). Better reliability compression: model pruning with calibrated uncertainty estimation for mobile deep learning applications. MOST 2025. DOI: 10.1109/MOST65065.2025.00012

Mantowsky, S., Mualla, F., Bukhari, S., & Schneider, G. (2023). DNN pruning and its effects on robustness. DOI: 10.5220/0011651000003411

Mitra, P., Schwalbe, G., & Klein, N. (2024). Investigating calibration and corruption robustness of post-hoc pruned perception CNNs: an image classification benchmark study. CVPR Workshops, 3542–3552. DOI: 10.1109/CVPRW63382.2024.00358; arXiv:2405.20876

Paganini, M. (2020). Prune responsibly. arXiv preprint.

Pachón, C. G., Pedraza, C., & Ballesteros, D. (2025). PruneEnergyAnalyzer: an open-source tool for measuring energy consumption in pruned neural networks. Big Data and Cognitive Computing, 9(8), 200. DOI: 10.3390/bdcc9080200

Rodriguez, C., Degioanni, L., Kameni, L., Vidal, R., & Neglia, G. (2024). Evaluating the energy consumption of machine learning: systematic literature review and experiments. arXiv:2408.15128

Rojahn, M., & Grum, M. (2025). Green AI: a systematic review and meta-analysis of its definitions, lifecycle models, hardware and measurement attempts. arXiv:2511.07090

Shen, L., Edalati, A., Meyer, B. H., Gross, W. J., & Clark, J. J. (2024). Robustness to distribution shifts of compressed networks for edge devices. arXiv:2401.12014

Tran, C., Fioretto, F., & Kim, J. (2022). Pruning has a disparate impact on model accuracy. arXiv:2205.13574

Tripp, C., Perr-Sauer, J., Gafur, J., Nag, A., Purkayastha, A., Zisman, S., & Bensen, E. (2024). Measuring the energy consumption and efficiency of deep neural networks: an empirical analysis and design recommendations (BUTTER-E). arXiv:2403.08151; dataset DOI: 10.25984/2329316

von Rad, J., Cao, Y., & Geiger, A. (2026). UniComp: a unified evaluation of large language model compression via pruning, quantization and distillation. EMNLP 2026. arXiv:2602.09130

Williams, M., & Aletras, N. (2023). How does calibration data affect the post-training pruning and quantization of large language models? arXiv:2311.09755

Yuan, Z., Liu, J., Wu, J., Yang, D., Wu, Q., Sun, G., et al. (2023). Benchmarking the reliability of post-training quantization: a particular focus on worst-case performance. arXiv:2303.13003

Appendix A — Outstanding Verification Before Final Submission

- Read Mitra et al. (2024) in full via the open-access preprint (arXiv:2405.20876) and confirm whether calibration is reported in aggregate only; WP2's per-class extension depends on this.
- Complete and report the preliminary minimum-detectable-energy-difference pilot (§12).
- Confirm the published venue for Tran et al. (2022); the arXiv record is marked as a preprint under review.
- Obtain full text of de Paula et al. (2025) and confirm the absence of hardware ground truth, on which the §3.4 positioning depends.
- Verify DOI and publication status for Paganini (2020) and Blakeney et al. (2021), which lack DOIs in the retrieved records.