LayerShield-QAT: Accuracy-Preserving Integer Quantization for Physics-Grounded Edge Regression

Anonymous authors
Paper under double-blind review

Abstract

Integer accelerators can make scientific inversion practical at the edge, yet regression models are unusually vulnerable to output discretization: errors that leave a classifier's decision unchanged can invalidate a continuous physical estimate. We present LayerShield-QAT, a physics-grounded multi-task regression and quantization workflow for gamma-ray well-logging inversion on 8-bit neural accelerators. The model maps 11 detector-window measurements to five coupled formation and mud properties, using shared representations and task-weighted objectives to prioritize formation density and photoelectric factor. Layer-wise diagnostics identify the input and output transformations as the dominant sources of quantization error; LayerShield-QAT therefore quantizes the hidden representation to INT8 while retaining higher precision at the boundary layers. Training additionally injects systematic, random, and count-statistical perturbations to reduce the synthetic-to-field gap. On the reported synthetic benchmark, multi-task learning increases the fraction of predictions within application tolerances from 23.61% to 89.53% for formation density and from 19.82% to 88.15% for photoelectric factor, while reaching 98.18–99.79% on three auxiliary mud targets. Accuracy-preserved quantization and deployment on NPU SoC systems reports 3.5× CPU and 23× NPU speedups over a legacy C implementation, with stable field data trends. The results support a broader principle: for high-accuracy edge regression, precision should be allocated according to physical sensitivity and layer-wise error propagation rather than uniformly.

1 Introduction

Deep neural networks (DNNs) increasingly replace iterative inverse solvers when latency and power are constrained. Quantization is central to this transition because integer arithmetic reduces storage, bandwidth, and inference cost (Han et al., 2016; Jacob et al., 2018; Nagel et al., 2021). Most quantization evidence, however, concerns classification, where moderate perturbations often preserve the predicted class. Scientific regression is less forgiving: a small numerical displacement can exceed an operational tolerance even when global error metrics appear satisfactory.

We study this problem in gamma-ray density logging. Eleven count-derived measurements from three detectors must be inverted into formation density R_z, formation photoelectric factor $P_{ef,z}$, mudcake thickness H_{mc}, mud density R_m, and mud photoelectric factor $P_{ef,m}$ (Ellis & Singer, 2007). The outputs are coupled by scattering and attenuation physics, while field measurements contain calibration offsets, environmental variation, and count noise. The deployment target is 8-bit integer NPU family with C7x+MMA acceleration.

Our central hypothesis is that uniform INT8 quantization is poorly matched to high-accuracy regression. We combine: (i) multi-task learning that exploits physically shared structure; (ii) tolerance-aware task weighting; (iii) noise-aware synthetic training; and (iv) layer-selective quantization-aware training (QAT) driven by explicit error diagnostics. The

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

resulting LayerShield-QAT workflow keeps boundary transformations at higher precision and quantizes the hidden trunk. This strategy follows the broader observation that layer sensitivity is heterogeneous and that selected layers may require distinct precision, clipping, or calibration treatment (Choi et al., 2017; 2018; Tambe et al., 2020).

Contributions. (1) We formulate low-bit geological inversion as a coupled, tolerance-aware multi-task problem. (2) We introduce a layer-wise diagnostic that motivates selective boundary precision for regression. (3) We integrate systematic and count-statistical noise into synthetic training. (4) We provide field and NPU deployment evidence, including reported 3.5× CPU and 17× acceleration.

2 Related Work

Neural-network quantization. Integer-only inference commonly uses affine quantization with calibrated scales and zero points, while QAT introduces fake-quantization during training so parameters adapt to discretization (Jacob et al., 2018; Hubara et al., 2017). PTQ avoids retraining but depends strongly on representative calibration data and activation-range estimation (Banner et al., 2019; Google AI Edge, 2024). Methods based on clipping, equalization, bias correction, and loss-aware calibration reduce outlier and distribution-shift effects (Choi et al., 2018; Nagel et al., 2019; Nahshan et al., 2021). A broader review of integer quantization choices, granularity, calibration, and hardware implications is provided by Nagel et al. (2021).

Mixed and adaptive precision. Quantization sensitivity varies across layers. Prior work has investigated precision allocation near the quantization limit, adaptive number representations, and hardware-algorithm co-design (Choi et al., 2017; Tambe et al., 2020). Vector quantization and deep compression offer complementary reductions in representation cost (Gong et al., 2014; Han et al., 2016). Our work differs by focusing on continuous, physics-grounded multi-output regression and by using layer-wise output distortion to protect the input and prediction boundaries while retaining INT8 computation in the shared trunk.

3 Method

The extent of Compton scattering process, in which high-energy gamma rays smash with electrons in the formation, losing energy and scattering in various directions, is the basis for the measurement (Ellis & Singer, 2007). The attenuation of gamma rays as they pass through the material follows exponential law: $\Phi(x) = \Phi_0 \exp(-\mu\rho_b x)$, where Φ_0 is the initial gamma ray flux, μ is the Compton scattering cross-section and x is the material thickness.

3.1 Physics-grounded inversion

Bulk formation density follows the mixing relation $\rho_b = \phi\rho_{fl} + (1-\phi)\rho_{ma}$, where ρ_{fl} represents the fluid density (typically water or hydrocarbons), ρ_{ma} is the matrix density (the intrinsic rock density), and ϕ is the porosity. Mudcake layer inserts an additional attenuating layer, making the five targets statistically dependent. This motivates shared feature learning rather than five independent regressors.

3.2 Multi-task network

Let $x \in \mathbb{R}^{11}$ be normalized detector-window counts. A shared multilayer perceptron produces $h = f_\theta(x)$, followed by a formation head $g_f(h) = (\hat{R}_z, \hat{P}_{ef,z})$ and a mud head $g_m(h) = (\hat{H}_{mc}, \hat{R}_m, \hat{P}_{ef,m})$. Training minimizes

$$\mathcal{L} = \sum_{k=1}^{5} \lambda_k \, \mathbb{E}\left[(\hat{y}_k - y_k)^2 / s_k^2\right], \tag{1}$$

where s_k is a target scale and larger λ_k prioritize formation outputs. Target normalization prevents unit magnitude from implicitly determining optimization priority.

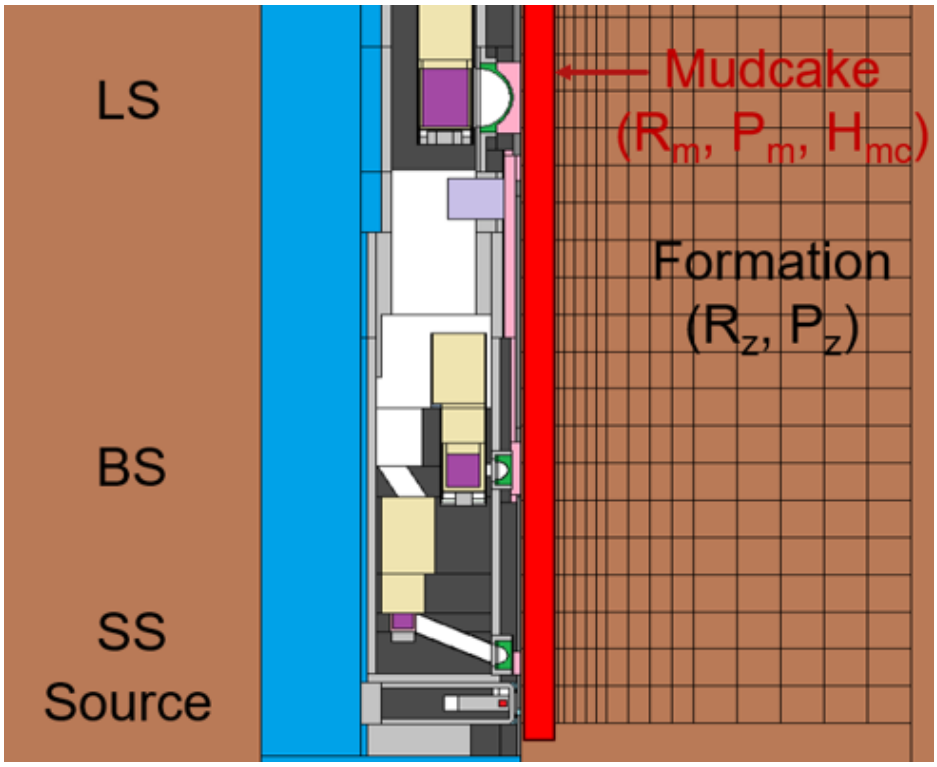


Figure 1: PEx logging geometry used in the source report. Detector windows encode coupled formation and mud responses.

3.3 Noise-aware synthetic training

For each synthetic detector vector, training samples combine a multiplicative systematic component, zero-mean (Gaussian distributed) random perturbations, and count-dependent statistical (Poisson distributed) noise. An approximately 1% systematic noise in amplitudes is used but all these noise contributions remain configuration parameters in the reported work.

3.4 Layer-selective quantization-aware training

Affine quantization maps x to $q = \text{clip}(\text{round}(x/\Delta) + z, q_{min}, q_{max})$. where Δ is the scale and z is the zero point. Integer inference then approximates real-valued operations using quantized weights and activations (Jacob et al., 2018; Nagel et al., 2021). QAT inserts fake-quantization in the forward pass and uses a straight-through approximation for gradients, allowing optimization to adapt to quantization noise (Jacob et al., 2018; Hubara et al., 2017).

We measure each layer's normalized distortion, $d_\ell = \text{RMSE}(a_\ell, \tilde{a}_\ell)/\Delta_\ell$. This diagnostic is consistent with sensitivity-driven precision allocation and layer-wise analysis used in low-precision inference (Choi et al., 2017; Tambe et al., 2020). The source diagnostics show elevated distortion at the first and last transformations, especially for $P_{ef,z}$ at the output ($d_\ell = 0.78$ versus 0.16 in hidden layers). LayerShield-QAT therefore leaves the first and final transformations unquantized or in accelerator-supported higher precision while applying INT8 QAT to the hidden trunk.

Outliers and activation-range mismatch can cause clipping, scale inflation, and bias shifts. Per-channel scaling, parameterized clipping, weight equalization, and bias correction are established remedies (Choi et al., 2018; Nagel et al., 2019; 2021); we treat these methods as baselines or complementary extensions rather than substitutes for boundary precision.

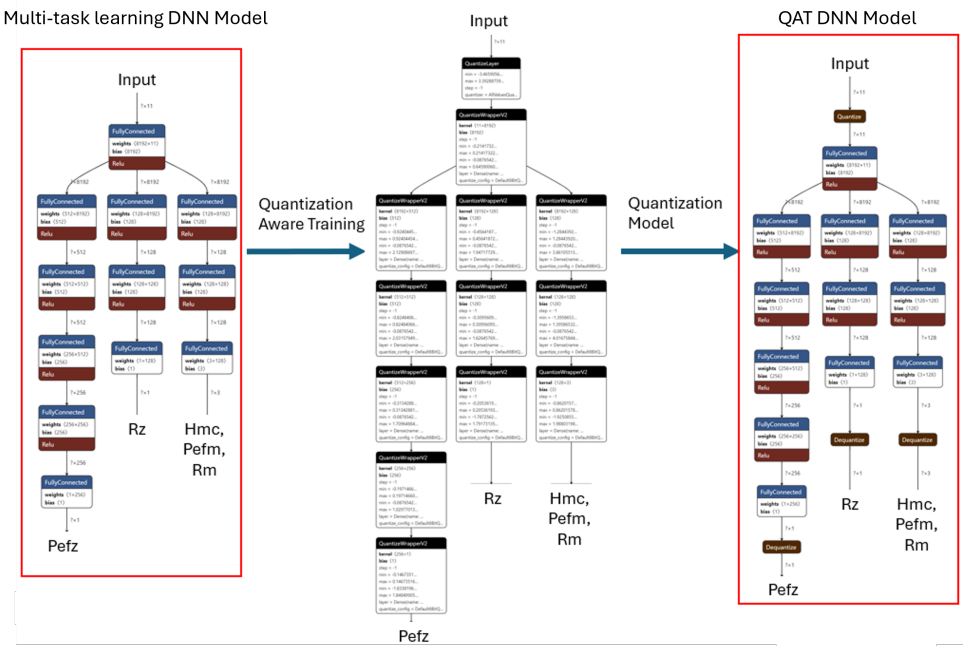


Figure 2: Reported comparison of floating-point, conventional QAT, and selective boundary-precision architectures.

Table 1: Percentage synthetic data respecting tolerance accuracy.

Target (tolerance)	MLP	MTL
R_z (< 0.001 g/cm^3)	23.61	89.53
$P_{ef,z}$ (< 0.01)	19.82	88.15
H_{mc} (< 0.04 in)	88.14	99.48
R_m (< 0.05 g/cm^3)	82.25	99.79
$P_{ef,m}$ (< 0.5)	76.86	98.18

4 Experimental Protocol

Data. Synthetic data is generated by the tool's forward modeling code to simulate five physical parameters and generate the corresponding 11 detector-window counts. 100,000 unique combinations of all five physical parameters are generated by sampling each parameter from a uniform distribution ranges of 1.8-3 g/cm^3, 1-10, 0-1 in., 1-2.5 g/cm^3 and 1-10, respectively. 80/20 training and testing split is used during learning phase. Field evaluation uses depth-indexed well logs.

Baselines. We compare the developed LayerShield-QAT method with conventional single-head MLP, the floating-point multi-task model, post-training INT8 quantization, conventional full-network QAT, alternative quantization strategies. Primary metrics are tolerance accuracy, MAE/RMSE per target, floating-point-to-quantized prediction drift, model size, latency, and frames per second.

Hardware. Models are converted through TensorFlow Lite and compiled with XXDL for NPU SoCs. CPU inference uses the Arm runtime; NPU inference uses C7x+MMA.

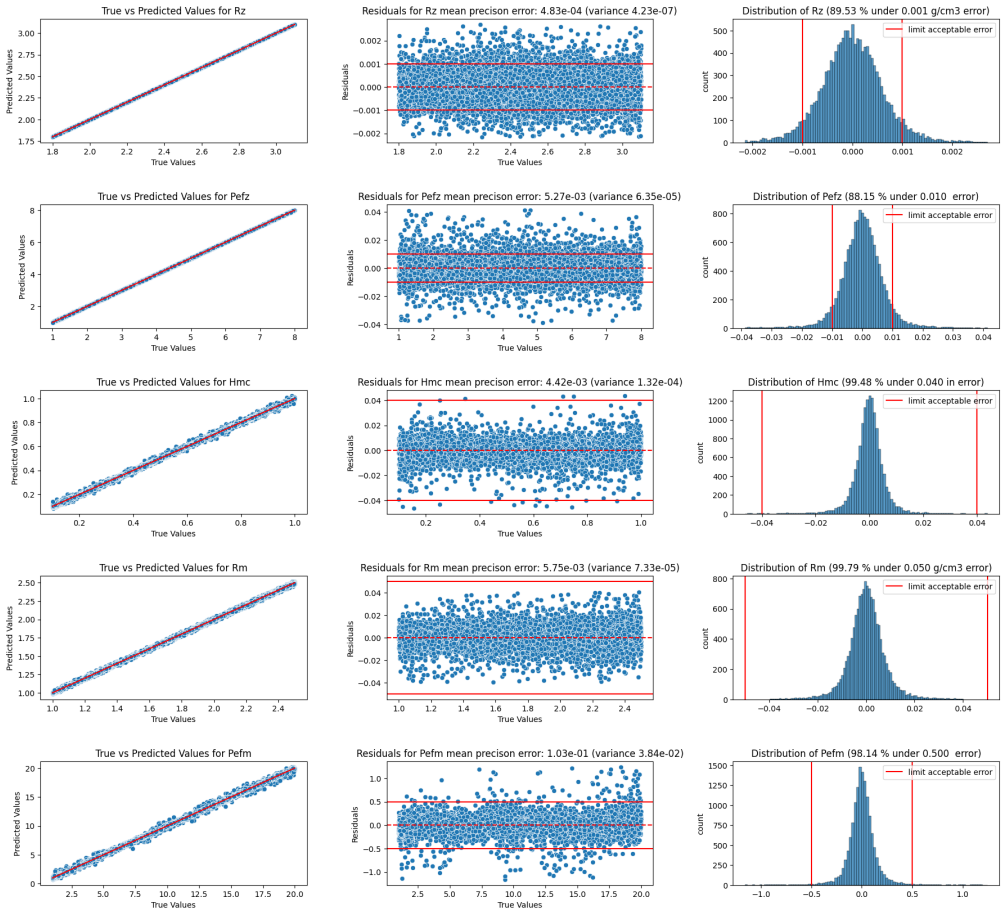


Figure 3: Synthetic data results: prediction agreement, residual structure (for red-lined threshold accuracy), and marginal residual distributions.

5 Results

5.1 Multi-task learning improves tolerance accuracy

Table 1 shows large gains for all five targets. The strongest absolute improvements occur on formation density (+65.92 percentage points) and formation photoelectric factor (+68.33 points), the two high-priority outputs. Auxiliary targets (mudcake parameters) reach 98.18–99.79%, consistent with shared mud/formation representations acting as useful inductive bias.

5.2 Selective precision addresses output sensitivity

Layers most impacted by quantization-induced distortions are frequently those with greater RMSE-to-scale ($d_\ell > 0.239$) ratios. Numerous investigations have verified this technique (Choi et al., 2017; Nagel et al., 2021; Tambe et al., 2020), highlighting the significance of layer-specific analysis in low-precision regimes. Layerwise diagnostics report $d_\ell = 0.25(0.49)$ for the first layer, 0.15(0.16) over hidden layers, and 0.29(0.78) for the final layer for $R_z(P_{ef,z})$, respectively. The hidden representation is comparatively robust to INT8 discretization, whereas preserving output resolution is critical. Reported comparisons show that selective boundary precision reduces quantization-induced error by more than 50% for the key outputs relative to the evaluated quantization alternatives.

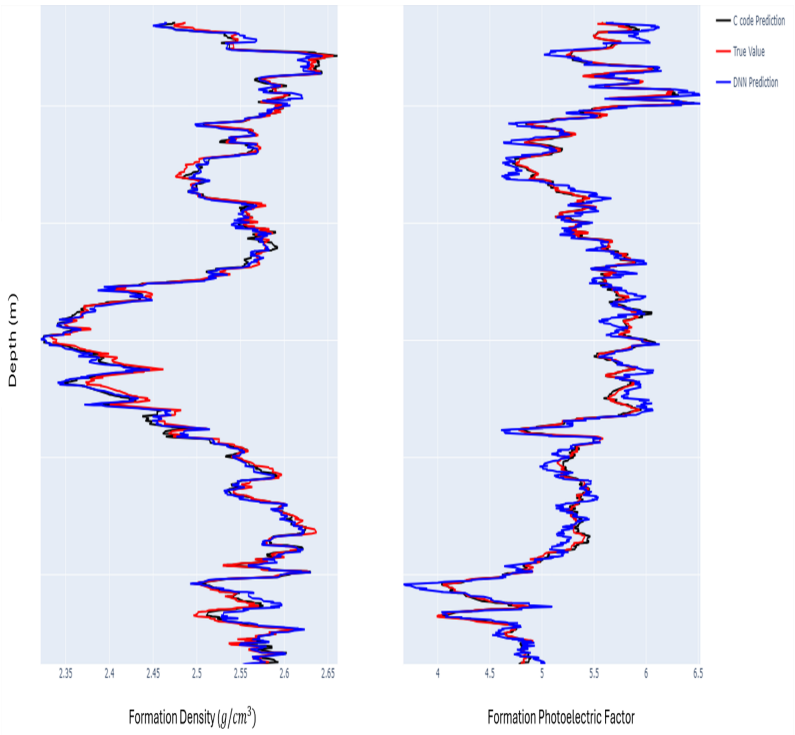


Figure 4: Field-log comparison across depth for reference measurements, legacy inversion, and LayerShield-QAT.

5.3 Field behavior and hardware acceleration

Depth plots show stable R_z tracking and consistent $P_{ef,z}$ trends, while also revealing a systematic field offset attributed in the report to missing temperature compensation. This is an important failure mode rather than evidence to suppress: quantization robustness does not remove calibration shift. Hardware tests report a 3.5× speedup for TFLite CPU inference and a 23× speedup for C7x+MMA NPU inference relative to legacy C code.

6 Discussion and Limitations

The evidence supports a simple design rule: spend precision where the physical inverse map is most sensitive. Boundary precision is inexpensive when most multiply-accumulate work remains INT8, and it protects both raw-input resolution and continuous outputs. This conclusion fits the broader literature on heterogeneous quantization sensitivity and adaptive precision (Choi et al., 2017; Tambe et al., 2020; Nagel et al., 2021). Multi-task learning complements this hardware choice by improving statistical efficiency before quantization.

7 Conclusion

LayerShield-QAT combines physics-grounded multi-task learning, noise-aware training, and layer-selective QAT for accurate edge regression. Reported results show substantial tolerance-accuracy gains over a conventional MLP and large NPU SoCs acceleration over legacy C code. More broadly, the work reframes mixed precision as sensitivity allocation: low precision is most effective when the model's physical and numerical error pathways determine where precision is retained.

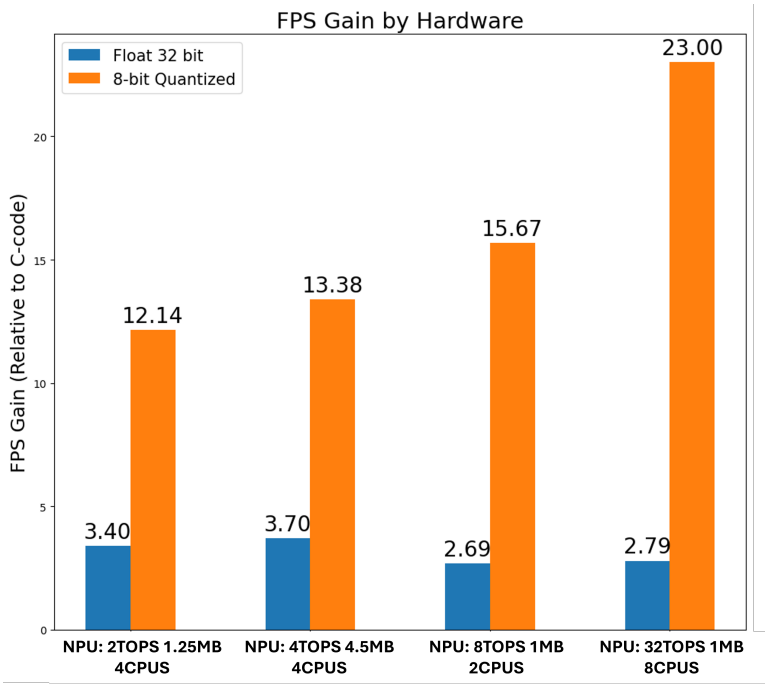


Figure 5: Reported FPS gain across NPU SoCs for floating-point and 8-bit deployment.

Ethics Statement

The application concerns subsurface measurement and embedded inference. No human-subject data are used. Model trained on synthetic data and tested on company's own well controlled environment data.

AI Use Statement

Generative AI assisted with restructuring and language editing of the authors' technical paper and with drafting.

References

Ron Banner, Yury Nahshan, Elad Hoffer, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Towards the limit of network quantization. *arXiv preprint arXiv:1612.01543*, 2017.

Darwin V. Ellis and Julian M. Singer. *Well Logging for Earth Scientists*. Springer, Dordrecht, 2 edition, 2007. doi: 10.1007/978-1-4020-4602-5.

Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. Compressing deep convolutional networks using vector quantization. In *arXiv preprint arXiv:1412.6115*, 2014.

Google AI Edge. Post-training quantization. `https://ai.google.dev/edge/litert/models/post_training_quantization`, 2024. Accessed 2024-12-31.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International Conference on Learning Representations*, 2016.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *Journal of Machine Learning Research*, 18(187):1–30, 2017.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2704–2713, 2018. doi: 10.1109/CVPR.2018.00286.

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. Data-free quantization through weight equalization and bias correction. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1325–1334, 2019.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

Yury Nahshan, Brian Chmiel, Chaim Baskin, Evgenii Zheltonozhskii, Ron Banner, Alex M. Bronstein, and Avi Mendelson. Loss aware post-training quantization. *Machine Learning*, 110:3245–3262, 2021. doi: 10.1007/s10994-021-06053-z.

Thierry Tambe, En-Yu Yang, Zishen Wan, Yuntian Deng, Vijay Janapa Reddi, Alexander Rush, David Brooks, and Gu-Yeon Wei. Algorithm-hardware co-design of adaptive floating-point encodings for resilient deep learning inference. In *57th ACM/IEEE Design Automation Conference*, pp. 1–6, 2020. doi: 10.1109/DAC18072.2020.9218516.

A Additional figures

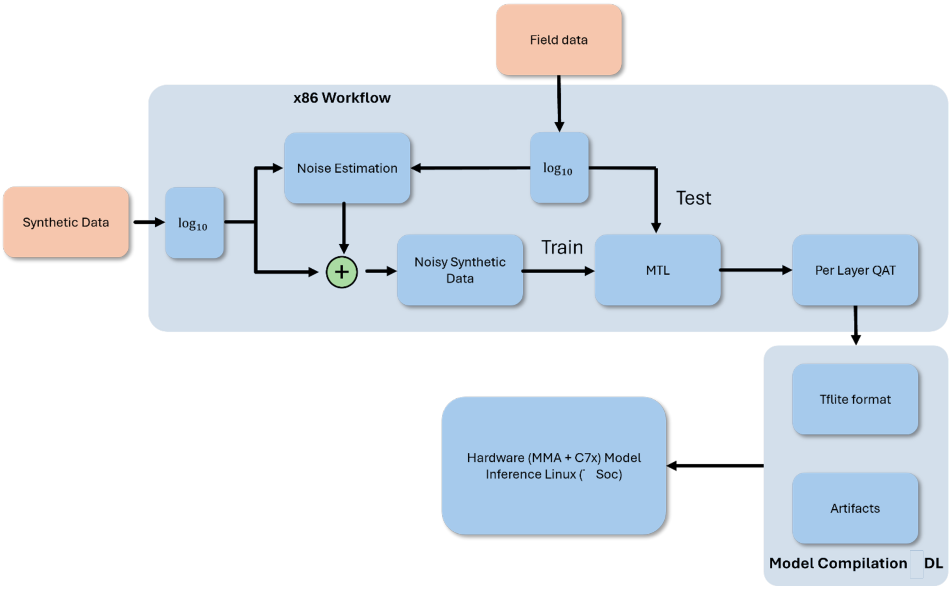


Figure 6: End-to-end training and deployment pipeline.