Sensitivity-Guided Mixed-Precision Quantization for Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Post-training quantization (PTQ) is essential for deploying large language models (LLMs) on resource-constrained hardware, but its impact on specific capabilities is poorly characterized. We present a capacity-controlled study of how weight-only quantization affects multi-step reasoning versus factual recall and commonsense inference, across six models (0.5B–8B; Qwen2.5, Qwen3, and Phi-3.5 families). We introduce the Reasoning Fragility Ratio (RFR), a diagnostic metric that quantifies capability-asymmetric degradation, and SGMPQ, a mixed-precision scheme that assigns 8-bit precision to the layers ranked most quantization-sensitive by a perturbation probe, with a capability-conditioned ratio as the ranking signal. Our evaluation uses real benchmarks (GSM8K, ARC-Challenge, HellaSwag, held-out perplexity) with Wilson confidence intervals and exact McNemar tests, and compares SGMPQ 8/4 (≈5.2 average bits) against capacity-matched controls: uniform W5/W6, a random mixed-precision control at the identical budget (replicated over five allocation seeds), and—decisively—two capability-agnostic sensitivity rankings at the same budget, plus GPTQ and AWQ at matched budgets. Three findings emerge. First, the quantizer dominates where it matters: GPTQ beats RTN by up to 63 accuracy points exactly where RTN breaks (W4 on the 3B base and Coder-3B models), rescuing Qwen2.5-Coder-3B, which RTN destroys outright, while the two quantizers converge on robust models at W5. Second, on robust models the bit budget masks placement: at equal ≈5.2 bits, sensitivity-guided placement is statistically indistinguishable from random placement on the two small instruct models, worse on the 3B base model, and better on the two largest models—though with five allocation seeds the permutation design (sign-flip floor p=0.0625) cannot reach conventional significance even if the advantage is real on every seed. Third, placement matters exactly where uniform quantization fails: on Coder-3B, both perturbation-probe-guided allocations tested preserve the model (59.4–64.2% GSM8K) where all five independently drawn random allocations collapse (0.4–2.4%; uniform W5 also rescues it), and the capability ratio itself is beneficial on one model (8B: +6.0 pp) but actively harmful on another (3B: −33.4 pp vs. the agnostic ranking). We release per-item results, sensitivity maps, and the full evaluation stack, and conclude that sensitivity-guided mixed precision is a targeted repair tool for quantization-brittle models—but no ranking signal we tested (capability ratio or agnostic) transfers safely across models, so any allocation must be validated against random-placement controls on the target model.

1 Introduction

Large language models are increasingly deployed under tight memory and compute budgets, and weight-only post-training quantization (PTQ) is the standard response. Modern PTQ methods—GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), OmniQuant (Shao et al., 2024)—achieve near-lossless compression on language-modeling benchmarks, but perplexity

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.

averages can hide capability-specific damage: a model may retain fluent factual recall while losing the ability to carry out multi-step reasoning.

This paper asks a targeted question: when quantization does damage a model, is the damage capability-asymmetric, and can precision be allocated to protect reasoning specifically? Multi-step reasoning requires precise intermediate computations across many token positions, whereas factual recall can often be retrieved with a single associative lookup; this structural difference motivates measuring the two separately and allocating precision accordingly.

We make three contributions:

1. Measurement. We profile how reasoning, factual recall, and commonsense degrade across models spanning 0.5B–8B at bit-widths W2–W4 and W8 (small-scale sweeps; Appendix B); the Reasoning Fragility Ratio (RFR) serves as a diagnostic summary of capability-asymmetric degradation (with the zero-denominator case handled explicitly), reported alongside raw accuracies throughout.
2. Method. We propose SGMPQ (Sensitivity-Guided Mixed-Precision Quantization), which ranks transformer blocks by a perturbation sensitivity probe—how much each block's removal from full precision shifts the model's representation on reasoning versus factual calibration text—and assigns 8-bit precision to the top 30% of layers, 4-bit to the rest (≈5.2 average bits). The ranking can be capability-conditioned (the ratio R_i) or capability-agnostic (pooled sensitivity); we evaluate both and find neither transfers safely across models (Sections 5.2–5.3).
3. Capacity-controlled evaluation. Because a higher average bit budget alone improves accuracy, all SGMPQ comparisons are made against capacity-matched controls: uniform W5 (5.0 bits), a Mixed-Random control that gives 8-bit to the same number of layers drawn uniformly at random (five independent allocation seeds), capability-agnostic sensitivity rankings at the identical budget, and GPTQ and AWQ at matched budgets. The evaluation uses 500-item GSM8K, 200-item ARC-Challenge, 200-item HellaSwag, and held-out perplexity, with Wilson 95% CIs, exact McNemar tests, and Holm correction.

The study yields a clear division of labor among three levers. The quantizer is the dominant lever in the regime where RTN fails: GPTQ repairs catastrophic W4 failures (+63.0 points on Coder-3B, +35.2 on the 3B base model) and rescues models that RTN destroys. The bit budget is the dominant lever on robust models, where moving from 4 to 5 average bits recovers most of the gap and makes sensitivity-guided placement statistically indistinguishable from random placement. Placement—SGMPQ's contribution—is decisive in a specific but practically important regime: models whose weight distributions make them brittle under uniform quantization. On Qwen2.5-Coder-3B, both perturbation-probe-guided assignments we test keep the model intact (59.4–64.2% GSM8K) where all five identically-budgeted random allocations collapse it ($\leq$ 2.4%)—but the capability-conditioned ratio is not the operative ingredient: a capability-agnostic ranking shares the rescue, and on the 3B base model the ratio actively harms (−33.4 pp vs. the agnostic ranking). We characterize precisely for which models placement helps, hurts, or is neutral, and show that sensitivity analysis is best understood as a complement to strong quantization arithmetic rather than a substitute for it.

2 Related Work

Post-training quantization. GPTQ (Frantar et al., 2023) uses approximate second-order information to quantize weights column-by-column with error compensation. AWQ (Lin et al., 2024) identifies salient weight channels from activation magnitudes. OmniQuant (Shao et al., 2024) learns clipping boundaries and supports mixed precision. SmoothQuant (Xiao et al., 2023) migrates quantization difficulty from activations to weights, and LLM.int8() (Dettmers et al., 2022) preserves outlier features via mixed-precision decomposition; compact formats combine low bit-widths with outlier handling (Dettmers et al., 2023b). LLM-QAT (Liu et al., 2024) extends quantization-aware training to LLMs without

original data. These methods optimize perplexity or task-averaged accuracy and do not distinguish reasoning from factual capabilities; our study measures that distinction directly.

Mixed-precision quantization. Prior mixed-precision work assigns bit-widths per layer via hardware-aware reinforcement learning (HAQ; Wang et al., 2019), Hessian-based sensitivity (HAWQ-V1/2/3; Dong et al., 2019; 2020; Yao et al., 2021), or differentiable search. QLoRA (Dettmers et al., 2023a) uses 4-bit NF4 formats for memory-efficient fine-tuning. These approaches rank layers by criteria that are partly task-derived (loss curvature is measured on a task loss; hardware cost is not), though we note HAWQ-V2's sensitivity criterion is conditioned on downstream task loss—a superset of ours in awareness terms; our distinction is the explicit capability decomposition and the matched-budget attribution methodology, not awareness per se. Closest in spirit to our budget-level finding, Dettmers & Zettlemoyer (2023) show that at 4-bit precision zero-shot performance is governed by bit scaling laws across model sizes, anticipating the regime where the bit budget dominates placement (our Finding 2). Our contribution is therefore not the first capability-aware sensitivity signal, but (i) a direct measurement of the reasoning-versus-factual axis under matched budgets, (ii) the demonstration—via capability-agnostic controls—that on fragile models the rescue comes from sensitivity-guidance itself rather than capability-conditioning, and (iii) the capacity-matched negative-result methodology. SGMPQ derives its allocation from a perturbation probe with no search over allocations.

LLM reasoning. Chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022) and process reward models (Lightman et al., 2024) establish that LLM reasoning proceeds through sequential intermediate computations whose correctness compounds. This sequential structure is the mechanical motivation for expecting reasoning to be disproportionately sensitive to weight perturbation, and for expecting generation-heavy tasks to amplify small numerical damage into large accuracy differences.

3 Method

3.1 Reasoning Fragility Ratio (RFR)

For a model quantized to b bits, the RFR of a reasoning task r relative to a factual task f is

$$\mathrm{RFR}_r(b) \;=\; \frac{\Delta^r_{\mathrm{acc}}(b)}{\Delta^f_{\mathrm{acc}}(b)} \;=\; \frac{\mathrm{Acc}_r^{\mathrm{FP16}} - \mathrm{Acc}_r^b}{\mathrm{Acc}_f^{\mathrm{FP16}} - \mathrm{Acc}_f^b}. \tag{1}$$

$\mathrm{RFR} > 1$ indicates that reasoning degrades more than factual recall under the same quantization level. When factual accuracy is exactly preserved ($\Delta^f_{\mathrm{acc}} = 0$) while reasoning drops, the ratio is mathematically undefined; we report such cases as "div" (denominator zero) rather than "∞", and always report raw accuracies alongside RFR so the underlying drops can be verified.

3.2 Capability-Specific Sensitivity Analysis

For each transformer block i we compute a sensitivity score separately on reasoning-calibration and factual-calibration data, using two complementary probes.

Perturbation sensitivity (used by SGMPQ). We quantize block i alone to b_l bits (all other blocks FP16) and measure the relative change in the model's internal representation on capability-specific prompts: final-block hidden states are mean-pooled over token positions, and the relative L2 distance between the full-precision and perturbed representations is averaged over prompts,

$$S^i_{\mathrm{pert}}(\mathrm{cap}) = \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{cap}}}\left[\frac{\left\|\bar{h}(x) - \bar{h}_{\mathrm{quant}(i)}(x)\right\|_2}{\left\|\bar{h}(x)\right\|_2 + 10^{-8}}\right], \tag{2}$$

where $\bar{h}(x) \in \mathbb{R}^d$ is the mean-pooled last-hidden-state vector. This is a representation-space distance, not a divergence between next-token distributions; per-block scores are normalized to $[0, 1]$ by dividing by the maximum across blocks. We select this probe because

Require: model $\mathcal{M}$ with N blocks; reasoning set $\mathcal{D}_r$; factual set $\mathcal{D}_f$; high-precision ratio k=0.3; bit-widths b_h=8, b_l=4

Step 1: per-block sensitivity profiling. For each block $i = 1, \ldots, N$: quantize block i to b_l bits (others FP16); measure relative-L2 representation shift $S^i_{\text{pert}}(\text{cap})$ for cap $\in \{r, f\}$ (Eq. 2).

Step 2: capability sensitivity ratio. $R_i \leftarrow \frac{S^i_{\text{pert}}(\text{reasoning})+\epsilon}{S^i_{\text{pert}}(\text{factual})+\epsilon}$, normalized to $[0, 1]$.

Step 3: precision assignment. $\mathcal{H} \leftarrow$ top-$\lceil k \cdot N \rceil$ blocks by R_i (reasoning-critical); assign b_h bits to $i \in \mathcal{H}$, b_l bits otherwise; return quantized model $\mathcal{M}_q$.

Algorithm 1: SGMPQ: Sensitivity-Guided Mixed-Precision Quantization

it directly measures the quantity quantization perturbs—the block's contribution to the representation—rather than a property of the loss surface (Appendix B.3 compares it with a gradient-based alternative).

Gradient sensitivity (comparison). For completeness we also record, for each block, the L2 norm of the gradient flowing into the block's input activations under the causal-LM loss on capability-specific text, averaged over calibration batches:

$$S^i_{\text{grad}}(\text{cap}) = \mathbb{E}_{x \sim \mathcal{D}_{\text{cap}}} \left[\left\| \frac{\partial \mathcal{L}_{\text{cap}}}{\partial a_i} \right\|_2 \right]. \tag{3}$$

Calibration composition. The probe uses no task benchmarks. Each capability's pool consists of 20 hand-written seed texts (reasoning: grade-school arithmetic word problems, logical-deduction puzzles, and multi-step chains; factual: short encyclopedic statements of common facts), replicated and shuffled with a fixed seed to yield 64 samples per capability; the perturbation measurement uses 32 per capability (truncated to 256 tokens) and gradient sensitivity uses 64. Both pools and the resulting per-layer sensitivity files are released, so the exact probe inputs can be inspected.

The capability sensitivity ratio of block i is

$$R_i = \frac{S^i(\text{reasoning}) + \epsilon}{S^i(\text{factual}) + \epsilon}, \tag{4}$$

normalized to $[0, 1]$ across blocks by dividing by the maximum ratio, with symmetric smoothing $\epsilon = 10^{-6}$ added to numerator and denominator.

3.3 SGMPQ: Sensitivity-Guided Mixed-Precision Quantization

SGMPQ assigns a two-level precision scheme (Algorithm 1):

- High precision (8-bit): the top-k layers by the ranking signal—by default the capability ratio R_i (reasoning-critical), with $k = \lceil 0.3 \times N \rceil$;
- Standard precision (4-bit): the remaining layers.

The first and last transformer blocks and the LM head stay at FP16 in all RTN-based configurations, and the GPTQ/AWQ arms quantize one additional linear per model (Section 3.4). We denote the configuration SGMPQ 8/4. Its average bit-width over quantized blocks is ≈ 5.2 bits (realized per-model averages 5.14–5.25; Appendix A, Table 1), which is higher than uniform W4 (4.0 bits). Because a higher budget alone can improve accuracy, every SGMPQ comparison in Section 5 is made against capacity-matched controls: uniform W5 (5.0 bits) and W6 (6.0 bits), a Mixed-Random control that assigns 8-bit to the same number of layers chosen uniformly at random (five independent seeds; Section 5.1), and capability-agnostic sensitivity rankings at the identical budget (Section 5.2), which isolate the contribution of capability-conditioning from that of sensitivity-guidance itself.

3.4 Quantization Implementation

Unless stated otherwise, quantization is round-to-nearest (RTN), per-group symmetric, group size 128; each transformer block's linear layers (attention projections and MLP) re-

ceive the same bit-width. We additionally implement GPTQ (Frantar et al., 2023) (sequential block-wise calibration, Hessian-based column updates, group size 128, 128 calibration sequences of length 512 drawn from a held-out split of Alpaca that excludes all perplexity-evaluation documents), both as uniform W4/W5 baselines and as a mixed-precision variant GPTQ-MP using SGMPQ's layer assignment. AWQ (Lin et al., 2024) is implemented with the same sequential per-block protocol, calibration corpus, and group size (scale search over a 20-point grid with duo-scaling, per-linear clipping search, asymmetric per-group quantization), in uniform W4/W5 and AWQ-MP variants. Per-model linear-layer counts differ by one between the RTN path (which skips one embedding-adjacent linear) and the GPTQ/AWQ path (e.g., 251 vs. 252 on the 3B-class models); spot-checks with and without the extra linear yield negligible deltas (Coder-3B W4: 6/500 both ways; Coder-3B W5: 54.0% vs. 54.8%; 8B W5: 88.8% vs. 89.0%—both deltas ≤ 0.8 pp, n.s., and the 251-variant reproduces the paper's uniform-W5 entries), confirming the gap is immaterial; both counts are recorded in the released run manifests.

4 Experimental Setup

Models. Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct, Qwen2.5-3B (base), Qwen2.5-Coder-3B-Instruct (Qwen Team, 2024), Qwen3-8B-Base (Qwen Team, 2025), and Phi-3.5-mini-instruct (3.8B; cross-family control) (Abdin et al., 2024).

Benchmarks.

- GSM8K (Cobbe et al., 2021): 500 problems from the official test split (deterministic seed-1234 subsample); greedy decoding, strict numeric-equivalence scoring of the final answer (marker-preferred extraction; tolerance $\max(10^{-4}, 10^{-3}|\text{gold}|)$).[1]
- ARC-Challenge (Clark et al., 2018): 200 items from the official test split (seed-1234 subsample); the model generates up to 16 tokens and the prediction is the leftmost standalone letter A–D in the response.
- HellaSwag (Zellers et al., 2019): 200 validation items (seed-1234 subsample); endings ranked by length-normalized continuation log-likelihood.[2]
- Perplexity: 100 held-out Alpaca (Taori et al., 2023) documents (excluded from all GPTQ/AWQ calibration), token-weighted cross-entropy over 512-token windows. Alpaca is a non-standard PPL corpus: these values serve as internal anchors across configurations of the same model and are not comparable to WikiText-2 literature values.

Accuracy metrics carry Wilson 95% CIs; paired comparisons use exact McNemar tests on per-item outcomes with bootstrap CIs on accuracy differences. Multiplicity control: the twelve primary McNemar tests (Appendix A, Table 2) are Holm-corrected within each model; six secondary per-model comparisons (Table 5, Appendix D) are Holm-corrected within their own per-model family. Claims resting on raw p-values in 0.02–0.05 that do not survive Holm correction are labeled marginal. Paired significance testing is performed on GSM8K; ARC-Challenge and HellaSwag agree directionally with the significant GSM8K contrasts (Appendix D) . Configurations. The degradation study spans W2A16–W8A16 (weight-only, activations FP16).[3] The capacity-matched comparison evaluates fourteen configurations per model: FP16; uniform W4/W5/W6; SGMPQ 8/4 (RTN); Mixed-Random (RTN, random assignment, five seeds); capability-agnostic sensitivity rankings (three models; Section 5.2); GPTQ W4, GPTQ W5, GPTQ-MP (SGMPQ assignment); AWQ W4, AWQ W5, AWQ-MP (SGMPQ assignment).

[1] GSM8K uses a zero-shot protocol whose instruction ends with "The answer is X" for strict extraction; absolute accuracies are therefore not directly comparable to leaderboard GSM8K numbers, but all comparisons in this paper use this fixed protocol.

[2] Our normalizer divides by token count of the continuation; the official harness normalizes by byte length. Absolute numbers may differ slightly from leaderboard values.

[3] The degradation study evaluates W2/W3/W4 on all six models and W8 on five of six (all but Phi-3.5).

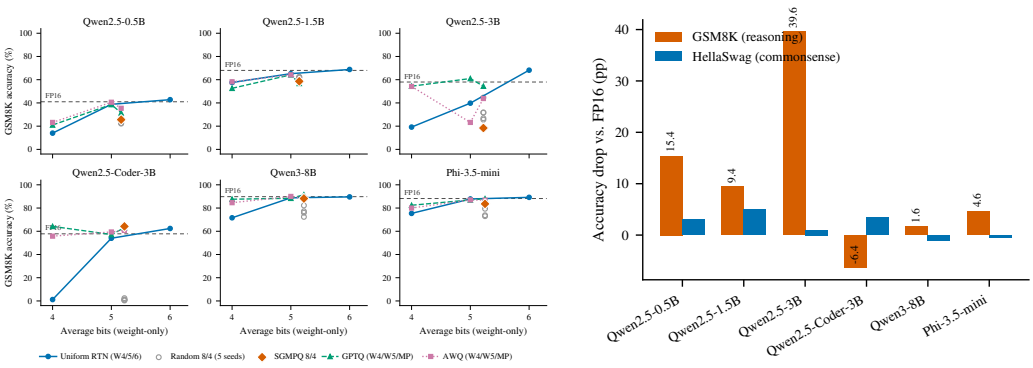

Figure 1: Left: GSM8K accuracy vs. average bits (capacity curves; the diamond is SGMPQ 8/4, open grey circles are the five Mixed-Random seeds). Right: reasoning (GSM8K) vs. commonsense (HellaSwag) degradation under SGMPQ 8/4 relative to FP16; most models lose more reasoning than commonsense at equal bit cost.

5 Results

We organize the results around three levers—quantizer, bit budget, and layer placement—and then examine the regime where placement is decisive.

Finding 1: Placement is indistinguishable from random on robust models, and budget explains why. At the matched ≈5.2-bit budget, SGMPQ 8/4 does not beat uniform W5 on the four robust models (0.5B: 25.6 vs. 38.8; 1.5B: 58.6 vs. 65.2; 3B: 18.4 vs. 39.8; Phi-3.5: 83.6 vs. 87.8 GSM8K). Against the Mixed-Random control at the identical bit distribution, SGMPQ is statistically indistinguishable on 0.5B (p=0.32), 1.5B (p=0.21), and Phi-3.5 (p=1.0), significantly worse on the 3B base model (−13.2 pp, Holm $p=1.1 \times 10^{-8}$), and significantly better on Coder-3B (+61.8 pp, Holm $p=6.9 \times 10^{-85}$) and 8B (+12.2 pp, Holm $p=6.9 \times 10^{-10}$) (Appendix A, Table 2). Section 5.1 replicates these single-draw contrasts with five allocation seeds per model and confirms every direction; Section 5.2 separates sensitivity-guidance from capability-conditioning.

Finding 2: The quantizer is the dominant lever where RTN fails. At 4.0 bits, GPTQ over RTN gains +35.2 pp GSM8K on the 3B base model (19.2→54.4), +63.0 pp on Coder-3B (1.2→64.2), +16.0 pp on 8B, and +7.0 pp on Phi-3.5, while reducing perplexity toward FP16 levels (Table 6, Appendix E). On the four robust models at W5 the two quantizers converge (deltas within 1 pp, n.s.), with two significant exceptions: GPTQ is better on the fragile 3B at W5 (+21.2 pp, $p=6.9 \times 10^{-15}$) and worse on 1.5B at W4 (−5.0 pp, p=0.049). Comparing GPTQ-W4 (4.0 bits) against RTN-based SGMPQ 8/4 (5.2 bits) per model: an exact tie on Coder-3B, statistical ties on Phi-3.5 and 8B, a large GPTQ win on 3B (+36.0 pp, $p=8 \times 10^{-35}$), and significant GPTQ losses on the two smallest instruct models (−4.6/−6.0 pp, both surviving Holm; Table 5). Second-order rounding is thus a stronger lever than layer selection under RTN in the collapsing regime, but at 4 bits it does not uniformly substitute for 1.2 extra bits.

Finding 3: The quantizer and the assignment compose selectively. GPTQ-MP (SGMPQ)—GPTQ quantization with SGMPQ's 8/4 assignment—is the numerically strongest configuration on 8B at the 500-item subsample (91.4 vs. FP16 89.8, above uniform W6's 89.6 at a lower budget), though its advantages over GPTQ-W5 and uniform W5 are marginal after Holm correction (Table 5; at the full split GPTQ-MP and GPTQ-W5 are indistinguishable, Limitations). At the budget-matched comparison (GPTQ-MP at 5.2 bits vs. uniform GPTQ-W5 at 5.0 bits), composition is worse on the three smallest or weakest models (0.5B: 31.4 vs. 38.6, Holm p=0.004; 1.5B: 56.8 vs. 64.2, Holm p=0.003; 3B: 54.4 vs. 61.0, p=0.010), better on Coder-3B (63.0 vs. 57.2, Holm p=0.031), and tied on Phi-3.5. Composition pays off reliably only on Coder-3B—the same fragile-model condition identified below.

Finding 4: On the fragile model, placement is decisive—but not capability-specific. On Coder-3B, uniform W4 RTN destroys the model (1.2% GSM8K, PPL 355), and SGMPQ 8/4 restores it to 64.2% (PPL 5.45)—equal to GPTQ-W4 at a strictly lower budget than uniform W5 (54.0). The five-seed random control (Section 5.1) shows this rescue is a property of the assignment, not the budget: all five independently drawn random allocations collapse (0.4–2.4%, mean 1.2 ± 0.8, PPL 319–469). The capability-agnostic control (Section 5.2) sharpens the attribution further: agnostic sensitivity rankings rescue the model nearly as well (59.4%), so the decisive factor is sensitivity-guidance itself, with the capability ratio contributing a small, statistically unconfirmed margin.

AWQ baseline. A second activation-aware quantizer reproduces both patterns (Table 8, Appendix G). AWQ behaves like GPTQ, not like RTN: at W4 it avoids the catastrophic RTN failures (Coder-3B: 55.8 vs. RTN's 1.2; 3B: 54.2 vs. 19.2), and on four of six models AWQ-W4 and GPTQ-W4 are statistically indistinguishable ($p \geq 0.086$). The SGMPQ assignment transfers across quantizers: AWQ-MP broadly tracks GPTQ-MP, and beats its RTN counterpart on 0.5B (+9.8 pp, p=3.4×10^{-5}) and 3B (+25.4 pp, p=5.8×10^{-22}). The budget story is unchanged—AWQ-W5 matches or beats AWQ-MP on five of six models—with one anomaly: on the 3B base model AWQ-W5 drops sharply (23.2 vs. AWQ-W4's 54.2, despite normal perplexity and ARC-C), and AWQ-MP is better by +20.6 pp (p=1.8×10^{-16}); coherent generations with near-random arithmetic point to an instability of AWQ's zero-point treatment at 5 bits. A four-calibration-seed replication confirms the drop is systematic but strongly calibration-dependent (GSM8K 23.2–49.8 across draws; Appendix I).

5.1 Five-Seed Replication of the Random Control

Because the Mixed-Random contrast above rests on one random allocation per model, we replicate it with five random assignments per model (seeds 0–4), each satisfying the same 30% 8-bit capacity constraint at the block level and evaluated on the full benchmark suite (Table 7, Appendix F).[4]

The replication confirms and sharpens the single-seed picture (Figure 1, left). On Coder-3B, all five random allocations collapse (GSM8K 0.4–2.4%, mean 1.2 ± 0.8), so SGMPQ's advantage over the random distribution is +63.0 pp. On the 3B base model, the reversal is confirmed at the distribution level: the random mean is 32.1 ± 7.5 vs. SGMPQ's 18.4 (−13.7 pp), and seed 2 reaches 44.6%. On 8B, the random mean is 79.3 ± 6.3 vs. SGMPQ 88.2 (+8.9 pp); one of five seeds is statistically indistinguishable from SGMPQ (p=0.88), so placement helps on average but is not overwhelming. On Phi-3.5, SGMPQ (83.6%) exceeds all five draws (range 73.0–83.4%, mean 78.3 ± 4.7), significantly for two ($p \leq 3.5 \times 10^{-6}$) and marginally for a third (p=0.031, n.s. after a ×5 multiplicity correction), for an average advantage of +5.3 pp. On the two small instruct models the comparison is neutral (0.5B: random mean 24.1 ± 1.9 with SGMPQ at 25.6 inside the range, min p=0.11; 1.5B: −2.4 pp vs. the random mean).

In sum, the value of sensitivity-guided placement is size- and fragility-dependent: a large rescue on the fragile Coder-3B, a harmful reversal on the mid-size 3B base model, moderate benefits on the two largest models tested (8B: +8.9 pp; Phi-3.5: +5.3 pp over the random means), and neutrality on the small instruct models. With only five allocation seeds, the exact paired permutation test over seed-level deltas (sign-flip support 2^5=32, floor p=0.0625) bounds what this design can establish: the Coder-3B rescue is qualitatively unambiguous (all five seeds collapse; SGMPQ beats every seed), while the 8B and Phi-3.5 advantages (p=0.125 and p=0.0625) do not establish distribution-level significance at α=0.05.

5.2 Capability-Agnostic Placement Control

The preceding contrasts establish that sensitivity-guided placement beats random placement on fragile models, but not that capability-conditioning is what drives the rescue: any signal concentrating 8-bit precision on quantization-sensitive layers could suffice. We therefore

[4]Seeds 3–4 on 8B and Coder-3B realize one fewer 8-bit linear than SGMPQ (76 vs. 77 of 251; their draws include block 0, whose first linear is skipped); block-level average bits are identical (5.22).

run two capability-agnostic rankings at the identical 30%/8-4 budget on the three models where placement matters most: Agnostic-Sum (top-30% by combined sensitivity S_r+S_f) and Agnostic-R (top-30% by reasoning sensitivity S_r alone, no ratio), reusing the same probe measurements as SGMPQ (Table 9, Appendix H).

Three results follow (McNemar tests in Table 10). (i) The Coder-3B rescue is not capability-specific. Both agnostic rankings avoid the collapse that kills all five random allocations (59.4% vs. SGMPQ's 64.2%; +4.8 pp for the ratio, raw p=0.043, n.s. after Holm correction). The ranking sets barely overlap—2 of 11 eight-bit layers shared—so the rescue does not hinge on one canonical layer set: what matters is protecting sufficiently many of the most quantization-sensitive layers, not reasoning-specificity. (On Coder-3B the two agnostic rankings select identical top-30% sets—one allocation; Appendix H.) (ii) The ratio can actively harm. On the 3B base model the ratio ranking is −33.4 pp worse than Agnostic-Sum (p=5.5 × 10^{-32}) and −13.7 pp worse than the random mean; the mechanism is quantified in Section 5.3. (iii) Where the ratio helps, the gain is real but modest. On 8B the ratio adds +6.0 pp over Agnostic-Sum (p=4.4 × 10^{-4}, surviving Holm), yet SGMPQ still does not beat uniform W5 (88.2 vs. 88.8). The operative ingredient behind the fragile-model rescue is thus sensitivity-guided placement, but the ranking signal itself is model-dependent in effect: the capability ratio harms 3B, and a replication finds the agnostic-sum ranking collapsing on Phi-3.5 (51.2% GSM8K vs. SGMPQ's 83.6% and a 78.3% random mean; paired McNemar p=1.9 × 10^{-33}, b/c=183/21) despite the most stable probe in the study (Section 5.3); a calibration-seed replication on 8B moves both arms together (88.2→85.6, 82.2→80.4), preserving the ratio's margin.

5.3 Probe Calibration-Pool Stability

The allocation conditions on a single calibration-pool draw, so we replicate the probe itself: K=5 independent calibration pools (order-permuted seed lists, seeded replicate-and-shuffle of the same 20 seed texts; 32 probe samples per pool), recomputing the full probe and top-30% set per pool. Across all six models, the mean Jaccard of each resampled top-30% set against the paper's ranking is high everywhere except the 3B base model (0.5B: 0.761; 1.5B: 0.513; 3B: 0.329; Coder-3B: 0.805; 8B: 0.672; Phi-3.5: 0.897). On Coder-3B the set is largely stable (0.805, ∼9/11 layers retained) and the ranking replicates across calibration resamples; the rescue is inferred to follow. On the 3B base model the ranking is unstable (0.329, ∼5–6/11 layers retained): the ratio signal does not separate from calibration noise, and forcing it into the allocation damages the model—explaining the reversal of Sections 5.1 and 5.2 as probe noise amplified by the ratio's division by small factual sensitivities. Phi-3.5 cautions that stability alone does not certify a signal: its probe is the most stable in the study (ratio ranking 0.897; the agnostic-sum ranking is in fact perfectly stable—identical top-30% set across all five pools), yet its agnostic-sum allocation collapses (Section 5.2), so a stable probe certifies reproducibility, not that the protected layers are the right ones. Two caveats: the pools vary shuffle and composition of a fixed 20-text set (subsampling noise, not sensitivity to a wholly different pool), and rank overlap is a proxy; quantifying outcome spread across resampled pools is a natural next step. The practical pre-deployment check is thus to verify ranking stability across calibration resamples before trusting any capability-conditioned allocation.

6 Analysis

6.1 Budget, Quantizer, Placement: An Empirical Ordering

Synthesizing the capacity-matched evaluation, three levers determine post-quantization reasoning quality, in decreasing order of effect size on our testbed:

1. The quantizer. Moving from RTN to GPTQ at fixed budget helps enormously exactly where RTN breaks (+63.0 pp on Coder-3B at W4, +16.0 pp on 8B; Table 6); on the four robust models at W5 the quantizers converge (deltas within 1 pp, n.s.). Wherever uniform quantization causes catastrophic damage, the quantizer is the single largest lever in the study.

2. The bit budget. At fixed quantizer, moving from 4 to 5 average bits recovers most of the remaining gap on robust models (uniform W5 vs. W4: +24.8 pp on 0.5B, +7.6 pp on 1.5B, +20.6 pp on 3B).
3. Layer selection. Conditional on budget and quantizer, sensitivity-guided placement matters decisively for fragile models (Coder-3B: +61.8 pp over random under RTN; all five random seeds collapse), is statistically indistinguishable from random on robust models, and is harmful on the 3B base model (−13.7 pp vs. the five-seed random mean). The capability-agnostic control attributes the fragile-model rescue to sensitivity-guidance rather than capability-conditioning, and the ratio's margin over the agnostic ranking flips sign across models (+6.0 pp on 8B, −33.4 pp on 3B).

The practical reading is that sensitivity-guided placement is a conditional lever: not a first-order lever for models that tolerate uniform quantization, but the only placement lever that keeps brittle models functional at a given budget. The capability ratio is a refinement whose sign is model-dependent, and the 3B case warns against a robust/fragile taxonomy: it is neither robust (uniform W4 costs it 39 points) nor rescued by guidance, because its probe is unstable; whether placement helps a fragile-looking model depends on probe stability (Section 5.3), not fragility alone.

6.2 Why Is Reasoning Sometimes More Fragile?

Two regularities recur. First, at matched bit cost, reasoning degrades more than commonsense for most models (Figure 1, right): under SGMPQ 8/4, GSM8K drops exceed HellaSwag drops on five of six models, widening at lower bit-widths. Part of this asymmetry is confounded with evaluation format: GSM8K is generative, HellaSwag likelihood-ranked, so we measure a property of the capability-plus-format pair, not of reasoning alone. Second, the asymmetry is governed by overall fragility: the models that lose reasoning disproportionately (Coder-3B, the 3B base model at W4) are precisely those whose perplexity explodes.

Direct measurement supports a representation-level account. On real prompts (50 reasoning + 50 factual per condition), the reasoning/factual ratio of hidden-state drift under W4 is mildly above 1 on 0.5B (1.17), 1.5B (1.10), Phi-3.5 (1.28), and 8B (1.21), and mildly below 1 on 3B (0.83) and Coder-3B (0.94); attention-divergence ratios cluster around 1.0. Quantization damage at W4 is thus close to capability-neutral in representation space; large behavioral asymmetries arise where damage crosses a model's viability threshold, past which generation-heavy tasks collapse wholesale. Consistent with this, on 60 real GSM8K problems the fragile Coder-3B at W4 abandons chains early (generation length contracts from 1016 to 380 characters; arithmetic-bearing tokens vanish from 100% to 6.7% of responses) rather than producing subtly wrong arithmetic. Separately, W2A16 collapses every model (perplexity $> 10^6$) and only the largest survives W3 (Appendix B).

6.3 Coder-3B: Anatomy of a Fragile Model

Qwen2.5-Coder-3B shares its parameter count with the base 3B model yet behaves qualitatively differently: uniform W4 RTN collapses it (1.2% GSM8K, PPL 355 vs. 5.51 at FP16), random 8/4 allocation collapses it in five of five draws, and its reasoning chains abandon arithmetic early (Section 6.2). Two independent repairs restore it: protecting the layers that SGMPQ's sensitivity probe identifies (64.2%), or compensating errors within them via GPTQ (64.2% at W4). That both repairs work—while random placement fails in every draw—localizes the damage to quantization-sensitive layers: protecting enough of them suffices, since the ratio and agnostic rankings share only 2 of 11 eight-bit layers yet both rescue. Our best hypothesis: code training induces outlier-rich weight distributions that RTN cannot represent, so code-specialized models require error-compensating quantizers or sensitivity-aware placement.

6.4 Limitations

SGMPQ's ranking comes from one perturbation metric at one calibration scale; alternative signals are a promising direction for further gains (Appendix B.3), and the capability-

conditioned ranking is itself calibration-sensitive (stable on Coder-3B, unstable on 3B; Section 5.3). GSM8K/ARC/HellaSwag leave code, long-context, and multilingual capabilities untested—including Coder-3B's specialist domain—and inferential claims rest on GSM8K and inherit its format confound (Section 6.2). The 500-item GSM8K estimates replicate on the full 1319-item split within 0.4 pp (8B: 88.4%, Phi-3.5: 83.5%); the same run's GPTQ arms (8B: GPTQ-MP 90.3%, GPTQ-W5 89.8%) are statistically indistinguishable (McNemar p=0.55; Finding 3 rescoped to the subsample); the five-seed design caps distribution-level evidence at permutation $p \geq 0.0625$; contrasts are Holm-corrected within model; perplexity uses 100 held-out documents without dispersion. Our AWQ implementation measures weight-selection logic rather than kernel speedups (OmniQuant omitted), and profiling the memory/latency/kernel overhead of mixed-precision layouts is a natural next step. The central negative result spans six models from three families; broader evidence across model families would strengthen its generality.

7 Conclusion

We presented a capacity-controlled quantization study (six LLMs, 0.5B–8B). First, the quantizer matters most where quantization fails: GPTQ beats RTN by up to 63 points. Second, on robust models, bit budget masks placement: guided allocation matches random and harms 3B. Third, for fragile models, placement is decisive, and sensitivity-guidance—not capability-conditioning—is the operative ingredient: both probe-guided allocations preserve Coder-3B; all five random draws collapse. But no ranking is a safe default: validate any allocation against random-placement controls on the target model.

Reproducibility Statement

All experiments use publicly released model checkpoints (Qwen2.5, Qwen3, Phi-3.5) and public benchmarks (GSM8K, ARC-Challenge, HellaSwag, Alpaca). Benchmark subsets are deterministic seed-1234 draws from the official splits; random-allocation seeds are enumerated in Section 5.1, and the capability-agnostic control arms, probe-resampling pools, and the AWQ calibration-seed replication are specified in Sections 5.2–5.3 and Appendix I. We release per-item predictions for every configuration, per-layer sensitivity files with the exact calibration pools, the quantization and evaluation code, and scripts that regenerate every table and figure from the released artifacts. The code will be made publicly available upon acceptance.

Ethics Statement

This study evaluates publicly available open-weight models on public benchmarks and involves no human subjects or personal data. Quantization changes model behavior; deployments that adopt aggressive quantization should re-validate capability retention on their own workloads, particularly for reasoning-heavy applications where silent degradation could mislead users. Our results are intended to make such degradation measurable and preventable, not to enable covert capability reduction.

AI Usage Statement

AI tools were used for code development assistance, manuscript editing, and literature search. All experimental design, execution, and quantitative analysis were performed by the authors, who take full responsibility for the work.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018.

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

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. In International Conference on Machine Learning (ICML), 2023.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In Advances in Neural Information Processing Systems (NeurIPS), 2023a.

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, 2023b.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

Zhen Dong, Zhewei Yao, Yaohui Cai, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. In Advances in Neural Information Processing Systems (NeurIPS), 2020.

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

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In International Conference on Learning Representations (ICLR), 2024.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. In Conference on Machine Learning and Systems (MLSys), 2024.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. Llm-qat: Data-free quantization aware training for large language models. In Findings of the Association for Computational Linguistics: ACL 2024, 2024.

Qwen Team. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

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

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In International Conference on Machine Learning (ICML), 2023.

Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael W Mahoney, and Kurt Keutzer. Hawq-v3: Dyadic neural network quantization. In International Conference on Machine Learning (ICML), 2021.

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 (ACL), 2019.

A Main Results and Paired Significance Tests

Table 1 reports the full twelve-configuration grid; Table 2 reports the twelve primary paired McNemar tests referenced throughout Section 5.

Table 1: Main results: accuracy (%; Wilson 95% CI for GSM8K) and perplexity across twelve configurations and six models. Bits = average bits over quantized blocks.

Model	Config	Bits	GSM8K↑	ARC-C↑	HellaSwag↑	PPL↓
Qwen2.5-0.5B-Inst	FP16	–	41.0 [36.8, 45.4]	52.5	48.0	7.21
	Uniform W4	4.00	14.0 [11.2, 17.3]	40.0	42.5	11.15
	Uniform W5	5.00	38.8 [34.6, 43.1]	48.5	46.5	7.98
	Uniform W6	6.00	42.8 [38.5, 47.2]	49.0	46.0	7.56
	SGMPQ 8/4	5.17	25.6 [22.0, 29.6]	36.0	45.0	10.31
	Mixed-Random	5.17	23.4 [19.9, 27.3]	42.0	46.0	9.46
	GPTQ W4	4.00	21.0 [17.7, 24.8]	24.5	45.5	7.86
	GPTQ W5	5.00	38.6 [34.4, 42.9]	41.5	47.5	7.45
	GPTQ-MP (SGMPQ)	5.17	31.4 [27.5, 35.6]	40.0	48.0	7.73
	AWQ W4	4.00	23.2 [19.7, 27.1]	30.5	46.5	8.92
	AWQ W5	5.00	40.6 [36.4, 45.0]	46.0	47.5	7.66
	AWQ-MP (SGMPQ)	5.17	35.4 [31.3, 39.7]	42.5	45.0	8.23
Qwen2.5-3B (base)	FP16	–	58.0 [53.6, 62.2]	82.5	68.5	4.82
	Uniform W4	4.00	19.2 [16.0, 22.9]	68.5	66.0	6.74
	Uniform W5	5.00	39.8 [35.6, 44.2]	80.0	71.0	5.50
	Uniform W6	6.00	68.2 [64.0, 72.1]	82.5	70.5	4.92
	SGMPQ 8/4	5.22	18.4 [15.2, 22.0]	71.5	67.5	6.61
	Mixed-Random	5.22	31.6 [27.7, 35.8]	69.5	66.0	6.48
	GPTQ W4	4.00	54.4 [50.0, 58.7]	80.5	69.0	4.95
	GPTQ W5	5.00	61.0 [56.7, 65.2]	82.0	69.0	4.85
	GPTQ-MP (SGMPQ)	5.22	54.4 [50.0, 58.7]	82.5	68.5	4.96
	AWQ W4	4.00	54.2 [49.8, 58.5]	81.0	71.0	6.27
	AWQ W5	5.00	23.2 [19.7, 27.1]	81.5	69.0	5.95
	AWQ-MP (SGMPQ)	5.22	43.8 [39.5, 48.2]	83.0	70.0	6.11
Qwen3-8B (base)	FP16	–	89.8 [86.8, 92.2]	92.5	76.0	3.60
	Uniform W4	4.00	71.6 [67.5, 75.4]	92.0	74.5	5.11
	Uniform W5	5.00	88.8 [85.7, 91.3]	93.5	75.5	3.82
	Uniform W6	6.00	89.6 [86.6, 92.0]	93.0	75.0	3.60
	SGMPQ 8/4	5.22	88.2 [85.1, 90.7]	92.0	77.0	4.78
	Mixed-Random	5.22	76.0 [72.1, 79.5]	91.0	73.5	4.61
	GPTQ W4	4.00	87.6 [84.4, 90.2]	92.0	73.5	3.75
	GPTQ W5	5.00	88.4 [85.3, 90.9]	93.5	74.0	3.64
	GPTQ-MP (SGMPQ)	5.22	91.4 [88.6, 93.6]	92.5	76.5	3.68
	AWQ W4	4.00	84.6 [81.2, 87.5]	91.5	71.5	3.95
	AWQ W5	5.00	90.0 [87.1, 92.3]	93.0	75.5	3.58
	AWQ-MP (SGMPQ)	5.22	88.6 [85.5, 91.1]	94.0	74.0	3.77
Qwen2.5-Coder-3B	FP16	–	57.8 [53.4, 62.1]	79.0	67.5	5.51
	Uniform W4	4.00	1.2 [0.6, 2.6]	0.5	37.0	355.35
	Uniform W5	5.00	54.0 [49.6, 58.3]	77.5	68.5	5.51
	Uniform W6	6.00	62.4 [58.1, 66.5]	78.5	66.0	5.56
	SGMPQ 8/4	5.22	64.2 [59.9, 68.3]	74.5	64.0	5.45
	Mixed-Random	5.22	2.4 [1.4, 4.1]	0.0	33.5	444.44
	GPTQ W4	4.00	64.2 [59.9, 68.3]	74.5	62.0	5.72
	GPTQ W5	5.00	57.2 [52.8, 61.5]	76.0	67.5	5.51
	GPTQ-MP (SGMPQ)	5.22	63.0 [58.7, 67.1]	74.5	65.0	5.59
	AWQ W4	4.00	55.8 [51.4, 60.1]	76.0	65.5	5.68
	AWQ W5	5.00	59.4 [55.0, 63.6]	77.5	67.5	5.46
	AWQ-MP (SGMPQ)	5.22	60.8 [56.5, 65.0]	75.0	68.0	5.56
Qwen2.5-1.5B-Inst	FP16	–	68.0 [63.8, 71.9]	73.5	68.5	5.34
	Uniform W4	4.00	57.6 [53.2, 61.9]	69.0	65.5	7.67
	Uniform W5	5.00	65.2 [60.9, 69.2]	71.5	67.5	5.72
	Uniform W6	6.00	68.8 [64.6, 72.7]	72.5	68.5	5.80
	SGMPQ 8/4	5.14	58.6 [54.2, 62.8]	69.5	63.5	7.13
	Mixed-Random	5.14	61.8 [57.5, 66.0]	70.0	66.0	6.76
	GPTQ W4	4.00	52.6 [48.2, 56.9]	68.5	69.5	7.10
	GPTQ W5	5.00	64.2 [59.9, 68.3]	71.5	68.5	6.46
	GPTQ-MP (SGMPQ)	5.14	56.8 [52.4, 61.1]	70.0	65.0	7.06
	AWQ W4	4.00	58.2 [53.8, 62.4]	69.0	64.0	6.27
	AWQ W5	5.00	64.2 [59.9, 68.3]	74.5	65.5	5.60
	AWQ-MP (SGMPQ)	5.14	59.6 [55.2, 63.8]	70.5	67.5	6.12
Phi-3.5-mini	FP16	–	88.2 [85.1, 90.7]	85.5	74.5	4.93
	Uniform W4	4.00	75.4 [71.4, 79.0]	82.5	74.5	5.92
	Uniform W5	5.00	87.8 [84.6, 90.4]	83.0	76.5	5.27
	Uniform W6	6.00	89.2 [86.2, 91.6]	86.0	74.0	5.00
	SGMPQ 8/4	5.25	83.6 [80.1, 86.6]	83.0	75.0	5.71
	Mixed-Random	5.25	83.4 [79.9, 86.4]	84.0	75.5	5.13
	GPTQ W4	4.00	82.4 [78.8, 85.5]	82.5	75.5	5.10
	GPTQ W5	5.00	87.0 [83.8, 89.7]	84.0	75.5	4.94
	GPTQ-MP (SGMPQ)	5.25	88.0 [84.9, 90.6]	84.5	72.5	5.13
	AWQ W4	4.00	80.0 [76.3, 83.3]	82.0	74.0	5.17
	AWQ W5	5.00	87.0 [83.8, 89.7]	84.5	76.5	5.11
	AWQ-MP (SGMPQ)	5.25	86.8 [83.6, 89.5]	83.0	75.0	4.88

Table 2: Paired significance tests (GSM8K, exact McNemar). Discordant pairs b/c: cases where the first configuration is wrong but the second right / vice versa. ΔAcc in percentage points with bootstrap 95% CI. Holm p: step-down correction within each model's test family.

Model	Comparison	Discordant b/c	p (exact)	p (Holm)	ΔAcc [95% boot CI]
Qwen2.5-0.5B-Inst	SGMPQ vs. U-W4	30/88	8.55e-08	1.71e-07	+11.6 [+7.2, +15.6]
Qwen2.5-0.5B-Inst	SGMPQ vs. Random-MP	45/56	0.32	0.32	+2.2 [-1.8, +6.2]
Qwen2.5-3B (base)	SGMPQ vs. U-W4	53/49	0.767	0.767	-0.8 [-4.8, +3.4]
Qwen2.5-3B (base)	SGMPQ vs. Random-MP	98/32	5.61e-09	1.12e-08	-13.2 [-17.6, -8.6]
Qwen3-8B (base)	SGMPQ vs. U-W4	21/104	2.1e-14	4.19e-14	+16.6 [+12.6, +20.8]
Qwen3-8B (base)	SGMPQ vs. Random-MP	20/81	6.93e-10	6.93e-10	+12.2 [+8.4, +16.0]
Qwen2.5-Coder-3B	SGMPQ vs. U-W4	1/316	2.38e-93	4.76e-93	+63.0 [+58.6, +67.2]
Qwen2.5-Coder-3B	SGMPQ vs. Random-MP	6/315	6.92e-85	6.92e-85	+61.8 [+57.2, +66.2]
Qwen2.5-1.5B-Inst	SGMPQ vs. U-W4	58/63	0.716	0.716	+1.0 [-3.4, +5.4]
Qwen2.5-1.5B-Inst	SGMPQ vs. Random-MP	81/65	0.214	0.429	-3.2 [-7.8, +1.6]
Phi-3.5-mini	SGMPQ vs. U-W4	36/77	0.000144	0.000288	+8.2 [+4.2, +12.2]
Phi-3.5-mini	SGMPQ vs. Random-MP	38/39	1	1	+0.2 [-3.2, +3.6]

B Small-Scale Supporting Analyses

This appendix reports supporting analyses on hand-written evaluation suites (49–51 items per capability, 20-document perplexity); all headline claims in the main text rest on the real-benchmark evaluation of Section 5.

B.1 Degradation Profiling Across Bit-Widths

Table 3 profiles accuracy on hand-written math, factual, and logic suites, together with perplexity and RFR, across bit-widths. Three patterns hold across models: (i) W2A16 collapses every capability on every model (perplexity $> 10^6$), consistent with the phase transition below 3 bits; (ii) at W4A16, several models preserve factual accuracy exactly while losing 2–16% on math (RFR reported as "div"), the capability-asymmetric signature that motivates SGMPQ; on Phi-3.5 the W4 math drop coincides with a slight factual gain (negative RFR), a reminder that the ratio assumes a non-increasing factual baseline; (iii) robustness rises with scale—the 8B model survives W3 (44% math, PPL 9.06) where the 3B model collapses (PPL 10,779)—and code-trained weights are markedly more brittle (Coder-3B collapses at W4: PPL 102 vs. 4.12 at FP16).

Table 3: Hand-written-suite degradation profile: accuracy (%) and perplexity across bit-widths. RFR=div indicates factual accuracy unchanged while reasoning degrades. PPL values $\geq$10K abbreviated.

Model	Bits	Math	Factual	Logic	PPL	RFR_{math}	$\text{RFR}_{\text{logic}}$
0.5B	FP16	44	89.8	68.6	4.32	–	–
	W4A16	32	79.6	52.9	5.28	1.18	1.54
	W3A16	10	4.1	9.8	62.6	0.40	0.69
	W2A16	0	8.2	2.0	14.5M	0.54	0.82
1.5B	FP16	72	95.9	86.3	3.64	–	–
	W4A16	56	95.9	78.4	4.12	div	div
	W3A16	0	10.2	3.9	35.7	0.84	0.96
	W2A16	0	6.1	3.9	8.2M	0.80	0.92
3B	FP16	80	95.9	90.2	3.66	–	–
	W4A16	70	95.9	82.4	6.32	div	div
	W3A16	0	0	0	10.8K	0.83	0.94
	W2A16	2	10.2	2.0	2.2B	0.91	1.03
Coder-3B	FP16	68	93.9	84.3	4.12	–	–
	W4A16	2	6.1	5.9	102.4	0.75	0.89
	W3A16	0	4.1	3.9	2036	0.76	0.90
	W2A16	0	2.0	0	1.1M	0.74	0.92
8B	FP16	96	93.9	92.2	3.54	–	–
	W4A16	94	93.9	92.2	4.93	div	1.00
	W3A16	44	51.0	70.6	9.06	1.21	0.50
	W2A16	0	4.1	2.0	4.8M	1.07	1.00
Phi-3.5	FP16	80	93.9	88.2	3.21	–	–
	W4A16	32	95.9	78.4	4.45	$-23.5^{\dagger}$	$-4.8^{\dagger}$
	W3A16	20	36.7	35.3	15.59	1.05	0.93
	W2A16	0	10.2	2.0	25M	0.96	1.03

B.2 Sensitivity Maps

Figure 2 shows per-layer capability sensitivity ratios R_i (Eq. 2 of the main text) with the resulting SGMPQ 8/4 assignment for two representative models. Reasoning-critical layers concentrate in the middle third of the network, and the ratio varies several-fold across layers, giving the ranking enough dynamic range to be informative; the top-ranked layers for each model are listed in the released sensitivity files.

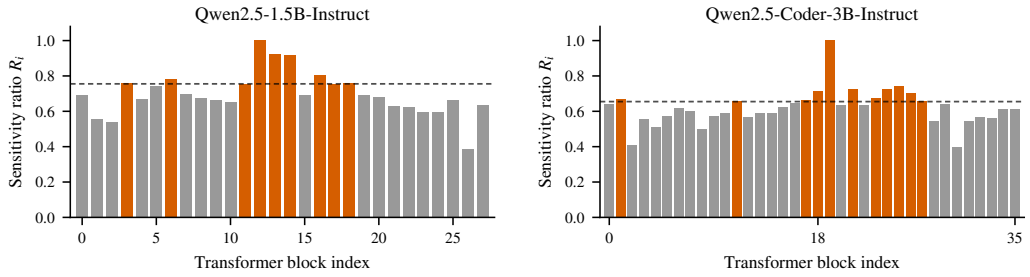

Figure 2: Per-layer capability sensitivity ratio R_i (bars; 8-bit layers highlighted) for Qwen2.5-1.5B-Instruct (left) and Qwen2.5-Coder-3B-Instruct (right). Layers above the dashed line receive 8-bit precision under SGMPQ 8/4.

B.3 Design-Choice Ablations

Table 4 summarizes three design-choice ablations on the hand-written suites.

High-precision ratio. Sweeping the 8-bit fraction from 10% to 50% shows that 30% is the sweet spot for the two models where placement matters most: Coder-3B exhibits a sharp threshold (2% math at 10–20% ratio vs. 70% at 30%), and the 3B base model peaks at 30% (86%). Very small models (0.5B) show no consistent benefit, and the 8B model is insensitive to the ratio (94% everywhere), consistent with its W4 robustness ceiling.

Assignment ablation. Comparing SGMPQ against five random assignments at the same 30% capacity on the hand-written suites foreshadows the scaled result: SGMPQ wins on Coder-3B (70% vs. a 26.8% random mean with three of five trials collapsed) and 3B, loses on 0.5B (18% vs. 26.8%), and ties on 8B.

Probe ablation. Perturbation-based sensitivity (Eq. 2) outperforms gradient-based ranking on the three larger models (e.g., Coder-3B: 70% vs. 2%—gradient ranking leaves the model collapsed), while the two probes tie on 8B and gradient ranking is mildly better on 0.5B. Directly measuring the representation shift caused by quantizing each block is the more faithful proxy for quantization damage, which is why SGMPQ uses it.

Table 4: Design-choice ablations on hand-written suites (math accuracy, %): high-precision ratio sweep, SGMPQ vs. five random assignments (random mean±std), and perturbation vs. gradient probe ranking.

	0.5B	1.5B	3B	Coder-3B	8B
Panel A: high-precision ratio sweep (8-bit fraction)					
10%	32	32	82	2	94
20%	28	32	80	2	94
30%	18	52	86	70	94
40%	20	62	80	68	94
50%	26	60	76	62	94
Panel B: SGMPQ vs. five random assignments (same 30% capacity)					
SGMPQ	18	52	86	70	94
Random mean±std	26.8±4.2	58.0±4.5	76.4±1.7	26.8±36.7	94.8±1.1
Panel C: probe ranking (perturbation vs. gradient)					
Perturbation	18	52	86	70	94
Gradient	26	40	68	2	94

B.4 Mechanistic Measurements

Hidden-state drift. Under W4, mean relative-L2 hidden-state drift (real prompts, 50 per condition) is mildly higher for reasoning inputs on 0.5B (R/F 1.17), 1.5B (1.10), Phi-3.5

(1.28), and 8B (1.21), and mildly lower on 3B (0.83) and Coder-3B (0.94); with n=50 these ratios carry substantial sampling uncertainty and should be read as indicative magnitudes. Drift grows steeply as bit-width falls (W2 produces 3–18× the W4 drift), and the layer of maximum drift is model-specific (near the output for 0.5B/1.5B/8B; near the input for 3B/Coder-3B at low bit-widths).

Attention divergence. Mean relative-L2 attention divergence over heads (100 prompts per condition) clusters tightly around 1.0 (ratios 0.83–1.22 across the five models with usable attention recordings). We find no consistent attention-level capability asymmetry.

Reasoning-chain integrity. On 60 real GSM8K problems with a 512-token budget: Coder-3B collapses at W4 (65.0%→3.3%; arithmetic-bearing tokens vanish from 100% to 6.7% of generations), 1.5B degrades moderately (71.7%→53.3%), 3B severely (41.7%→5.0%), 0.5B severely (41.7%→10.0%); Phi-3.5 degrades gracefully at W4 (85.0%→78.3%) but collapses at W3 (3.3%); 8B falls steeply at W4 under this protocol (91.7%→35.0%) despite mild perplexity damage. Broken models generate short, arithmetic-free outputs—early chain abandonment rather than subtly wrong arithmetic.

C Generation-Protocol Sensitivity of Base Models

The 8B model illustrates that quantization interacts with prompt format. Under the chain-measurement protocol of Appendix B.4 (unbatched generation, chat-template prompting, 1024-token prompt truncation), the W4 8B model falls from 91.7% to 35.0% chain success; under the GSM8K-500 protocol (batched generation with left-padding, raw completion prompting for base models, 512 new tokens), the same W4 model scores 71.6%. Inspecting per-item records shows the failure signature: at W4, 50 of the 60 chain responses degenerate into echoing the instruction preamble (0/60 at FP16), a failure mode that batched completion-style generation does not exhibit. Base models are known to be sensitive to prompt format, and quantization appears to amplify this sensitivity. All cross-configuration comparisons in this paper are made within a single fixed protocol, so this interaction affects no in-paper comparison; we report it as a methodology observation for practitioners evaluating quantized base models.

D Additional Paired Significance Tests

Table 5 reports the six secondary per-model McNemar comparisons (Holm-corrected within each model): the budget-matched GPTQ comparisons, SGMPQ vs. GPTQ-W4, and the AWQ contrasts. Directional agreement of ARC-Challenge and HellaSwag with the significant GSM8K contrasts is verified in the released per-item records; the 3B GSM8K reversal vs. the random control does not transfer to ARC/HellaSwag, and all inferential claims are restricted to GSM8K.

E RTN vs. GPTQ at Matched Budgets

Table 6 lists the full RTN-vs-GPTQ comparison at each matched budget (W4, W5, and the 5.2-bit mixed configuration) on all three benchmarks.

F Five-Seed Random-Control Details

Table 7 lists the per-seed results of the five-seed Mixed-Random replication (Section 5.1).

G AWQ Baseline Details

AWQ is implemented with the same sequential per-block protocol as the GPTQ arms: identical calibration corpus (128 shuffled Alpaca texts, seed 42, perplexity documents excluded),

Table 5: Additional paired tests (GSM8K, exact McNemar, Holm-corrected within each model): budget-matched GPTQ comparisons, SGMPQ vs. GPTQ-W4, and AWQ contrasts. Negative Δ favors the second configuration.

Model	Comparison	Discordant b/c	p (exact)	p (Holm)	ΔAcc (pp)	95% boot CI
Qwen2.5-0.5B-Inst	GPTQ-MP vs. GPTQ-W5	79/43	0.00143	0.00571	-7.2	[-11.4, -2.8]
Qwen2.5-0.5B-Inst	GPTQ-MP vs. U-W5	84/47	0.00156	0.00571	-7.4	[-12.0, -2.8]
Qwen2.5-0.5B-Inst	SGMPQ vs. GPTQ-W4	43/66	0.0346	0.0692	+4.6	[+0.4, +8.6]
Qwen2.5-0.5B-Inst	AWQ-W4 vs. U-W4	40/86	5.1e-05	0.000255	+9.2	[+4.8, +13.4]
Qwen2.5-0.5B-Inst	AWQ-W4 vs. GPTQ-W4	45/56	0.32	0.32	+2.2	[-1.8, +6.4]
Qwen2.5-0.5B-Inst	AWQ-MP vs. AWQ-W4	39/100	2.33e-07	1.4e-06	+12.2	[+7.8, +16.6]
Qwen2.5-3B (base)	GPTQ-MP vs. GPTQ-W5	94/61	0.00993	0.0199	-6.6	[-11.6, -1.8]
Qwen2.5-3B (base)	GPTQ-MP vs. U-W5	54/127	5.83e-08	2.33e-07	+14.6	[+9.6, +20.0]
Qwen2.5-3B (base)	SGMPQ vs. GPTQ-W4	209/29	8.35e-35	4.18e-34	-36.0	[-41.0, -30.8]
Qwen2.5-3B (base)	AWQ-W4 vs. U-W4	23/198	6.61e-36	3.96e-35	+35.0	[+30.0, +40.0]
Qwen2.5-3B (base)	AWQ-W4 vs. GPTQ-W4	83/82	1	1	-0.2	[-5.0, +4.6]
Qwen2.5-3B (base)	AWQ-MP vs. AWQ-W4	107/55	5.36e-05	0.000161	-10.4	[-15.2, -5.6]
Qwen3-8B (base)	GPTQ-MP vs. GPTQ-W5	11/26	0.0201	0.0803	+3.0	[+0.6, +5.4]
Qwen3-8B (base)	GPTQ-MP vs. U-W5	12/25	0.047	0.141	+2.6	[+0.2, +5.0]
Qwen3-8B (base)	SGMPQ vs. GPTQ-W4	27/30	0.791	0.791	+0.6	[-2.2, +3.6]
Qwen3-8B (base)	AWQ-W4 vs. U-W4	40/105	6.49e-08	3.9e-07	+13.0	[+8.4, +17.6]
Qwen3-8B (base)	AWQ-W4 vs. GPTQ-W4	41/26	0.0864	0.173	-3.0	[-6.2, +0.2]
Qwen3-8B (base)	AWQ-MP vs. AWQ-W4	19/39	0.0119	0.0596	+4.0	[+1.0, +7.0]
Qwen2.5-Coder-3B	GPTQ-MP vs. GPTQ-W5	53/82	0.0156	0.0469	+5.8	[+1.2, +10.2]
Qwen2.5-Coder-3B	GPTQ-MP vs. U-W5	50/95	0.000232	0.00116	+9.0	[+4.2, +13.8]
Qwen2.5-Coder-3B	SGMPQ vs. GPTQ-W4	74/74	1	1	+0.0	[-4.6, +4.6]
Qwen2.5-Coder-3B	AWQ-W4 vs. U-W4	2/275	3.17e-79	1.9e-78	+54.6	[+50.2, +59.2]
Qwen2.5-Coder-3B	AWQ-W4 vs. GPTQ-W4	110/68	0.00203	0.00813	-8.4	[-13.8, -3.2]
Qwen2.5-Coder-3B	AWQ-MP vs. AWQ-W4	56/81	0.0399	0.0799	+5.0	[+0.4, +9.8]
Qwen2.5-1.5B-Inst	GPTQ-MP vs. GPTQ-W5	84/47	0.00156	0.00779	-7.4	[-11.6, -3.0]
Qwen2.5-1.5B-Inst	GPTQ-MP vs. U-W5	81/39	0.000158	0.000948	-8.4	[-12.6, -4.2]
Qwen2.5-1.5B-Inst	SGMPQ vs. GPTQ-W4	63/93	0.0199	0.0798	+6.0	[+0.8, +10.8]
Qwen2.5-1.5B-Inst	AWQ-W4 vs. U-W4	72/75	0.869	1	+0.6	[-4.2, +5.4]
Qwen2.5-1.5B-Inst	AWQ-W4 vs. GPTQ-W4	68/96	0.0347	0.104	+5.6	[+0.4, +10.8]
Qwen2.5-1.5B-Inst	AWQ-MP vs. AWQ-W4	63/70	0.603	1	+1.4	[-2.8, +6.0]
Phi-3.5-mini	GPTQ-MP vs. GPTQ-W5	26/31	0.597	1	+1.0	[-2.0, +4.0]
Phi-3.5-mini	GPTQ-MP vs. U-W5	25/26	1	1	+0.2	[-2.6, +3.2]
Phi-3.5-mini	SGMPQ vs. GPTQ-W4	38/44	0.581	1	+1.2	[-2.4, +4.8]
Phi-3.5-mini	AWQ-W4 vs. U-W4	57/80	0.0598	0.299	+4.6	[+0.4, +9.0]
Phi-3.5-mini	AWQ-W4 vs. GPTQ-W4	48/36	0.23	0.919	-2.4	[-6.0, +1.2]
Phi-3.5-mini	AWQ-MP vs. AWQ-W4	21/55	0.000121	0.000724	+6.8	[+3.4, +10.0]

group size 128, per-block scale search (20-point grid, duo-scaling) followed by per-linear clipping search and asymmetric per-group quantization. Table 8 lists the full results for the three AWQ arms on all six models.

H Capability-Agnostic Control Details

Table 9 reports the capability-agnostic placement control at the identical 30%/8-4 budget; Table 10 the paired exact McNemar tests vs. SGMPQ (Holm-corrected within each model across the two tests). The two agnostic arms coincide exactly on Coder-3B (identical top-30% sets, hence identical accuracies); on 3B and 8B they differ in one layer each. Ranking overlaps against the ratio ranking: 2/11 shared layers on Coder-3B, 1/11 on 3B, 7/11 on 8B (released in the per-run manifests).

I AWQ-W5 Calibration-Seed Replication

The AWQ-W5 depression on the 3B base model (23.2% GSM8K vs. AWQ-W4's 54.2) is a single-cell result obtained with one calibration draw (Alpaca shuffle seed 42). To characterize its dependence on the calibration draw, we replicate uniform AWQ-W5 on this model with three additional calibration seeds (43, 44, 45) under the identical protocol. Results: GSM8K 28.2% (seed 43), 43.0% (44), and 49.8% (45), with ARC-C (80.0–82.0), HellaSwag (68.0–71.0), and perplexity (5.16–5.69) essentially unchanged across draws. Two observations follow. First, AWQ-W5 underperforms AWQ-W4 (54.2%) on all four calibration draws, so the W5 anomaly is systematic on this model rather than a one-off fluke. Second, the magnitude is strongly calibration-dependent (range 23.2–49.8, spread 26.6 pp), consistent

Table 6: RTN vs. GPTQ at matched bit budgets (GSM8K/ARC-C/HellaSwag, %). GPTQ dominates RTN in the catastrophic regime (Coder-3B: 1.2% → 64.2% GSM8K at W4); the quantizers converge in the graceful regime.

Model	Method	Bits	GSM8K	ARC-C	HellaSwag
Qwen2.5-0.5B-Inst	RTN	4.00	14.0	40.0	42.5
	GPTQ	4.00	21.0	24.5	45.5
Qwen2.5-0.5B-Inst	RTN	5.00	38.8	48.5	46.5
	GPTQ	5.00	38.6	41.5	47.5
Qwen2.5-0.5B-Inst	RTN	5.17	25.6	36.0	45.0
	GPTQ	5.17	31.4	40.0	48.0
Qwen2.5-3B (base)	RTN	4.00	19.2	68.5	66.0
	GPTQ	4.00	54.4	80.5	69.0
Qwen2.5-3B (base)	RTN	5.00	39.8	80.0	71.0
	GPTQ	5.00	61.0	82.0	69.0
Qwen2.5-3B (base)	RTN	5.22	18.4	71.5	67.5
	GPTQ	5.22	54.4	82.5	68.5
Qwen3-8B (base)	RTN	4.00	71.6	92.0	74.5
	GPTQ	4.00	87.6	92.0	73.5
Qwen3-8B (base)	RTN	5.00	88.8	93.5	75.5
	GPTQ	5.00	88.4	93.5	74.0
Qwen3-8B (base)	RTN	5.22	88.2	92.0	77.0
	GPTQ	5.22	91.4	92.5	76.5
Qwen2.5-Coder-3B	RTN	4.00	1.2	0.5	37.0
	GPTQ	4.00	64.2	74.5	62.0
Qwen2.5-Coder-3B	RTN	5.00	54.0	77.5	68.5
	GPTQ	5.00	57.2	76.0	67.5
Qwen2.5-Coder-3B	RTN	5.22	64.2	74.5	64.0
	GPTQ	5.22	63.0	74.5	65.0
Qwen2.5-1.5B-Inst	RTN	4.00	57.6	69.0	65.5
	GPTQ	4.00	52.6	68.5	69.5
Qwen2.5-1.5B-Inst	RTN	5.00	65.2	71.5	67.5
	GPTQ	5.00	64.2	71.5	68.5
Qwen2.5-1.5B-Inst	RTN	5.14	58.6	69.5	63.5
	GPTQ	5.14	56.8	70.0	65.0
Phi-3.5-mini	RTN	4.00	75.4	82.5	74.5
	GPTQ	4.00	82.4	82.5	75.5
Phi-3.5-mini	RTN	5.00	87.8	83.0	76.5
	GPTQ	5.00	87.0	84.0	75.5
Phi-3.5-mini	RTN	5.25	83.6	83.0	75.0
	GPTQ	5.25	88.0	84.5	72.5

with the hypothesized instability of AWQ's asymmetric zero-point treatment at 5 bits: the scale-search stage lands on different clipping solutions per draw, and some land worse than others. The single-cell number in Table 8 should therefore be read as one draw from a high-variance distribution whose upper range still falls short of AWQ-W4.

Table 7: Five-seed Mixed-Random control (GSM8K/ARC-C/HellaSwag in %, PPL, average bits). p (McN): exact McNemar p-value of each seed vs. SGMPQ 8/4 on paired GSM8K items.

Model	Seed	GSM8K↑	ARC-C↑	HellaSwag↑	PPL↓	AvgBits	p (McN)
Qwen2.5-0.5B-Inst	0	23.4	42.0	46.0	9.46	5.17	0.32
Qwen2.5-0.5B-Inst	1	22.6	41.5	45.0	10.27	5.17	0.16
Qwen2.5-0.5B-Inst	2	22.2	40.0	46.0	10.65	5.17	0.11
Qwen2.5-0.5B-Inst	3	26.6	37.5	44.5	10.33	5.17	0.71
Qwen2.5-0.5B-Inst	4	25.6	40.0	45.5	10.69	5.17	1
	mean±std (5 seeds)	24.1±1.9	40.2±1.8	45.4±0.7	10.28±0.49	–	
Qwen2.5-3B (base)	0	31.6	69.5	66.0	6.48	5.22	5.6e-09
Qwen2.5-3B (base)	1	26.8	72.0	65.0	6.36	5.22	9e-05
Qwen2.5-3B (base)	2	44.6	74.0	67.0	6.58	5.22	7.2e-23
Qwen2.5-3B (base)	3	25.6	77.5	66.5	5.80	5.22	0.00096
Qwen2.5-3B (base)	4	31.8	79.0	64.5	5.84	5.22	1.3e-07
	mean±std (5 seeds)	32.1±7.5	74.4±3.9	65.8±1.0	6.21±0.37	–	
Qwen3-8B (base)	0	76.0	91.0	73.5	4.61	5.22	6.9e-10
Qwen3-8B (base)	1	88.6	91.5	75.0	4.60	5.22	0.88
Qwen3-8B (base)	2	72.4	91.5	74.5	4.79	5.22	3.7e-15
Qwen3-8B (base)	3	82.2	91.5	75.0	4.22	5.22	0.0011
Qwen3-8B (base)	4	77.2	91.5	76.0	4.65	5.22	1.7e-08
	mean±std (5 seeds)	79.3±6.3	91.4±0.2	74.8±0.9	4.57±0.21	–	
Qwen2.5-Coder-3B	0	2.4	0.0	33.5	444.44	5.22	6.9e-85
Qwen2.5-Coder-3B	1	1.2	1.0	37.0	318.75	5.22	2.4e-93
Qwen2.5-Coder-3B	2	0.4	0.5	33.0	327.52	5.22	1.9e-96
Qwen2.5-Coder-3B	3	1.0	1.5	37.0	469.37	5.22	1.3e-90
Qwen2.5-Coder-3B	4	0.8	2.5	36.5	372.61	5.22	2.4e-92
	mean±std (5 seeds)	1.2±0.8	1.1±1.0	35.4±2.0	386.54±67.98	–	
Qwen2.5-1.5B-Inst	0	61.8	70.0	66.0	6.76	5.14	0.21
Qwen2.5-1.5B-Inst	1	62.0	68.0	64.5	7.00	5.14	0.17
Qwen2.5-1.5B-Inst	2	58.6	70.0	65.0	6.94	5.14	1
Qwen2.5-1.5B-Inst	3	59.8	71.0	66.5	7.22	5.14	0.67
Qwen2.5-1.5B-Inst	4	62.6	72.0	67.5	7.12	5.14	0.08
	mean±std (5 seeds)	61.0±1.7	70.2±1.5	65.9±1.2	7.01±0.17	–	
Phi-3.5-mini	0	83.4	84.0	75.5	5.13	5.25	1
Phi-3.5-mini	1	79.4	83.0	73.5	5.49	5.25	0.031
Phi-3.5-mini	2	73.8	84.5	75.0	5.45	5.25	1.9e-06
Phi-3.5-mini	3	73.0	82.5	73.5	5.69	5.25	3.5e-06
Phi-3.5-mini	4	81.8	81.0	75.0	6.28	5.25	0.36
	mean±std (5 seeds)	78.3±4.7	83.0±1.4	74.5±0.9	5.61±0.43	–	

Table 8: AWQ baseline: GSM8K/ARC-C/HellaSwag in %, perplexity on the 100-document Alpaca set. AWQ-MP uses SGMPQ's 8/4 layer assignment.

Model	Method	Avg bits	GSM8K	ARC-C	HellaSwag	PPL
0.5B	AWQ W4	4.0	23.2	30.5	46.5	8.92
0.5B	AWQ W5	5.0	40.6	46.0	47.5	7.66
0.5B	AWQ-MP (SGMPQ)	5.2	35.4	42.5	45.0	8.23
1.5B	AWQ W4	4.0	58.2	69.0	64.0	6.27
1.5B	AWQ W5	5.0	64.2	74.5	65.5	5.60
1.5B	AWQ-MP (SGMPQ)	5.2	59.6	70.5	67.5	6.12
3B	AWQ W4	4.0	54.2	81.0	71.0	6.27
3B	AWQ W5	5.0	23.2	81.5	69.0	5.95
3B	AWQ-MP (SGMPQ)	5.2	43.8	83.0	70.0	6.11
Coder-3B	AWQ W4	4.0	55.8	76.0	65.5	5.68
Coder-3B	AWQ W5	5.0	59.4	77.5	67.5	5.46
Coder-3B	AWQ-MP (SGMPQ)	5.2	60.8	75.0	68.0	5.56
8B	AWQ W4	4.0	84.6	91.5	71.5	3.95
8B	AWQ W5	5.0	90.0	93.0	75.5	3.58
8B	AWQ-MP (SGMPQ)	5.2	88.6	94.0	74.0	3.77
Phi-3.5	AWQ W4	4.0	80.0	82.0	74.0	5.17
Phi-3.5	AWQ W5	5.0	87.0	84.5	76.5	5.11
Phi-3.5	AWQ-MP (SGMPQ)	5.2	86.8	83.0	75.0	4.88

Table 9: Capability-agnostic placement control at the identical 30%/8-4 budget (GSM8K %, ARC-C %, HellaSwag %, PPL, average bits). Mixed-Random is the five-seed mean±std.

Model	GSM8K↑	ARC-C↑	HellaSwag↑	PPL↓	AvgBits	
Qwen2.5-Coder-3B	FP16	57.8	79.0	67.5	5.51	–
	SGMPQ 8/4 (ratio)	64.2	74.5	64.0	5.45	5.22
	Agnostic-Sum (S_r+S_f)	59.4	74.5	65.0	5.61	5.22
	Agnostic-R (S_r only)	59.4	74.5	65.0	5.61	5.22
	Mixed-Random (5 seeds)	1.2±0.8	1.1±1.0	35.4±2.0	386.54±67.98	5.22
	Uniform W5	54.0	77.5	68.5	5.51	5.00
Qwen2.5-3B (base)	FP16	58.0	82.5	68.5	4.82	–
	SGMPQ 8/4 (ratio)	18.4	71.5	67.5	6.61	5.22
	Agnostic-Sum (S_r+S_f)	51.8	77.5	69.5	5.26	5.22
	Agnostic-R (S_r only)	48.8	79.5	68.0	5.32	5.22
	Mixed-Random (5 seeds)	32.1±7.5	74.4±3.9	65.8±1.0	6.21±0.37	5.22
	Uniform W5	39.8	80.0	71.0	5.50	5.00
Qwen3-8B (base)	FP16	89.8	92.5	76.0	3.60	–
	SGMPQ 8/4 (ratio)	88.2	92.0	77.0	4.78	5.22
	Agnostic-Sum (S_r+S_f)	82.2	92.5	76.5	4.07	5.22
	Agnostic-R (S_r only)	82.4	92.5	76.0	4.07	5.22
	Mixed-Random (5 seeds)	79.3±6.3	91.4±0.2	74.8±0.9	4.57±0.21	5.22
	Uniform W5	88.8	93.5	75.5	3.82	5.00

Table 10: SGMPQ vs. capability-agnostic placements (GSM8K, exact McNemar, Holm-corrected within each model). Positive Δ favors SGMPQ.

Model	Comparison	Disc. b/c	p (exact)	p (Holm)	ΔAcc [95% CI]
Qwen2.5-Coder-3B	SGMPQ vs. Agnostic-Sum	53/77	0.0433	0.0865	+4.8 [+0.4, +8.8]
Qwen2.5-Coder-3B	SGMPQ vs. Agnostic-R	53/77	0.0433	0.0865	+4.8 [+0.4, +8.8]
Qwen2.5-3B (base)	SGMPQ vs. Agnostic-Sum	195/28	5.47e-32	1.09e-31	-33.4 [-38.8, -28.4]
Qwen2.5-3B (base)	SGMPQ vs. Agnostic-R	180/28	2.26e-28	2.26e-28	-30.4 [-35.4, -25.4]
Qwen3-8B (base)	SGMPQ vs. Agnostic-Sum	20/50	0.00044	0.000881	+6.0 [+2.8, +9.4]
Qwen3-8B (base)	SGMPQ vs. Agnostic-R	20/49	0.000636	0.000881	+5.8 [+2.6, +9.0]