AN α-LAW FOR SCALE MULTIPLIERS IN LOW-BIT POST-TRAINING QUANTIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preserving language-model accuracy below four bits per weight motivates understanding how quantizers allocate their limited representable values. We ask whether post-training quantizers with different optimization procedures choose scales that follow a shared, predictive relationship. We compare their choices through a per-group multiplier α that measures scale contraction relative to the group's largest weight magnitude. For AutoRound, GPTQ, and AWQ, its median follows $\alpha^* = q_{\max}/(q_{\max} + c)$, where $q_{\max}$ is the largest positive integer quantization code and c is fitted per method and configuration. A symmetric clipped mean-squared-error approximation interprets this contraction through the trade-off between clipping and rounding error. Across seven models from six families at two, three, and four bits per weight, the law achieves in-sample $R^2 > 0.97$. Constants fitted on models with at most nine billion parameters predict the sampled median α of Llama-3.1-70B with absolute errors below 0.016 for AutoRound and 0.008 for GPTQ at two and three bits per weight. Among the four tested methods, larger fitted c corresponds to higher two-bit accuracy averaged across the seven models. For Llama-3.1-70B quantized to two bits per weight, AutoRound and its ablation without rounding iterations have similar sampled median scale multipliers yet retain 91% and 45% of the half-precision reference model's ten-task mean accuracy. The α-law reveals shared structure in method-specific scale choices and supports prediction across bit widths and, for AutoRound and GPTQ, to a held-out 70B model.

1 INTRODUCTION

Post-training quantization (PTQ) methods at two and three bits per weight differ widely in the accuracy they retain at a matched bit width (Shao et al., 2024; Cheng et al., 2025). Preserving accuracy below four bits motivates studying how quantizers allocate their limited representable values. In per-group quantization, a block of weights shares a scale that sets its numerical range. For the uniform quantizers studied here, contracting the range reduces the spacing between representable values while clipping extreme weights, trading clipping error against rounding error (Banner et al., 2019). Shared block scales also underpin the Microscaling (MX) formats MXFP8, MXFP6, and MXFP4 (Darvish Rouhani et al., 2023), and NVIDIA's NVFP4 (Alvarez et al., 2025).

Existing methods optimize quantization through different procedures. GPTQ uses approximate second-order information to compensate for weight-quantization error (Frantar et al., 2023), while AWQ uses activation statistics to rescale salient weight channels (Lin et al., 2024). OmniQuant learns clipping thresholds and examines their distributions (Shao et al., 2024); AutoRound optimizes rounding through signed gradients (Cheng et al., 2024), with scale pre-tuning in its SignRound V2 implementation (Cheng et al., 2025). We ask whether their normalized scale choices follow a common relationship across bit widths, whether it predicts held-out widths and larger models, and how it relates to downstream accuracy.

We study the per-group scale multiplier $\alpha_g \equiv s_g\, q_{\max}/M_g$, where s_g is the quantization scale, M_g is the group's largest absolute weight value in the space where the grid is defined, and $q_{\max}$ is the largest positive integer code. The reference scale $M_g/q_{\max}$ corresponds to $\alpha_g{=}1$. This normalization makes scale contraction comparable across methods and bit widths. We find that the median follows $\alpha^{*}{=}q_{\max}/(q_{\max}{+}c)$ for AutoRound, GPTQ, and AWQ. For each method and configuration, a single

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.

fitted constant c, which we call the c-index, captures how the median multiplier changes with bit width.

We measure α across seven models spanning six model families (Llama-2-7B, Phi-3-mini-4k-instruct, Qwen3-1.7B, Mistral-7B-Instruct-v0.1, Qwen3-8B, Qwen2.5-7B-Instruct, and Gemma-2-9B) at two, three, and four bits per weight, denoted W2, W3, and W4, and then test the fitted law on Llama-3.1-70B. The scale analysis requires checkpoint scales and half-precision (FP16) reference weights, but no downstream evaluation. Our contributions are as follows.

1. **A shared scale law across methods and bit widths.** The law achieves in-sample $R^2 > 0.97$ for AutoRound, GPTQ, and AWQ across the seven models at W2–W4. A symmetric clipped mean-squared-error approximation interprets the contraction through the trade-off between clipping and rounding error (Section 3.1).
2. **Scale predictions generalize to held-out bit widths and a larger model.** For AutoRound and GPTQ, fitting W3 and W4 predicts W2 median α on the same models (Table 18). Constants fitted on models with at most nine billion parameters predict the sampled median α of Llama-3.1-70B with absolute errors below 0.016 for AutoRound and 0.008 for GPTQ at W2 and W3 (Section 4).
3. **Scale contraction relates to accuracy without determining it.** Among AutoRound, GPTQ, AWQ, and AR0 (AutoRound with scale pre-tuning but no rounding iterations), larger fitted c corresponds to higher W2 accuracy averaged across the seven models. This association also holds within the base-model and instruction-tuned subsets (Section 3). For Llama-3.1-70B quantized to two bits per weight, AutoRound and AR0 have sampled median α values of 0.377 and 0.372 yet retain 91% and 45% of the half-precision reference model's ten-task mean accuracy, respectively. The full optimization procedure changes both per-group scales and accuracy (Section G).

We also test whether calibration changes produce detectable differences in intermediate activations through activation patching between Pile calibration and α-elicit calibration (αEC), which augments Pile with benchmark-adjacent training text (Section 2). Patching substitutes one checkpoint's layer output into the other; the flip rate measures the fraction of tokens whose predicted next token changes. At W2, this rate is 4.4–9.5 times lower for AutoRound than for GPTQ on each of the seven models, and the gap narrows at W3 (Section 3). This descriptive comparison has no control for calibration-seed variation across the seven models. A separate controlled analysis on three models detects a calibration-content effect beyond seed variation only for GPTQ at W3 (Section B).

Related work. Second-order compression (Hassibi & Stork, 1992; Frantar & Alistarh, 2022), activation-aware rescaling (Xiao et al., 2023), and learned rounding (Nagel et al., 2020; Li et al., 2021) provide distinct ways to preserve accuracy at low precision. AutoRound's sign-gradient updates also connect to signSGD (Bernstein et al., 2018). Our analysis compares normalized scale choices across AutoRound, GPTQ, and AWQ and tests their predictability across bit widths. This focus complements work that changes weight representations through incoherence-based transforms (Chee et al., 2023), additive or lattice codebooks (Egiazarian et al., 2024; Tseng et al., 2024), and sparse or non-uniform formats (Dettmers et al., 2024; Kim et al., 2024). Surveys (Gholami et al., 2021; Zhu et al., 2024) and the LLMC benchmarking toolkit (Gong et al., 2024) provide broader coverage of quantization and compression methods.

Calibration data introduces a separate source of variation. Williams & Aletras (2024) show that calibration-domain choice changes downstream accuracy at W2 and W3. We use activation patching to compare checkpoints calibrated on different text, drawing on causal mediation and abstraction (Vig et al., 2020; Geiger et al., 2023), transformer-circuit analysis (Elhage et al., 2021; Olsson et al., 2022; Conmy et al., 2023), and knowledge localization (Meng et al., 2022).

2 Experimental Design

We fit scale summaries across bit widths, test W3/W4-to-W2 prediction on the same models, and evaluate transfer to a held-out 70B model. Accuracy evaluation and activation patching examine how scale choices relate to model behavior.

Weight-space quantities. For each group g of 128 weights within an output channel, we read the checkpoint scale s_g and compute $M_g = \max_{i \in g} |w_i|$ from the corresponding reference weights

in the space where the grid is defined, including AWQ's channel rescaling. The *scale multiplier* is $\alpha_g \equiv s_g q_{\max}/M_g$, with $q_{\max}=1, 3, 7$ at W2, W3, and W4. Thus $\alpha_g=1$ denotes the reference scale $M_g/q_{\max}$, and smaller positive values denote contraction; the W2 endpoint convention is in Section K.5. The small-model fits use the mean of per-matrix group medians for AutoRound, GPTQ, and AR0, and the median transformed-space clipping ratio for AWQ. At 70B, we pool retained groups from sampled matrices to compute each checkpoint's median. Coefficients of variation describe within-matrix dispersion.

Models and methods. The α-law panel has seven models from six families, four base and three instruction-tuned. Llama-2-7B (Touvron et al., 2023), Phi-3-mini-4k-instruct (Abdin et al., 2024), Qwen3-1.7B (Yang et al., 2025), Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), Qwen3-8B (Yang et al., 2025), Qwen2.5-7B-Instruct (Qwen et al., 2024), and Gemma-2-9B (Gemma Team et al., 2024) are each quantized to W2, W3, and W4 with group size 128. Llama-3.1-70B (Grattafiori et al., 2024) is held out for Section 4. We fit AutoRound (AR) (Cheng et al., 2024), GPTQ (Frantar et al., 2023), and AWQ (Lin et al., 2024) on all seven models. AR0 scale fits cover Llama-2-7B, Phi-3-mini, and Qwen3-1.7B; its W2 accuracy evaluation covers all seven. AutoRound uses the SignRound V2 implementation, and AR0 runs its scale pre-tuning search with `iter=0`, without rounding optimization (Cheng et al., 2025). AR0 is called RTN-opt in Jiang et al. (2026). Round-to-nearest (RTN-naive) and OmniQuant (Shao et al., 2024) provide additional scale-family comparisons. The fits use seed 42; W2 accuracy averages four calibration seeds ($42, 123, 456, 789$). Patching and 70B seed counts accompany their results.

Calibration data. The Pile-only controls use a 128-sequence anchor from Pile-10k (Gao et al., 2020) or its approximately 200-sequence size-matched extension. The *α-elicit calibration* (αEC) condition is the FragCal-DC set of Jiang et al. (2026). It combines the 128-sequence Pile anchor with 32 sequences from MMLU-auxiliary (Hendrycks et al., 2021), 32 from pooled ARC-Challenge, ARC-Easy (Clark et al., 2018), and OpenBookQA (Mihaylov et al., 2018), and 16 from HellaSwag (Zellers et al., 2019), all from training splits. Short examples are concatenated before sampling. Whole documents are retained after a 2048-token minimum-length filter, leaving approximately 200 of the nominal 208 sequences. The supplementary builder fixes the sources, formatting, and seeds. At 70B, AutoRound and AR0 use the same filtered Pile-10k sequences for each matched seed. Counts refer to sequences, whose lengths vary by source; mixture proportions are fixed rather than ablated.

Downstream evaluation. Downstream accuracy is the mean over ten lm-evaluation-harness tasks (Gao et al., 2023), namely ARC-Challenge and ARC-Easy (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), LAMBADA (Paperno et al., 2016), MMLU (Hendrycks et al., 2021), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), TruthfulQA-mc1 (Lin et al., 2022), and WinoGrande (Sakaguchi et al., 2021). We call this mean at W2 the *W2 accuracy*; *retention* is reserved for the same mean divided by the FP16 mean, used in Section 4.

Activation-level probes. At each decoder layer, a residual-stream patching harness (Vig et al., 2020; Meng et al., 2022) substitutes a Pile-calibrated model's residual into its αEC-calibrated counterpart on the same input, method, width, and seed. The *pathway-flip rate* is the fraction of tokens whose argmax next-token prediction changes relative to the unpatched recipient. The seven-model comparison is descriptive (§3); a nonzero rate alone does not establish a calibration-content effect. The controlled analysis on Llama-2-7B, Phi-3-mini-4k-instruct, and Qwen3-1.7B compares these sweeps with within-variant seed swaps on 130 prompts per sweep (Sections A and B).

3 An α-Law for Scale-Optimizing PTQ

3.1 Parametric Form and Scale Index

Figure 1 and the detailed taxonomy in Table 8 place the seven methods in the plane spanned by median α and the Spearman correlation ρ between each method's scales and group absmax. Two families appear, split by median α. The *absmax family* (RTN-naive and OmniQuant's exported scale) keeps median α near 1, although OmniQuant's ρ is low (0.56 at W2) because its learned clipping reorders groups without shrinking them on average. The *α-law family* (AutoRound, AR0, GPTQ, AWQ) moves α below 1 by an amount that depends on the bit width and on the method, and the

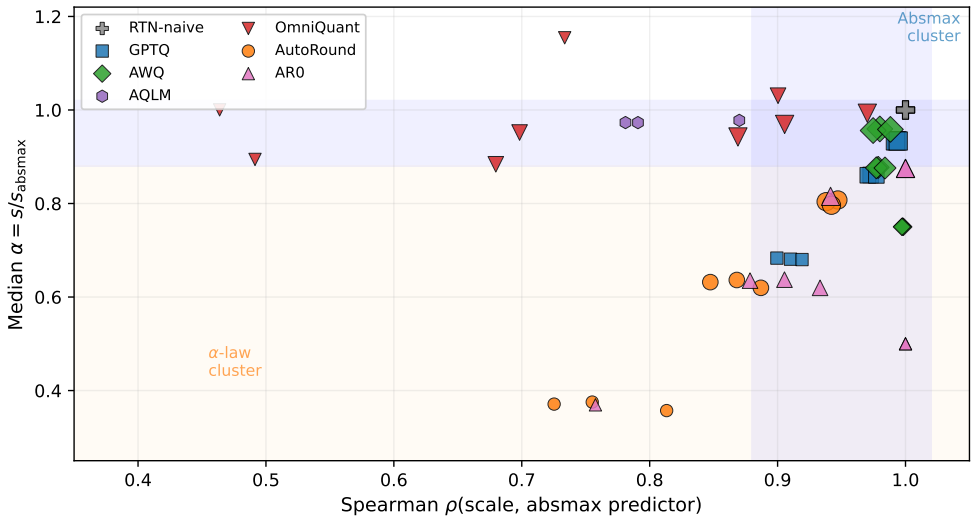


Figure 1: Scale-family scatter of ρ(scale, absmax) against median α for Llama-2-7B, Phi-3-mini-4k-instruct, and Qwen3-1.7B at W2–W4. RTN-naive and OmniQuant's exported scales keep median α near 1. AutoRound, AR0, GPTQ, and AWQ exhibit width-dependent contraction; per-method fits are in Figure 2. AQLM markers (W2 only) show a codebook quantizer's scale parameter for completeness without interpretation.

dependence has a common form. AQLM is a codebook quantizer with no uniform-grid scale in the sense above; it appears in Figure 1 for completeness only and is not placed in either family or fitted to the law.

Absmax family. RTN-naive has ρ=1 and $\mathrm{CV}(\alpha)$=0 by construction. Its α equals 1 under the positive-saturation convention s=absmax/$q_{\max}$ used on the ≤9B panel and $q_{\max}/2^{b-1}$ under the negative-half convention s=absmax/2^{b-1} used by the 70B export; the two are a constant rescale of the same degenerate field. OmniQuant optimizes two clipping boundaries and a rounded zero point. Its exported scale stays near absmax, but its native optimized scale normalized by the original group absmax follows the law on Llama-2-7B (R^2=0.903) and Gemma-2-9B (R^2=0.934). OmniQuant's fit is thus coordinate-sensitive and weaker than the fits for AutoRound, GPTQ, and AWQ.

α-law family. GPTQ ($\alpha \in [0.64, 0.93]$), AWQ ($\alpha \in [0.75, 0.96]$), AutoRound ($\alpha \in [0.36, 0.80]$), and AR0 follow the α-law with different constants c. AWQ first applies a per-channel importance scaling $\mathbf{W}' = \mathbf{W} \cdot \mathrm{diag}(\mathbf{s})$ and then quantizes in the transformed space, so we report its ρ and α_t in that space, where the grid is defined; the pre-scaling is absorbed into the activations. Original-space ρ for AWQ is lower (0.75–0.92) because the importance scaling redistributes weight magnitudes before clipping.

Derivation. We interpret scale contraction through the trade-off between rounding and clipping error. Consider a symmetric uniform quantizer with group absmax M, largest code $q_{\max}$, and clip threshold $T = \alpha M$. For Gaussian weights with standard deviation σ, the approximate mean squared error (MSE) is

$$\mathrm{MSE}_\lambda(T) = \lambda \cdot \underbrace{\frac{\Delta^2}{12} P_{\text{in}}}_{\text{rounding}} + \underbrace{2\sigma^2 \left[(1+a^2)\bar{\Phi}(a) - a\,\phi(a)\right]}_{\text{clipping}}, \tag{1}$$

where $\Delta = T/q_{\max}$ is the grid step, $a = T/\sigma$, $P_{\text{in}} = 2\Phi(a)-1$, and $\lambda \geq 1$ is an assumed weight on dense rounding error relative to sparse clipping error. Here ϕ and Φ are the standard normal density and cumulative distribution functions, and $\bar{\Phi} = 1 - \Phi$. Writing $T = M - c \cdot \Delta$ gives

$$\alpha^* = \frac{q_{\max}}{q_{\max} + c(\lambda)}. \tag{2}$$

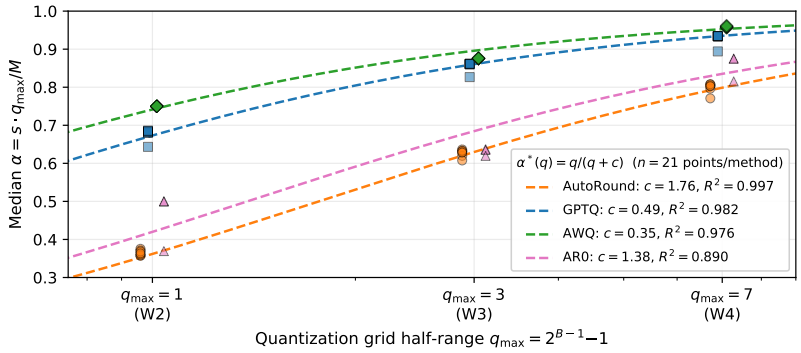


Figure 2: Method-specific α-law fits across 7 models (6 families) $\times$ 3 bit widths (n=21 points per method; n=9 for AR0). AutoRound (c=1.76), GPTQ (c=0.49), and AWQ (c=0.35) follow the law with $R^2 > 0.97$. AR0 (AutoRound with `iter=0`) reaches c=1.38 through scale search alone. Markers are jittered horizontally for visibility.

This change of variables expresses scale contraction in grid steps. The empirical claim is that a single fitted c describes the observed medians across bit widths for a given method and setting. Using the Gaussian group-maximum approximation $M/\sigma \approx 2.828$ for groups of 128 weights (Section D), minimizing Equation (1) at λ=1 gives c=1.29, 1.34, and 1.00 at W2, W3, and W4, so a plain per-group MSE optimum does not produce a constant c; a single $\lambda \approx 1.9$ approximately reproduces AutoRound's observed c=1.75 at all three widths. The separately fitted per-width λ values are 1.96, 1.67, and 2.03 (Section D). This approximate agreement is not an estimate of the rounding weight in the implemented W2 dominant-tail quantizer (Section K.5). We fit c empirically for each method and configuration and use it as an index of the α-field.

Cross-method fits. We fit c per method on the seven-model panel at three bit widths (n=21 cells per method, n=9 for AR0; Figure 2). AutoRound, GPTQ, and AWQ have in-sample panel $R^2 > 0.97$. OmniQuant gives the weaker fit reported above in its native-scale coordinate. The law is established here for per-group scalar quantizers at W2–W4.

AutoRound's block-level iterative optimization has the largest, c=1.76, close to two grid steps of clipping headroom traded for rounding precision. GPTQ's single-pass column-wise Hessian update gives c=0.49, about half a step, and AWQ's importance-scaled clip search gives c=0.35, closest to absmax. AR0 performs only AutoRound's importance-weighted scale grid search, with no SignSGD rounding iterations, and still lands in the α-law family (c=1.38, n=9, R^2=0.89; Section C). AR0's scale search alone is thus enough to leave the absmax family; whether it is enough for downstream accuracy is the subject of Section 4.

Mistral-7B is a systematic outlier for GPTQ, with lower α at every bit width. Excluding Mistral raises the GPTQ fit from R^2=0.982 to 0.999 and the AutoRound fit from 0.997 to 0.999, leaves AWQ unchanged at 0.976, and does not affect AR0, which has no Mistral cell. The outlier is concentrated in GPTQ.

Refitting five alternative single-parameter monotone forms (constant, linear, $\tanh$, exponential, power-law) on the same 21-cell panel gives at best R^2=0.82, against R^2>0.97 for Equation (2) on AutoRound, GPTQ, and AWQ (Section I).

Held-out width prediction. In a separate same-model width holdout, fitting only the original May W3/W4 cells predicts W2 on those models with maximum absolute median-α errors of 0.0155 for AutoRound and 0.0246 for GPTQ; AWQ reaches 0.0367 and the nine-cell AR0 cohort reaches 0.1112 (Table 18). Figure 3 separates these width-holdout errors from the new-model test in Section 4.

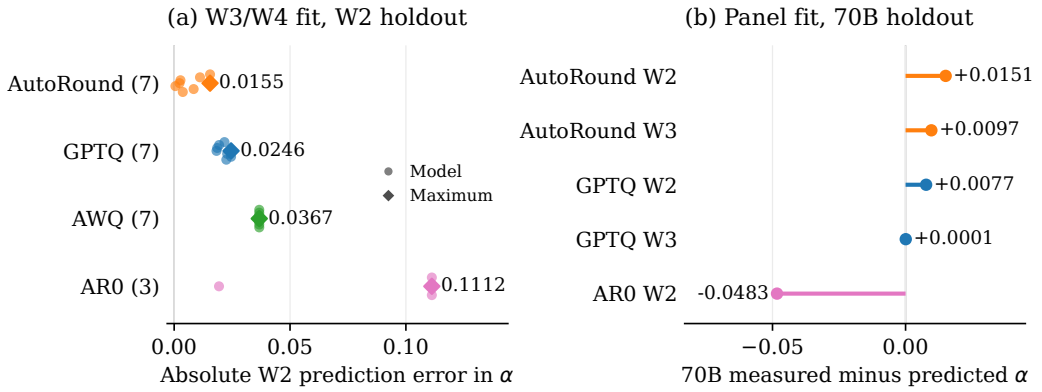


Figure 3: Held-out scale predictions at group size 128. (a) Absolute W2 errors after fitting W3/W4 on the same models. Circles show individual models; diamonds and labels show maxima. Model counts are in parentheses. AWQ uses transformed-space α; the other methods use reference-weight normalization. (b) Signed Llama-3.1-70B residuals using each method's own W2–W4 panel fit. Each point is a mean of checkpoint medians over nine sampled layers, with 8, 1, 8, 2, and 7 checkpoints from top to bottom. AutoRound and GPTQ transfer within 0.016 and 0.008; AR0's W2 residual is -0.048. Full results are in Tables 2 and 18.

3.2 Accuracy and Activation-Level Comparisons

c tracks method-mean W2 accuracy on the panel. Across 28 (method, model) cells, a per-cell regression of W2 accuracy on c gives a slope of 15.0 pp per unit c (Pearson r=0.87, R^2=0.76). The conditional standard error of 1.6 pp treats the cells as independent despite only four distinct method-level c values; it does not measure uncertainty over methods. We read the slope as an association across the four method means with within-method spread. AWQ (31.0%, per-model range 29.5–32.2%) and GPTQ (31.8%, range 29.9–36.4%, with Gemma-2-9B at 36.4%) sit near chance at W2; AR0 recovers part of the gap (40.5%, range 31.1–49.1%) and AutoRound most of it (54.3%, range 43.0–60.3%), each on seven models. AR0 has the widest within-method spread, and on Qwen3-1.7B it is level with GPTQ. Splitting the panel into base (n=4) and instruction-tuned (n=3) subpanels preserves the ordering AWQ≤GPTQ<AR0<AR in both; single-pass methods shift by under 2 pp between subpanels, AR0 by about 1 pp, and AutoRound by about 4 pp (instruct above base), without any rank inversion. AutoRound's W2 accuracy is also stable to the calibration corpus, moving by at most 3.6 pp across three non-Pile recipes on six panel models (Section J). The per-cell breakdown is in Section H.

Pathway-flip rate scaling. Across 14 Pile-to-αEC cells (7 models × 2 widths), GPTQ's median pathway-flip rate across layers is 4.4–9.5 times AutoRound's at W2 on every model, and 1.5–5.3 times at W3 (Table 1; per-layer curves for three models in Figure 8, Section F). The ratio decreases from W2 to W3 in five of seven models, increasing for Llama-2-7B and Qwen3-1.7B. The law predicts a smaller gap between the two methods' scale multipliers at higher bit widths, since a larger $q_{\max}$ shrinks the difference between α^* at c=1.76 and at c=0.49 (at W4 the medians are 0.80 and 0.93). The smaller scale gap is consistent with the flip-rate trend, although this comparison provides neither a quantitative mapping nor evidence of an effect above seed-swap noise across all seven models. The per-layer profile is shared between the two methods ($\rho \geq 0.78$ in 12 of 14 cells; the exceptions are Phi-3 and Qwen3-1.7B at W2), with a lower median under AutoRound in every cell.

4 Validation at 70B

The α-law and the c-index were fit on seven models of at most 9B parameters. As a separate new-model check, we test whether its fitted constants predict the sampled median α of held-out Llama-3.1-70B. We quantize Llama-3.1-70B (80 layers, group size 128) with AutoRound, GPTQ,

Table 1: Median across layers of the Pile-to-αEC calibration-contrast pathway-flip rate under AutoRound and GPTQ, their ratio, and the Spearman correlation of the two per-layer profiles, for 7 models at W2 and W3. All cells use seed 42; each method uses 130 prompts for Llama-2-7B, Phi-3-mini-4k, and Qwen3-1.7B, and 250 prompts for the other four models.

Model	Width	AR flip	GPTQ flip	GPTQ/AR	$\rho_{\mathrm{AR,GPTQ}}$
Llama-2-7B	W2	0.154	0.700	4.55	+0.93
Mistral-7B	W2	0.122	0.810	6.64	+0.98
Phi-3-mini-4k	W2	0.104	0.985	9.48	+0.53
Qwen3-1.7B	W2	0.204	0.888	4.36	+0.16
Qwen3-8B	W2	0.130	0.672	5.17	+0.93
Qwen2.5-7B	W2	0.096	0.824	8.58	+0.79
Gemma-2-9B	W2	0.060	0.374	6.23	+0.98
Llama-2-7B	W3	0.062	0.327	5.27	+0.99
Mistral-7B	W3	0.058	0.176	3.03	+1.00
Phi-3-mini-4k	W3	0.077	0.273	3.55	+0.94
Qwen3-1.7B	W3	0.046	0.238	5.17	+0.99
Qwen3-8B	W3	0.088	0.130	1.48	+0.98
Qwen2.5-7B	W3	0.066	0.114	1.73	+0.98
Gemma-2-9B	W3	0.040	0.068	1.70	+0.94

Table 2: Llama-3.1-70B sampled median α against each method's own panel fit (AutoRound c=1.763622, AR0 c=1.381732, GPTQ c=0.485675). Predictions and residuals are computed before rounding. AutoRound and GPTQ residuals are below 0.016 and 0.008, respectively; AR0's is -0.048. Means and standard deviations summarize checkpoint medians pooled across available Pile and αEC conditions. RTN-naive is listed in Table 14.

Method	Bits	Checkpoints	α (mean$\pm$std)	Predicted	Δ
AutoRound	W2	8	0.377 ± 0.001	0.362	$+0.015$
AutoRound	W3	1	$0.639\pm$—	0.630	$+0.010$
AR0	W2	7	0.372 ± 0.000	0.420	-0.048
GPTQ	W2	8	0.681 ± 0.001	0.673	$+0.008$
GPTQ	W3	2	0.861 ± 0.000	0.861	$< +0.001$

AWQ, AR0, and RTN-naive at W2 and W3; the full inventory of α-field checkpoints and downstream runs is in Section G. We report the median α across groups in nine evenly spaced layers of the 80. The α-law prediction is tested for AutoRound and GPTQ at W2 and W3 and for AR0 at W2 using each method's own panel fit; RTN-naive is a degenerate absmax field by construction. Downstream accuracy is reported for all five methods at W2 and for AutoRound, AWQ, and GPTQ at W3.

The sampled median α extrapolates from the $\leq$9B panel. Per-method c fitted on the panel predicts the median α over the sampled 70B layers within 0.016 for AutoRound and within 0.008 for GPTQ at both W2 and W3 (Table 2). The predictions use the unrounded panel fits, c=1.763622 and c=0.485675, respectively. AR0's nine-cell fit (c=1.38) predicts 0.420 at W2, against the observed 0.372, a residual of -0.048. Its own fitted constant therefore does not transfer to this 70B scale field. Using AutoRound's fitted c=1.763622 instead gives 0.362 and a residual of $+0.010$ for AR0. This cross-method comparison concerns their sampled medians, not a held-out prediction from AR0's fitted c. The standard deviation across checkpoint medians is at most 0.001, pooling up to eight checkpoints across seeds and calibration conditions.

Similar sampled medians, distinct per-group scales. AR0 has a sampled 70B-W2 median α of 0.372, close to full AutoRound's 0.377, yet retains only 45% of FP16 accuracy against AutoRound's 91% (Table 3). We also pair their checkpoint scales by group over 63 projections in nine sampled layers under Pile calibration. Across $60{,}162{,}048$ groups per matched seed, scale signs agree everywhere. Despite similar medians, paired per-group scales differ by a median of 3.88% at seed 123 and 3.81% at seed 456, measured as $\big||s_{g,\mathrm{AR0}}|/|s_{g,\mathrm{AR}}| - 1\big|$. Both runs use the same filtered Pile-10k calibration data; enabling the full SignSGD procedure changes per-group scales and raises retention by 46 pp. The comparison does not isolate the effect of rounding decisions at fixed scales.

Table 3: Llama-3.1-70B downstream accuracy at group size 128 (ten-task mean). W2 uses 4 Pile-10k seeds for AutoRound, AR0, and GPTQ and 1 seed for AWQ and RTN-naive. W3 uses 1 Pile-10k seed. Retention is the ten-task mean divided by the FP16 mean. AutoRound and AR0 have sampled W2 median α values of 0.377 and 0.372 (Table 2) and distinct paired per-group scales, yet differ by 46 pp in retention.

		W2		W3	
Method	Small-model c	Mean	Retention	Mean	Retention
FP16 (base)	—	0.7215	100%	0.7215	100%
AutoRound	1.76	0.659	91%	0.707	98%
AR0	1.38	0.321	45%	—	—
GPTQ	0.49	0.367	51%	0.680	94%
AWQ	0.35	0.322	45%	0.698	97%
RTN-naive	—	0.297	41%	—	—

Table 4: Robustness of the W2–W4 law and its fitted coefficient. All rows use group size 128 except the G64/G128/G256 sweeps. Corpus changes are relative to Pile; solver changes use the main-panel setting; tuning compares one matched base/instruct pair. AR denotes AutoRound.

Setting and model	Method	R^2	Coefficient behavior
Corpora, Llama-2-7B	GPTQ	≈ 0.9995	change $\leq 0.078\%$
Solver, Llama-2-7B	GPTQ	0.9969–0.9996	change $\leq 3.4\%$
Groups, Llama-2-7B	GPTQ	0.9994–0.9995	span 0.27%
Groups, Mistral-7B	AR	≥ 0.998	G64 to G256 $+26.5\%$
Mixtral-8x7B	GPTQ	0.9993	$c = 0.4705$
Mixtral-8x7B	AR	0.9980	$c = 1.69$
Tuning, Mistral-7B	AR	> 0.9996	change -0.8%

Small-model c ordering breaks at 70B-W2. The two methods with $c<0.5$ behave differently at 70B. AWQ (c=0.35) collapses to 0.322 (45%) with LAMBADA at 0.000, the signature of RTN-naive (0.297, 41%). GPTQ (c=0.49) keeps 0.367 (51%) under Pile and 0.390 (54%) under αEC, with LAMBADA at 0.14–0.21; it is degraded but functional. AR0 (c=1.38) sits at 0.321, level with AWQ, so between AWQ, GPTQ, and AR0 the small-model c ordering does not transfer to 70B accuracy. AutoRound, the one method with both a large c and rounding iterations, retains 91%; every other tested method retains 41–54%. At W3, AutoRound (0.707), AWQ (0.698), and GPTQ (0.680) all retain 94–98% (single seed), so their accuracy gap narrows from W2 to W3, as it does on the small-model panel.

5 Discussion

Robustness of the law and of c. The functional form persists across the tested corpora, solver settings, group sizes, and Mixtral projections (Table 4). Coefficient stability depends on the method and configuration. GPTQ's c varies little across group sizes 64, 128, and 256, whereas AutoRound's c rises with group size on both Llama-2-7B and Mistral-7B. The law therefore describes a shared form of scale selection with configuration-dependent contraction. Full controls and fits are in Section K.

The Mixtral fits cover all 896 stored attention and expert projections; GPTQ's expert-only fit gives c=0.4704. For GPTQ, top-1 agreement with FP16 expert routing over 11,122 matched tokens rises from 63.0% at W2 to 88.5% at W3 and 93.9% at W4.

Calibration content at the activation level. Calibration composition has been reported to change downstream accuracy at W2 and W3 (Williams & Aletras, 2024). We asked whether it also leaves a per-layer signature in the residual stream, using the calibration contrast of Section 2 on Llama-2-7B, Phi-3-mini-4k-instruct, and Qwen3-1.7B at three bit widths for GPTQ and AutoRound (Section B). A calibration-content effect above the seed-swap noise floor appears in one method–width combination, GPTQ at W3, and replicates on all three models. AutoRound has no cell significant at $p<0.05$ on

Llama-2; its W2 and W3 results are marginal (p=0.067 and 0.061) on 130 matched prompts. W4 is at noise for both methods, and GPTQ at W2 is below the usable-accuracy floor. Activation patching tests divergence at one layer and one position, so the AutoRound null does not show that AutoRound is insensitive to calibration.

Practical use. The c-index gives a common coordinate for comparing the scale contraction chosen by different quantizers. A reader can extract normalized scales from checkpoints and reference weights, fit c at W3 and W4, and predict the median scale multiplier at W2 without downstream task evaluation. The width holdout in Table 18 tests this workflow, and Table 2 tests transfer to a larger model. Repeating the fit under a changed configuration separates stability of the functional form from changes in scale contraction, as the group-size experiments show. The panel's method-mean accuracy ordering supplies a downstream association; prospective accuracy-based method selection remains untested.

Limitations. The calibration-contrast patching study covers three models; the uncontrolled Pile-to-αEC flip-rate comparison covers seven models with 130 or 250 prompts per method and cell, and the α-law itself needs only checkpoint reads. At 70B each method's own fitted c is tested for AutoRound and GPTQ at W2 and W3 and for AR0 at W2, where AR0 misses by 0.048; AWQ and RTN-naive at W2 have a single seed, and W3 downstream cells have one seed per calibration condition. The GPTQ W3 scale summary has two fields, one from each calibration condition. OmniQuant and AQLM were not run at 70B. The MoE extension covers one Mixtral configuration. Three of the seven panel models are instruction-tuned; the matched Mistral pair bounds the tuning effect in one setting only. The panel uses group size 128; the G64 and G256 runs show the form persists but AutoRound's c depends on group size. The W2 quantizer tested selects the sign of an asymmetric endpoint per group and has no independent zero point (Section K); general affine quantization with a learned zero point is untested.

6 Conclusion

We find shared predictive structure in the scale choices of low-bit quantization methods with different optimization procedures. The α-law describes this structure for AutoRound, GPTQ, and AWQ through a configuration-dependent measure of scale contraction. Held-out tests support prediction across bit widths and, for AutoRound and GPTQ, transfer to Llama-3.1-70B. At 70B and two bits per weight, AutoRound and AR0 have similar sampled median multipliers but different per-group scales and a 46-percentage-point gap in retention of the FP16 reference model's ten-task mean accuracy.

AI use statement

In this work, we used generative AI tools to draft sections of the paper, to edit author-written text for grammar, spelling, and wording, and to assist with plotting and data-formatting scripts. We have not used generative AI tools for research ideation, experimental design, or generation of results. Large language models are the subject of quantization and analysis in this study, not a tool used to develop the method. We have reviewed all AI-assisted work, and AI-assisted scripts were verified against the recorded artifacts by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

Hardware, software, calibration-data provenance, and the scale-extraction and activation-patching harnesses are inventoried in §L; sample-size and seed accounting is given in §A; the symmetric clipped-MSE approximation is in §D. The deterministic build script for the αEC calibration set (§2) is included in the supplementary code. Code and quantized checkpoints will be released upon acceptance.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint*

arXiv:2404.14219, 2024.

Eduardo Alvarez, Omri Almog, Eric Chung, Simon Layton, Dusan Stosic, Ronny Krashinsky, and Kyle Aubrey. Introducing NVFP4 for efficient and accurate low-precision inference. NVIDIA Technical Blog, June 2025.

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper/2019/hash/c0a62e133894cdce435bcb4a5df1db2d-Abstract.html.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569. PMLR, 2018. URL https://proceedings.mlr.press/v80/bernstein18a.html.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*, 2020.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, volume 36, 2023. doi: 10.52202/075280-0196. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0df38cd13520747e1e64e5b123a78ef8-Abstract-Conference.html.

Wenhua Cheng, Weiwei Zhang, Haihao Shen, Yiyang Cai, Xin He, Kaokao Lv, and Yi Liu. Optimize weight rounding via signed gradient descent for the quantization of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11332–11350. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.662. URL https://aclanthology.org/2024.findings-emnlp.662/.

Wenhua Cheng, Weiwei Zhang, Heng Guo, Haihao Shen, and Zaner Ma. SignRoundV2: Toward closing the performance gap in extremely low-bit post-training quantization for LLMs. *arXiv preprint arXiv:2512.04746*, 2025. URL https://arxiv.org/abs/2512.04746.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *NAACL-HLT*, pp. 2924–2936, 2019. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

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.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Bita Darvish Rouhani, Ritchie Zhao, Ankit More, Mathew Hall, Alireza Khodamoradi, Summer Deng, et al. Microscaling data formats for deep learning. *arXiv preprint arXiv:2310.10537*, 2023. URL https://arxiv.org/abs/2310.10537.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. SpQR: A sparse-quantized representation for near-lossless LLM weight compression. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.03078.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. In *International Conference on Machine Learning (ICML)*, volume 235, pp. 12284–12303, 2024. URL https://proceedings.mlr.press/v235/egiazarian24a.html.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. doi: 10.52202/068431-0323. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/1caf09c9f4e6b0150b06a07e77f2710c-Abstract-Conference.html.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, December 2023. URL https://zenodo.org/records/10256836.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *arXiv preprint arXiv:2301.04709*, 2023. URL https://arxiv.org/abs/2301.04709.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*, 2021. URL https://arxiv.org/abs/2103.13630.

Ruihao Gong, Yang Yong, Shiqiao Gu, Yushi Huang, Chengtao Lv, Yunchen Zhang, Dacheng Tao, and Xianglong Liu. LLMC: Benchmarking large language model quantization with a versatile compression toolkit. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 132–152. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-industry.12. URL https://aclanthology.org/2024.emnlp-industry.12/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal Brain Surgeon. In *Advances in Neural Information Processing Systems*, volume 5, 1992. URL https://proceedings.neurips.cc/paper_files/paper/1992/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Yehong Jiang, Yen-Kuang Chen, and Xinmin Tian. A method-class divide in sub-4-bit quantization: Iterative vs. single-pass sensitivity to data composition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. Accepted to the NeurIPS 2026 Main Track as a poster; proceedings metadata pending.

Sehoon Kim, Coleman Richard Charles Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23901–23923, 2024. URL https://proceedings.mlr.press/v235/kim24f.html.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021. URL https://arxiv.org/abs/2102.05426.

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. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 6, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2381–2391, 2018. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? Adaptive rounding for post-training quantization. In *International Conference on Machine Learning (ICML)*, volume 119, pp. 7197–7206, 2020. URL https://proceedings.mlr.press/v119/nagel20a.html.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1144. URL https://aclanthology.org/P16-1144/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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. doi: 10.1145/3474381.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2308.13137.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with Hadamard incoherence and lattice codebooks. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48630–48656, 2024. URL https://proceedings.mlr.press/v235/tseng24a.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Miles Williams and Nikolaos Aletras. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10100–10118. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.544. URL https://aclanthology.org/2024.acl-long.544/.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning (ICML)*, volume 202, pp. 38087–38099, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the Association for Computational Linguistics (ACL)*, 2019.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577, 2024. doi: 10.1162/tacl_a_00704. URL https://aclanthology.org/2024.tacl-1.85/.

A Sample-Size and Seed Accounting

Table 5 fixes the per-cell sample sizes for the Llama-2-7B calibration-content activation-patch contrast. Each method×width cell has four effect calibration seeds and four noise seed-pair sweeps; W3 has two additional noise seed pairs. Phi-3 and Qwen3 use four effect seeds and four noise pairs at every width. The controlled sweeps use 130 patching prompts per seed or seed pair, distinct from the 200/208 calibration examples used to construct the quantized checkpoints.

Table 5: Llama-2-7B controlled activation-patch per-cell sample sizes. "Effect" = Pile→αEC sweeps; "Noise" = within-variant seed-swap sweeps.

Method	Width	Effect seeds	Noise seed-pairs	Prompts/seed
GPTQ	W2	4	4	130
GPTQ	W3	4	6	130
GPTQ	W4	4	4	130
AR	W2	4	4	130
AR	W3	4	6	130
AR	W4	4	4	130

B Activation-Level Effect of Calibration Content

This section reports the calibration contrast of Section 2 on Llama-2-7B, Phi-3-mini-4k-instruct, and Qwen3-1.7B for GPTQ and AutoRound at W2, W3, and W4. For each cell, the *effect* sweep patches residuals between the Pile-calibrated and αEC-calibrated models and the *noise* sweep patches between two seeds of the same calibration set. Two summaries are used. The first counts changes from the recipient's unpatched top-1 prediction across all 130 prompts. We select the layer with the largest seed-averaged effect rate and apply a paired t-test to prompt-level rates averaged over effect seeds and noise pairs (Table 6, Llama-2-7B). The reported p-values are unadjusted for selecting that layer using the same prompts. The second uses patched answer-error rates on the αEC-correct/Pile-wrong prompt subset and the both-correct control subset, fixed using each model's GPTQ W3 seed-42 pair (50 and 20 prompts, respectively). For each sweep type, Δ is the former rate minus the latter, averaged over prompts and seeds or seed pairs. We evaluate $r{=}\Delta_{\text{eff}}/\Delta_{\text{noise}}$ at the layer with the most-negative mean effect on the answer logit in the disagreement subset (Table 7, all three models). *Base-wrong* is the unpatched recipient's answer-error rate on that subset, averaged over effect seeds; a value near 1 marks a baseline too degraded for the contrast to be informative.

Table 6: Peak-layer calibration pathway-flip rate on Llama-2-7B, effect against noise, with unadjusted paired prompt-level p-values. W4 cells are regime controls.

Method	**Width**	**Effect flip**	**Noise flip**	**Ratio**	**p-value**	**Status**
GPTQ	W2	0.752	0.754	1.00	0.93	ns (floor)
GPTQ	**W3**	**0.412**	**0.291**	**1.41**	$\mathbf{2.3 \times 10^{-6}}$	***
GPTQ	W4	0.102	0.098	1.04	0.73	ns (regime ctrl)
AR	W2	0.265	0.215	1.23	0.067	ns
AR	W3	0.154	0.123	1.25	0.061	ns
AR	W4	0.077	0.077	1.00	1.00	ns (regime ctrl)

Table 7: Control-adjusted answer-error ratio r at the layer of most-negative mean answer-logit change, with the peak layer index and base-wrong, for three models. GPTQ W3 is the only method–width combination with r well above 1 on every model; GPTQ W2 has base-wrong ≥ 0.98 on all three and is uninformative.

		Llama-2-7B (32 L)			Phi-3-mini (32 L)			Qwen3-1.7B (28 L)		
Method	Width	peak L	r	base-wrong	peak L	r	base-wrong	peak L	r	base-wrong
GPTQ	W2	—	0.85	0.98	0	—	1.00	5	—	1.00
GPTQ	W3	17	**1.37**	0.33	31	**2.37**	0.40	26	**5.26**	0.43
GPTQ	W4	—	0.95	0.34	30	1.10	0.36	23	0.94	0.30
AR	W2	—	0.82	0.59	30	1.14	0.65	27	0.76	0.68
AR	W3	—	0.87	0.32	31	1.26	0.39	7	0.85	0.33
AR	W4	—	1.07	0.31	14	0.92	0.34	25	0.95	0.31

On Llama-2-7B, GPTQ W3 gives a layer-31 effect rate of 0.412 against a noise rate of 0.291 (ratio 1.41, $p{=}2.3{\times}10^{-6}$); no other cell reaches $p{<}0.05$. The separate control-adjusted ratio is $r{=}1.37$ at layer 17. AutoRound W2 and W3 have marginal results ($p{=}0.067$ and $p{=}0.061$, respectively). The GPTQ W3 signal replicates on Phi-3-mini ($r{=}2.37$) and Qwen3-1.7B ($r{=}5.26$), while the twelve other informative cells across the three models fall in $r{\in}[0.76, 1.26]$. The peak layer differs by architecture (L=17 of 32 on Llama-2-7B, 31 of 32 on Phi-3, 26 of 28 on Qwen3-1.7B). No AutoRound cell reaches $p{<}0.05$ in the Llama-2 test, and GPTQ W4 is at noise on all three models; the marginal AutoRound results are not evidence of absence. Failure to detect a per-layer residual-stream difference between calibration variants does not rule out weight-space effects on downstream accuracy.

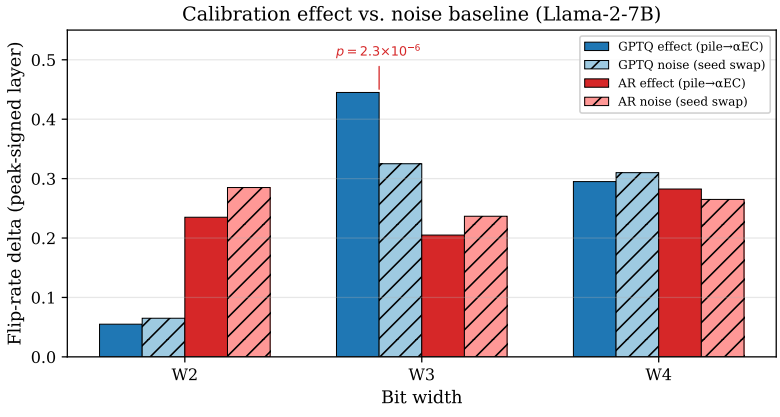


Figure 4: Control-adjusted answer-error rates on Llama-2-7B, effect (filled) against noise (hatched), at each cell's layer of most-negative mean answer-logit change. The annotated p-value tests the separate raw pathway-flip rates in Table 6, not these bars.

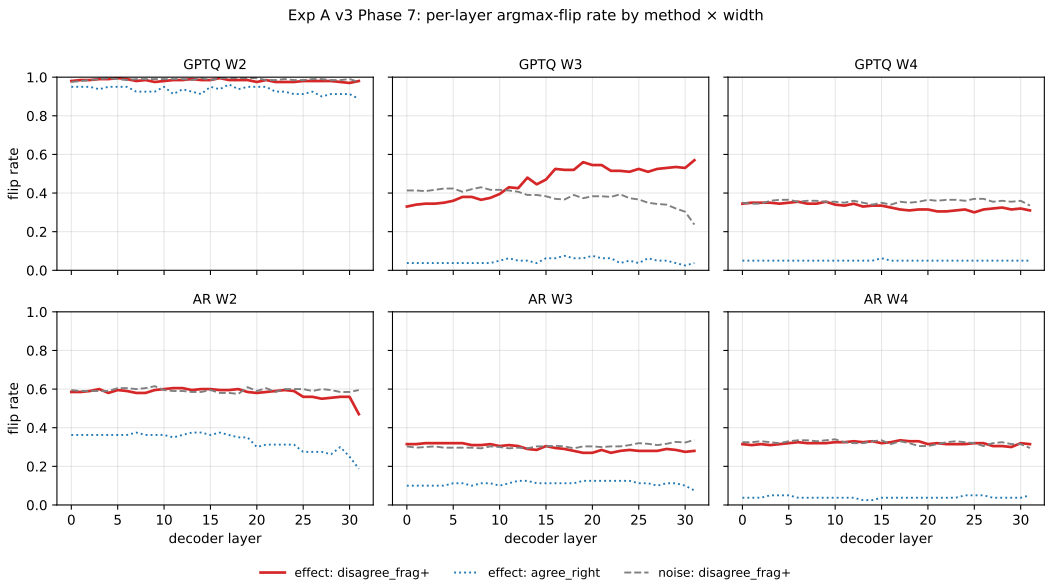


Figure 5: Llama-2-7B per-layer patched answer-error rates by method×width. Solid and dashed curves show effect and noise on the αEC-correct/Pile-wrong subset; dotted curves show effect on the both-correct control. GPTQ W3 shows the cross-model excess summarized in Table 7.

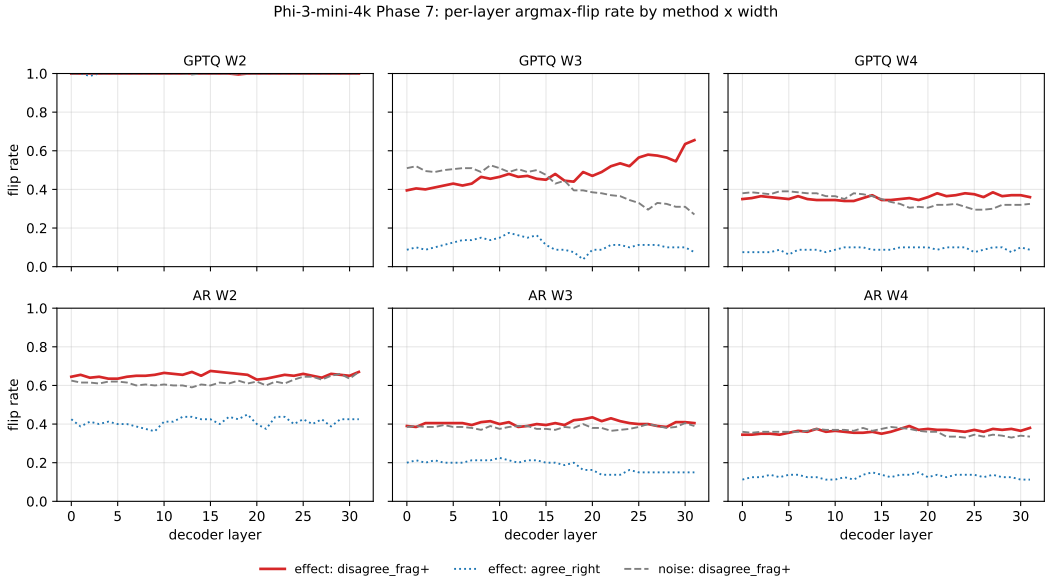


Figure 6: Phi-3-mini-4k-instruct, same conventions as Figure 5.



Figure 7: Qwen3-1.7B, same conventions as Figure 5.

C Per-Method Scale-Family Detail Tables

Table 8: Scale-family taxonomy, averaged over Llama-2-7B, Phi-3-mini-4k-instruct, and Qwen3-1.7B, the three models with per-method detail in Section C; the α-law fits in Figure 2 use the full seven-model panel. RTN-naive and OmniQuant (exported scale) keep median α near 1. GPTQ, AWQ, AR0, and AutoRound move α below 1 by a method-specific amount that shrinks as the bit width grows. Per-method detail is in Section C.

Method	Type	W	ρ_{absmax}	α_{med}	CV(α)
RTN-naive	baseline	W2	1.000	1.000	0.000
		W3	1.000	1.000	0.000
		W4	1.000	1.000	0.000
GPTQ	single-pass	W2	0.910	0.681	0.101
		W3	0.973	0.860	0.050
		W4	0.993	0.934	0.024
AWQ	single-pass	W2	0.998	0.750	0.030
		W3	0.980	0.876	0.048
		W4	0.981	0.958	0.035
AR0	scale-only	W2	0.919	0.456	0.044
		W3	0.906	0.631	0.083
		W4	0.980	0.855	0.019
AutoRound	iterative-round	W2	0.765	0.368	0.132
		W3	0.867	0.629	0.092
		W4	0.942	0.803	0.061
OmniQuant	iterative-clip	W2	0.563	1.016	0.214
		W3	0.759	0.955	0.129
		W4	0.915	0.968	0.074

C.1 GPTQ Absmax Correspondence

Table 9: Spearman ρ between GPTQ scales and absmax predictor $M_g/q_{\max}$. $\rho \geq 0.899$ in all 9 cells.

Family	W2 ($q_{\max} = 1$)	W3 ($q_{\max} = 3$)	W4 ($q_{\max} = 7$)
Llama-2-7B	0.899	0.973	0.993
Phi-3-mini	0.910	0.971	0.992
Qwen3-1.7B	0.919	0.977	0.994

C.2 AUTOROUND α-FIELD

Table 10: AutoRound α-field statistics across 3 families $\times$ 3 widths. α_{med} follows $q_{\max}/(q_{\max}+1.75)$ to within 0.012.

Family	W	$q_{\max}$	α_{med}	CV(α)	$q/(q+1.75)$
Llama	W2	1	0.375	0.128	0.364
Llama	W3	3	0.636	0.088	0.632
Llama	W4	7	0.808	0.057	0.800
Phi-3	W2	1	0.371	0.127	0.364
Phi-3	W3	3	0.632	0.089	0.632
Phi-3	W4	7	0.804	0.058	0.800
Qwen3	W2	1	0.357	0.141	0.364
Qwen3	W3	3	0.620	0.099	0.632
Qwen3	W4	7	0.796	0.070	0.800

C.3 AWQ IN TRANSFORMED SPACE

Table 11: AWQ clip ratio $\alpha_t = c_g/\max_{\text{group}}|W'|$ in the transformed weight space, with clipping threshold c_g and transformed-group absmax M'_g. Here α_t averages matrix-level group medians; the law fits take their median. CV and ρ are averaged across the same matrices with learned clipping thresholds, with $\rho \geq 0.97$.

Family	W	q	α_t	CV(α_t)	$\rho(c_g,\, M'_g)$
Llama	W2	1	0.751	0.024	0.998
Llama	W3	3	0.877	0.044	0.979
Llama	W4	7	0.959	0.034	0.980
Phi-3	W2	1	0.750	0.021	0.998
Phi-3	W3	3	0.876	0.047	0.977
Phi-3	W4	7	0.956	0.036	0.975
Qwen3	W2	1	0.750	0.046	0.997
Qwen3	W3	3	0.876	0.052	0.984
Qwen3	W4	7	0.958	0.035	0.988

D SYMMETRIC CLIPPED-MSE APPROXIMATION

Stationarity condition. This stationarity condition describes the symmetric model in Equation (1), not the dominant-tail W2 quantizer used in the experiments (Section K.5). Setting $d\,\text{MSE}_\lambda/dT = 0$ in Eq. 1 gives

$$\frac{\lambda\, a}{6\, q_{\max}^2}[P_{\text{in}} + a\,\phi(a)] = 4\big[\phi(a) - a\,\bar{\Phi}(a)\big]\,, \tag{3}$$

where $\bar{\Phi}(a) = 1 - \Phi(a)$. For Gaussian groups of size n=128, the mean normalized maximum is $b_n = \mathbb{E}[M_g]/\sigma \approx 2.828$. At λ=1, $\alpha^*_{\text{MSE}} = a^*/b_n \approx 0.437/0.691/0.875$ at $q_{\max}$=1/3/7, over-predicting observed α by $\sim$7 pp. The Gaussian assumption is a within-group simplification. The per-group partition and M_g rescaling do not establish that the groups are Gaussian. The $\sim$7 pp over-prediction across $q_{\max}$ marks a mismatch between this approximation and the observed medians; the empirical c-fit does not identify its cause.

Grid-step sacrifice. The effective grid-step sacrifice is

$$c(q) = q_{\max}\left(\frac{b_n}{a^*} - 1\right). \tag{4}$$

At the per-group MSE optimum (λ=1), c_{MSE} is $1.29/1.34/1.00$ at q=1/3/7. Per-group MSE alone therefore does not give a constant c.

Block-level coupling. AR minimizes a block-level loss $\|\mathbf{X}(\mathbf{w}-\mathbf{w}_q)\|^2$ coupling all groups through activations $\mathbf{X}$. We model a possible preference for reducing dense rounding errors over sparse clipping errors with the factor λ; the block loss alone does not determine that factor.

Table 12: Per-group MSE-optimal (λ=1) vs. rounding-upweighted vs. observed AR α. Each α_λ uses the listed width-specific $\lambda_{c=1.75}$; $\Delta = \alpha_\lambda - \alpha_{\text{obs}}$.

q	α_{MSE}	c_{MSE}	$\lambda_{c=1.75}$	α_λ	α_{obs}	Δ	$q/(q{+}1.75)$
1	0.437	1.29	1.96	0.364	0.367	-0.003	0.364
3	0.691	1.34	1.67	0.632	0.629	$+0.003$	0.632
7	0.875	1.00	2.03	0.800	0.802	-0.002	0.800

The fitted λ per bit width in this symmetric model is $1.96/1.67/2.03$ (mean 1.89, range < 0.4). A single $\lambda{\approx}1.9$ approximately matches AutoRound's observed c at all three widths; it is not a measured weight for the implemented dominant-tail W2 quantizer.

CV scaling. CV(α) decreases monotonically with $q_{\max}$:

$$\text{CV}(\alpha) \approx \frac{0.63}{q_{\max} + 4.2}, \tag{5}$$

with $R^2 = 0.92$. Multi-seed confirmation shows $\sigma(\text{CV}) < 0.0003$.

E Multi-Seed α-Law Robustness

Table 13: AutoRound α-field statistics on Llama-2-7B, Pile calibration, G128, and four seeds per width. Each seed summarizes 224 matrices by the median of their group medians and the median of their within-matrix CVs. Entries report means and population standard deviations across seeds. Inter-seed $\sigma(\alpha_{\text{med}}) < 0.0005$.

Width	$q_{\max}$	α_{med}	σ_{seed}	CV(α)	σ_{seed}(CV)	$q/(q{+}1.75)$
W2	1	0.3741	0.00048	0.1140	0.00028	0.3636
W3	3	0.6336	0.00017	0.0813	0.00024	0.6316
W4	7	0.8042	0.00010	0.0538	0.00017	0.8000

The individual Llama-2-7B α_{med} values, indexed by seed, are

- W2, α_{42}=0.3740, α_{123}=0.3734, α_{456}=0.3744, α_{789}=0.3746 (range 0.0012)
- W3, α_{42}=0.6334, α_{123}=0.6334, α_{456}=0.6337, α_{789}=0.6338 (range 0.0004)
- W4, α_{42}=0.8044, α_{123}=0.8041, α_{456}=0.8041, α_{789}=0.8042 (range 0.0003)

F SCALE-FAMILY ρ-HEATMAP AND PER-LAYER FLIP PROFILES

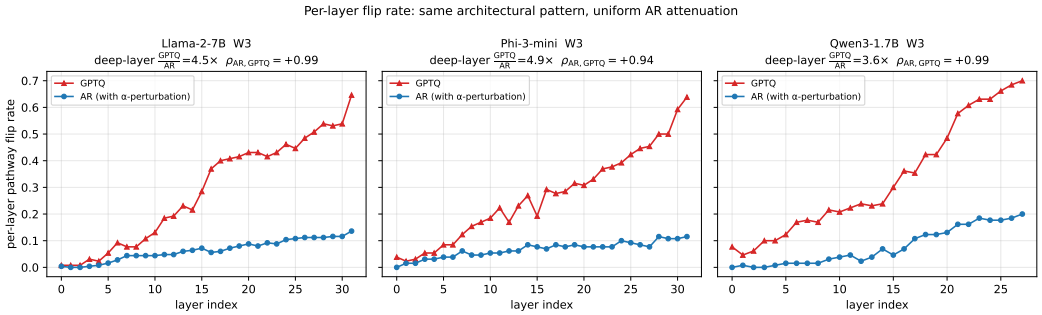


Figure 8: Per-layer Pile-to-αEC pathway-flip rate under AutoRound and GPTQ at W3 on Llama-2-7B, Phi-3-mini-4k-instruct, and Qwen3-1.7B. The plotted Llama-2 AutoRound curve uses 250 prompts; its GPTQ curve and both methods' Phi-3 and Qwen3 curves use 130. The Llama-2 table row instead uses the 130-prompt AutoRound sweep. The per-layer profiles are correlated at $\rho \geq 0.94$ and the GPTQ level is 3.6–5.3 times the AutoRound level; the remaining panel models are summarized in Table 1.

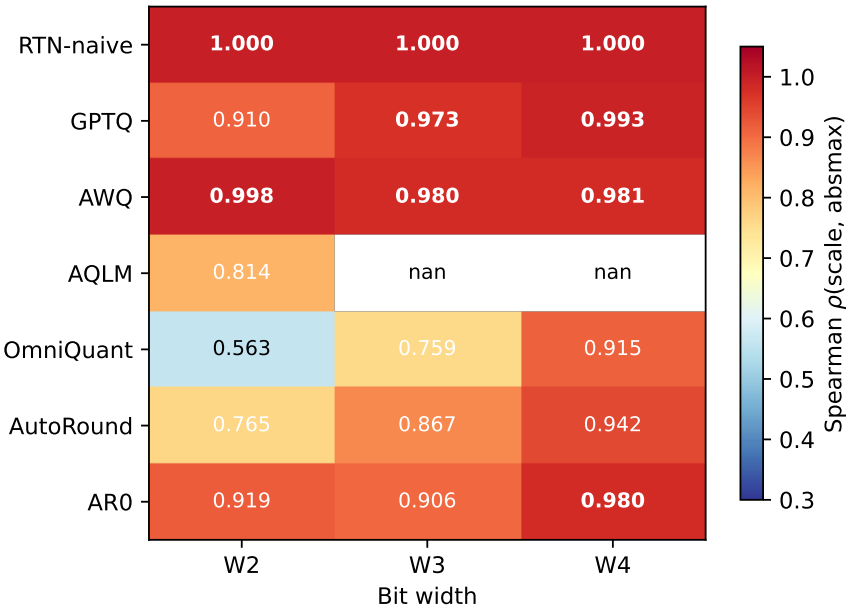


Figure 9: Spearman ρ(scale, absmax) by method and bit width (family-averaged). Family labels refer to median α; ρ varies within each family, from 0.56 (OmniQuant) to 1.00 (RTN-naive) at W2 in the absmax family. α-law-family methods diverge most from absmax at W2.

G 70B SCALE-FAMILY VALIDATION

To test whether the α-law generalizes beyond the $\leq$9B panel, we quantize Llama-3.1-70B (80 layers, G128) with AutoRound, GPTQ, AWQ, AR0, and RTN-naive at W2 and W3, under Pile and αEC calibration (Section 2), with up to four random seeds per condition. The sampled α-field inventory contains 27 checkpoints; the ten-task inventory contains 34 evaluation runs, including a separate

AutoRound W2 Pile-128 control for which no α-field was extracted. For AR0 at W2, seven scale fields were extracted across calibration conditions, while eight downstream runs were evaluated.

G.1 α-FIELD AT 70B

Nine evenly spaced layers are sampled per checkpoint, selecting each scale tensor once through the checkpoint index. We retain groups with $M_g > 10^{-10}$ and $0.01 < \alpha_g < 2$, using the original signed scales. Each checkpoint median pools these retained positive-alpha groups across sampled matrices, whereas the small-model AutoRound, GPTQ, and AR0 fits average matrix medians (Section 2). The paired-scale analysis below includes both signs.

Table 14: α-field summary at 70B scale. Predictions use each method's unrounded panel fit, c=1.763622 for AutoRound, c=1.381732 for AR0, and c=0.485675 for GPTQ; residuals precede rounding. RTN-naive is the absmax reference. AR0's fitted c overpredicts its observed W2 median by 0.048. Means and standard deviations summarize checkpoint medians pooled across available Pile and αEC conditions. A trimmed view appears in the main body as Table 2.

Method	Bits	Checkpoints	α (mean±std)	Predicted	Δ
AutoRound	W2	8	0.377 ± 0.001	0.362	+0.015
AutoRound	W3	1	0.639±—	0.630	+0.010
AR0	W2	7	0.372 ± 0.000	0.420	-0.048
GPTQ	W2	8	0.681 ± 0.001	0.673	+0.008
GPTQ	W3	2	0.861 ± 0.000	0.861	$< +0.001$
RTN-naive	W2	1	0.500±—[1]	absmax cluster	

AutoRound and GPTQ match their own fitted predictions within 0.016 and 0.008, respectively. AR0's own c=1.381732 predicts 0.420 at W2 and misses the observed 0.372 by 0.048. AutoRound's fitted c=1.763622 would predict 0.362 for AR0, a separate cross-method comparison with a residual of +0.010. The AR0 c=1.38 fit is reproduced from nine May model-width cells on Llama-2-7B, Phi-3-mini, and Qwen3-1.7B at W2–W4. Those cells average rounded per-module median α values. RTN-naive produces a degenerate field (α=0.50, CV=0). The standard deviation across checkpoint medians, pooling seeds and calibration conditions, is at most 0.001.

Matched AutoRound and AR0 W2 scales. For Pile seeds 123 and 456, we align the 63 sampled projection scale tensors by group using each checkpoint's safetensors index. Each pair contains 60,162,048 nonzero groups, including both scale signs; signs agree in every group at both seeds. The median of $\left||s_{g,\mathrm{AR0}}|/|s_{g,\mathrm{AR}}| - 1\right|$ is 3.88% at seed 123 and 3.81% at seed 456, with 61.1% and 61.9% of groups, respectively, differing by at most 5%. The archived AutoRound and AR0 median-α extractor scans physical shards and counts seven numerically identical duplicate scale tensors from layer 79. Each such archive entry includes 3,351,996 duplicate positive groups alongside 30,213,608 index-selected positive groups. The median tables above and the paired analysis use one index-selected copy per tensor. The original shard-weighted seed-mean medians were 0.376082 (AutoRound W2), 0.638596 (AutoRound W3), and 0.370483 (AR0 W2), compared with 0.376913, 0.639464, and 0.371554 after index selection. GPTQ has no duplicate sampled scales, and its medians are unchanged. The paired calculation is reproduced by `compare_70b_scales.py` in the supplementary code. The Pile runs share model, seed, group size, and filtered Pile-10k calibration data. AutoRound uses 200 rounding iterations while AR0 uses zero; the full procedure changes both per-group scales and downstream accuracy.

[1]The 70B RTN-naive checkpoint is exported by the auto-round library, which uses the negative-half scale convention s=absmax$(W)/2^{b-1}$ and yields α=$q_{\max}/2^{b-1}$=0.5 at W2. The ≤9B panel uses the positive-saturation RTN (s=absmax$(W)/q_{\max}$), which yields α=1. Both are degenerate absmax fields with ρ=1 and CV(α)=0; the difference is a constant rescale of the same scale geometry and does not affect cluster membership.

G.2 Downstream Evaluation at 70B

Table 15 lists every 70B downstream cell. At W2, AutoRound, AR0, and GPTQ have four Pile-10k seeds each; GPTQ additionally has four αEC seeds; AWQ and RTN-naive have one seed. AutoRound also has a separate single-seed Pile-128 control. At W3, AutoRound, AWQ, and GPTQ have one Pile-10k seed and one αEC seed. Cross-seed standard deviation at W2 is 0.003 for AutoRound, 0.002 for AR0, and 0.006 (Pile) or 0.010 (αEC) for GPTQ.

Table 15: Llama-3.1-70B ten-task mean by method, width, and calibration set. Retention is relative to the FP16 mean of 0.7215. The Pile-128 row is a separate control, not part of the Pile-10k/αEC comparison. LAMBADA is the per-task accuracy on `lambada_openai`, the task that collapses first at W2.

Method	Bits	Calibration	n	Mean	Retention	LAMBADA
AutoRound	W2	Pile-10k	4	0.659 ± 0.003	91%	0.75–0.76
AutoRound	W2	αEC	4	0.668 ± 0.003	93%	0.75–0.76
AutoRound	W2	Pile-128 control	1	0.654	91%	0.764
AR0	W2	Pile-10k	4	0.321 ± 0.002	45%	0.05–0.06
AR0	W2	αEC	4	0.324 ± 0.001	45%	0.06
GPTQ	W2	Pile-10k	4	0.367 ± 0.006	51%	0.14–0.17
GPTQ	W2	αEC	4	0.390 ± 0.010	54%	0.15–0.21
AWQ	W2	Pile-10k	1	0.322	45%	0.000
AWQ	W2	αEC	1	0.321	44%	0.000
RTN-naive	W2	—	1	0.297	41%	0.003
AutoRound	W3	Pile-10k	1	0.707	98%	0.790
AutoRound	W3	αEC	1	0.711	98%	0.787
AWQ	W3	Pile-10k	1	0.698	97%	0.768
AWQ	W3	αEC	1	0.699	97%	0.763
GPTQ	W3	Pile-10k	1	0.680	94%	0.755
GPTQ	W3	αEC	1	0.682	94%	0.748

At W2, AWQ and RTN-naive collapse with LAMBADA at or near 0; GPTQ is degraded but functional, with LAMBADA between 0.14 and 0.21 and a $+2.3$ pp gain from αEC calibration. AR0's sampled median α is 0.372, compared with AutoRound's 0.377, and it retains 45%, level with AWQ, with LAMBADA at 0.06. With the full SignSGD rounding procedure enabled, AutoRound retains 91% and has LAMBADA at 0.75; the sampled per-group scales agree in sign but differ in magnitude (Section G). At W3 all three evaluated methods retain 94–98% and the αEC calibration changes the mean by at most 0.4 pp.

H c-Index vs. Extreme-Bit W2 Accuracy

Table 16 reports per-(method, model) W2 Pile-calibration accuracy on the ten-task suite, averaged over 4 calibration seeds. A per-cell linear regression of W2 accuracy on c across n=28 cells (4 methods $\times$ 7 models) yields slope 15.0 ± 1.6 pp per unit c (Pearson r=0.87, R^2=0.76). The 1.6 pp is the regression's conditional standard error when the 28 cells are treated as independent. The cells share only four method-level c values, so this standard error does not quantify uncertainty over methods; the 28 cells describe within-method spread. AR0 cells for Llama-2-7B, Mistral-7B, Qwen2.5-7B-Instruct, Qwen3-1.7B, and Qwen3-8B were scored with the `auto_round` evaluation harness; all other cells used the same lm-eval-harness configuration through a separate driver. On matched checkpoints the two harnesses differ by $+0.3$ pp on average (Jiang et al., 2026).

Table 16: W2 Pile-calibration accuracy by method and model (ten-task mean accuracy, %; mean over 4 seeds). Column "Mean" is per-method overall.

Method	c	Llama-2-7B	Mistral-7B	Q2.5-7B-Inst	Qwen3-1.7B	Qwen3-8B	Phi-3-mini	Gemma-2-9B	Mean
AWQ	0.35	29.5	29.9	29.8	32.0	32.1	31.7	32.2	31.0
GPTQ	0.49	31.9	31.0	31.3	30.5	31.7	29.9	36.4	31.8
AR0	1.38	41.3	46.3	38.2	31.1	38.7	38.9	49.1	40.5
AR	1.76	52.6	56.0	60.3	43.0	57.0	52.5	58.5	54.3

The ordering AR $>$ AR0 $>$ GPTQ $\approx$ AWQ matches the c-index ordering. AWQ and GPTQ, the two methods with the smallest c, sit near chance at W2; AR0 and AutoRound preserve substantial accuracy.

I Functional-Form Selection for the α-Law

We compare the Hill-type α-law with five other one-parameter forms on the same per-method panel (n=21 cells, 7 models $\times$ {W2, W3, W4}; AR, GPTQ, and AWQ separately). We assess each form on (i) boundary-condition compliance ($\alpha \to 1$ as $q_{\max} \to \infty$, recovering the absmax limit; the $q_{\max} \to 0$ side is included as a mathematical regularity condition rather than a physical regime, since $q_{\max}$=1 at W2 is the smallest grid in our panel) and (ii) joint R^2 across all 21 cells per method (Table 17).

Table 17: Joint fit of six candidate single-parameter forms to the $(q_{\max}, \alpha_{\text{med}})$ panel (n=21 cells per method). "Limits" column marks whether the form satisfies the $q_{\max} \to 0$ and $q_{\max} \to \infty$ boundary conditions. Among limit-respecting candidates, Hill is the only form in this set with R^2>0.95 on all three methods; exponential, tanh, and power-law fits saturate too early, too late, or diverge. AWQ is fit in the transformed space (§C.3). Negative R^2 indicates the form fits worse than predicting the per-method mean.

Form	Expression	Limits	c_{AR}	R^2_{AR}	R^2_{GPTQ}	R^2_{AWQ}
constant	$\alpha = c$	no	0.60	0.000	0.000	0.000
linear	$\alpha = c\, q_{\max}$	no	0.13	−0.29	−12.1	−22.8
tanh	$\alpha = \tanh(q_{\max}/c)$	yes	4.08	0.636	0.350	0.287
exponential	$\alpha = 1 - e^{-q_{\max}/c}$	yes	3.09	0.817	0.516	0.406
power-law	$\alpha = 1 - q_{\max}^{-c}$	no	0.87	−0.39	−12.6	−24.5
Hill (chosen)	$\boldsymbol{\alpha = q_{\max}/(q_{\max} + c)}$	yes	**1.76**	**0.997**	**0.982**	**0.976**

The constant and linear forms violate the limits and fit poorly ($R^2 \leq 0$), as expected, since a constant α ignores the bit-width gradient and a linear α exceeds 1 at higher $q_{\max}$. The exponential and tanh forms respect the limits but cannot match both the slow-saturation regime at small $q_{\max}$ and the near-asymptote at $q_{\max}$=7 with a single c. The exponential reaches R^2=0.82 on AR but only 0.52 on GPTQ and 0.41 on AWQ, tanh saturates too quickly to track the W3→W4 plateau (R^2=0.64 AR, 0.35 GPTQ, 0.29 AWQ), and the power-law form diverges as $q_{\max} \to 0$ and underfits all three panels. The Hill form is the only form in our candidate set that achieves R^2>0.95 jointly on AR, GPTQ, and AWQ ($c_{\text{AWQ}} \approx 0.35$ in the transformed space; §C.3). The May E1b model-width inputs, fit code, and per-cell residuals for the four-method α-law figure are included in the supplementary code.

Held-out W2 on the original May cells. We fit c using only W3 and W4 on the saved May E1b cells used for Figure 2, then predict W2 on the same models without fitting to their W2 values (Table 18). This is a within-panel width check, not a new-model test. The source cells and per-model errors are included with the fit script in the supplementary code.

Table 18: W3/W4-to-W2 holdout on the original May E1b model-width cells. The error is measured minus predicted median α; mean and maximum columns report its absolute value over W2 models. W3/W4 fits use 14 cells for each of AutoRound, GPTQ, and AWQ, and six for AR0.

Method	W2 models	$c_{\mathrm{W3,W4}}$	Mean error	Max. error
AutoRound	7	1.7786	0.0063	0.0155
GPTQ	7	0.5116	0.0211	0.0246
AWQ	7	0.4020	0.0367	0.0367
AR0	3	1.5720	0.0806	0.1112

AutoRound and GPTQ have smaller W2 holdout errors in this panel. The seven saved AWQ W2 transformed-α values are all 0.750. Two of the three saved AR0 W2 cells are 0.500, producing its larger holdout error. These results use the original May inputs and do not replace AR0's nine-cell all-width fit.

J Calibration-Recipe Robustness for AR

To check whether AR's W2 accuracy depends on the specific choice of calibration corpus, we re-quantized AR at W2g128 on six panel models with three additional calibration recipes that differ from Pile in source distribution, namely arXiv-OOD (out-of-domain academic prose), diversity-DC (diversity-curated DC subset), and LSA-DC (LSA-balanced DC subset). Each (model, recipe) cell uses four calibration seeds; the ten-task mean accuracy (mean over seeds) is reported in Table 19.

Table 19: AR W2g128 ten-task accuracy by calibration recipe (mean over four seeds; 12 runs per model across three recipes). "Range" is $\max - \min$ across the three recipes; σ is the cross-recipe standard deviation. Gemma-2-9B was not run.

Model	arXiv-OOD	diversity-DC	LSA-DC	Range	σ
Llama-2-7B	52.31%	54.05%	54.34%	2.03 pp	1.10 pp
Mistral-7B	56.86%	58.88%	59.41%	2.55 pp	1.35 pp
Phi-3-mini	51.86%	55.42%	55.24%	3.56 pp	2.01 pp
Qwen3-1.7B	42.30%	43.14%	43.91%	1.61 pp	0.81 pp
Qwen3-8B	55.60%	55.06%	56.84%	1.78 pp	0.91 pp
Q2.5-7B-Inst	59.22%	57.82%	59.13%	1.40 pp	0.78 pp

The cross-recipe range is at most 3.6 pp (Phi-3-mini) and below 2.6 pp on the other five models; the per-model cross-recipe σ never exceeds 2.1 pp. For the three models where a Pile-200 baseline at s=42 is available (Mistral-7B-Inst 57.18%, Qwen3-8B 55.93%, Q2.5-7B-Inst 58.68%), the Pile accuracy sits inside the alt-recipe band. This experiment varies only the calibration corpus for AR at W2; it does not test whether the c-index ordering itself ($c_{\mathrm{AR}}>c_{\mathrm{GPTQ}}>c_{\mathrm{AWQ}}$) is recipe-invariant, which would require parallel GPTQ and AWQ sweeps.

K Additional Robustness Experiments

The AutoRound runs in this section use the SignRound V2 implementation (Cheng et al., 2025), labeled SR-V2 in the tables; its G128 constant matches the AutoRound c=1.76 of the main panel.

K.1 Calibration Corpora and GPTQ Settings

The matched calibration-corpus experiment uses Llama-2-7B, G128, W2–W4, seed 42, `sym=true`, `desc_act=false`, and `damp_percent=0.05`. Each corpus supplies 128 effective rows and 124,215 non-padding Llama tokens, matching the archived Pile baseline. C4 and WikiText-103 training text is concatenated and packed to that budget. The quantization script uses `seqlen=2048` as a 12,288-character ceiling, not a fixed token length. All extractions cover 224 modules without skipped tensors. The fitted c span is 0.000370, or 0.078% of the Pile value.

Table 20: Llama-2-7B GPTQ G128 fits under three calibration corpora.

Corpus	α_{W2}	α_{W3}	α_{W4}	c	R^2
Pile	.679064	.860243	.933894	.477009	.999535
C4	.679268	.860250	.933894	.476677	.999511
WikiText-103	.679268	.860294	.933894	.476639	.999517

A separate Pile sweep uses the same model, seed, sample count, length setting, group size, and widths. Its reference is the Pile fit above (c=0.477009), with activation ordering disabled and damping 0.05. Enabling `desc_act` at damping 0.05 changes c by -3.4%, while damping 0.01 and 0.10 change it by -1.5% and $+0.4\%$ with activation ordering disabled. All four fits retain R^2=0.9969–0.9996. The matched W2 ARC-Challenge scores are 0.2594 with activation ordering, 0.2406 with damping 0.01, and 0.2457 with damping 0.10. Fit persistence does not imply identical downstream accuracy.

K.2 Group-Size Dependence

Table 21: Fitted c across group sizes. The law remains strong for both methods, but coefficient stability is method dependent.

Model and method	Metric	G64	G128	G256
Llama-2-7B GPTQ	c	0.4757	0.4770	0.4759
Llama-2-7B GPTQ	R^2	0.9994	0.9995	0.9994
Llama-2-7B SR-V2	c	1.5125	1.76	1.9004
Mistral-7B SR-V2	c	1.5439	1.7363	1.9525
Mistral-7B SR-V2	R^2	0.9999	> 0.999	0.998

GPTQ's coefficient spans 0.27% over the tested group sizes. Mistral SR-V2 rises by 26.5% from G64 to G256, independently reproducing the configuration dependence observed on Llama-2-7B. At W2, Mistral ARC-Challenge changes by 0.2 points for SR-V2, while Llama-2-7B GPTQ scores 0.2875 at G64 and 0.2602 at G256. Coefficient stability and downstream stability are distinct measurements.

K.3 Mixtral Projection Fits and Routing

The Mixtral-8x7B extraction covers all 896 stored quantized attention and expert projections and all eight experts, with no skipped tensors.

Table 22: Mixtral-8x7B projection-scale fits at G128.

Method and scope	c	R^2	Max. residual
GPTQ, all projections	0.4705	0.9993	0.0031
GPTQ, experts only	0.4704	0.9993	0.0031
SR-V2, all projections	1.69	0.9980	0.0126

The matched GPTQ routing experiment covers 11,122 token positions across 32 layers. Top-1 FP16 routing agreement is 63.0%, 88.5%, and 93.9% at W2, W3, and W4. Complete top-2 disjoint rates fall from 6.92% to 0.48% and 0.16%, while mean layer-wise expert-load total variation falls from 10.04% to 2.23% and 1.14%. These measurements establish routing stability as precision rises.

K.4 Matched Tuning and OmniQuant Coordinates

A controlled Mistral-7B G128 pair gives c=1.7363 for the base model and c=1.7225 for the instruction-tuned model, a -0.8% difference. Both fits have R^2>0.9996 and cover all 224 projections. Instruction tuning changes c by under 1% in this matched setting.

OmniQuant learns upper and lower clipping factors and a rounded zero point. In a recovered native-state Pile seed-42 matrix, normalizing its optimized scale by original group absmax gives R^2=0.9034 on Llama-2-7B. Separate upper and lower tails give 0.9171 and 0.9199, while normalized half-span fails. The native optimized-scale coordinate replicates on Gemma-2-9B with R^2=0.934. The law is therefore coordinate-sensitive for OmniQuant and its fit remains weaker than for AutoRound, GPTQ, and AWQ.

K.5 Asymmetry Scope

The tested W2 implementation clamps codes to $[-2, 1]$ and selects the sign of each group scale from the dominant weight tail. In real-valued notation it therefore chooses either $\{-2, -1, 0, 1\}|s_g|$ or $\{-1, 0, 1, 2\}|s_g|$. The W2 fits directly cover this constrained asymmetric endpoint choice. They do not cover a general affine quantizer

$$\widehat{w} = \Delta(q - z), \qquad B_- = -z\Delta, \qquad B_+ = (2^b - 1 - z)\Delta, \tag{6}$$

with an independently learned zero point z.

L Reproducibility Inventory

All experiments were run on Intel Data Center GPU Max 1550 nodes with 12 tiles per node. Quantization and evaluation code is based on open-source implementations of each method (cited in §2). Evaluation uses lm-evaluation-harness (Gao et al., 2023) on the ten tasks listed in Section 2. The activation-patching harness is 179 lines of Python, uses `register_forward_hook`, and performs no custom dequantization. Scale extraction reads checkpoints with `safetensors.safe_open()` and numpy and requires no GPU forward passes. The supplementary `fit_alpha_law.py` reproduces the original panel fits, width holdouts, and canonical 70B predictions from included summaries, and generates Figure 3. The bundle includes the 70B sampling manifest, archived summaries, tested dependency versions, and relocation and extraction self-tests. Tensor-level audits accept a user-supplied data root and require the original reference weights and quantized checkpoint shards. All calibration data are drawn from training splits only. Code and checkpoints will be released upon acceptance.