RETHINKING NUMERICAL STABILITY IN HESSIAN-BASED PTQ: ACCUMULATION-ORDER SENSITIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-Training Quantization (PTQ) is widely used to reduce the computational and memory costs of large language models. Hessian-based PTQ methods such as GPTQ estimate empirical second-order statistics from calibration samples to guide weight quantization. Although the sum of these statistics is invariant to sample order in exact arithmetic, floating-point non-associativity can yield different empirical Hessians depending solely on the order in which the same samples are accumulated. While calibration sample selection is commonly treated as an important experimental condition, accumulation order is typically left uncontrolled. To our knowledge, we provide the first systematic evidence that reordering fixed calibration samples alone affects quantization and downstream performance in Hessian-based PTQ, with variability comparable to calibration resampling. We isolate this effect by varying only accumulation order across Qwen and Llama models, multiple Hessian-based PTQ methods, and weight bit-widths. In some settings, order-induced performance variation exceeds inter-method performance gaps and even reverses method rankings. Further analysis shows that order-dependent perturbations in the empirical Hessian are amplified through subsequent Hessian-dependent operations, ultimately altering quantization decisions. As a practical control, numerical stabilization incorporating compensated summation substantially suppresses order-induced variation. These findings highlight the need to account for accumulation order when evaluating and comparing Hessian-based PTQ methods. Our code is available at `https://anonymous.4open.science/r/ReSQ-B3EF`.

1 INTRODUCTION

As large language models grow in size, reducing their memory and computational costs becomes increasingly important for efficient deployment (Kaplan et al., 2020; Grattafiori et al., 2024; Zhu et al., 2024). Quantization addresses these costs by representing weights and activations at lower precision (Dettmers et al., 2022; Yao et al., 2022; Xiao et al., 2023; Shao et al., 2024). Post-training quantization (PTQ) is particularly attractive since it reduces model precision without full retraining (Chee et al., 2023; Dettmers et al., 2023; Lin et al., 2024). Among PTQ approaches, Hessian-based methods use second-order information about layer reconstruction error to guide quantization and compensate for quantization errors. In this work, we focus on these methods such as GPTQ (Frantar et al., 2022), GPTAQ (Li et al., 2025), BoA (Kim et al., 2024), and TurboBoA (Kim et al., 2026).

These methods construct empirical Hessian statistics from the activations of a small calibration set to characterize the curvature of their reconstruction objectives (Frantar et al., 2022). To limit activation memory usage in large-scale models, these statistics are typically accumulated one calibration sample or small batch at a time. Although the exact sum is independent of the order of its terms, floating-point addition is non-associative because intermediate sums are rounded to finite precision (Goldberg, 1991). Therefore, changing only the accumulation order can produce different empirical Hessians. Batching does not necessarily eliminate order dependence, because the reductions within each matrix multiplication are also subject to floating-point rounding. These differences can propagate through matrix inversion and Cholesky factorization, affecting subsequent quantization decisions and error compensation. Yet accumulation order is rarely reported as an experimental condition, even when the calibration sample selection is specified.

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.

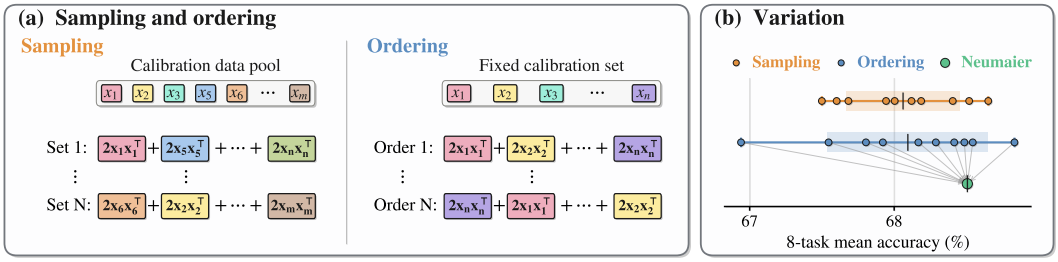

Figure 1: **Experimental factors and performance variation.** (a) *Sampling* uses independently selected calibration sets, whereas *Ordering* reorders the contributions from a fixed calibration set. (b) For Qwen3-14B quantized with GPTQ at INT3, *Ordering* induces accuracy variation comparable to *Sampling*, while Neumaier removes the observed variation across accumulation orders. Each point shows one run's eight-task mean accuracy.

We compare performance variation under two factors, *Sampling* and *Ordering*, as illustrated in Figure 1(a). Under *Sampling*, each run uses an independently sampled calibration set of the same size, whereas under *Ordering*, all runs use the same calibration samples but accumulate their Hessian contributions in different orders. In every run, we quantize the model and compute the unweighted mean zero-shot accuracy across the same eight language understanding and commonsense reasoning benchmarks widely used to evaluate LLM quantization (Section 3.1). For each factor, we quantify performance variation as the standard deviation of the eight-task mean accuracy across runs.

Figure 1(b) illustrates this comparison for Qwen3-14B with GPTQ at INT3, showing substantial run-to-run variation under both *Ordering* and *Sampling*. Across Qwen3-0.6B, Qwen3-8B, and Qwen3-14B (Yang et al., 2025), Llama-3.2-1B (Meta, 2024), Llama-3.1-8B (Grattafiori et al., 2024), and Llama-2-13B (Touvron et al., 2023) and four Hessian-based PTQ methods, *Ordering* alone produces performance variation comparable to *Sampling*. We observe this sensitivity with both 4-bit and 3-bit weights (Section 3, Figure 2). To examine its numerical origin, we conduct fixed-activation diagnostics on Llama-3.2-1B and observe approximately $84\times$ geometric-mean amplification of order-induced Hessian perturbations through GPTQ's damped inverse computation (Section 4.2, Figure 4(b)). Comparisons of complete quantization runs further show that changing accumulation order changes quantized integer weights and layer outputs (Section 4.3, Figure 4(c)).

This sensitivity matters when comparing quantization methods. Reliable comparisons between methods require accounting for sources of experimental variability (Bouthillier et al., 2021). When the performance gap between two PTQ methods is comparable to the variation induced by accumulation order, a single run is insufficient to establish a consistent advantage. In our experiments, some method rankings reverse across accumulation orders (Section 3.2). Fixing the calibration samples and quantization hyperparameters is therefore not always sufficient to obtain a stable comparison. These observations motivate numerical stabilization of Hessian accumulation to reduce order-induced variability and make comparisons between PTQ methods more reliable.

To reduce this order sensitivity, we incorporate Neumaier compensated summation (Neumaier, 1974) into FP32 Hessian accumulation and compare it with standard FP32 and FP64 accumulation. Across the evaluated methods and settings, Neumaier summation significantly reduces order-induced variability and, in several cases, yields no observed variation at the reported precision. For example, Figure 1(b) shows that Neumaier eliminates the observed *Ordering* variation for Qwen3-14B with GPTQ at INT3. For GPTQ on Llama-3.1-8B at INT3 using one RTX A6000 per run, the fused Neumaier compensated implementation takes $1.05\times$ the FP32 quantization runtime, compared with $10.05\times$ for FP64 accumulation (Section 5). We recommend incorporating Neumaier compensated summation into Hessian accumulation as a practical step toward reducing accumulation-order sensitivity in Hessian-based PTQ.

Our contributions are as follows:

- We show that, even with a fixed calibration set, accumulation order alone can induce substantial performance variation in Hessian-based PTQ, comparable in magnitude to that caused by calibration *Sampling* (Section 3).

- We analyze how order-dependent numerical differences in the empirical Hessian are amplified through Hessian-dependent operations and translate into different quantization outcomes (Section 4).
- We demonstrate that numerical stabilization incorporating Neumaier compensated summation substantially reduces accumulation-order variability, with low quantization-time overhead (Section 5).

2 Background and Related Work

2.1 Hessian-Based PTQ

A starting point for Hessian-based PTQ is the layer-wise reconstruction objective used in GPTQ (Frantar et al., 2022):

$$\min_{\widehat{W}} \left\| WX - \widehat{W}X \right\|_F^2, \tag{1}$$

where $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ denotes the original full-precision weight matrix, $\widehat{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ the quantized weight matrix, and $X \in \mathbb{R}^{d_{\text{in}} \times N}$ the matrix of input activations collected at the previous quantized layer from the calibration set. Here, d_{in} and d_{out} are the input and output dimensions of the layer, respectively, and N is the total number of input activation vectors. The curvature of this objective is characterized by the empirical Hessian

$$H \;=\; 2\,XX^\top \;=\; 2\sum_{i=1}^{N} x_i x_i^\top, \tag{2}$$

where $H \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$ and $x_i \in \mathbb{R}^{d_{\text{in}}}$ is the i-th column of X.

Hessian-based PTQ builds on the second-order optimization principles of Optimal Brain Surgeon (Hassibi & Stork, 1992), which were subsequently adapted for quantization in Optimal Brain Compression (Frantar & Alistarh, 2022) and extended to large language models in GPTQ (Frantar et al., 2022). GPTQ quantizes the weight matrix column by column and compensates for the resulting quantization error in the remaining unquantized columns using a damped inverse of the empirical Hessian in Equation 2. GPTAQ (Li et al., 2025) instead adopts the asymmetric objective $\|W\widetilde{X} - \widehat{W}X\|_F^2$, where $\widetilde{X}$ and X denote input activations from the original unquantized model and the partially quantized model, respectively. This objective accounts for the input mismatch caused by quantization of preceding layers. BoA and TurboBoA (Kim et al., 2024; 2026) extend the reconstruction scope beyond individual layers to account for interactions within blocks.

Despite these differences, these methods rely on empirical second-order statistics computed from calibration activations. We study how accumulation-order-dependent numerical differences in these statistics affect subsequent Hessian-dependent operations and quantization outcomes. Our work focuses on Hessian-based PTQ and does not cover methods that avoid constructing these statistics.

2.2 Performance Sensitivity in PTQ

Prior work on calibration in PTQ has primarily focused on how the calibration set is constructed. Williams & Aletras (2024) examine how data source, calibration set size, and sequence length affect quantization performance. He et al. (2026) study how calibration data affect capabilities after compression and propose data curation based on representativeness and diversity in activation space. Other approaches generate calibration data from the model itself to reduce reliance on external data (Williams et al., 2025), or select sequences that collectively activate outlier channels that random *Sampling* may miss (Shihab et al., 2026). MaCa (Son et al., 2026) incorporates multiple sequence lengths and per-sequence normalization into Hessian estimation. These studies address the choice of calibration data and how their contributions are represented in the calibration statistics.

Our work instead asks whether changing only the accumulation order of a fixed calibration set can affect quantization outcomes. In exact arithmetic, reordering a fixed set of Hessian contributions leaves their sum unchanged. However, in finite precision, the computed Hessian can depend on accumulation order, potentially affecting quantization decisions and downstream performance.

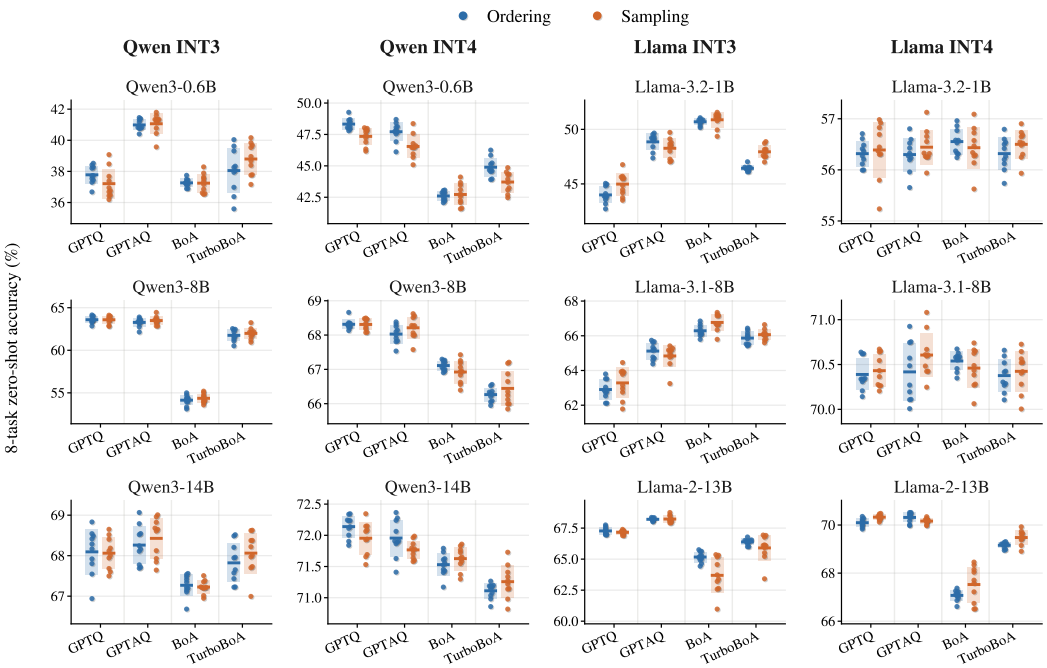

Figure 2: ***Ordering* versus *Sampling* variability in 3- and 4-bit weight-only quantization.** Columns show Qwen3 at 3/4 bits, then Llama at 3/4 bits, with model size increasing downward. Blue denotes reordered fixed calibration data, and orange denotes independently sampled sets. Dots show per-run eight-task mean zero-shot accuracy, bars show run means, and boxes span ± 1 sample standard deviation ($n = 10$ unless indicated).

2.3 Floating-Point Accumulation and Error Mitigation

Floating-point summation introduces rounding errors that can accumulate and make the computed result sensitive to the order of addition. Established approaches to reducing these errors include higher-precision accumulation and compensated summation (Higham, 2002). Compensated summation methods, such as Kahan summation and Neumaier's correction, track rounding errors during accumulation and incorporate a correction into the computed sum (Kahan, 1965; Neumaier, 1974). Reproducible summation and exact accumulation have also been studied to obtain consistent results across different accumulation orders in parallel and distributed systems (Demmel & Nguyen, 2013; Neal, 2015; Xie et al., 2025).

We investigate Neumaier compensated summation as a practical way to reduce accumulation-order sensitivity in Hessian-based PTQ, evaluating its effects on both numerical differences in the accumulated statistics and downstream performance variation.

3 Accumulation Order Sensitivity in Hessian-Based PTQ

3.1 Experimental Setup

We compare two sources of performance sensitivity under matched quantization configurations and evaluation protocols. For *Sampling*, we use ten independently sampled calibration sets of the same size. For *Ordering*, we fix one calibration set and evaluate ten permutations of its accumulation order. All other settings are held fixed within each comparison.

We evaluate six models: Qwen3-0.6B, Qwen3-8B, and Qwen3-14B, along with Llama-3.2-1B, Llama-3.1-8B, and Llama-2-13B. Unless otherwise stated, we use 128 calibration sequences from WikiText-2, each containing exactly 2,048 tokens after tokenization. We evaluate weight-only quantization with 3- and 4-bit weights, denoted INT3 and INT4, respectively, while retaining 16-bit ac-

Table 1: ***Ordering* and *Sampling* variability across calibration corpora and set sizes.**

		8-task Acc. (%)		WikiText-2 PPL ↓		C4 PPL ↓	
Corpus	**Sequences**	**Ord.**	**Samp.**	**Ord.**	**Samp.**	**Ord.**	**Samp.**
WikiText-2	128	62.90±0.58	63.28±0.87	8.162±0.025	8.151±0.026	14.629±0.068	14.666±0.066
WikiText-2	256	62.86±0.40	63.43±0.91	8.116±0.027	8.163±0.066	14.317±0.048	14.299±0.071
SlimPajama	128	64.25±0.33	63.86±0.90	8.788±0.041	8.793±0.049	13.768±0.037	13.791±0.066
SlimPajama	256	64.28±0.61	64.46±0.64	8.681±0.039	8.691±0.042	13.640±0.053	13.671±0.099
C4	128	64.70±0.57	64.51±0.71	8.972±0.057	8.997±0.092	13.694±0.045	13.707±0.049
C4	256	64.52±0.56	64.51±0.62	8.870±0.045	8.938±0.161	13.519±0.033	13.567±0.054

tivations. Our experiments are based on the publicly released implementations of GPTQ, GPTAQ, BoA, and TurboBoA.[1] Implementation details are described in Appendix D.2.

We evaluate zero-shot performance on eight widely used language understanding and commonsense reasoning benchmarks: ARC-Challenge and ARC-Easy (Clark et al., 2018), BoolQ (Clark et al., 2019), LAMBADA (Paperno et al., 2016), WinoGrande (Sakaguchi et al., 2021), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), and PIQA (Bisk et al., 2020). We report the unweighted mean accuracy across these eight tasks and its sample standard deviation across runs.

The run definitions, seed lists, metric choices, and evaluation configuration are provided in Appendix A.1. Appendix A.2 describes the controls used to isolate accumulation order and distinguishes the fixed-activation diagnostics from complete quantization runs, in which later activations can inherit differences from earlier quantized modules.

3.2 Observed Sensitivity across Models, Methods, and Bit Widths

***Ordering* alone can induce performance variation comparable to *Sampling*.** Figure 2 shows that changing only the accumulation order, while keeping all calibration samples fixed, produces performance variation comparable to or greater than that from calibration resampling in several configurations. For example, for Qwen3-14B with GPTQ at INT3, the standard deviation of eight-task mean accuracy is 0.56 under *Ordering* and 0.39 under *Sampling*, making *Ordering* variation approximately $1.4\times$ as large. Across all tested configurations, the *Ordering* standard deviation exceeds its *Sampling* counterpart in 5 of 24 configurations at INT4 and 8 of 24 at INT3. The relative magnitude depends on the model, method, and bit width, and the sensitivity is not confined to small models. For example, at INT3, *Ordering* variation exceeds *Sampling* variation for TurboBoA on Qwen3-0.6B, GPTQ on Llama-2-13B, and GPTQ and BoA on Qwen3-14B. These results show that accumulation order is a consequential source of performance variation in Hessian-based PTQ, even when calibration-set composition is fixed. Detailed results are provided in Appendix F.

Method rankings can reverse across accumulation orders. The *Ordering* variation observed above can directly affect comparisons between quantization methods. Figure 3 compares GPTQ and GPTAQ on Qwen3-8B at INT3, pairing their eight-task mean accuracies by *Ordering* seed across ten seeds. Although GPTAQ was proposed as a subsequent method that extends GPTQ to improve quantization performance, their relative ranking is not consistent across accumulation orders: GPTQ scores higher than GPTAQ in seven of the ten seed-matched pairs, while GPTAQ scores higher in the remaining three. Such reversals also occur at INT4. On Llama-3.1-8B, GPTQ scores higher than GPTAQ in three of the ten pairs. Thus, the method that appears better in a single run can depend on the accumulation order, even when the calibration samples and quantization configuration remain fixed.

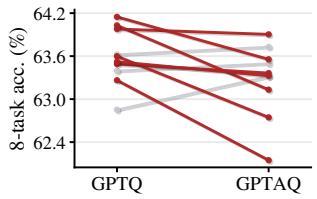

Figure 3: **Rank changes across accumulation orders (Qwen3-8B, INT3).** Red lines indicate rank reversals with same seed.

***Ordering* variability persists across calibration corpora and sample sizes.** We conduct additional experiments on Llama-3.1-8B with GPTQ at INT3 using calibration sets of 128 and 256 sequences from WikiText-2, SlimPajama, and C4. We follow the same *Ordering* and *Sampling* protocols, keeping the sequence length and quantization settings fixed within each comparison. Table 1 shows that changing only the accumulation order produces nonzero downstream variation in all six conditions. For C4 and for SlimPajama with 256 sequences, the *Ordering* standard deviation is close to the *Sampling* standard deviation. Increasing the calibration set to 256 sequences does not eliminate this

[1]Official implementations: GPTQ, GPTAQ, BoA, and TurboBoA.

sensitivity: the *Ordering* standard deviation increases on SlimPajama and remains similar on C4. Therefore, the *Ordering* sensitivity is neither limited to WikiText-2 nor eliminated by doubling the calibration-set size.

Task-level results provide further context. Williams & Aletras (2024) report particularly pronounced variation across calibration sets on BoolQ. Our results likewise show large BoolQ variation in smaller models. However, *Ordering*-induced variation in eight-task mean accuracy is also observed in larger models, including configurations where it is comparable to or greater than *Sampling* variation. Appendix B.2 provides the task-level analysis.

4 Numerical Origin of Ordering Variability

How can changing only the accumulation order of contributions from a fixed calibration set produce a different quantized model? We investigate this question in three steps. First, Section 4.1 shows how finite precision accumulation introduces order-dependent rounding errors as the Hessian is accumulated. Next, Section 4.2 measures how these rounding errors are amplified by damped inversion under fixed activations. Finally, Section 4.3 examines changes in integer weights and module outputs across complete quantization runs and compares them with changes caused by *Sampling*. We focus on GPTQ as a representative method, since other methods follow the same Hessian-based error-compensation framework.

4.1 Order-Dependent Hessian Perturbation

Although the empirical Hessian in Eq. 2 is order independent in exact arithmetic, its finite precision computation need not be. Let $x_i \in \mathbb{R}^{d_{\text{in}}}$ be an input activation vector, N the number of outer-product contributions, and π a permutation of these contributions. At the level of individual outer-product contributions, order-dependent accumulation can be expressed as

$$S_\pi^0 = 0, \qquad S_\pi^k = \mathrm{fl}\Big(S_\pi^{k-1} + 2x_{\pi(k)}x_{\pi(k)}^\top\Big), \qquad \widehat{H}_\pi = S_\pi^N = H + \Delta H_\pi, \tag{3}$$

where S_π^k is the partial sum after k contributions and $\mathrm{fl}(\cdot)$ denotes finite precision evaluation. Here H is the exact-arithmetic Hessian, $\widehat{H}_\pi$ its computed estimate, and ΔH_π the accumulation error. Each addition rounds according to the incoming contribution and the current partial sum, which itself depends on the order of the preceding additions. This makes ΔH_π order dependent, as reviewed in Section 2.3 and detailed in Appendix C.1.

We compare two accumulation orders π and π' of a fixed calibration set using the relative difference

$$\rho(A_\pi, A_{\pi'}) = \frac{2\,\|A_\pi - A_{\pi'}\|_F}{\|A_\pi\|_F + \|A_{\pi'}\|_F}, \tag{4}$$

where A_π is the matrix quantity computed under order π and $\|\cdot\|_F$ denotes the Frobenius norm. To enable comparisons across modules and stages of quantization with different matrix scales, we normalize the Frobenius difference by the mean of the two norms. We use ρ to compare the Hessian and its damped inverse across accumulation orders and analyze the differences during quantization.

Empirical characterization of Hessian perturbations. Figure 4(a) shows, for one Hessian entry for the layer-0 query projection (`q_proj`) in Llama-3.2-1B, the difference between FP32 and FP64 partial sums accumulated in the same order. The curves separate early and remain different after accumulation, with absolute partial-sum errors of up to 2.5×10^{-3}. At the matrix level, the mean relative difference between the accumulated FP32 Hessians across five disjoint pairs of accumulation orders, $\rho(\widehat{H}_\pi, \widehat{H}_{\pi'})$, is 2.06×10^{-7}. Thus, changing only the accumulation order produces small but measurable differences in the accumulated Hessians.

4.2 Amplification of Hessian Perturbations

Perturbation amplification. GPTQ introduces diagonal damping to stabilize inversion of the empirical Hessian, which can be singular or poorly conditioned. Under order π, it inverts $\widehat{H}_\pi + \lambda_\pi I$, where $\lambda_\pi = \frac{p}{d_{\text{in}}}\,\mathrm{tr}(\widehat{H}_\pi)$. We set $p = 10^{-2}$, following the default damping setting in the official

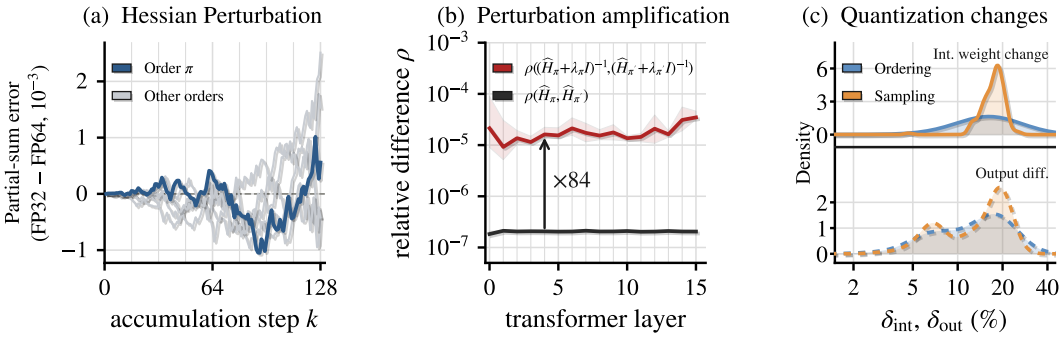

Figure 4: **Numerical effect of accumulation order** (a) Order-dependent rounding error in Hessian entry $(0, 0)$ for the layer-0 query projection (`q_proj`). Each curve shows the difference between FP32 and FP64 partial sums accumulated in the same order. (b) Relative Hessian and damped-inverse perturbations across depth. Lines and bands show geometric means and interquartile ranges across input locations after averaging differences over five *Ordering* pairs. (c) Integer-weight (solid) and module-output (dashed) changes under *Ordering* (blue) and *Sampling* (orange), shown as kernel density estimates of $\log_{10} \delta$ (each curve integrates to one over the logarithmic axis).

GPTQ implementation. Even with damping, small Hessian perturbations can be amplified by inversion. For the same pair of orders, we quantify amplification by

$$\mathcal{A} = \frac{\rho\Big(\big(\widehat{H}_\pi + \lambda_\pi I\big)^{-1}, \big(\widehat{H}_{\pi'} + \lambda_{\pi'} I\big)^{-1}\Big)}{\rho\big(\widehat{H}_\pi, \widehat{H}_{\pi'}\big)}, \tag{5}$$

which measures the ratio of the relative damped-inverse difference to the relative Hessian difference. We use the same activations from the unquantized model for all accumulation orders, isolating this sensitivity from activation differences caused by quantization in earlier layers.

On Llama-3.2-1B, we analyze input-Hessian at four locations per transformer layer: the shared input to the query, key, and value projections, the input to the output projection, the shared input to the gate and up projections, and the input to down projection. This gives 64 locations across 16 layers. As shown in Figure 4(b), the geometric mean of $\mathcal{A}$ over these locations and five disjoint *Ordering* pairs is approximately 84. Amplification varies across modules and is not monotonic with depth. The pair-averaged input Hessian differences span only $1.86\times$ across all 64 locations, whereas the corresponding inverse differences span more than two orders of magnitude (see Appendix C.2). Although damping stabilizes Hessian inversion, increasing its strength does not consistently reduce *Ordering* variability in our sweep, as analyzed in Appendix C.3.

Cross-method comparison. Under fixed activations on Llama-3.2-1B, we compare the input-Hessian accumulation rules of GPTQ, GPTAQ, BoA, and TurboBoA, followed by a common damped-inverse computation. Across the same 64 locations and five *Ordering* pairs per location, the geometric-mean amplification ranges from $61.50\times$ to $83.62\times$ across methods. BoA and TurboBoA yield identical results because they share the input accumulation rule in this diagnostic. Appendix C.4 details the protocol and per-method results.

4.3 From Perturbations to Quantization Changes

The upper-triangular Cholesky factor U of the damped inverse Hessian under order π satisfies $U^\top U = (\widehat{H}_\pi + \lambda_\pi I)^{-1}$. GPTQ uses this factor to compensate the remaining weights for the error caused by quantizing each column. The quantization of column j and the subsequent weight updates are given by

$$\widehat{W}_{:,j} = \mathcal{Q}(W_{:,j}), \qquad W_{:,k} \leftarrow W_{:,k} - \big(W_{:,j} - \widehat{W}_{:,j}\big)\frac{U_{jk}}{U_{jj}}, \quad k > j, \tag{6}$$

where W is the current weight matrix incorporating error compensation from previously quantized columns, $W_{:,j}$ denotes its jth column, and $\mathcal{Q}$ maps weights to the quantization grid. A change in

U_{jk}/U_{jj} changes the weights presented to subsequent rounding steps. Crossing a rounding boundary changes the stored integer and the error propagated to later columns. We measure the resulting changes below.

Integer-weight changes and their output effect. We compare a pair of runs a, b within each condition, *Ordering* or *Sampling*. For each run $r \in \{a, b\}$, $\widehat{W}_r^{(m)} \in \mathbb{R}^{d_{\text{out}}^{(m)} \times d_{\text{in}}^{(m)}}$ denotes the quantized weight matrix of module m, $Z_r^{(m)}$ denotes its stored integer codes, and $X_r^{(m)}$ denotes the input activations captured during a forward pass of the quantized model produced by run r on its calibration set. Here, $d_{\text{out}}^{(m)}$ and $d_{\text{in}}^{(m)}$ are the output and input dimensions of module m. We define the fraction of integer weights that differ between the two runs as

$$\delta_{\text{int}}^{(m)} = \frac{1}{d_{\text{out}}^{(m)} d_{\text{in}}^{(m)}} \sum_{i=1}^{d_{\text{out}}^{(m)}} \sum_{j=1}^{d_{\text{in}}^{(m)}} \mathbf{1}\Big[(Z_a^{(m)})_{ij} \neq (Z_b^{(m)})_{ij}\Big], \tag{7}$$

where $\mathbf{1}[\cdot]$ is the indicator function. Scales and zero points are identical between the two runs in each *Ordering* pair. Therefore, differences in their integer weights correspond to different choices on the same quantization grid.

To assess the output effect of these weight differences, we also measure the relative module-output difference on a common input activation matrix X as

$$\delta_{\text{out}}^{(m)}(X) = \rho\Big(\widehat{W}_a^{(m)} X, \widehat{W}_b^{(m)} X\Big), \tag{8}$$

where ρ is defined in Eq. 4. For each pair, we compute Eq. 8 on both $X_a^{(m)}$ and $X_b^{(m)}$, using the same input for the two weight matrices in each comparison. We report the average of these two relative differences as $\delta_{\text{out}}^{(m)}$, expressed as a percentage. This treats both runs' inputs equally and measures the output difference between the two quantized weight matrices, rather than reconstruction error relative to the unquantized model.

We evaluate five disjoint *Ordering* pairs and five disjoint *Sampling* pairs on Llama-3.2-1B at INT3. These are complete quantization runs. Unlike the fixed-activation diagnostic in Section 4.2, the input activations at later modules can differ because earlier modules may have been quantized differently. The measured weight changes therefore include the effects of both local Hessian differences and differences propagated from earlier quantized modules.

Figure 4(c) shows distributions across all 112 quantized modules in layers 0 to 15, using each module's average over the five run pairs. Averaged equally over modules, the integer weight mismatch rates are 14.9% for *Ordering* and 17.5% for *Sampling*. Together with the identical scales and zero points within each *Ordering* pair, these results show that accumulation order changes integer-code assignments on the same quantization grid. The corresponding mean relative module-output differences are 12.7% for *Ordering* and 14.1% for *Sampling*. Appendix C.5 reports results by projection and layer. Thus, changing only the accumulation order of a fixed calibration set produces changes comparable in magnitude to those caused by *Sampling* in this experiment, both in stored integer weights and in module outputs evaluated on common inputs.

5 Precision Intervention and Practical Stabilization

Section 4 shows how order-dependent differences in Hessian accumulation propagate through subsequent computations and affect quantization outcomes. We investigate whether reducing accumulation error can mitigate the resulting *Ordering* variability. We evaluate stabilized implementations across quantization methods and experimental settings, and report their memory and runtime costs.

5.1 Neumaier Compensated Summation

In standard FP32 accumulation, each Hessian contribution is directly added to the running sum, without explicitly accounting for rounding errors introduced by finite precision addition. Neumaier compensated summation instead maintains an additional FP32 correction buffer that tracks low-order information lost during accumulation and adds the correction to the final sum. This reduces

Table 2: ***Ordering* variation with stabilized accumulation at INT3.** Mean accuracy and σ_{ord} across 10 orders. Neumaier gives $\sigma_{\mathrm{ord}} = 0.00$ in 14 of 16 settings.

		FP32		Neumaier		FP64	
Model	**Method**	**Acc.**	$\boldsymbol{\sigma_{\mathrm{ord}}}$	**Acc.**	$\boldsymbol{\sigma_{\mathrm{ord}}}$	**Acc.**	$\boldsymbol{\sigma_{\mathrm{ord}}}$
Qwen3-0.6B	GPTQ	37.78	0.58	37.26	**0.00**	37.50	0.01
	GPTAQ	40.99	0.34	41.22	**0.00**	40.30	0.48
	BoA	37.26	0.27	37.13	**0.00**	37.61	**0.00**
	TurboBoA	38.05	1.42	39.46	**0.00**	39.09	**0.00**
Qwen3-8B	GPTQ	63.59	0.39	63.42	**0.00**	63.81	0.44
	GPTAQ	63.27	0.51	63.54	**0.00**	63.23	0.58
	BoA	54.14	0.58	53.87	0.02	53.49	0.48
	TurboBoA	61.75	0.66	62.68	**0.00**	61.90	0.11
Llama-3.2-1B	GPTQ	43.99	0.77	43.27	**0.00**	44.25	0.33
	GPTAQ	48.85	0.75	48.95	**0.00**	48.42	0.80
	BoA	50.69	0.29	50.85	**0.00**	50.49	0.07
	TurboBoA	46.43	0.26	47.12	**0.00**	46.06	**0.00**
Llama-3.1-8B	GPTQ	62.90	0.58	63.00	**0.00**	63.44	0.47
	GPTAQ	65.12	0.48	65.51	**0.00**	64.99	0.44
	BoA	66.29	0.32	66.75	0.13	66.55	0.28
	TurboBoA	65.87	0.37	65.80	**0.00**	66.03	0.03

accumulation error while retaining FP32 working precision. Algorithm 1 in Appendix D.1 provides the detailed procedure. We apply Neumaier compensated summation to GPTQ, GPTAQ, BoA, and TurboBoA, with accompanying adjustments to normalization and reduction operations where needed. Method-specific implementation details and pseudocode are provided in Appendix D.2.

5.2 Stability and Computational Cost

We compare baseline FP32 accumulation, Neumaier compensated summation in FP32, and FP64 accumulation followed by FP32 Hessian-dependent computation.

***Ordering* variability.** Table 2 shows that Neumaier reduces *Ordering* variation in all 16 INT3 settings, with $\sigma_{\mathrm{ord}} = 0.00$ in 14 cases. In the remaining two BoA settings, σ_{ord} decreases from 0.58 to 0.02 on Qwen3-8B and from 0.32 to 0.13 on Llama-3.1-8B. FP64 accumulation reduces rounding error but does not eliminate order dependence because floating-point addition remains non-associative. Neumaier instead preserves low-order information lost during accumulation through a correction term, substantially reducing the remaining order-dependent error. In our experiments, the quantized weights are bit-identical across all ten orders in the 14 settings with $\sigma_{\mathrm{ord}} = 0.00$.

Cost. We measure the runtime and memory costs of the fused Neumaier implementation (Appendix D.2) for GPTQ on Llama-3.1-8B at INT3, using a fixed calibration set of 128 samples. For each condition, we run three quantization jobs concurrently, each on a separate RTX A6000 GPU. Across three runs, Neumaier takes 1046.72 ± 8.89 s, compared with 1000.73 ± 17.70 s for FP32 (mean $\pm$ std.). Its mean accumulation time is $1.09\times$ that of FP32, while its mean total quantization runtime is $1.05\times$. FP64 requires $10.05\times$ the mean total quantization runtime of FP32. Measured peak GPU memory is 5.43 GiB for Neumaier, compared with 4.79 GiB for FP32 and 6.68 GiB for FP64. The additional cost is incurred only during offline quantization and introduces no inference overhead. Detailed results are provided in Appendix D.3.

6 Conclusion

In this work, we show that changing only the accumulation order of a fixed calibration set can induce substantial performance variation in Hessian-based PTQ, with a magnitude comparable to calibration resampling in some settings. We observe this sensitivity across the evaluated models, quantization methods, and bit widths. Our analysis traces its numerical origin to order-dependent differences in FP32 Hessian accumulation, which propagate through Hessian-dependent operations and alter quantization decisions and final weights. Neumaier compensated summation substantially reduces this order-induced variability, providing a practical means of numerical stabilization. These findings highlight the need to report accumulation order alongside calibration sample selection and to account for order-induced variability when comparing Hessian-based PTQ methods. We recommend incorporating Neumaier compensated summation into Hessian accumulation to reduce this source of variability. Further discussion and limitations are provided in Appendix E.

AI USE STATEMENT

During the final review of this manuscript, we used generative AI tools to assist with language editing and to review the clarity and consistency of the manuscript. The authors reviewed the suggested corrections and take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

Appendix A specifies the experimental settings, accumulation-order seeds, evaluation protocol, and controls used to isolate accumulation-order sensitivity. Appendix C describes the numerical diagnostics and aggregation procedures, and Appendix D provides the method-specific stabilization procedures and runtime measurements. Complete results are reported in Appendix F. The accompanying code is available through the anonymous repository linked in the abstract.

REFERENCES

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. *Proceedings of machine learning and systems*, 3:747–769, 2021.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. Quip: 2-bit quantization of large language models with guarantees. *Advances in neural information processing systems*, 36:4396–4429, 2023.

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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

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.

James Demmel and Hong Diep Nguyen. Fast reproducible floating-point summation. In *2013 IEEE 21st Symposium on Computer Arithmetic*, pp. 163–172. IEEE, 2013.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35: 30318–30332, 2022.

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. *arXiv preprint arXiv:2306.03078*, 2023.

Elias Frantar 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*, 2022.

David Goldberg. What every computer scientist should know about floating-point arithmetic. *ACM computing surveys (CSUR)*, 23(1):5–48, 1991.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 5, 1992.

Bowei He, Lihao Yin, Hui-Ling Zhen, Shuqi Liu, Han Wu, Xiaokun Zhang, Mingxuan Yuan, and Chen Ma. Preserving llm capabilities through calibration data curation: From analysis to optimization. *Advances in Neural Information Processing Systems*, 38:58531–58572, 2026.

Nicholas J Higham. *Accuracy and stability of numerical algorithms*. SIAM, 2002.

William Kahan. Pracniques: further remarks on reducing truncation errors. *Communications of the ACM*, 8(1):40, 1965.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Junhan Kim, Ho-young Kim, Eulrang Cho, Chungman Lee, Joonyoung Kim, and Yongkweon Jeon. Boa: Attention-aware post-training quantization without backpropagation. *arXiv preprint arXiv:2406.13474*, 2024.

Junhan Kim, Yeo Jeong Park, Seungwoo Son, Chungman Lee, Ho-young Kim, Joonyoung Kim, and Yongkweon Jeon. Turboboa: Faster and exact attention-aware quantization without backpropagation. *arXiv preprint arXiv:2602.04929*, 2026.

Yuhang Li, Ruokai Yin, Donghyun Lee, Shiting Xiao, and Priyadarshini Panda. Gptaq: Efficient finetuning-free quantization for asymmetric calibration. *arXiv preprint arXiv:2504.02692*, 2025.

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. *Proceedings of machine learning and systems*, 6:87–100, 2024.

Meta. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models, 2024. URL `https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/`.

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 *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Radford M Neal. Fast exact summation using small and large superaccumulators. *arXiv preprint arXiv:1505.05571*, 2015.

Arnold Neumaier. Rundungsfehleranalyse einiger verfahren zur summation endlicher summen. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 54(1):39–51, 1974.

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 *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1525–1534, 2016.

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.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Gao Peng, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 45472–45496, 2024.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Coverage-based calibration for post-training quantization via weighted set cover over outlier channels. *arXiv preprint arXiv:2604.24008*, 2026.

Seungwoo Son, Ingyu Seong, Junhan Kim, Hyemi Jang, and Yongkweon Jeon. On the importance of a multi-scale calibration for quantization. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1186–1190. IEEE, 2026.

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

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, 2024.

Miles Williams, George Chrysostomou, and Nikolaos Aletras. Self-calibration for language model quantization and pruning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 10149–10167, 2025.

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*, pp. 38087–38099. PMLR, 2023.

Peichen Xie, Xian Zhang, and Shuo Chen. Repdl: Bit-level reproducible deep learning training and inference. *arXiv preprint arXiv:2510.09180*, 2025.

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.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in neural information processing systems*, 35:27168–27183, 2022.

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.

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.

APPENDIX

TABLE OF CONTENTS

A Additional Experimental Details and Controls **14**

A.1 Experimental Setup 14

A.2 Controls for Isolating Accumulation Order 14

B Additional Results on Accumulation-Order Sensitivity **15**

B.1 Ordering versus Sampling Variability 15

B.2 Task-Level Ordering Variability 16

C Additional Analysis of the Numerical Origin of Ordering Variability **17**

C.1 Floating-Point Accumulation Error 17

C.2 Module-Level Hessian Perturbations and Amplification 17

C.3 Sensitivity to Damping 19

C.4 Cross-Method Input-Hessian Sensitivity 20

C.5 Module-Level Integer-Weight Changes and Module-Output Differences 21

D Additional Details on Precision Intervention and Practical Stabilization **22**

D.1 Neumaier Compensated Summation 22

D.2 Method-Specific Implementations 23

D.3 Runtime and Memory Costs 25

E Discussion and Limitations **26**

E.1 Discussion 26

E.2 Limitations 27

F Complete Result Tables **27**

A Additional Experimental Details and Controls

A.1 Experimental Setup

Accumulation orders. The *Ordering* experiments use 10 permutations generated from the fixed random integer seeds 777, 1024, 2024, 3407, 5000, 7777, 10042, 27182, 31415, and 424242. These ten *Ordering* seed identifiers are shared across model–method configurations. Pairwise method comparisons match runs by seed within the same model and bit width.

Calibration sets. The primary *Ordering* and *Sampling* comparisons use WikiText-2 for calibration. For *Ordering* comparisons, we fix a set of 128 calibration samples and vary only their accumulation order, whereas for *Sampling* comparisons, we resample 128 calibration samples from the same corpus. *Sampling* results use ten independently sampled calibration sets of 128 sequences, with random seeds 1 through 10. The additional Llama-3.1-8B study in Table 1 applies the same protocol to WikiText-2, SlimPajama, and C4 calibration sets of 128 or 256 sequences. The *Ordering* intervention permutes calibration contributions, while the fixed-activation diagnostics hold the activation tensors themselves fixed as described in Appendix A.2. Eq. 3 uses vector outer products to express the sum, and implementations accumulate matrix-product contributions, whose grouping matters for floating-point reduction.

Quantization and evaluation. We evaluate weight-only quantization at INT4 and INT3. Unless otherwise specified, the damping percentage is 10^{-2}, following the default in the original GPTQ implementation. We evaluate zero-shot performance on ARC-Challenge, ARC-Easy, BoolQ, LAMBADA, WinoGrande, HellaSwag, OpenBookQA, and PIQA, and report the unweighted mean accuracy across the eight tasks. Perplexity is additionally measured on WikiText-2 and C4 using the same evaluation protocol across runs. The eight-task mean uses `acc` for ARC-Easy, BoolQ, LAMBADA, and WinoGrande, and `acc_norm` for ARC-Challenge, HellaSwag, OpenBookQA, and PIQA. We first compute this unweighted mean within each run, then compute its sample standard deviation across runs with denominator $n - 1$. We evaluate all models using lm-evaluation-harness 0.4.11, Transformers 4.57.6, and PyTorch 2.10.0 with CUDA 12.8.

A.2 Controls for Isolating Accumulation Order

Only the summation order changes in the diagnostics. In the controlled measurements in Sections 4.1 and 4.2, we use the same activation tensors from the unquantized model across all accumulation orders. We vary only the order in which their contributions are accumulated into the empirical Hessian, isolating accumulation-order effects from changes in the input activations. All FP32 accumulation results use FP32 matrix multiplication with TF32 disabled.

In complete quantization runs, however, later activations can differ because earlier modules have already been quantized. Therefore, the comparison of complete quantization runs in Section 4.3 includes both the local Hessian perturbation and differences inherited from earlier quantization decisions.

Repeating a fixed order produces no observed variation. With the calibration data and accumulation order held fixed, repeated quantization and evaluation runs produce no observed differences in the quantized tensors or reported evaluation metrics in our execution environment. This control supports attributing the variation across accumulation orders to the ordering intervention rather than to run-to-run nondeterminism under a fixed order.

B Additional Results on Accumulation-Order Sensitivity

B.1 Ordering versus Sampling Variability

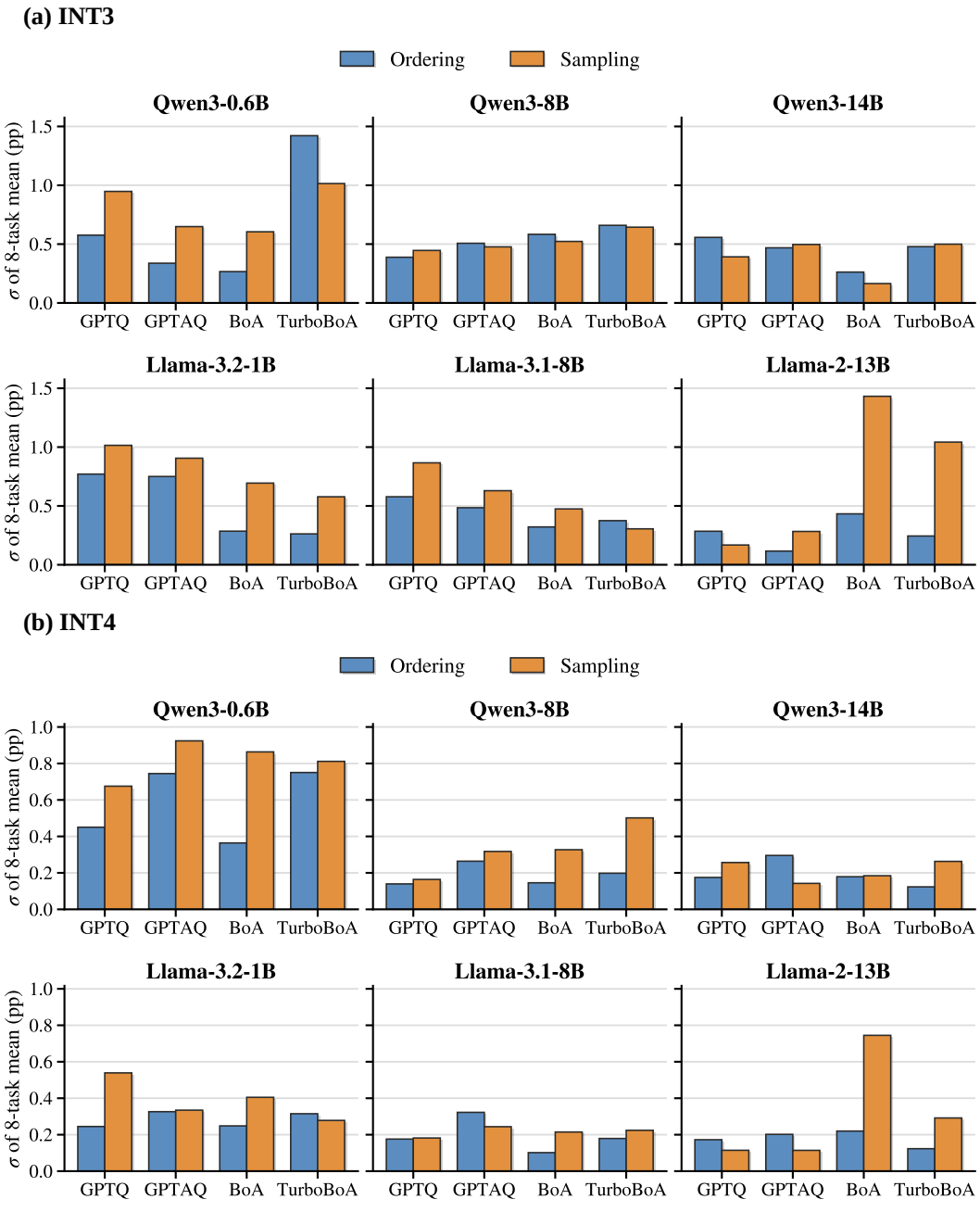

Figure 5: ***Ordering* and *Sampling* variability at INT3 and INT4.** INT3 is shown above and INT4 below. Bars show the sample standard deviation of eight-task mean accuracy across *Ordering* (blue) and *Sampling* (orange), in percentage points. The six model panels share a vertical scale within each bit width, but the scales differ between INT3 and INT4.

B.2 Task-Level Ordering Variability

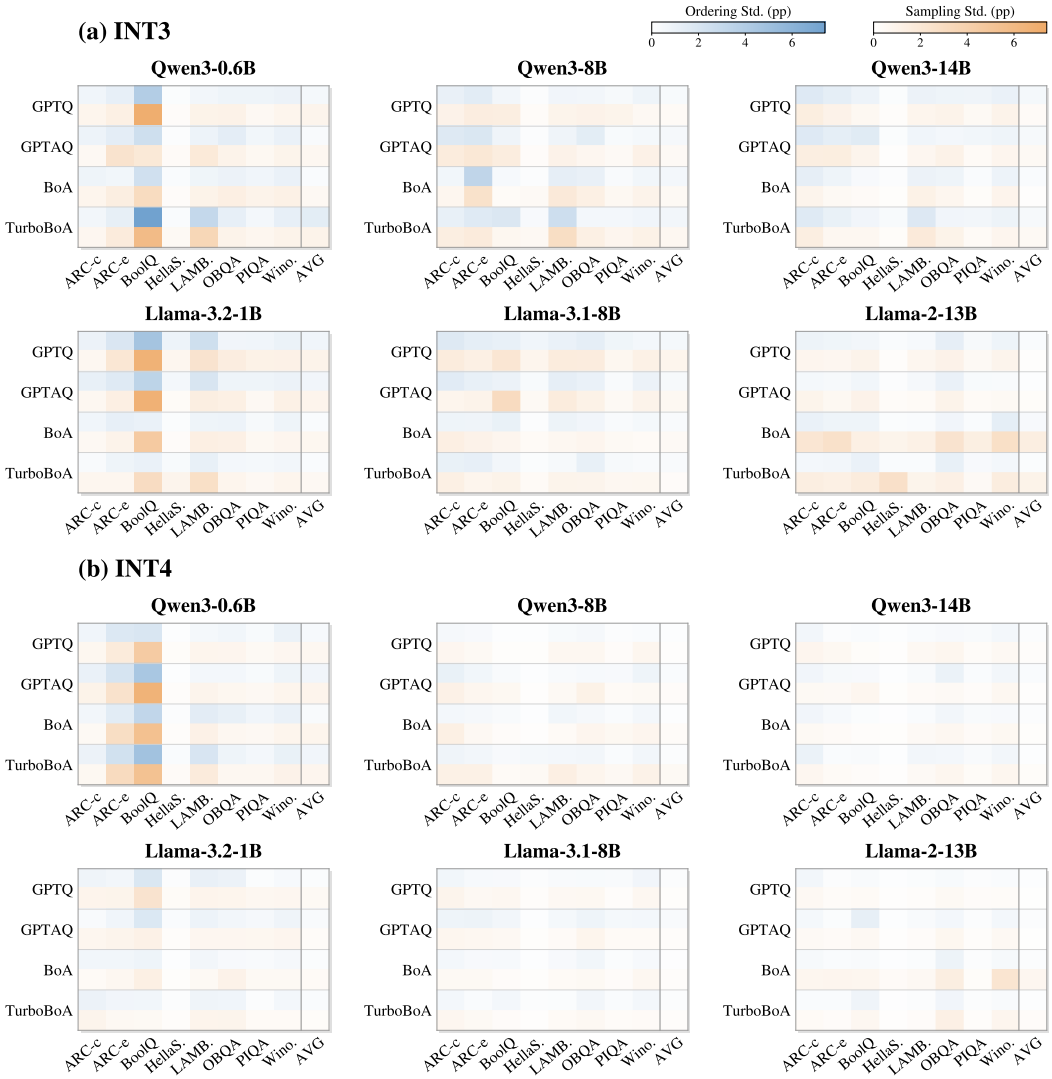

Figure 6: **Task-level standard deviations at INT3 and INT4.** INT3 is shown above and INT4 below. For each method, the upper row shows *Ordering* (blue) and the lower row *Sampling* (orange). Columns identify the eight tasks and their per-run mean (AVG). Color intensity indicates the sample standard deviation in percentage points, with shared numerical limits across all panels.

Task-level *Ordering* sensitivity is more pronounced in smaller models. At INT3 on Qwen3-0.6B with GPTAQ, the BoolQ *Ordering* standard deviation is 2.5 percentage points, compared with 0.34 points for the eight-task mean (Table 12). This variability appears less pronounced in the larger models we evaluate. Appendix F provides the complete aggregate and per-task results at both bit widths.

Figure 6 shows task-level variability across all eight tasks, models, and methods at INT3 and INT4. Each cell reports the sample standard deviation across runs in percentage points (pp). The AVG column takes the standard deviation of the eight-task mean computed within each run, rather than the mean of the task-level standard deviations. A common linear color scale is used across both bit widths, all models, and both experiments.

C Additional Analysis of the Numerical Origin of Ordering Variability

C.1 Floating-Point Accumulation Error

A floating-point addition aligns two operands to a common exponent and rounds the result to the working precision. For two positive normalized values with exponents $e_s \geq e_x$, let

$$\delta_e = e_s - e_x$$

denote the exponent gap. The aligned exact sum can be written as

$$s + x = \big(M_s + M_x 2^{-\delta_e}\big) 2^{e_s - t + 1}, \tag{9}$$

where t is the significand precision and M_s and M_x are the integer significands.

As δ_e increases, fewer bits of the smaller operand contribute at the working precision. In recursive Hessian accumulation, the running partial sum changes after every contribution. The rounding error introduced by a contribution therefore depends on where that contribution appears in the accumulation order.

For scalar floating-point addition under rounding to nearest and in the absence of overflow or underflow,

$$\mathrm{fl}(a + b) = (a + b)(1 + \varepsilon), \qquad |\varepsilon| \leq u, \tag{10}$$

where u is the unit roundoff. For FP32, $t = 24$ and $u = 2^{-24} \approx 6 \times 10^{-8}$.

The matrix recursion in Eq. 3 applies this rounding elementwise. Changing accumulation order changes the sequence of partial matrices and can therefore change the resulting perturbation ΔH_π. This is consistent with the relative Hessian differences on the order of 10^{-7} observed in FP32 and the substantially smaller differences observed in FP64.

C.2 Module-Level Hessian Perturbations and Amplification

Accumulation and inversion across all layers. Figure 7 compares order-dependent Hessian differences before and after damped inversion across all 16 layers of Llama-3.2-1B. We examine four distinct input Hessians per layer: the shared input to `q/k/v`, the input to `o`, the shared input to `gate/up`, and the input to `down`, giving 64 input locations in total. The diagnostic uses 128 fixed WikiText-2 calibration sequences, each containing 2,048 stored tokens. Activations are captured from the unquantized model and held fixed across all accumulation orders. For each sequence, GPTQ contributions are computed in FP32 as $2XX^\top$, with TF32 disabled, and accumulated in the specified sequence order.

Each bar reports the arithmetic mean of the symmetric relative Frobenius differences over five disjoint *Ordering* pairs. Panel (a) shows differences in the accumulated Hessian $\widehat{H}$, while panel (b) shows differences after the damped-inverse computation:

$$G_\pi = \widehat{H}_\pi + \lambda_\pi I, \qquad \lambda_\pi = \frac{p}{d_{\mathrm{in}}} \operatorname{tr}(\widehat{H}_\pi).$$

The pair-averaged Hessian differences range from 1.25×10^{-7} to 2.32×10^{-7}, whereas the corresponding inverse differences range from 8.73×10^{-7} to 3.22×10^{-4}. Repeated execution with the same input and order produces identical Hessian tensors in every tested case. These results were obtained under the FP32 setting described above.

Neumaier and FP64 references. The bars in Figure 7 compare two accumulation orders within the same numerical mode (FP32, Neumaier, or FP64). They measure order-dependent differences rather than error relative to the exact sum. For a fixed Hessian entry, let x_k denote the contribution from calibration sequence k, and let $S = \sum_{k=1}^{n} x_k$ denote the exact sum, which is independent of accumulation order.

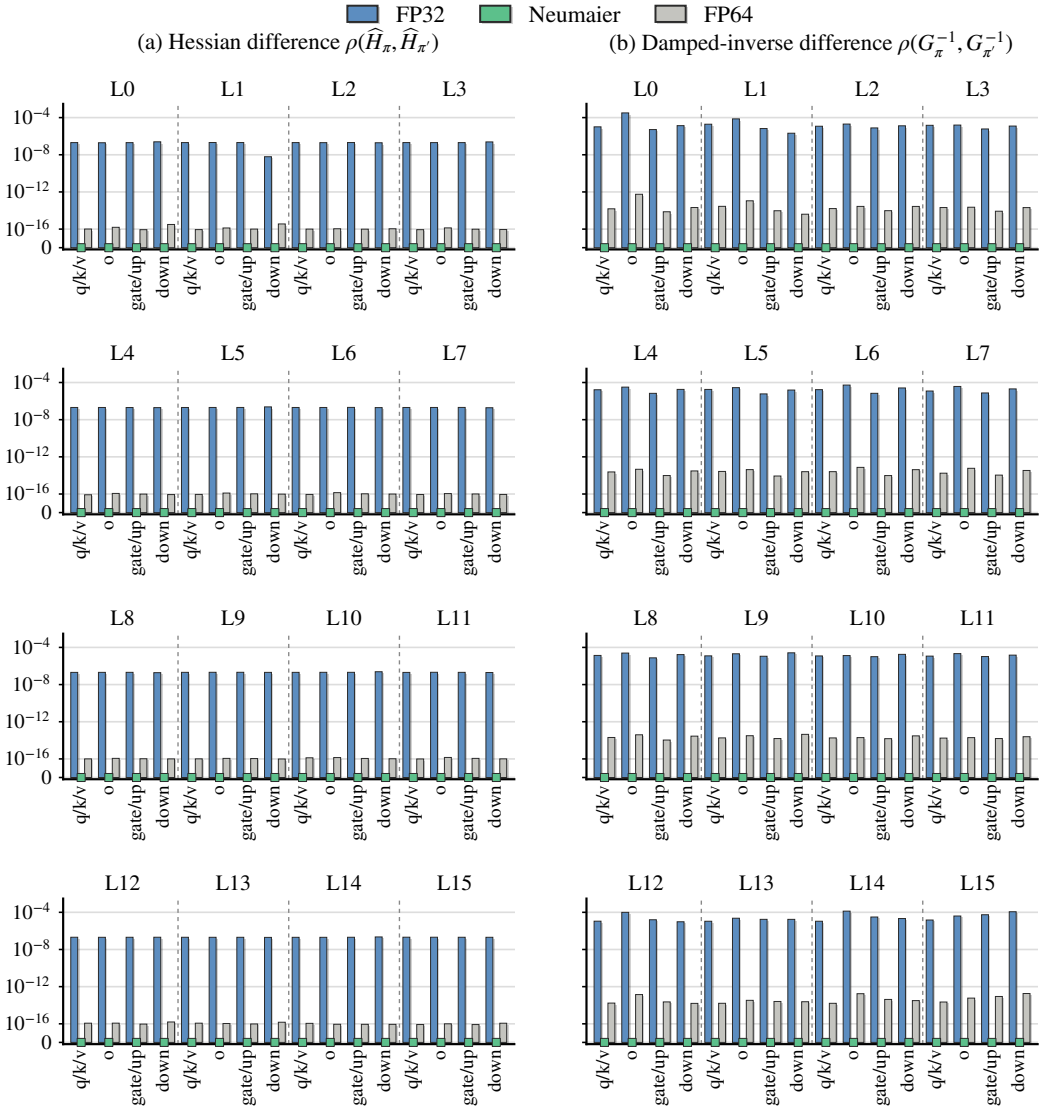

Figure 7: **Order-dependent Hessian perturbations and inverse amplification across all layers of Llama-3.2-1B.** We examine four input locations per layer across all 16 layers, giving 64 locations in total. Using fixed activations from 128 WikiText-2 calibration sequences of 2,048 tokens each, we compute GPTQ contributions as $2XX^\top$ in FP32 with TF32 disabled. Each bar reports the mean symmetric relative Frobenius difference over five disjoint *Ordering* pairs: (a) accumulated Hessians and (b) their damped inverses, with damping $\lambda_\pi = \frac{p}{d_{\text{in}}} \operatorname{tr}(\widehat{H}_\pi)$ and $p = 10^{-2}$. Both panels use the same logarithmic scale. The geometric mean of the 320 location–pair amplification ratios is $83.62\times$. Dashed lines separate layers. The four input locations within each layer correspond to `q/k/v`, `o`, `gate/up`, and `down`.

Uncompensated accumulation. Under accumulation order π, step k computes $s_k = \mathrm{fl}(s_{k-1} + x_{\pi(k)})$. Define the local rounding error as $e_k = (s_{k-1} + x_{\pi(k)}) - s_k$. Under round-to-nearest arithmetic and in the absence of overflow or underflow, $|e_k| \le u|s_{k-1} + x_{\pi(k)}|$, where u is the unit roundoff. Recursively, $s_n = S - \sum_{k=1}^{n} e_k$. Because each e_k depends on the current partial sum s_{k-1}, changing the accumulation order changes the sequence of rounding errors and can therefore change the final result. The same argument applies to FP64 accumulation. FP64 reduces the magnitude of the individual rounding errors because its unit roundoff is much smaller, but floating-point addition remains non-associative. Consequently, higher precision reduces order dependence without guaranteeing its elimination. This is consistent with the small but nonzero FP64 differences in Figure 7, which are approximately 10^{-16} for the accumulated Hessian and 10^{-14} after the damped inverse.

Neumaier compensation. The local error e_k defined above is precisely the information that uncompensated accumulation discards. Neumaier summation instead retains this residual in a separate correction variable. For $t_k = \mathrm{fl}(s_{k-1} + x_{\pi(k)})$, it computes $r_k = (s_{k-1} - t_k) + x_{\pi(k)}$ if $|s_{k-1}| \ge |x_{\pi(k)}|$, and $r_k = (x_{\pi(k)} - t_k) + s_{k-1}$ otherwise. Under the usual round-to-nearest assumptions, this branchwise correction recovers the low-order information lost in the addition (Higham, 2002). The residuals are accumulated in a correction variable c, and the final compensated result is formed as $s_n + c$. Thus, whereas uncompensated accumulation discards each local rounding residual, Neumaier carries these residuals forward and adds them back at the end. The correction accumulation itself is still performed in finite precision, so compensated summation does not in general guarantee order independence. Empirically, however, the remaining differences are sufficiently small that the accumulated statistics become identical across orders in nearly all evaluated settings.

Full-depth aggregation. Let $a_{m,p}$ denote the amplification in Eq. 5 at input location m for *Ordering* pair p. Across $M = 64$ locations and $P = 5$ disjoint pairs, the geometric mean is

$$\mathcal{A}_{\mathrm{GM}} = \exp\left(\frac{1}{MP}\sum_{m=1}^{M}\sum_{p=1}^{P}\log a_{m,p}\right) = 83.62.$$

Each location-pair combination contributes equally in log space. This is the GPTQ result reported in the cross-method diagnostic in Appendix C.4 and annotated in Figure 4(b), where it is rounded to approximately $84\times$. It summarizes multiplicative amplification across the diagnostic. It does not imply that every location amplifies its input perturbation by this factor.

Variability across input locations. To summarize each location separately, we divide its pair-averaged inverse difference by its pair-averaged Hessian difference. These ratios range from 3.76 at the layer-1 down projection to 1511.82 at the layer-0 attention output projection. This ratio of arithmetic means differs from the geometric mean of individual location-pair amplification factors defined above. Across all 64 locations, the largest pair-averaged Hessian difference is only $1.86\times$ the smallest, yet the inverse differences vary substantially more. These results show that, under the tested conditions, similarly sized Hessian perturbations can lead to substantial inverse differences across input locations.

The fixed-activation protocol isolates accumulation and inversion from input changes inherited from earlier quantized modules. These measurements characterize GPTQ input-Hessian accumulation followed by the common damped inverse on Llama-3.2-1B.

C.3 Sensitivity to Damping

Increasing damping does not monotonically reduce *Ordering* variability. We examine whether increasing the damping percentage reduces σ_{ord} in this configuration. We test this by sweeping the damping percentage p over two orders of magnitude, from 0.003 to 0.3, around the default $p = 10^{-2}$ used by every other experiment in this work. For each value, we repeat the *Ordering* experiment using all ten permutations specified in Appendix A.1 and report the same eight-task mean accuracy as the main tables.

Across the tested damping values in Table 3, σ_{ord} ranges from 0.22 to 0.41 percentage points and does not decrease monotonically as p increases. Mean accuracy varies by 0.25 percentage points

Table 3: **Sensitivity to damping** (GPTQ, Llama-3.2-1B, INT4, FP32 accumulation). Mean, range, and standard deviation of eight-task accuracy over the ten accumulation orders of Appendix A.1; † marks the default $p = 10^{-2}$ used in all other experiments.

p	acc	min	max	σ_{ord}
0.003	56.31	55.87	56.63	0.22
0.01†	56.32	55.99	56.71	0.24
0.03	56.45	55.55	56.98	0.41
0.1	56.41	55.69	56.96	0.36
0.3	56.20	55.55	56.55	0.27

across the sweep. Thus, increasing damping does not consistently reduce the observed *Ordering* variability in this configuration. Each estimate is based on ten runs, so these results do not establish that the underlying variability is insensitive to damping.

Table 4: **Cross-method input-Hessian sensitivity with a common damped inverse** (64 input locations × 5 *Ordering* pairs per method).

Method	Mean $\rho(\widehat{H}_\pi, \widehat{H}_{\pi'})$	Mean $\rho(G_\pi^{-1}, G_{\pi'}^{-1})$	GM amp.
GPTQ	2.06×10^{-7}	2.67×10^{-5}	$83.62\times$
GPTAQ	3.36×10^{-7}	3.17×10^{-5}	$61.50\times$
BoA	3.27×10^{-7}	3.21×10^{-5}	$62.40\times$
TurboBoA	3.27×10^{-7}	3.21×10^{-5}	$62.40\times$

C.4 Cross-Method Input-Hessian Sensitivity

We compare method-specific input-Hessian accumulation under fixed activations, followed by a common damped-inverse computation. The diagnostic uses Llama-3.2-1B and the stored token IDs of 128 WikiText-2 calibration sequences, each containing exactly 2,048 tokens after tokenization. We hold the unquantized BF16 activations fixed and accumulate Hessian in FP32 with TF32 disabled. We analyze 64 distinct input locations across all 16 layers: four per layer, represented by `q_proj`, `o_proj`, `gate_proj`, and `down_proj`. The query, key, and value projections share one input location, and the gate and up projections share another. The ten *Ordering* seeds form five disjoint pairs: $(777, 1024)$, $(3407, 5000)$, $(7777, 10042)$, $(27182, 31415)$, and $(2024, 424242)$.

GPTQ uses the contribution-scaling fix described in Appendix D.2. GPTAQ uses its per-sequence running-mean accumulation, while BoA and TurboBoA use their shared per-token input-covariance accumulation. For every method and accumulation order π, we form $G_\pi = \widehat{H}_\pi + \lambda_\pi I$, where $\lambda_\pi = \frac{p}{d_{\text{in}}} \operatorname{tr}(\widehat{H}_\pi)$ and $p = 10^{-2}$, and compute its inverse through Cholesky factorization.

Relative differences use ρ as defined in Eq. 4. Table 4 reports arithmetic means of these differences and the geometric mean of the inverse-to-Hessian difference ratios over all 320 location-pair combinations per method. All input differences are nonzero, so every combination contributes to the amplification statistic.

The geometric-mean amplification ranges from $61.50\times$ to $83.62\times$, a maximum-to-minimum ratio of approximately 1.36. BoA and TurboBoA give identical results because they share the input accumulation rule in this diagnostic. These measurements support sensitivity of the common inverse to perturbations produced by the tested accumulation rules.

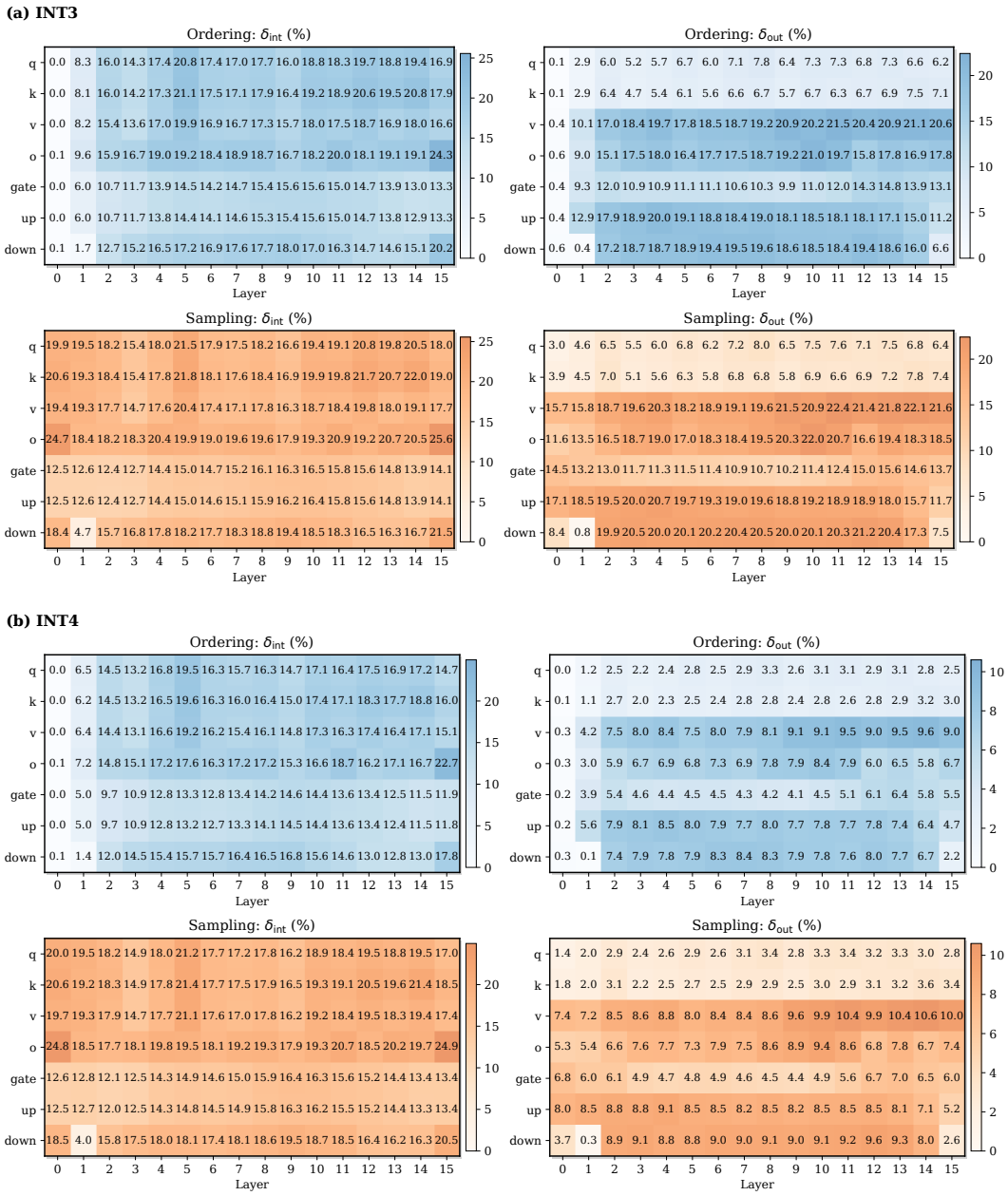

Figure 8: **Module-level integer-weight changes and module-output differences** (Llama-3.2-1B, GPTQ) at **(a)** INT3 and **(b)** INT4. Within each panel, the top row shows *Ordering* (blue) and the bottom row *Sampling* (orange); the left column shows δ_{int} and the right column δ_{out}. Columns within each heatmap identify layers and rows identify projections. Each cell reports the mean over five disjoint pairs, in percent. *Ordering* and *Sampling* share numerical color limits within each metric and bit width. All layers 0–15 are included (112 modules); panel (a) corresponds to Figure 4(c).

C.5 Module-Level Integer-Weight Changes and Module-Output Differences

Figure 8 resolves the quantities summarized in Figure 4(c) and Section 4.3 by layer and projection. Both the INT3 and INT4 panels show all 112 modules in layers 0–15. Each cell averages five disjoint run pairs within its experiment. The metrics follow Eq. 7 and 8.

Averaging over all 112 modules gives INT4 *Ordering* and *Sampling* means of 13.6% and 17.2% for δ_{int}, and 5.3% and 6.2% for δ_{out}. At INT3, the corresponding means are 14.9%, 17.5%, 12.7%,

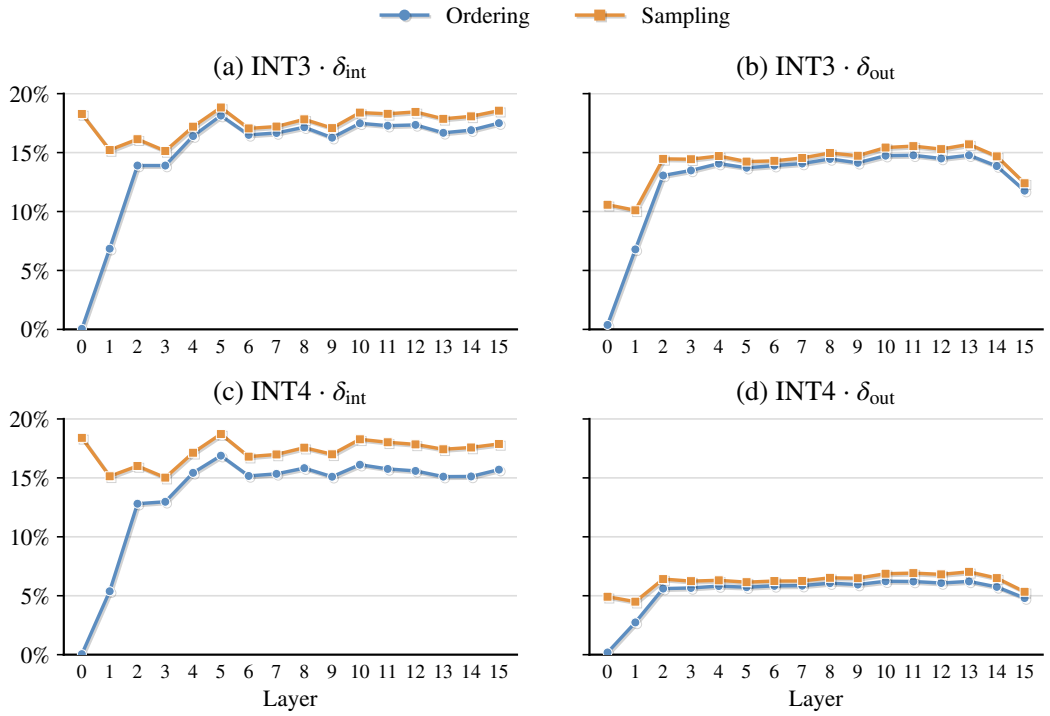

Figure 9: **Layer-wise integer-weight and module-output differences under *Ordering* and *Sampling* on Llama-3.2-1B.** Panels (a,b) show INT3 and panels (c,d) show INT4. The left column reports δ_{int}, and the right column reports δ_{out}. Each point is the arithmetic mean over seven projections and five disjoint run pairs within the corresponding experiment. For δ_{out}, both weight matrices are evaluated on shared inputs, retaining the average over the two input sets within each pair.

and 14.1%, matching Section 4.3. These averages assign equal weight to each module rather than weighting by its number of parameters.

Layer-wise patterns. Figure 9 summarizes the same measurements by layer. Each point averages the seven projections (`q`, `k`, `v`, `o`, `gate`, `up`, and `down`) and the five run pairs within the corresponding experiment. Averaging the 16 layer values therefore recovers the full-depth module means reported above.

For both precisions, the layer-averaged *Ordering* differences are small at layer 0 and increase over the first few layers. *Sampling* already produces appreciable differences at layer 0 and remains above *Ordering* in the displayed layer means. Beyond the early layers, the two conditions exhibit broadly similar patterns, with a smaller gap for δ_{out} than at layer 0. The INT3 output differences are also consistently larger than their INT4 counterparts across the displayed layers.

D Additional Details on Precision Intervention and Practical Stabilization

D.1 Neumaier Compensated Summation

Neumaier compensated summation, implemented as shown in Algorithm 1, reduces the rounding error introduced during Hessian accumulation while retaining FP32 working precision. Neumaier maintains a correction matrix C in addition to the running Hessian accumulator H.

Algorithm 1 Neumaier compensated Hessian accumulation

$H \leftarrow 0, \quad C \leftarrow 0$
for each calibration batch **do**
  $\Delta \leftarrow 2XX^\top$
  $T \leftarrow H + \Delta$
  $Mask \leftarrow (|H| \geq |\Delta|)$
  $C[Mask] \leftarrow C[Mask] + ((H[Mask] - T[Mask]) + \Delta[Mask])$
  $C[\neg Mask] \leftarrow C[\neg Mask] + ((\Delta[\neg Mask] - T[\neg Mask]) + H[\neg Mask])$
  $H \leftarrow T$
end for
return $H + C$

For entries where $|H| \geq |\Delta|$, the correction tracks low-order information lost from Δ during the addition. For the complementary entries, the symmetric expression tracks the loss from H. Both H and C remain in FP32, and the final $H+C$ is used in the subsequent Hessian-dependent computation.

The method-specific stabilized configurations defined in Appendix D.2 substantially reduce *Ordering* variability. At INT3, the standard deviation of eight-task mean accuracy is 0.00 at the reported precision in twenty of the twenty-four measured cells, and is reduced in the remainder. In the four settings with residual σ_{ord}, namely BoA on Qwen3-8B, Llama-3.1-8B, and Llama-2-13B and TurboBoA on Llama-2-13B, an order-dependent reduction outside the compensated accumulation path remains; in all other settings, the quantized weights are bit-identical across orders.

D.2 Method-Specific Implementations

We adapt Neumaier compensated summation to the accumulation, normalization, and reduction structure of each method, as illustrated in Figure 10.

GPTQ. We form each Hessian contribution as $2(XX^\top)$ instead of $(\sqrt{2}X)(\sqrt{2}X)^\top$. Applying the factor of two after the matrix product avoids the rounding introduced by scaling the inputs with the inexact FP32 representation of $\sqrt{2}$. Multiplication by two is exact in binary floating-point arithmetic when the result is representable, although the matrix product itself still incurs rounding. This reformulation is independent of Neumaier compensation, which addresses rounding during the subsequent accumulation.

We use the same post-multiplication scaling in both the naïve and compensated GPTQ conditions. The naïve update retains only the rounded running sum, whereas the compensated update additionally tracks the rounding residual in an FP32 correction buffer.

GPTAQ. To avoid repeated rescaling of the running mean, we accumulate the unnormalized statistic and apply normalization once before the Hessian-dependent computation. For GPTAQ, BoA, and TurboBoA, these restructurings are required to place the compensation and are treated as part of the method-specific stabilized configuration.

BoA and TurboBoA. Both input and output covariance collectors use Neumaier compensation for accumulation across batches. In Figure 10(c), `cov = step(cov, p, comp)` applies the same compiled Neumaier update as in panel (a), with a separate accumulator and correction buffer for each covariance. The compensated sum is normalized once after all batches. The subsequent reference unrotation implementation reduces the position-wise matrices Z_i using `mean(dim=0)`. Our stabilized implementation sums the weighted contributions $z_i = w_i Z_i$ using a fixed pairwise reduction tree and divides the result by $\sum_i w_i$. For equal-length calibration sequences, the position weights w_i are constant, so this weighted mean agrees with the reference mean in exact arithmetic. This reduction across positions is separate from the compensated covariance accumulation across batches.

Before

(a) GPTQ: scaling and compensated accumulation

```
def accumulate(inp, H):
    x = sqrt(2) * inp.float()
    delta = x.matmul(x.t())
    return H + delta
```

After

```
def neumaier_step(H, delta, C):
    T = H + delta
    C += torch.where(
        H.abs() >= delta.abs(),
        (H - T) + delta,
        (delta - T) + H
    )
    return T

step = torch.compile(neumaier_step)
# Each batch: sqrt2 fix
x = inp.float()
delta = 2.0 * x.matmul(x.t())
H = step(H, delta, C)

# Before quantization
H = H + C
```

(b) GPTAQ: accumulate first, normalize once

```
def process_batch(self, inp, b):
    n = self.nsamples
    self.H *= n / (n + b)
    self.nsamples += b
    x = sqrt(2 / (n + b)) * inp.float()
    self.H += x.matmul(x.t())
```

```
def process_batch(self, inp, b):
    self.nsamples += b
    x = inp.float()
    delta = 2.0 * x.matmul(x.t())
    self.H = step(
        self.H, delta, self.C_H
    )

def finalize(self):
    self.H = (
        self.H + self.C_H
    ) / self.nsamples
```

(c) BoA / TurboBoA: covariance and position reductions

```
# Covariance collector (XXT / YYT)
x = data.float()
p = x @ x.transpose(-1, -2)
nb = x.shape[-1]
cov = cov * (n / (n + nb))
n += nb
cov = cov + (2.0 / n) * p
# After all batches (YYT shown)
YYT = cov

def unrotate(YYT, R):
    Rt = R.transpose(-1, -2)
    z = Rt @ YYT @ R
    return z.mean(dim=0)
```

```
# Covariance collector (XXT / YYT)
x = data.float()
p = 2.0 * (x @ x.transpose(-1, -2))
cov = step(cov, p, comp)
n += x.shape[-1]
# After all batches (YYT shown)
YYT = (cov + comp) / n

def tree_sum(x):
    while x.shape[0] > 1:
        n = x.shape[0] // 2
        pairs = x[:2*n:2] + x[1:2*n:2]
        x = torch.cat((pairs, x[2*n:]), 0)
    return x[0]

def unrotate(YYT, R, w):
    Rt = R.transpose(-1, -2)
    wc = w[:, None, None, None]
    z = (Rt @ YYT @ R) * wc
    return tree_sum(z) / w.sum()
```

Shading marks the modified operations.

Figure 10: Method-specific placement of numerical stabilization. *Before* shows the reference implementations. In (a), the post-multiplication scaling ($2.0\,XX^\top$) is also used in the naïve GPTQ condition, so the naïve and Neumaier conditions differ only in compensated accumulation.

Masked Neumaier and Fused Neumaier. Neumaier accumulation maintains a running sum S and a correction C that tracks rounding errors from successive additions. For an incoming contribution Δ, the update is

$$T = S + \Delta, \qquad C \leftarrow C + \begin{cases} (S - T) + \Delta, & |S| \geq |\Delta|, \\ (\Delta - T) + S, & |S| < |\Delta|, \end{cases} \qquad S \leftarrow T.$$

The comparison and correction are applied elementwise. The corrected result is obtained by combining S and C after accumulation.

A direct masked implementation first constructs the Boolean mask $|S| \geq |\Delta|$. It then uses Boolean indexing to select the entries belonging to each branch, computes their corrections, and writes the updates back to the corresponding positions. These operations introduce additional gather and scatter operations, temporary tensors, and kernel launches. The resulting overhead reflects how the update is implemented, in addition to the arithmetic required by compensation.

The fused implementation evaluates the elementwise comparison and correction within a fused kernel. Each entry selects the appropriate correction without constructing separate Boolean-indexed subsets. Combining these operations reduces intermediate memory traffic and kernel-launch overhead while preserving the Neumaier update. The matrix multiplication that produces Δ is unchanged, so this optimization concerns compensation between completed Hessian contributions. The masked and fused implementations produce bit-identical quantized weights in the configurations where they are directly compared. Appendix D.3 reports their measured runtime costs.

D.3 Runtime and Memory Costs

Compensation adds one FP32 Hessian-sized correction buffer. For module m, let d_{in}^m and d_{out}^m denote its input and output dimensions, respectively. The weight matrix is $W_m \in \mathbb{R}^{d_{\text{out}}^m \times d_{\text{in}}^m}$, and the empirical Hessian has shape $d_{\text{in}}^m \times d_{\text{in}}^m$. The Hessian-dependent computation also maintains a working copy of the weights. Let

$$M_H = \sum_m \left(d_{\text{in}}^m\right)^2, \qquad M_W = \sum_m d_{\text{out}}^m d_{\text{in}}^m, \qquad \alpha = \frac{M_H}{M_H + M_W}. \tag{11}$$

The naïve FP32 working set is $M_H + M_W$, whereas compensated accumulation requires $2M_H + M_W$. Its relative memory cost is therefore

$$\frac{2M_H + M_W}{M_H + M_W} = 1 + \alpha. \tag{12}$$

On Llama-3.1-8B, $\alpha = 0.58$, yielding a $1.58\times$ analytical working-set ratio for Neumaier compensation, compared with $2.00\times$ when both the Hessian-side buffers and working weight copy are carried in FP64. FP64 accumulation followed by an FP32 solver also has a $1.58\times$ analytical ratio under this accounting; these ratios are distinct from measured peak GPU memory.

The fused implementation substantially reduces accumulation overhead. Across three complete quantization runs per condition in Table 5, the mean accumulation-stage times are 229.87 s for FP32, 950.98 s for masked Neumaier, and 251.06 s for fused FP32 Neumaier. These correspond to $1.00\times$, $4.14\times$, and $1.09\times$ the FP32 accumulation time, respectively. These measurements cover all accumulation calls in each run. The masked and fused implementations produce bit-identical quantized weights in the configurations where they are directly compared.

Table 5: **Memory and runtime costs for GPTQ on Llama-3.1-8B, INT3, one RTX A6000.** Memory is the analytical working-memory ratio (Equation 12), and Peak is the maximum PyTorch-allocated GPU memory measured across runs. Accum. and Solver are runtime factors for Hessian accumulation and for the subsequent Hessian-dependent computation, relative to FP32 and excluding evaluation. Total time is the mean $\pm$ sample standard deviation over n runs.

Condition	n	Total time (s)	Memory (×)	Peak (GiB)	Total (×)	Accum. (×)	Solver (×)
Naïve FP32	3	1000.73 ± 17.70	1.00	4.79	1.00	1.00	1.00
Neumaier, masked	3	1735.42 ± 27.31	1.58	11.37	1.73	4.14	1.00
Neumaier, fused	3	$\mathbf{1046.72 \pm 8.89}$	**1.58**	**5.43**	**1.05**	**1.09**	**1.01**
FP64 accumulation only	3	10055.95 ± 38.78	1.58	6.68	10.05	40.31	1.01
FP64 throughout	3	10713.47 ± 39.32	2.00	9.60	10.71	40.33	2.26

Full-run cost. In Table 5, fused Neumaier takes 1046.72 ± 8.89 s, compared with 1000.73 ± 17.70 s for naïve FP32, giving a ratio of mean total times of $1.05\times$. Its peak GPU memory is 5.43 versus 4.79 GiB for FP32. FP64 accumulation-only and FP64 throughout have mean total runtime ratios of $10.05\times$ and $10.71\times$, respectively. Fused FP32 Neumaier has lower measured total runtime and peak memory than either FP64 configuration in this benchmark. Its mean total time is approximately 5% higher than that of naïve FP32, with the end-to-end difference also reflecting variable overhead outside accumulation and the solver. All costs arise during offline quantization, exclude evaluation, and do not affect inference.

E Discussion and Limitations

E.1 Discussion

Method comparisons should account for accumulation order. The rank changes in Section 3.2 show that one accumulation order may not be enough when methods have similar accuracy. Reporting results over the same set of permutations shows how often one method outperforms another. When calibration sets are randomly sampled, accumulation order may also vary unless it is controlled separately, so the observed variation can reflect both sample selection and accumulation order. The calibration set, accumulation order, and numerical settings should therefore be controlled and reported. Fixing one order makes a result repeatable in the same environment, but a method that performs better under that order may not perform better under other orders.

Numerical stabilization reduces variation but does not guarantee higher accuracy. Reducing accumulation error changes the estimated Hessian and can alter subsequent quantization decisions, but does not guarantee higher accuracy. FP64 accumulation and Neumaier compensation follow different numerical paths and can therefore produce different quantized models. The FP64 results in Section 5.2 illustrate this difference, while Neumaier provides a practical way to reduce ordering variation. The outcome also depends on which operations use higher precision or compensation, so we distinguish FP64 accumulation from using higher precision in later computations.

Method rankings may differ across model families. We observe different rankings on Qwen3 and Llama, suggesting that a method that performs better on one family may not perform better on another. One possible explanation is that some recent Hessian-based PTQ methods, such as TurboBoA, have been developed and evaluated mainly on Llama-family models (Kim et al., 2026), so their design choices may not transfer equally to other model families. Qwen3 also differs architecturally from Llama, for example through query–key normalization in attention (Yang et al., 2025). These differences could affect how quantization errors propagate and how well each method accounts for them. However, our experiments do not isolate the effect of model architecture or the development context of each method. We therefore cannot attribute the observed ranking differences to query–key normalization or to a specific weakness of layer-wise PTQ. Comparisons across model families are needed before drawing broader conclusions about which method performs better.

E.2 LIMITATIONS

Implementation and hardware dependence. Neumaier compensation reduces rounding error when accumulating matrix-product contributions, but does not correct rounding within each matrix product. Order-dependent reductions outside the compensated path remain in several BoA settings and one TurboBoA setting at INT3 (Appendix D.1). All experiments use Ampere-family GPUs, and the runtime and memory measurements in Appendix D.3 cover GPTQ on Llama-3.1-8B at INT3 using an RTX A6000. The measured overhead may therefore differ across methods, models, hardware, and software environments.

Limited comparison of stabilization methods. Our experiments focus on Neumaier compensation rather than a broad comparison of alternatives such as pairwise reduction and other compensated summation schemes. To isolate the effect of compensation, we keep the product computation and other numerically equivalent components unchanged wherever possible, so that differences primarily reflect how partial results are accumulated. This design is intended to study the source of ordering variability rather than to identify the best stabilization algorithm or implementation. A broader comparison of summation schemes, together with kernel-level optimization of their runtime and memory costs, is left for future work.

F COMPLETE RESULT TABLES

This section reports aggregate and per-task results at INT4 and INT3, grouped by model family. All runs use WikiText-2 calibration, with perplexity evaluated on WikiText-2 and C4. *Ord.* denotes *Ordering*, while *Samp.* denotes *Sampling*. *Neu.* denotes Neumaier.

Each metric uses ten runs. Accuracy is reported in percent, and the eight-task mean accuracy is the unweighted mean defined in Appendix A.1. Standard deviations are computed across runs with denominator $n - 1$. The σ ratio is the *Ordering* standard deviation divided by the *Sampling* standard deviation for the same metric, using unrounded values. In the per-task tables, subscripts give standard deviations.

Green highlights Neu. rows with ten runs for every reported metric and zero dispersion at the displayed precision. For the per-task tables, this condition applies to every task and the eight-task mean. Green rows are bit-identical across all ten orders. Grey marks unquantized references.

GPTAQ perplexity on Llama-3.2-1B. GPTAQ on Llama-3.2-1B yields unstable perplexity at both bit widths (WikiText-2 PPL 1043 ± 1220 at INT3, 26.6 ± 7.7 at INT4 under *Ordering*) while its zero-shot accuracy remains stable (48.85 ± 0.75 at INT3). The same implementation produces stable perplexity on the other five models (Tables 6–9), so the instability is specific to this method–model pair rather than to the *Ordering* protocol. Neumaier eliminates the observed across-order variation in this setting, but perplexity remains high (PPL 207). We report these values for completeness.

Table 6: **INT4 aggregate results for Qwen3.**

Model	Method	EXP	8-Task (%) acc ↑	min	max	std ↓	σ ratio	WikiText-2 PPL mean / std ↓	σ ratio	C4 PPL mean / std ↓	σ ratio
Qwen3-0.6B	FP16 (original)	—	50.17	—	—	—	—	20.96	—	30.34	—
	GPTQ	Ord.	48.32	47.79	49.26	0.45	0.67	25.22 / 0.139	0.32	37.04 / 0.267	0.60
		Samp.	47.34	46.17	48.03	0.68		25.38 / 0.440		37.28 / 0.443	
		Neu.	48.24	48.24	48.24	0.00		25.22 / 0.000		36.64 / 0.000	
	GPTAQ	Ord.	47.72	46.12	48.70	0.74	0.81	23.36 / 0.044	0.52	34.82 / 0.153	0.64
		Samp.	46.54	45.09	48.36	0.92		23.42 / 0.084		35.11 / 0.241	
		Neu.	47.72	47.72	47.72	0.00		23.40 / 0.000		34.97 / 0.000	
	BoA	Ord.	42.60	42.07	43.08	0.36	0.42	28.13 / 0.441	0.74	43.07 / 0.729	**1.30**
		Samp.	42.72	41.58	44.11	0.86		28.04 / 0.592		42.94 / 0.562	
		Neu.	43.23	43.23	43.23	0.00		28.81 / 0.000		43.58 / 0.000	
	TurboBoA	Ord.	44.89	43.91	46.26	0.75	0.92	25.67 / 0.639	**1.21**	38.90 / 0.821	**1.30**
		Samp.	43.71	42.47	44.86	0.81		25.60 / 0.527		39.01 / 0.634	
		Neu.	43.06	43.06	43.06	0.00		25.10 / 0.000		38.13 / 0.000	
Qwen3-8B	FP16 (original)	—	69.08	—	—	—	—	9.73	—	15.43	—
	GPTQ	Ord.	68.31	68.13	68.66	0.14	0.85	9.92 / 0.028	**1.03**	15.76 / 0.046	**1.53**
		Samp.	68.31	68.07	68.50	0.16		9.92 / 0.028		15.82 / 0.030	
		Neu.	68.44	68.44	68.44	0.00		9.85 / 0.000		15.69 / 0.000	
	GPTAQ	Ord.	68.03	67.53	68.38	0.26	0.83	9.98 / 0.020	0.66	15.87 / 0.030	0.61
		Samp.	68.21	67.58	68.62	0.32		9.96 / 0.031		15.86 / 0.050	
		Neu.	68.59	68.59	68.59	0.00		9.96 / 0.000		15.83 / 0.000	
	BoA	Ord.	67.11	66.91	67.30	0.15	0.45	10.66 / 0.042	0.57	17.55 / 0.050	0.39
		Samp.	66.92	66.39	67.43	0.33		10.63 / 0.074		17.54 / 0.126	
		Neu.	67.03	67.01	67.14	0.04		10.60 / 0.002		17.57 / 0.022	
	TurboBoA	Ord.	66.26	65.95	66.56	0.20	0.39	10.97 / 0.127	0.61	17.49 / 0.173	0.60
		Samp.	66.45	65.85	67.21	0.50		10.83 / 0.207		17.27 / 0.291	
		Neu.	66.35	66.35	66.35	0.00		11.07 / 0.000		17.69 / 0.000	
Qwen3-14B	FP16 (original)	—	72.51	—	—	—	—	8.65	—	13.82	—
	GPTQ	Ord.	72.14	71.84	72.35	0.17	0.68	8.82 / 0.021	0.90	14.18 / 0.028	**1.01**
		Samp.	71.95	71.53	72.35	0.26		8.81 / 0.023		14.17 / 0.028	
		Neu.	72.20	72.20	72.20	0.00		8.84 / 0.000		14.23 / 0.000	
	GPTAQ	Ord.	71.96	71.41	72.37	0.30	**2.08**	8.82 / 0.008	**1.02**	14.09 / 0.020	0.60
		Samp.	71.77	71.58	71.99	0.14		8.82 / 0.008		14.14 / 0.033	
		Neu.	71.94	71.94	71.94	0.00		8.80 / 0.000		14.06 / 0.000	
	BoA	Ord.	71.53	71.17	71.79	0.18	0.97	9.19 / 0.064	**1.39**	15.05 / 0.057	0.79
		Samp.	71.63	71.30	71.86	0.18		9.17 / 0.046		14.98 / 0.072	
		Neu.	71.75	71.58	71.77	0.06		9.10 / 0.025		14.95 / 0.023	
	TurboBoA	Ord.	71.11	70.86	71.27	0.12	0.47	9.16 / 0.017	0.58	14.85 / 0.058	**1.05**
		Samp.	71.26	70.82	71.73	0.26		9.18 / 0.029		14.89 / 0.055	
		Neu.	71.23	71.23	71.23	0.00		9.14 / 0.000		14.85 / 0.000	

Table 7: **INT4 aggregate results for Llama.**

Model	Method	EXP	8-Task (%) acc ↑	min	max	std ↓	σ ratio	WikiText-2 PPL mean / std ↓	σ ratio	C4 PPL mean / std ↓	σ ratio
Llama-3.2-1B	FP16 (original)	—	58.28	—	—	—	—	9.76	—	14.02	—
	GPTQ	Ord.	56.32	55.99	56.71	0.24	0.45	10.58 / 0.012	0.10	16.41 / 0.042	0.54
		Samp.	56.39	55.24	56.98	0.54		10.63 / 0.118		16.46 / 0.079	
		Neu.	56.60	56.60	56.60	0.00		10.55 / 0.000		16.45 / 0.000	
	GPTAQ	Ord.	56.30	55.65	56.80	0.33	0.98	26.59 / 7.699	0.78	43.29 / 11.485	0.80
		Samp.	56.44	55.94	57.13	0.33		28.96 / 9.927		44.15 / 14.357	
		Neu.	56.49	56.49	56.49	0.00		16.05 / 0.000		25.62 / 0.000	
	BoA	Ord.	56.55	56.24	56.96	0.25	0.61	10.53 / 0.008	0.91	16.16 / 0.016	0.54
		Samp.	56.43	55.63	57.09	0.40		10.53 / 0.008		16.17 / 0.029	
		Neu.	56.31	56.31	56.31	0.00		10.53 / 0.000		16.19 / 0.000	
	TurboBoA	Ord.	56.32	55.74	56.79	0.32	**1.13**	10.50 / 0.006	**1.02**	16.28 / 0.020	0.75
		Samp.	56.50	55.93	56.90	0.28		10.50 / 0.006		16.25 / 0.027	
		Neu.	56.47	56.47	56.47	0.00		10.50 / 0.000		16.26 / 0.000	
Llama-3.1-8B	FP16 (original)	—	71.77	—	—	—	—	6.24	—	9.54	—
	GPTQ	Ord.	70.39	70.14	70.64	0.18	0.97	6.61 / 0.007	**1.00**	10.62 / 0.011	0.62
		Samp.	70.43	70.21	70.67	0.18		6.62 / 0.007		10.63 / 0.018	
		Neu.	70.14	70.14	70.14	0.00		6.60 / 0.000		10.63 / 0.000	
	GPTAQ	Ord.	70.42	70.01	70.93	0.32	**1.32**	6.76 / 0.051	0.81	10.86 / 0.041	**1.05**
		Samp.	70.61	70.25	71.08	0.24		6.75 / 0.063		10.84 / 0.039	
		Neu.	70.79	70.79	70.79	0.00		6.82 / 0.000		10.91 / 0.000	
	BoA	Ord.	70.54	70.35	70.67	0.10	0.47	6.64 / 0.005	**1.20**	10.69 / 0.011	0.99
		Samp.	70.46	70.06	70.74	0.21		6.64 / 0.004		10.69 / 0.011	
		Neu.	70.59	70.55	70.72	0.07		6.63 / 0.001		10.68 / 0.003	
	TurboBoA	Ord.	70.38	70.11	70.66	0.18	0.80	6.65 / 0.004	0.67	10.76 / 0.009	0.49
		Samp.	70.42	70.00	70.73	0.22		6.65 / 0.005		10.75 / 0.018	
		Neu.	70.40	70.40	70.40	0.00		6.64 / 0.000		10.75 / 0.000	
Llama-2-13B	FP16 (original)	—	70.50	—	—	—	—	4.88	—	6.73	—
	GPTQ	Ord.	70.10	69.83	70.36	0.17	**1.51**	5.02 / 0.002	0.14	6.97 / 0.004	0.16
		Samp.	70.33	70.12	70.48	0.11		5.01 / 0.013		6.96 / 0.025	
		Neu.	70.09	70.09	70.09	0.00		5.01 / 0.000		6.97 / 0.000	
	GPTAQ	Ord.	70.32	69.97	70.52	0.20	**1.77**	4.96 / 0.002	**2.00**	6.89 / 0.003	0.86
		Samp.	70.16	69.97	70.35	0.11		4.96 / 0.001		6.89 / 0.003	
		Neu.	70.41	70.41	70.41	0.00		4.96 / 0.000		6.89 / 0.000	
	BoA	Ord.	67.07	66.61	67.37	0.22	0.30	5.06 / 0.006	0.58	6.92 / 0.001	0.63
		Samp.	67.53	66.51	68.46	0.74		5.07 / 0.010		6.93 / 0.002	
		Neu.	66.98	66.86	67.17	0.11		5.05 / 0.001		6.92 / 0.001	
	TurboBoA	Ord.	69.16	68.95	69.31	0.12	0.42	5.00 / 0.001	0.23	6.92 / 0.002	0.94
		Samp.	69.48	68.91	69.93	0.29		5.00 / 0.005		6.93 / 0.003	
		Neu.	69.33	69.33	69.33	0.00		4.99 / 0.000		6.93 / 0.000	

Table 8: **INT3 aggregate results for Qwen3.**

Model	Method	EXP	8-Task (%) acc ↑	min	max	std ↓	σ ratio	WikiText-2 PPL mean / std ↓	σ ratio	C4 PPL mean / std ↓	σ ratio
Qwen3-0.6B	FP16 (original)	—	50.17	—	—	—	—	20.96	—	30.34	—
	GPTQ	Ord.	37.78	36.68	38.52	0.58	0.61	53.05 / 1.471	0.58	82.75 / 3.553	0.86
		Samp.	37.21	36.19	39.08	0.95		53.57 / 2.546		83.82 / 4.146	
		Neu.	37.26	37.26	37.26	0.00		54.39 / 0.000		82.27 / 0.000	
	GPTAQ	Ord.	40.99	40.41	41.46	0.34	0.52	34.80 / 0.515	0.47	56.34 / 1.482	0.67
		Samp.	41.06	39.57	41.80	0.65		35.23 / 1.094		57.15 / 2.215	
		Neu.	41.22	41.22	41.22	0.00		34.42 / 0.000		55.94 / 0.000	
	BoA	Ord.	37.26	36.89	37.75	0.27	0.44	60.47 / 2.252	0.71	116.17 / 8.264	0.93
		Samp.	37.24	36.53	38.29	0.60		60.93 / 3.174		119.25 / 8.891	
		Neu.	37.13	37.13	37.13	0.00		59.35 / 0.000		116.77 / 0.000	
	TurboBoA	Ord.	38.05	35.59	40.04	1.42	**1.40**	47.43 / 3.835	0.66	79.00 / 5.559	0.62
		Samp.	38.80	37.16	40.17	1.01		48.01 / 5.847		82.95 / 8.916	
		Neu.	39.46	39.46	39.46	0.00		47.12 / 0.000		80.11 / 0.000	
Qwen3-8B	FP16 (original)	—	69.08	—	—	—	—	9.73	—	15.43	—
	GPTQ	Ord.	63.59	62.85	64.15	0.39	0.87	11.25 / 0.087	**1.44**	18.06 / 0.149	**1.55**
		Samp.	63.58	62.81	64.12	0.45		11.24 / 0.060		18.07 / 0.096	
		Neu.	63.42	63.42	63.42	0.00		11.22 / 0.000		17.99 / 0.000	
	GPTAQ	Ord.	63.27	62.15	63.90	0.51	**1.06**	11.15 / 0.053	**1.28**	18.14 / 0.130	**1.55**
		Samp.	63.49	62.85	64.45	0.48		11.12 / 0.041		18.12 / 0.084	
		Neu.	63.54	63.54	63.54	0.00		11.08 / 0.000		18.04 / 0.000	
	BoA	Ord.	54.14	53.11	55.00	0.58	**1.12**	16.41 / 0.323	**1.32**	31.52 / 0.840	**1.16**
		Samp.	54.36	53.59	55.20	0.52		15.97 / 0.244		30.68 / 0.722	
		Neu.	53.87	53.82	53.87	0.02		16.17 / 0.027		30.92 / 0.163	
	TurboBoA	Ord.	61.75	60.53	62.55	0.66	**1.02**	15.33 / 0.327	0.77	26.97 / 0.700	0.70
		Samp.	61.98	60.95	63.25	0.64		15.03 / 0.423		26.67 / 1.004	
		Neu.	62.68	62.68	62.68	0.00		15.19 / 0.000		27.11 / 0.000	
Qwen3-14B	FP16 (original)	—	72.51	—	—	—	—	8.65	—	13.82	—
	GPTQ	Ord.	68.09	66.94	68.83	0.56	**1.42**	9.69 / 0.052	**1.26**	15.71 / 0.052	0.80
		Samp.	68.06	67.50	68.65	0.39		9.69 / 0.041		15.72 / 0.065	
		Neu.	68.51	68.51	68.51	0.00		9.66 / 0.000		15.78 / 0.000	
	GPTAQ	Ord.	68.26	67.69	69.07	0.47	0.94	9.70 / 0.024	0.57	15.70 / 0.062	**1.40**
		Samp.	68.43	67.64	69.01	0.50		9.69 / 0.041		15.72 / 0.045	
		Neu.	68.19	68.19	68.19	0.00		9.68 / 0.000		15.72 / 0.000	
	BoA	Ord.	67.27	66.68	67.56	0.26	**1.59**	10.61 / 0.063	0.96	18.73 / 0.108	0.63
		Samp.	67.23	66.95	67.51	0.16		10.52 / 0.066		18.71 / 0.172	
		Neu.	67.08	67.08	67.08	0.00		10.59 / 0.000		18.66 / 0.000	
	TurboBoA	Ord.	67.82	67.22	68.52	0.48	0.96	10.93 / 0.086	0.53	19.14 / 0.189	0.61
		Samp.	68.06	66.99	68.63	0.50		10.98 / 0.162		19.27 / 0.307	
		Neu.	68.03	68.03	68.03	0.00		10.87 / 0.000		18.92 / 0.000	

Table 9: **INT3 aggregate results for Llama.**

Model	Method	EXP	8-Task (%) acc ↑	min	max	std ↓	σ ratio	WikiText-2 PPL mean / std ↓	σ ratio	C4 PPL mean / std ↓	σ ratio
Llama-3.2-1B	FP16 (original)	—	58.28	—	—	—	—	9.76	—	14.02	—
	GPTQ	Ord.	43.99	42.71	45.05	0.77	0.76	15.60 / 0.101	0.31	31.28 / 1.039	**1.20**
		Samp.	44.98	43.50	46.77	1.02		15.64 / 0.329		30.57 / 0.865	
		Neu.	43.27	43.27	43.27	0.00		15.45 / 0.000		30.08 / 0.000	
	GPTAQ	Ord.	48.85	47.37	49.66	0.75	0.83	1043.54 / 1220.266	**2.61**	581.25 / 463.649	**1.83**
		Samp.	48.27	47.02	49.72	0.91		517.38 / 466.965		319.88 / 253.972	
		Neu.	48.95	48.95	48.95	0.00		207.06 / 0.000		188.48 / 0.000	
	BoA	Ord.	50.69	50.16	51.07	0.29	0.41	13.25 / 0.114	0.28	22.21 / 0.064	0.16
		Samp.	50.88	49.36	51.57	0.69		13.34 / 0.411		22.74 / 0.406	
		Neu.	50.85	50.85	50.85	0.00		13.27 / 0.000		22.40 / 0.000	
	TurboBoA	Ord.	46.43	46.10	47.03	0.26	0.45	13.27 / 0.020	0.11	23.75 / 0.062	0.14
		Samp.	47.94	47.01	48.88	0.58		13.20 / 0.179		23.20 / 0.429	
		Neu.	47.12	47.12	47.12	0.00		13.27 / 0.000		23.78 / 0.000	
Llama-3.1-8B	FP16 (original)	—	71.77	—	—	—	—	6.24	—	9.54	—
	GPTQ	Ord.	62.90	62.12	63.80	0.58	0.67	8.16 / 0.025	0.97	14.63 / 0.068	**1.03**
		Samp.	63.28	61.78	64.46	0.87		8.15 / 0.026		14.67 / 0.066	
		Neu.	63.00	63.00	63.00	0.00		8.20 / 0.000		14.58 / 0.000	
	GPTAQ	Ord.	65.12	64.38	65.74	0.48	0.77	9.76 / 0.612	**1.01**	16.74 / 0.544	0.76
		Samp.	64.84	63.25	65.42	0.63		9.60 / 0.606		16.94 / 0.714	
		Neu.	65.51	65.51	65.51	0.00		9.28 / 0.000		17.09 / 0.000	
	BoA	Ord.	66.29	65.79	66.85	0.32	0.68	7.63 / 0.016	0.63	13.68 / 0.022	0.86
		Samp.	66.77	65.80	67.35	0.47		7.63 / 0.025		13.66 / 0.025	
		Neu.	66.75	66.37	66.80	0.13		7.63 / 0.000		13.70 / 0.000	
	TurboBoA	Ord.	65.87	65.41	66.44	0.37	**1.23**	7.69 / 0.010	0.89	13.84 / 0.033	0.82
		Samp.	66.06	65.60	66.64	0.31		7.68 / 0.011		13.82 / 0.040	
		Neu.	65.80	65.80	65.80	0.00		7.69 / 0.000		13.89 / 0.000	
Llama-2-13B	FP16 (original)	—	70.50	—	—	—	—	4.88	—	6.73	—
	GPTQ	Ord.	67.27	66.95	67.75	0.28	**1.70**	5.41 / 0.007	0.82	7.81 / 0.007	0.69
		Samp.	67.16	66.89	67.40	0.17		5.41 / 0.008		7.81 / 0.010	
		Neu.	67.25	67.25	67.25	0.00		5.40 / 0.000		7.81 / 0.000	
	GPTAQ	Ord.	68.21	68.00	68.36	0.12	0.41	5.33 / 0.007	**1.25**	7.72 / 0.009	0.71
		Samp.	68.23	67.91	68.76	0.28		5.32 / 0.006		7.72 / 0.013	
		Neu.	68.25	68.25	68.25	0.00		5.34 / 0.000		7.71 / 0.000	
	BoA	Ord.	65.17	64.43	65.76	0.43	0.30	5.66 / 0.022	0.41	7.64 / 0.007	0.98
		Samp.	63.69	60.98	65.36	1.43		5.67 / 0.054		7.63 / 0.007	
		Neu.	65.59	65.48	65.90	0.13		5.68 / 0.002		7.64 / 0.002	
	TurboBoA	Ord.	66.38	65.99	66.78	0.24	0.23	5.47 / 0.012	0.23	7.81 / 0.014	0.29
		Samp.	65.90	63.42	66.94	1.04		5.46 / 0.053		7.81 / 0.049	
		Neu.	66.16	66.03	66.17	0.04		5.44 / 0.001		7.79 / 0.001	

Table 10: **INT4 per-task results for Qwen3.**

Model	Method	EXP	ARC-c	ARC-e	BoolQ	HellaS.	LAMB.	OBQA	PIQA	Wino.	AVG
Qwen3-0.6B	FP16	—	34.2	60.6	64.0	47.3	39.9	31.4	67.5	56.4	50.17
	GPTQ	Ord.	$32.0_{0.8}$	$56.7_{1.8}$	$66.4_{2.0}$	$43.8_{0.1}$	$37.6_{0.5}$	$31.0_{0.7}$	$64.9_{0.4}$	$54.1_{1.0}$	$48.32_{0.45}$
		Samp.	$31.6_{0.7}$	$55.1_{1.7}$	$62.8_{4.5}$	$43.7_{0.3}$	$36.8_{0.9}$	$30.5_{0.8}$	$64.4_{0.5}$	$53.8_{0.7}$	$47.34_{0.68}$
		Neu.	$31.8_{\mathbf{0.0}}$	$57.6_{\mathbf{0.0}}$	$63.6_{\mathbf{0.0}}$	$43.8_{\mathbf{0.0}}$	$38.8_{\mathbf{0.0}}$	$31.4_{\mathbf{0.0}}$	$65.5_{\mathbf{0.0}}$	$53.5_{\mathbf{0.0}}$	$48.24_{\mathbf{0.00}}$
	GPTAQ	Ord.	$31.1_{1.1}$	$56.2_{2.1}$	$61.3_{4.4}$	$44.6_{0.1}$	$38.0_{0.6}$	$30.9_{0.6}$	$65.2_{0.4}$	$54.4_{0.9}$	$47.72_{0.74}$
		Samp.	$30.7_{1.0}$	$54.1_{2.5}$	$54.7_{6.6}$	$44.6_{0.2}$	$38.5_{0.9}$	$30.0_{0.7}$	$65.4_{0.6}$	$54.4_{0.8}$	$46.54_{0.92}$
		Neu.	$32.3_{\mathbf{0.0}}$	$56.4_{\mathbf{0.0}}$	$57.9_{\mathbf{0.0}}$	$44.6_{\mathbf{0.0}}$	$37.3_{\mathbf{0.0}}$	$31.4_{\mathbf{0.0}}$	$66.3_{\mathbf{0.0}}$	$55.6_{\mathbf{0.0}}$	$47.72_{\mathbf{0.00}}$
	BoA	Ord.	$27.6_{0.7}$	$46.4_{1.5}$	$58.7_{3.2}$	$40.3_{0.2}$	$24.9_{1.5}$	$27.7_{1.2}$	$62.2_{0.8}$	$53.0_{1.1}$	$42.60_{0.36}$
		Samp.	$27.6_{0.5}$	$47.2_{2.9}$	$57.7_{5.4}$	$40.3_{0.2}$	$26.1_{1.2}$	$28.2_{0.7}$	$61.7_{0.6}$	$53.0_{0.8}$	$42.72_{0.86}$
		Neu.	$28.3_{\mathbf{0.0}}$	$48.3_{\mathbf{0.0}}$	$59.3_{\mathbf{0.0}}$	$39.7_{\mathbf{0.0}}$	$25.3_{\mathbf{0.0}}$	$29.2_{\mathbf{0.0}}$	$61.8_{\mathbf{0.0}}$	$53.9_{\mathbf{0.0}}$	$43.23_{\mathbf{0.00}}$
	TurboBoA	Ord.	$29.0_{1.0}$	$47.1_{2.4}$	$58.6_{4.8}$	$42.2_{0.3}$	$34.2_{2.0}$	$28.9_{0.9}$	$64.3_{0.6}$	$54.8_{1.1}$	$44.89_{0.75}$
		Samp.	$29.3_{0.6}$	$47.0_{3.1}$	$48.7_{5.3}$	$42.4_{0.4}$	$33.9_{1.7}$	$29.1_{0.7}$	$64.6_{0.7}$	$54.8_{1.1}$	$43.71_{0.81}$
		Neu.	$28.6_{\mathbf{0.0}}$	$44.1_{\mathbf{0.0}}$	$47.9_{\mathbf{0.0}}$	$42.2_{\mathbf{0.0}}$	$32.7_{\mathbf{0.0}}$	$29.6_{\mathbf{0.0}}$	$64.6_{\mathbf{0.0}}$	$54.8_{\mathbf{0.0}}$	$43.06_{\mathbf{0.00}}$
Qwen3-8B	FP16	—	56.5	83.4	86.6	74.8	64.3	41.6	77.6	67.8	69.08
	GPTQ	Ord.	$54.6_{0.5}$	$82.0_{0.4}$	$86.7_{0.2}$	$74.1_{0.2}$	$62.8_{0.4}$	$42.1_{0.5}$	$76.4_{0.2}$	$67.8_{0.3}$	$68.31_{0.14}$
		Samp.	$55.1_{0.7}$	$81.7_{0.5}$	$86.6_{0.2}$	$74.2_{0.2}$	$62.9_{0.6}$	$41.4_{0.6}$	$76.7_{0.3}$	$67.8_{0.8}$	$68.31_{0.16}$
		Neu.	$53.2_{\mathbf{0.0}}$	$81.0_{\mathbf{0.0}}$	$87.2_{\mathbf{0.0}}$	$74.2_{\mathbf{0.0}}$	$62.9_{\mathbf{0.0}}$	$43.8_{\mathbf{0.0}}$	$76.9_{\mathbf{0.0}}$	$68.4_{\mathbf{0.0}}$	$68.44_{\mathbf{0.00}}$
	GPTAQ	Ord.	$53.5_{1.1}$	$81.1_{0.7}$	$86.6_{0.3}$	$73.8_{0.2}$	$62.3_{0.3}$	$41.4_{0.5}$	$76.8_{0.4}$	$68.7_{0.9}$	$68.03_{0.26}$
		Samp.	$54.2_{1.2}$	$82.1_{0.8}$	$86.1_{0.6}$	$73.9_{0.2}$	$62.4_{0.5}$	$41.7_{1.1}$	$76.8_{0.4}$	$68.5_{0.5}$	$68.21_{0.32}$
		Neu.	$54.6_{\mathbf{0.0}}$	$82.2_{\mathbf{0.0}}$	$86.4_{\mathbf{0.0}}$	$73.7_{\mathbf{0.0}}$	$62.8_{\mathbf{0.0}}$	$41.8_{\mathbf{0.0}}$	$77.4_{\mathbf{0.0}}$	$69.9_{\mathbf{0.0}}$	$68.59_{\mathbf{0.00}}$
	BoA	Ord.	$53.7_{0.6}$	$80.7_{0.5}$	$86.1_{0.3}$	$71.1_{0.1}$	$63.0_{0.7}$	$39.9_{0.5}$	$76.5_{0.3}$	$65.9_{0.6}$	$67.11_{0.15}$
		Samp.	$53.0_{1.3}$	$80.1_{0.6}$	$85.9_{0.3}$	$71.2_{0.2}$	$63.4_{0.4}$	$39.9_{0.8}$	$76.2_{0.6}$	$65.6_{0.8}$	$66.92_{0.33}$
		Neu.	$52.2_{0.2}$	$80.2_{0.2}$	$86.3_{\mathbf{0.0}}$	$70.9_{0.1}$	$63.9_{0.4}$	$40.0_{0.1}$	$76.3_{0.1}$	$66.3_{\mathbf{0.0}}$	$67.03_{0.04}$
	TurboBoA	Ord.	$52.4_{0.8}$	$79.6_{0.7}$	$86.4_{0.4}$	$70.7_{0.5}$	$58.7_{0.6}$	$39.8_{0.7}$	$75.7_{0.3}$	$66.8_{0.7}$	$66.26_{0.20}$
		Samp.	$52.1_{1.2}$	$79.8_{1.2}$	$86.3_{0.5}$	$71.1_{0.5}$	$59.3_{1.3}$	$39.5_{0.9}$	$76.0_{0.6}$	$67.5_{1.1}$	$66.45_{0.50}$
		Neu.	$52.0_{\mathbf{0.0}}$	$80.2_{\mathbf{0.0}}$	$86.0_{\mathbf{0.0}}$	$70.8_{\mathbf{0.0}}$	$59.3_{\mathbf{0.0}}$	$39.6_{\mathbf{0.0}}$	$76.3_{\mathbf{0.0}}$	$66.5_{\mathbf{0.0}}$	$66.35_{\mathbf{0.00}}$
Qwen3-14B	FP16	—	60.8	84.2	89.3	78.8	67.8	46.4	79.8	73.0	72.51
	GPTQ	Ord.	$60.8_{0.7}$	$83.9_{0.2}$	$88.9_{0.3}$	$78.1_{0.2}$	$67.3_{0.3}$	$45.9_{0.4}$	$79.8_{0.4}$	$72.5_{0.6}$	$72.14_{0.17}$
		Samp.	$60.3_{0.8}$	$83.7_{0.6}$	$88.7_{0.3}$	$78.2_{0.2}$	$67.2_{0.6}$	$45.5_{0.8}$	$79.6_{0.3}$	$72.4_{0.6}$	$71.95_{0.26}$
		Neu.	$61.0_{\mathbf{0.0}}$	$84.4_{\mathbf{0.0}}$	$89.2_{\mathbf{0.0}}$	$78.1_{\mathbf{0.0}}$	$67.5_{\mathbf{0.0}}$	$46.2_{\mathbf{0.0}}$	$79.4_{\mathbf{0.0}}$	$71.7_{\mathbf{0.0}}$	$72.20_{\mathbf{0.00}}$
	GPTAQ	Ord.	$60.3_{0.7}$	$83.7_{0.5}$	$88.5_{0.4}$	$78.0_{0.1}$	$66.8_{0.3}$	$46.2_{1.0}$	$79.9_{0.3}$	$72.3_{0.7}$	$71.96_{0.30}$
		Samp.	$59.7_{0.5}$	$83.6_{0.5}$	$88.5_{0.8}$	$78.0_{0.2}$	$66.6_{0.4}$	$46.0_{0.5}$	$79.9_{0.3}$	$71.8_{0.7}$	$71.77_{0.14}$
		Neu.	$60.4_{\mathbf{0.0}}$	$84.5_{\mathbf{0.0}}$	$88.3_{\mathbf{0.0}}$	$78.0_{\mathbf{0.0}}$	$67.0_{\mathbf{0.0}}$	$45.8_{\mathbf{0.0}}$	$79.6_{\mathbf{0.0}}$	$72.0_{\mathbf{0.0}}$	$71.94_{\mathbf{0.00}}$
	BoA	Ord.	$60.7_{0.7}$	$83.9_{0.5}$	$89.2_{0.2}$	$76.2_{0.1}$	$67.1_{0.3}$	$44.8_{0.7}$	$78.6_{0.3}$	$71.8_{0.4}$	$71.53_{0.18}$
		Samp.	$60.5_{0.5}$	$84.0_{0.4}$	$89.1_{0.2}$	$76.2_{0.2}$	$67.8_{0.5}$	$44.1_{0.6}$	$79.0_{0.4}$	$72.3_{0.6}$	$71.63_{0.18}$
		Neu.	$61.5_{0.2}$	$83.9_{0.1}$	$89.0_{0.1}$	$76.3_{\mathbf{0.0}}$	$67.1_{\mathbf{0.0}}$	$45.1_{0.3}$	$79.0_{0.1}$	$72.0_{\mathbf{0.0}}$	$71.75_{0.06}$
	TurboBoA	Ord.	$59.3_{1.0}$	$83.2_{0.3}$	$89.0_{0.4}$	$77.2_{0.2}$	$65.0_{0.7}$	$44.5_{0.6}$	$78.9_{0.3}$	$71.8_{0.7}$	$71.11_{0.12}$
		Samp.	$59.4_{0.7}$	$83.3_{0.4}$	$89.0_{0.3}$	$77.1_{0.2}$	$65.4_{0.5}$	$44.8_{0.8}$	$78.9_{0.4}$	$72.1_{0.8}$	$71.26_{0.26}$
		Neu.	$59.0_{\mathbf{0.0}}$	$83.4_{\mathbf{0.0}}$	$89.2_{\mathbf{0.0}}$	$77.1_{\mathbf{0.0}}$	$65.3_{\mathbf{0.0}}$	$45.2_{\mathbf{0.0}}$	$79.0_{\mathbf{0.0}}$	$71.7_{\mathbf{0.0}}$	$71.23_{\mathbf{0.00}}$

Table 11: **INT4 per-task results for Llama.**

Model	Method	EXP	ARC-c	ARC-e	BoolQ	HellaS.	LAMB.	OBQA	PIQA	Wino.	AVG
Llama-3.2-1B	FP16	—	37.4	66.2	63.6	64.2	62.3	37.2	74.9	60.5	58.28
	GPTQ	Ord.	$35.2_{0.8}$	$63.7_{0.7}$	$61.1_{1.9}$	$61.4_{0.2}$	$58.7_{1.1}$	$36.3_{1.0}$	$73.8_{0.2}$	$60.5_{0.4}$	$56.32_{0.24}$
		Samp.	$34.8_{0.9}$	$63.6_{1.0}$	$61.2_{2.4}$	$61.4_{0.3}$	$58.9_{0.9}$	$36.7_{0.9}$	$73.8_{0.7}$	$60.8_{0.7}$	$56.39_{0.54}$
		Neu.	$35.9_{\mathbf{0.0}}$	$63.6_{\mathbf{0.0}}$	$61.7_{\mathbf{0.0}}$	$61.7_{\mathbf{0.0}}$	$60.5_{\mathbf{0.0}}$	$35.8_{\mathbf{0.0}}$	$73.5_{\mathbf{0.0}}$	$60.1_{\mathbf{0.0}}$	$56.60_{\mathbf{0.00}}$
	GPTAQ	Ord.	$34.6_{0.3}$	$64.4_{0.8}$	$62.6_{1.8}$	$61.7_{0.3}$	$57.6_{0.9}$	$35.6_{0.7}$	$73.5_{0.5}$	$60.4_{0.8}$	$56.30_{0.33}$
		Samp.	$34.6_{0.8}$	$64.2_{0.9}$	$62.9_{1.2}$	$62.0_{0.2}$	$57.8_{0.7}$	$36.1_{0.7}$	$73.6_{0.7}$	$60.3_{0.8}$	$56.44_{0.33}$
		Neu.	$34.3_{\mathbf{0.0}}$	$64.3_{\mathbf{0.0}}$	$63.5_{\mathbf{0.0}}$	$62.0_{\mathbf{0.0}}$	$58.3_{\mathbf{0.0}}$	$35.6_{\mathbf{0.0}}$	$72.8_{\mathbf{0.0}}$	$61.2_{\mathbf{0.0}}$	$56.49_{\mathbf{0.00}}$
	BoA	Ord.	$36.2_{0.7}$	$65.3_{0.7}$	$61.7_{0.9}$	$61.4_{0.2}$	$57.8_{0.7}$	$36.3_{0.4}$	$74.0_{0.4}$	$59.6_{0.5}$	$56.55_{0.25}$
		Samp.	$35.6_{0.4}$	$64.9_{0.7}$	$62.0_{1.3}$	$61.2_{0.2}$	$57.5_{0.5}$	$36.6_{1.1}$	$73.9_{0.5}$	$59.8_{0.5}$	$56.43_{0.40}$
		Neu.	$35.2_{\mathbf{0.0}}$	$65.6_{\mathbf{0.0}}$	$61.5_{\mathbf{0.0}}$	$61.6_{\mathbf{0.0}}$	$56.9_{\mathbf{0.0}}$	$36.2_{\mathbf{0.0}}$	$73.7_{\mathbf{0.0}}$	$59.7_{\mathbf{0.0}}$	$56.31_{\mathbf{0.00}}$
	TurboBoA	Ord.	$35.6_{1.0}$	$64.9_{0.7}$	$61.6_{0.7}$	$61.1_{0.2}$	$57.2_{0.8}$	$36.7_{0.8}$	$74.1_{0.2}$	$59.5_{0.6}$	$56.32_{0.32}$
		Samp.	$35.5_{0.9}$	$65.5_{0.6}$	$62.8_{0.5}$	$60.8_{0.2}$	$56.7_{0.9}$	$36.8_{0.9}$	$74.0_{0.4}$	$59.8_{0.4}$	$56.50_{0.28}$
		Neu.	$36.0_{\mathbf{0.0}}$	$65.2_{\mathbf{0.0}}$	$62.6_{\mathbf{0.0}}$	$61.3_{\mathbf{0.0}}$	$57.1_{\mathbf{0.0}}$	$35.8_{\mathbf{0.0}}$	$74.0_{\mathbf{0.0}}$	$59.9_{\mathbf{0.0}}$	$56.47_{\mathbf{0.00}}$
Llama-3.1-8B	FP16	—	54.9	82.3	82.9	79.3	74.7	45.8	80.9	73.5	71.77
	GPTQ	Ord.	$52.7_{0.7}$	$80.7_{0.4}$	$81.4_{0.4}$	$77.8_{0.2}$	$73.5_{0.4}$	$43.8_{0.4}$	$80.0_{0.3}$	$73.1_{0.7}$	$70.39_{0.18}$
		Samp.	$52.6_{0.9}$	$80.5_{0.6}$	$81.6_{0.7}$	$77.9_{0.2}$	$73.6_{0.6}$	$44.2_{0.6}$	$79.6_{0.2}$	$73.5_{0.7}$	$70.43_{0.18}$
		Neu.	$53.0_{\mathbf{0.0}}$	$80.6_{\mathbf{0.0}}$	$81.1_{\mathbf{0.0}}$	$77.4_{\mathbf{0.0}}$	$73.0_{\mathbf{0.0}}$	$44.2_{\mathbf{0.0}}$	$80.2_{\mathbf{0.0}}$	$71.6_{\mathbf{0.0}}$	$70.14_{\mathbf{0.00}}$
	GPTAQ	Ord.	$52.4_{0.8}$	$80.7_{0.9}$	$81.8_{0.6}$	$78.2_{0.2}$	$73.4_{0.6}$	$44.5_{0.9}$	$79.6_{0.6}$	$72.8_{0.6}$	$70.42_{0.32}$
		Samp.	$53.1_{0.8}$	$80.9_{0.7}$	$81.6_{0.6}$	$78.2_{0.2}$	$73.7_{0.4}$	$44.2_{0.9}$	$80.2_{0.5}$	$73.0_{0.4}$	$70.61_{0.24}$
		Neu.	$53.2_{\mathbf{0.0}}$	$81.2_{\mathbf{0.0}}$	$80.3_{\mathbf{0.0}}$	$78.2_{\mathbf{0.0}}$	$73.5_{\mathbf{0.0}}$	$46.6_{\mathbf{0.0}}$	$80.5_{\mathbf{0.0}}$	$72.8_{\mathbf{0.0}}$	$70.79_{\mathbf{0.00}}$
	BoA	Ord.	$53.3_{0.6}$	$80.9_{0.4}$	$82.1_{0.5}$	$77.5_{0.1}$	$73.5_{0.4}$	$43.8_{0.7}$	$79.7_{0.3}$	$73.5_{0.5}$	$70.54_{0.10}$
		Samp.	$53.1_{0.5}$	$80.9_{0.5}$	$81.8_{0.3}$	$77.7_{0.2}$	$73.8_{0.4}$	$43.0_{0.5}$	$79.8_{0.3}$	$73.5_{0.7}$	$70.46_{0.21}$
		Neu.	$53.2_{0.1}$	$80.3_{0.1}$	$82.0_{\mathbf{0.0}}$	$77.7_{\mathbf{0.0}}$	$73.4_{\mathbf{0.0}}$	$44.2_{0.3}$	$79.7_{0.2}$	$74.2_{0.1}$	$70.59_{0.07}$
	TurboBoA	Ord.	$52.9_{0.6}$	$80.9_{0.3}$	$81.3_{0.6}$	$77.5_{0.2}$	$73.3_{0.3}$	$43.8_{0.6}$	$79.6_{0.3}$	$73.8_{0.6}$	$70.38_{0.18}$
		Samp.	$53.5_{0.7}$	$81.0_{0.5}$	$81.3_{0.3}$	$77.6_{0.1}$	$73.4_{0.4}$	$43.6_{0.6}$	$79.7_{0.3}$	$73.2_{0.4}$	$70.42_{0.22}$
		Neu.	$54.1_{\mathbf{0.0}}$	$80.9_{\mathbf{0.0}}$	$81.0_{\mathbf{0.0}}$	$77.3_{\mathbf{0.0}}$	$73.0_{\mathbf{0.0}}$	$43.6_{\mathbf{0.0}}$	$79.8_{\mathbf{0.0}}$	$73.5_{\mathbf{0.0}}$	$70.40_{\mathbf{0.00}}$
Llama-2-13B	FP16	—	48.8	79.0	82.2	79.6	76.5	45.0	80.4	72.5	70.50
	GPTQ	Ord.	$48.5_{0.6}$	$78.1_{0.2}$	$81.7_{0.3}$	$78.8_{0.2}$	$76.1_{0.2}$	$45.4_{0.4}$	$80.2_{0.2}$	$71.9_{0.3}$	$70.10_{0.17}$
		Samp.	$48.4_{0.6}$	$78.4_{0.3}$	$82.2_{0.4}$	$78.8_{0.1}$	$76.3_{0.2}$	$46.1_{0.5}$	$80.2_{0.3}$	$72.3_{0.3}$	$70.33_{0.11}$
		Neu.	$48.2_{\mathbf{0.0}}$	$77.7_{\mathbf{0.0}}$	$81.7_{\mathbf{0.0}}$	$78.8_{\mathbf{0.0}}$	$76.2_{\mathbf{0.0}}$	$45.8_{\mathbf{0.0}}$	$80.0_{\mathbf{0.0}}$	$72.5_{\mathbf{0.0}}$	$70.09_{\mathbf{0.00}}$
	GPTAQ	Ord.	$48.7_{0.6}$	$78.6_{0.3}$	$81.2_{1.3}$	$79.0_{0.1}$	$76.4_{0.3}$	$45.6_{0.6}$	$80.2_{0.2}$	$72.7_{0.6}$	$70.32_{0.20}$
		Samp.	$48.2_{0.5}$	$78.6_{0.4}$	$81.2_{0.6}$	$79.0_{0.1}$	$76.4_{0.3}$	$45.2_{0.7}$	$80.4_{0.2}$	$72.4_{0.6}$	$70.16_{0.11}$
		Neu.	$49.5_{\mathbf{0.0}}$	$78.7_{\mathbf{0.0}}$	$81.8_{\mathbf{0.0}}$	$78.9_{\mathbf{0.0}}$	$75.9_{\mathbf{0.0}}$	$46.0_{\mathbf{0.0}}$	$80.3_{\mathbf{0.0}}$	$72.1_{\mathbf{0.0}}$	$70.41_{\mathbf{0.00}}$
	BoA	Ord.	$48.4_{0.5}$	$77.4_{0.3}$	$75.9_{0.4}$	$77.0_{0.2}$	$75.2_{0.2}$	$38.5_{0.9}$	$79.8_{0.2}$	$64.3_{0.5}$	$67.07_{0.22}$
		Samp.	$47.6_{0.8}$	$76.7_{0.8}$	$78.3_{0.8}$	$77.4_{0.4}$	$75.7_{0.6}$	$39.4_{1.4}$	$79.7_{0.3}$	$65.5_{2.2}$	$67.53_{0.74}$
		Neu.	$48.0_{0.2}$	$76.8_{0.2}$	$76.7_{0.4}$	$77.0_{0.1}$	$74.9_{0.2}$	$39.5_{0.4}$	$79.4_{0.2}$	$63.5_{0.2}$	$66.98_{0.11}$
	TurboBoA	Ord.	$48.6_{0.3}$	$77.7_{0.3}$	$79.2_{0.8}$	$78.3_{0.1}$	$76.5_{0.2}$	$42.8_{0.7}$	$80.2_{0.2}$	$69.9_{0.3}$	$69.16_{0.12}$
		Samp.	$48.9_{0.5}$	$77.7_{0.3}$	$80.0_{0.7}$	$78.3_{0.1}$	$76.5_{0.3}$	$43.8_{1.3}$	$80.1_{0.2}$	$70.5_{0.8}$	$69.48_{0.29}$
		Neu.	$48.5_{\mathbf{0.0}}$	$78.3_{\mathbf{0.0}}$	$79.6_{\mathbf{0.0}}$	$78.3_{\mathbf{0.0}}$	$76.1_{\mathbf{0.0}}$	$43.6_{\mathbf{0.0}}$	$80.5_{\mathbf{0.0}}$	$69.9_{\mathbf{0.0}}$	$69.33_{\mathbf{0.00}}$

Table 12: **INT3 per-task results for Qwen3.**

Model	Method	EXP	ARC-c	ARC-e	BoolQ	HellaS.	LAMB.	OBQA	PIQA	Wino.	AVG
Qwen3-0.6B	FP16	—	34.2	60.6	64.0	47.3	39.9	31.4	67.5	56.4	50.17
	GPTQ	Ord.	$26.7_{0.8}$	$37.8_{1.2}$	$50.3_{3.9}$	$36.4_{0.2}$	$14.6_{0.7}$	$27.5_{0.8}$	$58.4_{0.9}$	$50.6_{1.0}$	$37.78_{0.58}$
		Samp.	$26.5_{0.9}$	$36.0_{1.3}$	$47.6_{7.0}$	$36.0_{0.3}$	$14.9_{1.1}$	$27.6_{1.1}$	$57.6_{0.6}$	$51.5_{0.8}$	$37.21_{0.95}$
		Neu.	$\mathbf{25.2_{0.0}}$	$\mathbf{35.7_{0.0}}$	$\mathbf{52.5_{0.0}}$	$\mathbf{36.1_{0.0}}$	$\mathbf{15.4_{0.0}}$	$\mathbf{27.0_{0.0}}$	$\mathbf{57.5_{0.0}}$	$\mathbf{48.8_{0.0}}$	$\mathbf{37.26_{0.00}}$
	GPTAQ	Ord.	$26.1_{0.9}$	$39.9_{1.4}$	$60.2_{2.5}$	$38.2_{0.2}$	$22.4_{0.9}$	$28.9_{1.4}$	$59.8_{0.7}$	$52.4_{1.0}$	$40.99_{0.34}$
		Samp.	$25.6_{0.6}$	$40.4_{2.4}$	$60.3_{1.9}$	$38.0_{0.4}$	$23.2_{1.8}$	$28.5_{1.0}$	$59.7_{0.6}$	$52.8_{1.0}$	$41.06_{0.65}$
		Neu.	$\mathbf{25.9_{0.0}}$	$\mathbf{40.2_{0.0}}$	$\mathbf{62.9_{0.0}}$	$\mathbf{38.0_{0.0}}$	$\mathbf{21.5_{0.0}}$	$\mathbf{29.4_{0.0}}$	$\mathbf{59.6_{0.0}}$	$\mathbf{52.3_{0.0}}$	$\mathbf{41.22_{0.00}}$
	BoA	Ord.	$24.0_{0.8}$	$34.3_{0.7}$	$59.9_{2.4}$	$32.6_{0.3}$	$13.9_{0.8}$	$26.5_{1.0}$	$56.3_{0.8}$	$50.6_{1.1}$	$37.26_{0.27}$
		Samp.	$24.0_{0.9}$	$34.7_{1.5}$	$58.1_{3.1}$	$32.6_{0.3}$	$14.7_{1.1}$	$27.2_{1.4}$	$55.8_{1.0}$	$50.9_{1.2}$	$37.24_{0.60}$
		Neu.	$\mathbf{22.2_{0.0}}$	$\mathbf{37.2_{0.0}}$	$\mathbf{55.5_{0.0}}$	$\mathbf{32.1_{0.0}}$	$\mathbf{14.2_{0.0}}$	$\mathbf{27.6_{0.0}}$	$\mathbf{56.2_{0.0}}$	$\mathbf{52.1_{0.0}}$	$\mathbf{37.13_{0.00}}$
	TurboBoA	Ord.	$24.7_{0.7}$	$37.4_{1.3}$	$52.6_{7.4}$	$35.9_{0.3}$	$14.2_{3.0}$	$28.3_{1.2}$	$58.9_{0.6}$	$52.4_{1.2}$	$38.05_{1.42}$
		Samp.	$23.9_{0.7}$	$37.4_{1.7}$	$57.8_{5.8}$	$35.8_{0.5}$	$16.3_{3.5}$	$27.9_{1.0}$	$58.7_{0.8}$	$52.6_{1.0}$	$38.80_{1.01}$
		Neu.	$\mathbf{25.0_{0.0}}$	$\mathbf{36.4_{0.0}}$	$\mathbf{61.5_{0.0}}$	$\mathbf{35.9_{0.0}}$	$\mathbf{16.5_{0.0}}$	$\mathbf{28.8_{0.0}}$	$\mathbf{58.5_{0.0}}$	$\mathbf{53.0_{0.0}}$	$\mathbf{39.46_{0.00}}$
Qwen3-8B	FP16	—	56.5	83.4	86.6	74.8	64.3	41.6	77.6	67.8	69.08
	GPTQ	Ord.	$47.8_{1.1}$	$74.4_{1.5}$	$82.7_{0.7}$	$68.8_{0.2}$	$57.7_{0.7}$	$39.1_{0.7}$	$74.7_{0.5}$	$63.6_{0.9}$	$63.59_{0.39}$
		Samp.	$48.1_{1.1}$	$74.7_{1.5}$	$82.6_{1.4}$	$68.9_{0.3}$	$57.0_{1.0}$	$38.8_{1.1}$	$74.3_{1.0}$	$64.2_{0.6}$	$63.58_{0.45}$
		Neu.	$\mathbf{46.7_{0.0}}$	$\mathbf{72.6_{0.0}}$	$\mathbf{83.4_{0.0}}$	$\mathbf{68.9_{0.0}}$	$\mathbf{58.2_{0.0}}$	$\mathbf{37.8_{0.0}}$	$\mathbf{74.7_{0.0}}$	$\mathbf{65.0_{0.0}}$	$\mathbf{63.42_{0.00}}$
	GPTAQ	Ord.	$45.8_{1.7}$	$73.8_{1.9}$	$83.6_{0.8}$	$69.5_{0.2}$	$55.8_{0.8}$	$38.4_{1.4}$	$74.5_{0.4}$	$64.8_{0.6}$	$63.27_{0.51}$
		Samp.	$46.9_{1.6}$	$75.1_{2.0}$	$82.3_{1.6}$	$69.4_{0.3}$	$56.0_{1.1}$	$38.3_{0.7}$	$75.0_{0.5}$	$64.9_{1.2}$	$63.49_{0.48}$
		Neu.	$\mathbf{46.8_{0.0}}$	$\mathbf{72.6_{0.0}}$	$\mathbf{83.2_{0.0}}$	$\mathbf{69.7_{0.0}}$	$\mathbf{57.3_{0.0}}$	$\mathbf{39.2_{0.0}}$	$\mathbf{75.0_{0.0}}$	$\mathbf{64.5_{0.0}}$	$\mathbf{63.54_{0.00}}$
	BoA	Ord.	$38.1_{0.7}$	$61.0_{3.3}$	$82.2_{0.4}$	$56.2_{0.2}$	$36.1_{1.3}$	$34.0_{1.1}$	$68.8_{0.4}$	$56.7_{0.7}$	$54.14_{0.58}$
		Samp.	$37.7_{0.8}$	$62.6_{2.5}$	$82.4_{0.4}$	$56.1_{0.5}$	$35.8_{1.9}$	$34.0_{1.3}$	$69.4_{0.6}$	$56.8_{1.0}$	$54.36_{0.52}$
		Neu.	$38.7_{0.3}$	$57.5_{0.1}$	$\mathbf{83.5_{0.0}}$	$\mathbf{56.4_{0.0}}$	$35.6_{0.2}$	$\mathbf{33.6_{0.0}}$	$68.8_{0.1}$	$56.8_{0.1}$	$53.87_{0.02}$
	TurboBoA	Ord.	$47.3_{1.2}$	$73.2_{1.7}$	$83.9_{1.9}$	$63.8_{0.4}$	$47.4_{2.6}$	$39.8_{0.6}$	$73.2_{0.6}$	$65.5_{0.8}$	$61.75_{0.66}$
		Samp.	$47.3_{1.4}$	$73.0_{1.7}$	$84.1_{0.7}$	$64.5_{0.6}$	$49.0_{2.8}$	$39.0_{1.3}$	$73.4_{0.8}$	$65.5_{1.1}$	$61.98_{0.64}$
		Neu.	$\mathbf{46.8_{0.0}}$	$\mathbf{75.1_{0.0}}$	$\mathbf{84.6_{0.0}}$	$\mathbf{63.9_{0.0}}$	$\mathbf{49.8_{0.0}}$	$\mathbf{41.8_{0.0}}$	$\mathbf{73.1_{0.0}}$	$\mathbf{66.4_{0.0}}$	$\mathbf{62.68_{0.00}}$
Qwen3-14B	FP16	—	60.8	84.2	89.3	78.8	67.8	46.4	79.8	73.0	72.51
	GPTQ	Ord.	$53.3_{1.7}$	$77.9_{1.3}$	$87.3_{0.9}$	$75.1_{0.2}$	$61.5_{1.0}$	$42.8_{0.9}$	$77.4_{0.8}$	$69.6_{1.1}$	$68.09_{0.56}$
		Samp.	$52.6_{1.5}$	$78.0_{1.1}$	$87.1_{0.6}$	$75.2_{0.2}$	$61.2_{0.9}$	$43.3_{1.0}$	$77.4_{0.5}$	$69.7_{1.0}$	$68.06_{0.39}$
		Neu.	$\mathbf{53.3_{0.0}}$	$\mathbf{79.4_{0.0}}$	$\mathbf{88.5_{0.0}}$	$\mathbf{75.1_{0.0}}$	$\mathbf{60.6_{0.0}}$	$\mathbf{43.4_{0.0}}$	$\mathbf{78.1_{0.0}}$	$\mathbf{69.6_{0.0}}$	$\mathbf{68.51_{0.00}}$
	GPTAQ	Ord.	$54.1_{1.7}$	$79.0_{1.3}$	$85.9_{1.6}$	$75.1_{0.3}$	$60.9_{0.8}$	$43.2_{0.6}$	$77.9_{0.7}$	$70.0_{0.9}$	$68.26_{0.47}$
		Samp.	$53.9_{1.4}$	$79.1_{1.5}$	$86.8_{0.9}$	$74.9_{0.2}$	$61.6_{0.7}$	$43.5_{1.1}$	$77.7_{0.4}$	$69.9_{0.8}$	$68.43_{0.50}$
		Neu.	$\mathbf{53.4_{0.0}}$	$\mathbf{77.9_{0.0}}$	$\mathbf{86.9_{0.0}}$	$\mathbf{75.3_{0.0}}$	$\mathbf{60.4_{0.0}}$	$\mathbf{45.2_{0.0}}$	$\mathbf{77.1_{0.0}}$	$\mathbf{69.3_{0.0}}$	$\mathbf{68.19_{0.00}}$
	BoA	Ord.	$53.8_{1.3}$	$79.3_{0.9}$	$87.2_{0.4}$	$69.1_{0.2}$	$62.3_{1.0}$	$41.3_{0.8}$	$76.2_{0.3}$	$69.0_{0.9}$	$67.27_{0.26}$
		Samp.	$53.7_{0.9}$	$79.6_{0.5}$	$87.4_{0.4}$	$69.3_{0.2}$	$61.5_{1.1}$	$41.2_{0.8}$	$76.7_{0.3}$	$68.3_{0.9}$	$67.23_{0.16}$
		Neu.	$\mathbf{52.4_{0.0}}$	$\mathbf{78.0_{0.0}}$	$\mathbf{87.6_{0.0}}$	$\mathbf{69.1_{0.0}}$	$\mathbf{62.3_{0.0}}$	$\mathbf{42.0_{0.0}}$	$\mathbf{75.8_{0.0}}$	$\mathbf{69.4_{0.0}}$	$\mathbf{67.08_{0.00}}$
	TurboBoA	Ord.	$54.9_{1.6}$	$79.6_{1.1}$	$87.5_{0.7}$	$71.2_{0.4}$	$59.1_{1.8}$	$42.6_{0.7}$	$77.2_{0.7}$	$70.6_{0.9}$	$67.82_{0.48}$
		Samp.	$55.3_{1.4}$	$80.6_{0.7}$	$87.9_{0.7}$	$71.5_{0.3}$	$59.2_{1.8}$	$42.8_{1.1}$	$77.0_{0.4}$	$70.2_{0.8}$	$68.06_{0.50}$
		Neu.	$\mathbf{56.1_{0.0}}$	$\mathbf{79.8_{0.0}}$	$\mathbf{88.4_{0.0}}$	$\mathbf{71.8_{0.0}}$	$\mathbf{58.2_{0.0}}$	$\mathbf{41.6_{0.0}}$	$\mathbf{77.2_{0.0}}$	$\mathbf{71.2_{0.0}}$	$\mathbf{68.03_{0.00}}$

Table 13: **INT3 per-task results for Llama.**

Model	Method	EXP	ARC-c	ARC-e	BoolQ	HellaS.	LAMB.	OBQA	PIQA	Wino.	AVG
Llama-3.2-1B	FP16	—	37.4	66.2	63.6	64.2	62.3	37.2	74.9	60.5	58.28
	GPTQ	Ord.	$27.0_{1.0}$	$47.0_{1.9}$	$55.3_{4.6}$	$46.3_{0.9}$	$29.3_{2.5}$	$30.1_{0.7}$	$61.9_{0.8}$	$55.1_{1.0}$	$43.99_{0.77}$
		Samp.	$27.9_{0.7}$	$49.3_{2.1}$	$55.6_{6.7}$	$47.7_{0.8}$	$31.0_{2.4}$	$29.9_{1.5}$	$63.5_{1.2}$	$54.9_{1.2}$	$44.98_{1.02}$
		Neu.	$27.9_{\mathbf{0.0}}$	$49.8_{\mathbf{0.0}}$	$47.2_{\mathbf{0.0}}$	$46.3_{\mathbf{0.0}}$	$30.0_{\mathbf{0.0}}$	$28.8_{\mathbf{0.0}}$	$61.6_{\mathbf{0.0}}$	$54.4_{\mathbf{0.0}}$	$43.27_{\mathbf{0.00}}$
	GPTAQ	Ord.	$31.0_{1.2}$	$55.8_{1.6}$	$56.5_{3.4}$	$53.1_{0.6}$	$37.1_{2.0}$	$33.7_{0.8}$	$67.5_{0.9}$	$56.0_{1.0}$	$48.85_{0.75}$
		Samp.	$29.9_{0.7}$	$54.3_{1.3}$	$54.0_{6.7}$	$53.3_{0.4}$	$36.7_{1.4}$	$33.8_{1.3}$	$67.0_{0.6}$	$57.1_{1.2}$	$48.27_{0.91}$
		Neu.	$30.0_{\mathbf{0.0}}$	$55.1_{\mathbf{0.0}}$	$60.2_{\mathbf{0.0}}$	$52.7_{\mathbf{0.0}}$	$37.7_{\mathbf{0.0}}$	$33.0_{\mathbf{0.0}}$	$66.9_{\mathbf{0.0}}$	$55.9_{\mathbf{0.0}}$	$48.95_{\mathbf{0.00}}$
	BoA	Ord.	$30.8_{0.8}$	$56.7_{1.2}$	$62.0_{1.1}$	$54.1_{0.2}$	$42.9_{0.9}$	$33.2_{0.9}$	$69.3_{0.5}$	$56.6_{0.9}$	$50.69_{0.29}$
		Samp.	$31.3_{0.6}$	$58.2_{1.0}$	$59.9_{4.6}$	$53.9_{0.3}$	$44.5_{1.4}$	$33.2_{1.3}$	$69.8_{0.6}$	$56.3_{0.8}$	$50.88_{0.69}$
		Neu.	$31.7_{\mathbf{0.0}}$	$56.7_{\mathbf{0.0}}$	$62.3_{\mathbf{0.0}}$	$53.8_{\mathbf{0.0}}$	$43.5_{\mathbf{0.0}}$	$33.2_{\mathbf{0.0}}$	$69.6_{\mathbf{0.0}}$	$56.0_{\mathbf{0.0}}$	$50.85_{\mathbf{0.00}}$
	TurboBoA	Ord.	$27.8_{0.3}$	$55.3_{0.9}$	$61.3_{1.0}$	$47.9_{0.4}$	$28.8_{1.2}$	$31.1_{0.6}$	$63.8_{0.6}$	$55.4_{0.7}$	$46.43_{0.26}$
		Samp.	$29.0_{0.8}$	$57.2_{0.8}$	$60.2_{3.0}$	$50.2_{0.9}$	$33.3_{2.7}$	$31.8_{0.9}$	$66.1_{0.6}$	$55.8_{0.9}$	$47.94_{0.58}$
		Neu.	$28.4_{\mathbf{0.0}}$	$56.0_{\mathbf{0.0}}$	$61.6_{\mathbf{0.0}}$	$48.2_{\mathbf{0.0}}$	$30.6_{\mathbf{0.0}}$	$32.0_{\mathbf{0.0}}$	$64.1_{\mathbf{0.0}}$	$56.0_{\mathbf{0.0}}$	$47.12_{\mathbf{0.00}}$
Llama-3.1-8B	FP16	—	54.9	82.3	82.9	79.3	74.7	45.8	80.9	73.5	71.77
	GPTQ	Ord.	$42.2_{1.7}$	$70.7_{1.3}$	$72.6_{1.2}$	$71.1_{0.8}$	$62.2_{1.1}$	$40.1_{1.3}$	$75.5_{0.8}$	$68.7_{0.6}$	$62.90_{0.58}$
		Samp.	$42.7_{1.7}$	$72.4_{1.3}$	$72.8_{2.1}$	$71.1_{0.9}$	$62.3_{1.6}$	$40.1_{1.6}$	$76.0_{0.7}$	$68.9_{1.2}$	$63.28_{0.87}$
		Neu.	$41.6_{\mathbf{0.0}}$	$72.9_{\mathbf{0.0}}$	$73.9_{\mathbf{0.0}}$	$70.0_{\mathbf{0.0}}$	$60.8_{\mathbf{0.0}}$	$39.8_{\mathbf{0.0}}$	$76.1_{\mathbf{0.0}}$	$68.9_{\mathbf{0.0}}$	$63.00_{\mathbf{0.00}}$
	GPTAQ	Ord.	$44.4_{1.6}$	$73.6_{1.2}$	$75.8_{1.0}$	$73.2_{0.4}$	$64.8_{1.3}$	$41.4_{1.1}$	$77.5_{0.3}$	$70.2_{1.0}$	$65.12_{0.48}$
		Samp.	$44.3_{0.8}$	$74.4_{1.0}$	$73.1_{3.1}$	$73.5_{0.5}$	$64.7_{1.6}$	$41.3_{1.1}$	$77.3_{0.5}$	$70.2_{1.0}$	$64.84_{0.63}$
		Neu.	$44.6_{\mathbf{0.0}}$	$72.5_{\mathbf{0.0}}$	$77.4_{\mathbf{0.0}}$	$73.6_{\mathbf{0.0}}$	$64.0_{\mathbf{0.0}}$	$43.2_{\mathbf{0.0}}$	$77.1_{\mathbf{0.0}}$	$71.7_{\mathbf{0.0}}$	$65.51_{\mathbf{0.00}}$
	BoA	Ord.	$45.6_{0.9}$	$75.8_{0.8}$	$78.8_{1.1}$	$73.3_{0.2}$	$66.8_{0.9}$	$40.7_{0.7}$	$77.6_{0.4}$	$71.7_{0.6}$	$66.29_{0.32}$
		Samp.	$46.7_{1.4}$	$76.4_{1.1}$	$79.8_{0.8}$	$73.2_{0.4}$	$67.9_{0.6}$	$41.0_{0.8}$	$77.1_{0.5}$	$72.1_{0.4}$	$66.77_{0.47}$
		Neu.	$45.2_{0.2}$	$76.4_{0.1}$	$79.2_{0.1}$	$73.6_{\mathbf{0.0}}$	$68.0_{0.1}$	$42.1_{0.4}$	$78.0_{0.3}$	$71.5_{0.1}$	$66.75_{0.13}$
	TurboBoA	Ord.	$43.8_{1.1}$	$74.7_{1.2}$	$79.2_{0.7}$	$72.4_{0.2}$	$65.7_{0.5}$	$41.5_{1.2}$	$77.4_{0.5}$	$72.2_{0.4}$	$65.87_{0.37}$
		Samp.	$44.9_{1.3}$	$75.4_{0.9}$	$78.8_{1.1}$	$72.4_{0.4}$	$66.7_{1.0}$	$41.7_{0.9}$	$77.0_{0.7}$	$71.6_{0.7}$	$66.06_{0.31}$
		Neu.	$43.2_{\mathbf{0.0}}$	$75.1_{\mathbf{0.0}}$	$79.4_{\mathbf{0.0}}$	$72.3_{\mathbf{0.0}}$	$66.9_{\mathbf{0.0}}$	$41.6_{\mathbf{0.0}}$	$77.3_{\mathbf{0.0}}$	$70.6_{\mathbf{0.0}}$	$65.80_{\mathbf{0.00}}$
Llama-2-13B	FP16	—	48.8	79.0	82.2	79.6	76.5	45.0	80.4	72.5	70.50
	GPTQ	Ord.	$45.5_{0.9}$	$75.8_{0.8}$	$78.9_{0.6}$	$75.5_{0.3}$	$72.4_{0.5}$	$43.4_{1.3}$	$77.8_{0.5}$	$68.8_{0.8}$	$67.27_{0.28}$
		Samp.	$45.2_{0.8}$	$75.5_{0.7}$	$79.3_{1.0}$	$75.6_{0.2}$	$72.0_{0.8}$	$43.1_{1.1}$	$77.9_{0.5}$	$68.6_{0.9}$	$67.16_{0.17}$
		Neu.	$45.6_{\mathbf{0.0}}$	$76.0_{\mathbf{0.0}}$	$78.4_{\mathbf{0.0}}$	$76.1_{\mathbf{0.0}}$	$72.8_{\mathbf{0.0}}$	$42.8_{\mathbf{0.0}}$	$77.4_{\mathbf{0.0}}$	$68.9_{\mathbf{0.0}}$	$67.25_{\mathbf{0.00}}$
	GPTAQ	Ord.	$45.6_{0.5}$	$76.4_{0.5}$	$79.5_{0.7}$	$76.5_{0.2}$	$73.7_{0.5}$	$44.0_{1.1}$	$79.2_{0.4}$	$70.7_{0.4}$	$68.21_{0.12}$
		Samp.	$46.2_{1.0}$	$76.6_{0.6}$	$79.4_{1.1}$	$76.7_{0.3}$	$73.6_{0.4}$	$44.2_{1.0}$	$78.9_{0.4}$	$70.1_{0.5}$	$68.23_{0.28}$
		Neu.	$45.6_{\mathbf{0.0}}$	$76.3_{\mathbf{0.0}}$	$79.9_{\mathbf{0.0}}$	$76.0_{\mathbf{0.0}}$	$73.5_{\mathbf{0.0}}$	$45.0_{\mathbf{0.0}}$	$78.6_{\mathbf{0.0}}$	$71.0_{\mathbf{0.0}}$	$68.25_{\mathbf{0.00}}$
	BoA	Ord.	$44.5_{1.1}$	$75.1_{0.9}$	$77.0_{0.8}$	$74.2_{0.2}$	$74.1_{0.3}$	$37.2_{0.7}$	$78.0_{0.3}$	$61.2_{1.3}$	$65.17_{0.43}$
		Samp.	$42.5_{2.1}$	$72.5_{2.6}$	$76.9_{1.3}$	$73.2_{1.0}$	$72.5_{1.2}$	$34.9_{2.3}$	$76.7_{1.4}$	$60.5_{2.7}$	$63.69_{1.43}$
		Neu.	$45.3_{0.5}$	$75.9_{0.4}$	$76.9_{0.6}$	$74.5_{0.2}$	$74.5_{0.2}$	$37.6_{0.2}$	$78.3_{0.2}$	$61.8_{0.5}$	$65.59_{0.13}$
	TurboBoA	Ord.	$43.7_{0.6}$	$74.5_{0.6}$	$77.9_{1.0}$	$74.2_{0.3}$	$74.4_{0.3}$	$40.5_{1.2}$	$77.4_{0.4}$	$68.4_{0.6}$	$66.38_{0.24}$
		Samp.	$43.6_{1.5}$	$74.0_{1.3}$	$78.2_{1.7}$	$73.2_{2.6}$	$74.5_{1.0}$	$39.2_{1.3}$	$77.4_{0.5}$	$67.2_{1.6}$	$65.90_{1.04}$
		Neu.	$44.1_{\mathbf{0.0}}$	$73.8_{\mathbf{0.0}}$	$76.2_{0.1}$	$74.3_{\mathbf{0.0}}$	$74.2_{\mathbf{0.0}}$	$40.3_{0.2}$	$77.2_{0.1}$	$69.2_{0.3}$	$66.16_{0.04}$