STE IS NOT A HEURISTIC: A VARIATIONAL JUSTIFICATION IN QUANTIZATION-AWARE TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

The straight-through estimator (STE) is widely used in quantization-aware training (QAT) and other discrete optimization problems in deep learning, but it is considered to be a convenient heuristic with limited theoretical justification. We show that STE admits a principled variational derivation: starting from a variational relaxation of the combinatorial quantization objective under stochastic rounding, STE emerges as the first-order Taylor approximation of the multivariate expansion around the MAP configuration, placing it on rigorous theoretical footing. This variational perspective naturally extends beyond uniform grids: replacing stochastic rounding with a Laplace categorical distribution that is annealed toward the hard solution concentrates gradient signal on uncertain assignments and enables joint learning of non-uniform codebooks whose grid points migrate to minimise reconstruction error, an extension we call non-uniform latent QAT (NUL-QAT). Evaluating on the Qwen3 family (0.6B–32B) at 3-, 2-, and 1.58-bit precision with block-wise QAT, we find that NUL-QAT outperforms STE-QAT on WikiText perplexity and on MMLU, GSM8K, WinoGrande, HellaSwag, and ARC-C accuracy, with the largest gains at extreme bit-widths, confirming that the variational framework captures weight-distribution structure that uniform grids cannot. A custom Triton kernel evaluates the Laplace expectation without materialising the K-fold probability tensor, and the learned non-uniform grids allow efficient inference with lookup table-based kernels such as FLUTE.

1 INTRODUCTION

Post-training quantization (PTQ) and quantization-aware training (QAT) have become standard tools for compressing large language models (LLMs) to 4-bit and, increasingly, to 3- and 2-bit representations (Frantar et al., 2023; Lin et al., 2024; Chen et al., 2024). Recent scaling-law analyses suggest that extremely low-bit quantization is not merely a curiosity: Liu et al. (2025) show that aggressively quantizing a large model to 2 or 3 bits consistently outperforms keeping a smaller model at higher precision on the Pareto frontier of size versus accuracy, making the sub-3-bit regime of special interest. At these aggressive bit-widths, however, the quantization grid is extremely coarse: a 2-bit codebook has only 4 representable values per group, and even 3-bit offers just 8. When the grid is *uniform* (e.g., an equally-spaced integer grid) a disproportionate amount of representational capacity is wasted on weight-space regions with low density, while the high-density core of the distribution is under-resolved.

Non-uniform quantization addresses this mismatch by placing grid points where they are most needed. The idea is well-established in signal processing (Lloyd, 1982) and has appeared in neural-network compression in various forms in Han et al. (2016) and Park et al. (2017). Furthermore, non-uniform codebooks are now supported by efficient inference libraries such as FLUTE (Guo et al., 2024) that perform lookup table-based dequantization inside the matrix-multiply loop, achieving near-uniform-quantization throughput for small codebooks ($K \leq 16$, or 4 bits and lower).

Straight-through estimation is the standard technique for QAT, but it is a heuristic that lacks a rigorous justification and is known to be inappropriate when the quantization grid is non-uniform because the gradient is fixed at 1 regardless of grid spacing (Yamamoto, 2021). In this paper, we present **NUL-QAT** (**N**on-**U**niform **L**atent **Q**uantization-**A**ware **T**raining), a method that places these heuristics on principled footing and extends them to learn non-uniform codebooks. Our starting point

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.

is a variational relaxation of the combinatorial quantization objective: by introducing a categorical distribution over grid assignments for each weight element, we obtain a differentiable upper bound on the discrete loss. A first-order multivariate Taylor expansion around the MAP configuration makes this bound tractable, and we show that the familiar STE emerges as the first-order approximation under stochastic rounding.

Our approach, NUL-QAT, then replaces stochastic rounding with a *Laplace categorical* distribution. Unlike stochastic rounding, which is always supported on exactly two grid points, the Laplace categorical assigns non-zero probability to every grid point and can be annealed to the hard solution through a temperature parameter. Because the codebook values are themselves learnable, the method naturally produces non-uniform grids without any separate codebook design step.

Our contributions are as follows:

- We derive a variational explanation for STE as a first-order approximation under stochastic rounding, placing this heuristic on more rigorous footing.
- We introduce the Laplace categorical distribution as a replacement for stochastic rounding: it provides full support over all grid points (Figure 1) and admits a temperature parameter that can anneal toward the hard solution.
- By using the Laplace categorical distribution, we can learn non-uniform quantization grids through gradient descent, whereas STE does not benefit from allowing the grid to be learned.
- We evaluate on Qwen3 models (0.6B-32B) at 3-, 2-, and 1.58-bit precision, showing that, compared to uniform QAT, the Laplace categorical with learned non-uniform grids provides modest gains at 3-bit and large gains at 2-bit and 1.58-bit.
- We implement efficient Triton kernels for the Laplace categorical forward and backward passes, enabling practical training speed.

Our code is available at `https://github.com/redacted-during-review`.

2 METHOD

Quantization-aware training is a discrete optimisation problem: assign each of N weight elements to one of K grid values so that the resulting low-bit network minimises a downstream loss. The combinatorial search has cost K^N and is NP-hard in general. The straight-through estimator (STE; Bengio et al., 2013) is a popular heuristic: use a continuous latent weight as a stand-in for the quantised one, round in the forward pass, and propagate gradient through the round as if it were the identity. STE is convenient and effective, but struggles with non-uniform codebooks.

We derive the STE from a more principled perspective. Replacing the discrete assignments with a categorical distribution over the grid produces a mean-field variational upper bound on the discrete loss, whose first-order Taylor expansion around the MAP configuration is tractable in a single forward pass. When the chosen distribution is stochastic rounding, the first-order expansion reproduces the STE gradient exactly, placing the heuristic on rigorous footing. We also propose the Laplace categorical distribution as an alternative to stochastic rounding. In Figure 2, we illustrate the forward and backward passes for NUL-QAT.

2.1 PROBLEM SETUP

A neural network has weight matrices $\mathbf{W} = (W_1, \ldots, W_N)$ and a differentiable loss $L(\mathbf{W})$. Each element of each W_ℓ is quantised to one of K levels via a per-group scale $s > 0$, zero point z, and grid $\boldsymbol{g} = (g_1, \ldots, g_K)$, with dequantisation $\mathrm{dequant}(v) = (v - z)\, s$. We suppress the layer index ℓ and the per-group decomposition for brevity, treating $(s, z, \boldsymbol{g})$ as the parameters of a single group whose elements are indexed by (i, j). Writing $q_{ij} \in \{1, \ldots, K\}$ for the assignment of element (i, j), the quantisation problem is

$$\min_{\{q_{ij}\},\, s,\, z,\, \boldsymbol{g}} L\big(\boldsymbol{W}(\{q_{ij}\}, s, z, \boldsymbol{g})\big), \tag{1}$$

where $\boldsymbol{W}(...)$ refers to the dequantized weight matrices. Solving this problem requires searching through K^N configurations for N weight elements and is NP-hard in general.

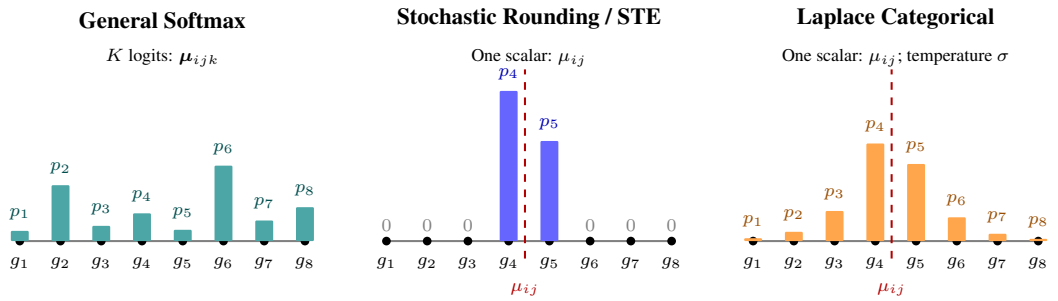


Figure 1: Assignment distributions for one weight element on an 8-point (3-bit) grid. **Left:** General softmax, uses one logit per grid point and can place mass independently of grid distance. Requires 8 logits per weight element, versus 1 for the other distributions. Very memory intensive. **Middle:** Stochastic rounding uses a scalar latent μ_{ij} between g_4 and g_5 and places mass only on these two bracketing points. **Right:** The Laplace categorical uses the same scalar latent and assigns non-zero probability to every grid point, with mass peaked near μ_{ij}.

2.2 Variational Relaxation

We use variational optimisation (Staines & Barber, 2012) to relax the discrete optimization problem. From variational optimization, we know that the minimum of L is bounded above by its expectation under any joint distribution over assignments. We treat the discrete assignments q_{ij} as random variables with a distribution over K classes so that the minimisation occurs in expectation, which provides a differentiable upper bound. Under the mean-field factorisation $P(\mathbf{q}) = \prod_{ij} \boldsymbol{p}_{ij}$, this upper bound is

$$\min_{\mathbf{q}\in\{1,\dots,K\}^N} L(\mathbf{W}) \;\le\; \mathcal{J} \;\equiv\; E_P(L(\mathbf{W})). \tag{2}$$

and $E_{p_{ij}}(L(\mathbf{W})) = \sum_{\mathbf{W}} L(\mathbf{W}) \prod_{ij} p_{ij,q_{ij}}$. The latent distribution for grid assignments is parameterized as $\boldsymbol{p}_{ij} = (p_{ij,1}, \dots, p_{ij,K})$ with a parameter vector $\boldsymbol{\mu}_{ij} = (\mu_{ij,1}, \dots, \mu_{ij,K})$:

$$p_{ij,k} \;=\; p_{ij,k}\big(\boldsymbol{\mu}_{ij};\, \boldsymbol{g}\big), \tag{3}$$

which forms a distribution over the K possible values of the quantization grid.

However, minimising the right-hand side in (2) over the distributional parameters $\{\boldsymbol{\mu}_{ij}\}$ is still computationally expensive and requires sampling. Instead, we make the bound tractable through a first-order Taylor expansion of $\mathcal{J}$ around the MAP configuration that collapses the expectation to a single forward pass.

Figure 2 summarizes the computation used throughout. The network is evaluated at the hard MAP weights in the forward pass, while the backward pass retains the Jacobian of the expected weights. Stochastic rounding and the Laplace categorical thus share the same hard forward computation but induce different gradients through their assignment distributions.

2.3 First-Order Expansion

Let $\hat{\mathbf{W}}$ denote the MAP configuration, which are the dequantised weights when each element is set to its mode $q^*_{ij} = \arg\max_k p_{ij,k}$, and $\hat{w}_{ij} = g_{q^*_{ij}}$ for the grid-space MAP value of element (i, j). Expand $\mathcal{J}$ to first order in the dequantised weights around $\hat{\mathbf{W}}$:

$$\begin{aligned} \mathcal{J} &\approx L(\hat{\mathbf{W}}) + \left\langle \nabla_{\mathbf{W}} L\big|_{\hat{\mathbf{W}}},\; E_P(\mathbf{W}) - \hat{\mathbf{W}} \right\rangle_F \\ &= L(\hat{\mathbf{W}}) + \sum_{ij} \frac{\partial L}{\partial W_{ij}}\bigg|_{\hat{\mathbf{W}}} \big(\bar{w}_{ij} - \hat{w}_{ij}\big) \end{aligned} \tag{4}$$

where

$$\bar{w}_{ij} = (E_{p_{ij}}(g_{q_{ij}}) - z)s = (\bar{g}_{ij} - z)s$$

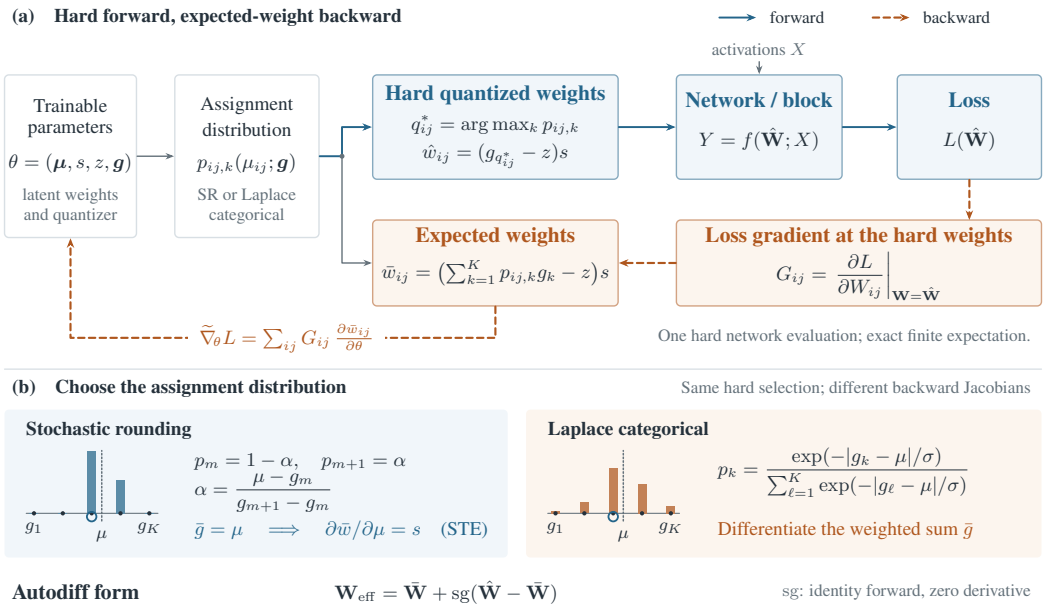


Figure 2: **Forward-backward passes for NUL-QAT.** The forward pass (solid blue) evaluates the network at the MAP assignment $\hat{\mathbf{W}}$; the backward pass (dashed orange) differentiates through the expected weight $\bar{\mathbf{W}} = \mathbb{E}_P[\mathbf{W}]$. This stop-gradient ('sg') construction is equivalent to differentiating the first-order Taylor expansion of the combinatorial loss at $\hat{\mathbf{W}}$ (Eq. 7): the expansion linearises around the hard forward pass, so its gradient flows through $\bar{\mathbf{W}}$ without requiring sampling or materialising the full probability tensor.

$$\bar{g}_{ij} \equiv \sum_k g_k \, p_{ij,k}$$
$$\hat{w}_{ij} = (g_{q^*_{ij}} - z)s$$

and $\bar{w}_{ij} - \hat{w}_{ij} = (\bar{g}_{ij} - g_{q^*_{ij}})\, s$. Therefore, the first-order approximation of the variational upper bound $\mathcal{J}$ is

$$\mathcal{J}^{(1)} \;=\; L(\hat{\mathbf{W}}) \;+\; \sum_{ij} G_{ij} \left(\bar{w}_{ij} - \hat{w}_{ij}\right), \tag{5}$$

with $G_{ij} \equiv \frac{\partial L}{\partial W_{ij}}|_{\hat{\mathbf{W}}}$. A single MAP evaluation to get the dequantized weights and a forward/backward pass yield $L(\hat{\mathbf{W}})$ and the gradients G_{ij}; no sampling is required.

2.4 Equivalence to STE under Stochastic Rounding

We show that if the assignment distribution is stochastic rounding, the first-order expansion of the variational upper bound $\mathcal{J}^{(1)}$ recovers the STE gradient exactly.

Stochastic rounding (SR). A single scalar latent $\mu_{ij,k} \equiv \mu_{ij} \in [\min g, \max g]$ places mass on the two grid points bracketing μ_{ij} in proportion to its distance from each. Letting m be the index for which $g_m \le \mu_{ij} \le g_{m+1}$,

$$p^{\text{SR}}_{ij,k} = \begin{cases} \dfrac{g_{m+1} - \mu_{ij}}{g_{m+1} - g_m} & k = m, \\ \dfrac{\mu_{ij} - g_m}{g_{m+1} - g_m} & k = m+1, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

SR has the convenient property that $\bar{g}_{ij} = \mu_{ij}$ identically, so the expected grid-space value is exactly the latent.

At first order, the stochastic rounding distribution has the property $\bar{g}_{ij} = \mu_{ij}$ identically, which means that

$$\begin{aligned} \mathcal{J}^{(1)} &= L(\hat{\mathbf{W}}) \; + \; s \sum_{ij} G_{ij}(\bar{g}_{ij} - g_{q^*_{ij}}) \\ &= L(\hat{\mathbf{W}}) \; + \; s \sum_{ij} G_{ij}(\mu_{ij} - g_{q^*_{ij}}). \end{aligned} \tag{7}$$

and the derivative of $\mathcal{J}^{(1)}$ reduces to

$$\frac{\partial \mathcal{J}^{(1)}}{\partial \mu_{ij}} \; = \; s\, G_{ij}. \tag{8}$$

Note that the derivative of $\hat{W}$ with respect to μ_{ij} is zero because the argmax operation is not differentiable.[1] Substituting $\mu_{ij} = w_{ij}/s + z$ recovers the canonical *straight-through estimator*,

$$\frac{\partial \mathcal{J}^{(1)}}{\partial w_{ij}} \; = \; G_{ij} \; = \; \frac{\partial L}{\partial w_{ij}}\bigg|_{\hat{\mathbf{W}}}. \tag{9}$$

Therefore, the STE can be understood as arising from a first-order approximation to a variational upper bound on the discrete optimization problem under stochastic rounding. A natural question is whether the choice of distribution matters, since any distribution over the K grid points could have been chosen instead.

Gradient corrections for s, z, $\boldsymbol{g}$. Differentiating (7) fully in the continuous parameters $\theta \in \{s, z, \boldsymbol{g}\}$ generates additional terms beyond those just computed; we show in Appendix B.2 that they are suppressed whenever the assignment distribution is peaked, and so retain only the gradient through $\bar{w}$ that (7) already supplies.

2.5 Laplace Alternative to Stochastic Rounding

We propose the Laplace categorical distribution as a better-behaved alternative to stochastic rounding. Similar to SR, it is parameterized by a single scalar latent $\mu_{ij,k} \equiv \mu_{ij}$. The Laplace categorical is the maximum-entropy distribution on $\{g_1, \ldots, g_K\}$ subject to a constraint on the expected absolute distance from μ_{ij}:

$$p^{\text{Lap}}_{ij,k} = \frac{\exp\big(-|g_k - \mu_{ij}|/\sigma\big)}{\sum_{l=1}^{K} \exp\big(-|g_l - \mu_{ij}|/\sigma\big)}. \tag{10}$$

and $\sigma > 0$ is the temperature that controls the sharpness of the distribution. Like SR, it is parameterised by a single grid-space parameter; unlike SR, it places non-zero mass on every grid point, and that mass is an explicit function of $\boldsymbol{g}$. Figure 1 contrasts a general softmax over the grid with SR and Lap, whose scalar latent straddles two grid points.

In the Appendix, we show that the Laplace categorical distribution does produce useful gradients for the grid values, while stochastic rounding does not. This means that learning the grid values is not practicable under STE, but is possible under the Laplace categorical. Our experimental results confirm that the Laplace categorical with learned non-uniform grids provides significant gains over STE at low bit-widths.

2.6 Initialisation

Uniform grid points are initialized using standard round-to-nearest (RTN) quantization. Non-uniform grid points are initialized by using an interative expectation-maximisation (EM) algorithm for the grid point values and scales/zero points (i.e., iterating between k-means clustering for the grid and scale/zero-point updates given the current grid).

[1] We do not use heuristics like the STE to pretend that non-differentiable operations are differentiable; instead, the STE gradient emerges naturally from the first-order expansion.

2.7 Training Kernel

A naive implementation of the Laplace forward pass materialises the full $N \times K$ probability tensor $p_{ij,k}$ before contracting it against $\boldsymbol{g}$ to obtain $\bar{w}_{ij}$. For $K{=}8$ this multiplies the memory footprint of every quantised parameter matrix by eight, and at the model scales of interest the resulting allocation is prohibitive. We fuse the softmax in (10), the contraction $\bar{w}_{ij} = \sum_k g_k p_k(\mu_{ij})$, and the matching backward pass for $\boldsymbol{\mu}$ and $\boldsymbol{g}$ into a single Triton kernel that streams over the K axis in registers and never writes the probability tensor to global memory. Peak memory then tracks the parameter count rather than K times that count, and training throughput is within a small constant of standard STE QAT. The kernel is used during training only; one can reuse the existing FLUTE lookup table library for efficient inference.

3 Experimental Setup

3.1 Models

We evaluate on the Qwen3 family of dense transformer models (Qwen Team, 2025) (Apache 2.0), spanning a wide range of scales: 0.6B, 1.7B, 4B, 8B, 14B, and 32B parameters. All models share the same architecture and differ only in depth and width. We quantize to 1.58-bit ($K{=}3$, ternary), 2-bit ($K{=}4$), and 3-bit ($K{=}8$) precision, covering the regime where codebook expressiveness matters most.

3.2 Training Protocol

We perform block-wise QAT in the style of EfficientQAT (Chen et al., 2024): each transformer block is calibrated independently against its frozen full-precision counterpart on the same activations, instantiating the generic loss $L(W)$ of §2 as the squared Frobenius reconstruction error

$$L(W) = \left\| f(W; X) - f(W^\star; X) \right\|_F^2, \tag{11}$$

where $f(\cdot\,; X)$ is the block's forward pass on calibration activations X and $W^\star$ collects the unquantised references.

We sampled 20,000 documents from the C4 dataset (Raffel et al., 2020) (ODC-BY) as calibration data, using sequences of length 4096. All methods share identical calibration data and training budgets to ensure a fair comparison; all experiments were run on NVIDIA A100 80 GB GPUs. Full hyperparameters (learning rate, optimiser, temperature schedule, group size) are listed in Appendix C. Note that we use a group size of 64 for all of our experiments for computational efficiency.

3.3 Evaluation

Perplexity. We report perplexity on WikiText-103 (Merity et al., 2016) (CC BY-SA 4.0), evaluated with a context length of 4096 tokens.

Downstream accuracy. We evaluate on five benchmarks:

- **MMLU** (Hendrycks et al., 2021) (MIT): a 57-subject multiple-choice benchmark covering STEM, humanities, social sciences, and professional domains. We report 5-shot accuracy.
- **GSM8K** (Cobbe et al., 2021) (MIT): grade-school math word problems requiring multi-step arithmetic reasoning. We report 4-shot flexible-match accuracy.
- **WinoGrande** (Sakaguchi et al., 2021) (CC BY): a commonsense reasoning benchmark posed as pronoun resolution over Winograd-style sentence pairs. We report 0-shot accuracy.
- **ARC-Challenge** (Clark et al., 2018) (CC BY-SA 4.0): the hard subset of the AI2 Reasoning Challenge, made up of grade-school natural science questions. We report 0-shot length-normalized accuracy (`acc_norm`).
- **HellaSwag** (Zellers et al., 2019) (MIT): a commonsense benchmark where the model picks the most plausible continuation of a context from adversarial candidates. We report 0-shot length-normalized accuracy (`acc_norm`).

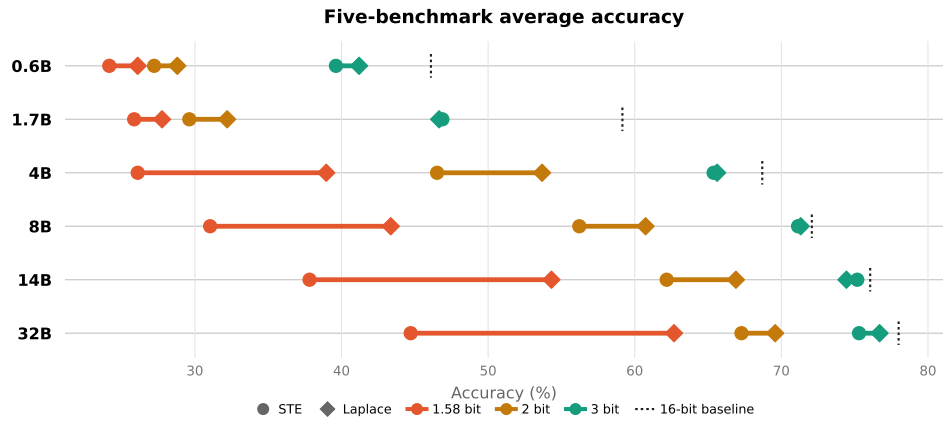


Figure 3: Five-benchmark average accuracy across MMLU, GSM8K, WinoGrande, ARC-Challenge, and HellaSwag for Qwen3 models at each quantization level. For 1.58- and 2-bit quantization grids, NUL-QAT outperforms STE-QAT by a wide margin, and the differences are smaller at 3 bits.

In our ablation studies, we show that the more general softmax distribution does not improve over the Laplace, and also show that STE does not benefit from a non-uniform codebook initialization.

3.4 Inference Deployment

NUL-QAT introduces no new inference machinery. The learned per-group quantization grid $(g_1, ..., g_K)$ can be exported directly as a FLUTE look-up table (Guo et al., 2024), and decoding proceeds through FLUTE's existing fused Triton kernels. Table 4 confirms that FLUTE matches the best uniform-quantisation kernels (e.g., MARLIN) and substantially exceeds the additive-codebook kernels of QTIP and AQLM.

3.5 Baselines

We compare two configurations, both using layer-wise QAT with identical hyperparameters:

1. **Uniform QAT (STE).** Standard uniform grid with the straight-through estimator. Grid values are fixed to the integers $\{0, 1, \ldots, 2^b - 1\}$ and only s and z are learned.
2. **NUL-QAT (our method).** Non-uniform grid with learnable $\boldsymbol{g}$, using the Laplace categorical for the forward pass, temperature annealing, and the STE gradient pass-through for all parameters.

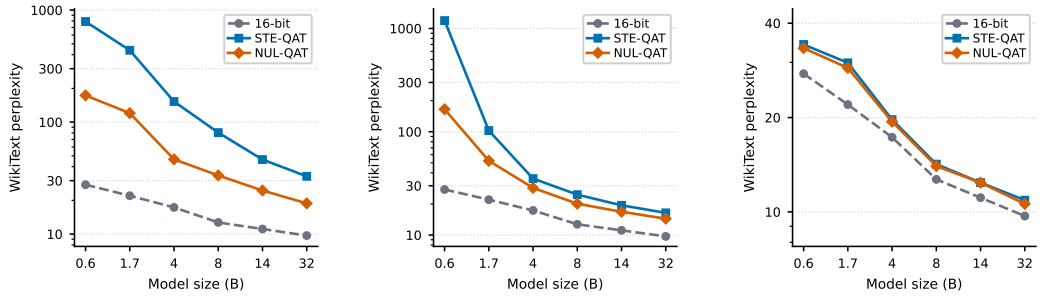
Comparing (1) and (2) isolates the effect of the Laplace categorical assignment distribution and learnable non-uniform codebooks. Both configurations are evaluated at 3-bit, 2-bit, and 1.58-bit precision.

4 Experiments

4.1 Downstream Accuracy

Figure 3 shows the average accuracy across all Qwen3 model sizes and bit-widths on the average across MMLU (5-shot), GSM8K (4-shot, flexible-match), WinoGrande (0-shot), ARC-Challenge (0-shot, length-normalized), and HellaSwag (0-shot, length-normalized). In our experiments, the cost of running NUL-QAT was modest, and the largest Qwen3-32B model was quantized within 10 hours on a 4x A100 80GB GPU node.

Non-uniform grids help on downstream tasks. NUL-QAT at 3-bit is often better than STE across tasks, but the difference is modest. At 1.58- and 2-bit, the gap is larger: the non-uniform grid is necessary to maintain reasonable downstream performance, particularly for the smaller models.



(a) 1.58-bit quantization (b) 2-bit quantization (c) 3-bit quantization

Figure 4: WikiText perplexity across Qwen3 model sizes. Each panel uses a logarithmic y-axis and compares the full-precision 16-bit reference with uniform STE-QAT and NUL-QAT.

In general, small models are more difficult to quantize regardless of the method used at lower bit-widths, and the benefits of using NUL-QAT in low-bit settings are more visible at 8B parameters and higher.

Implications for low-bit deployment. Taken together, these results suggest that NUL-QAT is especially valuable in the regime that Liu et al. (2025) identify as the Pareto frontier: aggressively quantizing a larger model to 2 or 3 bits rather than deploying a smaller model at higher precision. Uniform grids lead to performance degradation at these extreme bit-widths; learnable non-uniform grids help close much of the gap.

4.2 Perplexity Results

Figure 4 summarises perplexity on WikiText for all models, bit-widths, and methods.

Non-uniform grids help on perplexity. Across model sizes and bit-widths, NUL-QAT improves upon uniform QAT, confirming that a learnable codebook trained through the Laplace categorical captures weight-distribution structure that a uniform grid cannot. Similarly to when running downstream evaluations, the gains are especially pronounced at 1.58-bit and 2-bit, where the grid has only 4 or fewer values and grid placement is critical.

Scaling behaviour and the sub-3-bit frontier. For both methods, perplexity degrades as model size decreases. However, the gap between uniform QAT and NUL-QAT *widens* at smaller scales and lower bit-widths: at 32B the uniform baseline is competitive, while at 0.6B and 1.7B the non-uniform grid is essential for acceptable quality. This pattern aligns with the Pareto scaling laws of Liu et al. (2025), who show that 2- and 3-bit quantization occupies the optimal frontier for size-accuracy trade-offs, which is precisely the regime where NUL-QAT's learnable grid placement provides the largest gains.

5 Ablations

We ablate two design choices in NUL-QAT: the initialisation of the non-uniform codebook, and the choice of relaxation distribution.

5.1 Non-Uniform Initialisation Does Not Help STE

It is possible that the gains of NUL-QAT over STE-QAT stem primarily from the non-uniform grid initialisation rather than the Laplace categorical relaxation. Table 1 shows that non-uniform initialisation provides no meaningful improvement to STE-QAT: without the Laplace categorical to route gradients through the codebook, the grid points receive no learning signal and the non-uniform structure is wasted. We discuss this further in the Appendix with reference to the form of the gradient under STE.

Table 1: 3-bit WikiText perplexity (↓) for STE-QAT with uniform (RTN) versus non-uniform initialisation. Non-uniform (NU) initialisation alone does not improve STE.

Model	RTN init	NU init
Qwen3-0.6B	**34.2**	36.5
Qwen3-1.7B	**29.9**	31.1
Qwen3-4B	**19.7**	19.8
Qwen3-8B	**14.2**	15.3

Table 2: 3-bit WikiText perplexity (↓) under different distributions for NUL-QAT. Softmax does not improve over Laplace or stochastic rounding (SR), and its memory complexity scales exponentially with bits, making it impractical at higher bit-widths.

Model	Laplace	Softmax	SR
Qwen3-0.6B	**33.3**	33.9	34.2
Qwen3-1.7B	**28.8**	30.5	29.9
Qwen3-4B	**19.4**	19.8	19.7

5.2 Softmax Relaxation Does Not Improve over Laplace

As mentioned above, the use of stochastic rounding as the distribution leads to the standard STE, and our proposed Laplace categorical distribution is just one of any number of possible relaxations of the hard quantisation assignment. A more general alternative is the softmax distribution over grid points. Table 2 compares the Laplace categorical, softmax, and stochastic rounding relaxations on WikiText perplexity for 3-bit grids. The softmax distribution does not improve over Laplace or stochastic rounding at any model size. Moreover, the softmax relaxation is impractical at scale: computing the full probability tensor requires materialising K^n configurations, where K is the number of grid points and n the number of elements in the quantisation block, so memory usage grows exponentially with the number of bits. The Laplace categorical, by contrast, is linear in the number of bits and can be used at any bit-width, and effectively takes advantage of the structure of the quantisation problem, in that the Laplace categorical is unimodal and ordinal along the quantization grid.

6 Discussion and Conclusion

As LLMs grow ever larger, the need to better understand and improve our existing quantization techniques becomes more pressing. Our work provides a new way to interpret the straight-through estimator in quantization-aware training as a first-order Taylor approximation of a variational objective.

Based on this insight, we introduced a principled extension of STE for non-uniform layer-wise QAT. We are able to learn non-uniform codebooks by replacing stochastic rounding with a Laplace categorical distribution whose temperature is annealed toward the hard solution. Experiments on Qwen3 models from 0.6B to 32B at 3-, 2-, and 1.58-bit precision demonstrate meaningful improvements in perplexity and downstream accuracy over STE-based QAT with uniform grids, with the largest gains at the most extreme bit-widths where every codebook value counts.

There is a rich literature on quantization for LLMs and the properties of STE, and we review the prior work in this area in more detail in the Appendix.

Our current method learns individual grid values per group; an interesting future direction is learning small *shared codebooks* (e.g. product quantization or multi-dimensional codebooks), where the Laplace categorical could be extended to a multivariate distribution over codebook entries.

AI Use Statement

In this work, we used AI-assisted code generation to help write the Triton kernel for the Laplace categorical distribution and adapt the EfficientQAT implementation for our experiments. The generated code was reviewed and edited by the authors to ensure correctness and efficiency. We also used AI tools to assist in the editing of the text of the paper, including grammar and style suggestions.

Acknowledgments

Placeholder while under review.

REFERENCES

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Julian Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B*, 48(3):259–302, 1986.

Mengzhao Chen, Wenqi Shao, Peng Xu, Jiahao Wang, Peng Gao, Kaipeng Zhang, and Ping Luo. EfficientQAT: Efficient quantization-aware training for large language models. *arXiv preprint arXiv:2407.11062*, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized language models. *arXiv preprint arXiv:2305.14314*, 2023.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. *arXiv preprint arXiv:2401.06118*, 2024.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488, 2022.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2023.

Han Guo, William Brandon, Radostin Cholakov, Jonathan Ragan-Kelley, Eric Xing, and Yoon Kim. Fast matrix multiplications for lookup table-quantized llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12419–12433, 2024.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *Proc. ICLR*, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proc. ICLR*, 2021.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *Proc. ICLR*, 2017.

Sehoon Kim, Coleman Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. *arXiv preprint arXiv:2306.07629*, 2024.

Junghyup Lee, Dohyung Kim, and Bumsub Ham. Network quantization with element-wise gradient scaling. In *Proc. CVPR*, pp. 6448–6457, 2021.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. *MLSys*, 2024.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: Data-free quantization aware training for large language models. *arXiv preprint arXiv:2305.17888*, 2023.

Zechun Liu, Changsheng Zhao, Hanxian Huang, Sijia Chen, Jing Zhang, Jiawei Zhao, Scott Roy, Lisa Jin, Yunyang Xiong, Yangyang Shi, Lin Xiao, Yuandong Tian, Bilge Soran, Raghuraman Krishnamoorthi, Tijmen Blankevoort, and Vikas Chandra. ParetoQ: Improving scaling laws in extremely low-bit LLM quantization. *arXiv preprint arXiv:2502.02631*, 2025.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proc. ICLR*, 2017.

Vladimir Malinovskii, Andrei Panferov, Ivan Ilin, Han Guo, Peter Richtárik, and Dan Alistarh. Pushing the limits of large language model quantization via the linearity theorem. *arXiv preprint arXiv:2411.17525*, 2024.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Markus Nagel, Marios Fournarakis, Yelysei Bondarenko, and Tijmen Blankevoort. Overcoming oscillations in quantization-aware training. In *Proc. ICML*, 2022.

Eunhyeok Park, Junwhan Ahn, and Sungjoo Yoo. Weighted-entropy-based quantization for deep neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7197–7205, 2017.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Joe Staines and David Barber. Variational optimization. *arXiv preprint arXiv:1212.4507*, 2012.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. *arXiv preprint arXiv:2402.04396*, 2024.

Guoan Wang, Feiyu Wang, Zongwei Lv, Yikun Zong, and Tong Yang. HESTIA: A Hessian-Guided Differentiable Quantization-Aware Training Framework for Extremely Low-Bit LLMs. *arXiv preprint arXiv:2601.20745*, 2026.

Chen Xu et al. QA-LoRA: Quantization-aware low-rank adaptation. *arXiv preprint arXiv:2309.14717*, 2024.

Kohei Yamamoto. Learnable companding quantization for accurate low-bit neural networks. *arXiv preprint arXiv:2103.07156*, 2021.

Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-Sheng Hua. Quantization networks. In *Proc. CVPR*, 2019.

Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. Understanding straight-through estimator in training activation quantized neural nets. *arXiv preprint arXiv:1903.05662*, 2019.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

A Related Work

Post-training quantization and second-order methods. GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), QuIP# (Tseng et al., 2024), and SqueezeLLM (Kim et al., 2024) represent the state of the art in PTQ for LLMs. GPTQ descends from Optimal Brain Quantization (Frantar et al., 2022), which uses the inverse Hessian of the layer-wise reconstruction loss to optimally compensate unquantised weights after each rounding decision, which is a second-order approach that is effective but relies on greedy, per-element rounding. These PTQ methods operate without retraining and perform well at 4-bit precision, but degrade significantly at very low bit-widths (2–3 bit), motivating QAT-based approaches.

Scaling laws and the sub-3-bit frontier. Liu et al. (2025) demonstrate that aggressively quantising a large model to 2 or 3 bits consistently outperforms keeping a smaller model at higher precision on the Pareto frontier of size versus accuracy. This finding establishes the sub-3-bit regime as the high-impact frontier for deployment and motivates the development of more effective quantisation methods (including non-uniform grids) at these extreme bit-widths.

Quantization-aware training (QAT). Full-model (end-to-end) QAT (Liu et al., 2023) is prohibitively expensive for LLMs. EfficientQAT (Chen et al., 2024) and QLoRA-based approaches (Dettmers et al., 2023; Xu et al., 2024) reduce cost via block-wise or adapter-based training, making QAT practical at LLM scale. All of these methods rely on the STE for gradient flow through the rounding operator. The present work operates within the EfficientQAT block-wise paradigm; extending to end-to-end QAT is left to future work.

STE analysis and alternatives. The STE (Bengio et al., 2013) has been analysed by Yin et al. (2019) and Yang et al. (2019). LSQ (Esser et al., 2019) and EWGS (Lee et al., 2021) introduce gradient-scaling heuristics to improve scale optimisation. Nagel et al. (2022) analyse STE-induced oscillations. Our work provides a principled variational derivation of the STE and identifies the exact gradient correction terms it drops (Appendix B.2).

Differentiable relaxations for quantization. Gumbel–Softmax (Jang et al., 2017; Maddison et al., 2017) is the most common differentiable relaxation for categoricals; however, it introduces high-variance gradients when the logits are not well-initialised. Our Laplace categorical is deterministic in the forward pass and requires no sampling. Concurrently, Wang et al. (2026) propose HESTIA, which also replaces the hard rounding operator with a temperature-controlled softmax relaxation for extremely low-bit QAT. HESTIA uses a Gaussian (squared-distance) kernel, $p_k \propto \exp\big(-|g_k - \mu|^2/\sigma^2\big)$, whereas NUL-QAT uses a Laplace (absolute-distance) kernel, $p_k \propto \exp\big(-|g_k - \mu|/\sigma\big)$. In our experiments, the two kernel forms yield comparable results. The key difference is in scope: HESTIA fixes the quantisation grid to a uniform or prescribed set and focuses on Hessian-guided temperature annealing, while NUL-QAT treats the grid values as learnable parameters and jointly optimises non-uniform codebooks.

Non-uniform quantization and inference efficiency. Learned codebooks have been explored in vision models (Han et al., 2016; Yamamoto, 2021) but GPU-efficient inference was a bottleneck. Recently, libraries like FLUTE (Guo et al., 2024) provided fused Triton kernels that perform LUT-based dequantisation inside the matrix-multiply loop. Malinovskii et al. (2024) show that FLUTE achieves throughput on par with or exceeding uniform-quantisation kernels such as MARLIN, while substantially outperforming kernels for more complex codebook methods like AQLM (Egiazarian et al., 2024) and QTIP (Table 4). The performance gap arises because additive codebook methods (e.g. AQLM) amortise bit cost across large groups by reconstructing weight tiles as sums of codewords, but this additive reconstruction is too expensive to fuse efficiently into the matrix-multiply loop. NUL-QAT uses small per-group codebooks ($K \leq 16$) that can fit in the accelerator's SRAM, maintaining the throughput advantage of FLUTE while gaining the representational flexibility of non-uniform grids.

B Derivation Details

B.1 Equivalence of the Multivariate and Single-Site Expansions

The main text presents a multivariate first-order Taylor expansion of the loss around the full MAP configuration $\hat{W}$ (Eq. 7). An alternative derivation fixes every element except (i, j) at its MAP assignment, which is the single-site conditional in the style of iterated conditional modes (Besag, 1986), and expands only in the free variable. Because the first-order multivariate expansion is separable across elements (each element contributes an independent linear term), the per-element gradient of the multivariate objective with respect to $\bar{w}_{ij}$ is identical to the gradient obtained from the single-site conditional expansion. We use the multivariate framing in the main text because it makes the shared MAP forward pass explicit and avoids the notational overhead of the single-site conditional.

B.2 Gradient Corrections for Quantization Parameters

The first-order objective (Eq. 7) depends on the continuous quantization parameters $\theta \in \{s, z, \boldsymbol{g}\}$ through both the MAP configuration $\hat{W}$ and the dequantization formula. Differentiating by the product rule yields four terms:

$$\frac{\partial \mathcal{J}^{(1)}}{\partial \theta} = \underbrace{\frac{\partial \mathcal{L}(\hat{W})}{\partial \theta}}_{\text{(a)}} + \underbrace{\frac{\partial^2 \mathcal{L}}{\partial \hat{w}_{ij}\, \partial \theta}\bigg|_{\hat{W}} \cdot s\,(\bar{w}_{ij} - \hat{w}_{ij})}_{\text{(b)}} + \underbrace{\frac{\partial \mathcal{L}}{\partial \hat{w}_{ij}}\bigg|_{\hat{W}} \frac{\partial\big[(\bar{w}_{ij} - z)\cdot s\big]}{\partial \theta}}_{\text{(c)}} - \underbrace{\frac{\partial \mathcal{L}}{\partial \hat{w}_{ij}}\bigg|_{\hat{W}} \frac{\partial\big[(\hat{w}_{ij} - z)\cdot s\big]}{\partial \theta}}_{\text{(d)}}. \quad (12)$$

Standard QAT with the STE computes term (a) and folds the pass-through into term (d). Terms (b) and (c) are entirely dropped. Below we describe each term and argue that, in the regime relevant to NUL-QAT with a peaked Laplace categorical, terms (b), (c), and (d) are dominated by term (a).

Term (a): MAP loss gradient. This is the standard backpropagation gradient of the loss at the MAP configuration with respect to θ, computed via the STE pass-through. It is first-order, inexpensive, and provides the dominant gradient signal.

Term (b): mixed second derivative. This term involves $\partial^2 \mathcal{L}/(\partial \hat{w}_{ij}\, \partial \theta)$, which is the Hessian–vector product, modulated by the displacement $s(\bar{w}_{ij} - \hat{w}_{ij})$. It is expensive to compute (requiring second-order information) and is suppressed by two factors: (i) the displacement $(\bar{w}_{ij} - \hat{w}_{ij})$ is small whenever the assignment distribution is peaked, which is the case for the majority of weight elements after the initial training phase; and (ii) even when the displacement is non-negligible, the mixed partial is itself a second-order quantity that contributes at $O(\epsilon^2)$ in the Taylor remainder, where ϵ is the typical grid spacing. We omit it.

Terms (c) and (d): dequantization derivatives. Term (c) propagates the loss gradient through the *expected* dequantised value $(\bar{w}_{ij} - z) \cdot s$, while term (d) propagates through the *MAP* dequantised

value $(\hat{w}_{ij} - z) \cdot s$. Their difference is

$$\text{(c)} - \text{(d)} = \left.\frac{\partial \mathcal{L}}{\partial \hat{w}_{ij}}\right|_{\hat{W}} \cdot \frac{\partial\big[(\bar{w}_{ij} - \hat{w}_{ij}) \cdot s\big]}{\partial \theta}, \tag{13}$$

Because $\bar{w}_{ij}$ and $\hat{w}_{ij}$ are both grid-space values, the zero point z cancels in their difference: the correction (c) − (d) is *exactly zero* for $\theta = z$. For $\theta = s$, the derivative evaluates to

$$\frac{\partial\big[(\bar{w}_{ij} - \hat{w}_{ij}) \cdot s\big]}{\partial s} = \bar{w}_{ij} - \hat{w}_{ij},$$

which is proportional to the displacement $(\bar{w}_{ij} - \hat{w}_{ij})$ and is the same quantity that suppresses term (b). When the Laplace temperature is low (as it is for most of training after the warm-up phase), the distribution is sharply peaked around the mode so that $\bar{w}_{ij} \approx \hat{w}_{ij}$ for the vast majority of elements. The scale gradient correction therefore vanishes as $\sigma \to 0$, and (c) and (d) nearly cancel throughout training.

More precisely, summing over all N elements: the fraction of elements for which $|\bar{w}_{ij} - \hat{w}_{ij}|$ exceeds half a grid spacing decreases exponentially with $1/\sigma$. Fewer than 5% of elements contribute a non-negligible (c)−(d) residual at any given training step. The aggregate correction to the θ gradient is therefore a small perturbation of term (a).

We verified experimentally that including terms (c) and (d) (with term (b) approximated via diagonal curvature) does not improve convergence over the simpler approach of using term (a) alone with the STE pass-through for all parameters. We therefore recommend the standard STE pass-through for scale, zero-point, and grid gradients, and focus the methodological contribution on the choice of assignment distribution (Laplace categorical vs. stochastic rounding).

C Hyperparameters

Table 3 lists the hyperparameters used for all experiments.

Table 3: Hyperparameters for EfficientQAT Block-AP training.

Hyperparameter	**Value**
Optimizer	AdamW
Learning rate	7e-5
Batch size	16
Training steps / block	1250
Group size	64
Sequence length	4096
Calibration sequences	20000
σ_0 (initial temperature)	1.0
$\sigma_{\min}$ (final temperature)	1.0
Temperature schedule	Cosine
GPU	NVIDIA A100 80 GB

Annealing was applied to the Laplace categorical temperature σ, with a linear schedule from $\sigma_0 = 2.0$ to $\sigma_{\min} = 0.9$ over the course of training. We used a random sample from C4 for calibration and validation. Our code is based on a fork of the EfficientQAT repository, with modifications to support the Laplace categorical distribution in a memory-efficient Triton kernel.

D Throughput for Different Kernels

We note that FLUTE achieves much higher throughput than QTIP, AQLM, etc., since there is no need to materialize large additive codebooks. Table 4 compares the end-to-end throughput of Llama-3.1-8B at different bit-widths and batch sizes, taken from Malinovskii et al. (2024).

Table 4: End-to-end throughput (tok/s, ↑) for Llama-3.1-8B at different bit-widths and batch sizes. Kernels that require large additive codebooks like QTIP and AQLM are much slower than look-up table based kernels like FLUTE, even though the former achieve better compression. Taken from Malinovskii et al. (2024).

Batch Size	**1**	**4**	**16**
FP16	57	224	862
Kernel (wbits)	**2 / 3 / 4**	**2 / 3 / 4**	**2 / 3 / 4**
MARLIN	– / – / 133	– / – / 530	– / – / 1873
NF4	– / – / 31	– / – / 101	– / – / 399
AQLM	69 / – / –	81 / – / –	312 / – / –
QTIP	155 / 136 / 122	230 / 190 / 166	249 / 202 / 177
FLUTE	**173 / 150 / 139**	**687 / 592 / 548**	**2432 / 2122 / 1979**