THE MISALLOCATION OF PRECISION IN QUANTIZED LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixed-precision quantization must decide where to spend a fixed budget of full-precision weights, and the criteria in use are derived from activation statistics: GPTQ minimises layer-wise error under the Hessian $H{=}2XX^\top$, ResQ keeps that covariance's highest-variance subspace at 8 bits. We test the premise underneath, that numerical sensitivity tracks functional importance, by taking its natural per-layer summary, $\mathrm{Tr}(H_\ell)/d$, as a cross-layer allocation rule. In Llama-3.1-8B it is not merely uncorrelated with functional importance but *inverted*: Spearman $\rho{=}{-}0.497$ ($p{<}0.001$, $n{=}64$). The sub-layers it ranks highest sit in a functional dead zone; those it ranks lowest are load-bearing. The cost is measurable: at an identical 6/32 FP16 budget and identical attention/MLP composition, allocating by Hessian rank recovers $7\times$ less of the damage 4-bit quantization does to ARC-Challenge (7.7% vs. 53.8%, $p{=}0.002$), and leaves the model significantly worse than FP16 where the functionally-chosen budget does not. We establish this by treating 4-bit NF4 quantization as a controlled perturbation probe, analogous to lesion studies in neuroscience: quantize globally, restore full precision to one sub-layer, measure task recovery. A 64-sub-layer sweep localises the effect to *named sub-layers* rather than to components as classes. Against a do-nothing null measured on the same examples (ARC exactly 0/468) with BH-FDR correction, L13–14 attention clear the null on GSM8K and MATH but on neither measure of ARC, while L6–7 MLP clear it on all three; the component-level contrast is itself not significant ($p{=}0.107$). Split-half selection shows the ranking beats a random equal-budget subset on ARC ($+10.4$pp, 95% CI [$+3.7$, $+18.5$]) while carrying 4.6–12.2pp of selection optimism. The inversion replicates on Mistral-7B and under GPTQ, and the Llama-3.1-8B-derived layer set transferred to Mistral-7B unchanged recovers 47.1% of its ARC damage against 2.0% for Mistral-7B's own Hessian-top layers, though offset by collateral damage elsewhere, so it transfers as a diagnostic rather than an allocator. A surgical model keeping 6 of 32 sub-layers at full precision reaches FP16 equivalence at 42% memory, yet neither Hessian sensitivity nor weight-space SQNR ($\rho{=}{+}0.032$, *ns*) predicts which six.

1 INTRODUCTION

Mixed-precision quantization spends a small, fixed budget of full-precision weights on the parts of the network judged most worth protecting. Which parts those are is decided by a numerical criterion derived from activation statistics: GPTQ Frantar et al. (2023) minimises layer-wise error under the Hessian $H{=}2XX^\top$ at a uniform bit-width, and ResQ Saxena et al. (2025) keeps the highest-variance PCA subspace of that same covariance, about $1/8$ of the hidden dimension, at 8 bits within each layer. Neither ranks layers against one another. Both rest on a premise that is assumed rather than tested: that numerical sensitivity tracks functional importance. The premise is load-bearing the moment a budget is spread across layers, because that budget is finite: every sub-layer protected for the wrong reason is one that could have been protected for the right one. We therefore ask what happens when the statistic these methods share is used as the per-layer criterion that cross-layer allocation requires, summarised as $\mathrm{Tr}(H_\ell)/d = \mathbb{E}[\|\mathbf{x}_\ell\|^2]/d$.

The premise fails, and it fails in the worst available direction. The trace is not merely a weak predictor of functional importance but an *inverted* one ($\rho{=}{-}0.497$, $p{<}0.001$; Figure 1): the sub-layers it ranks

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.

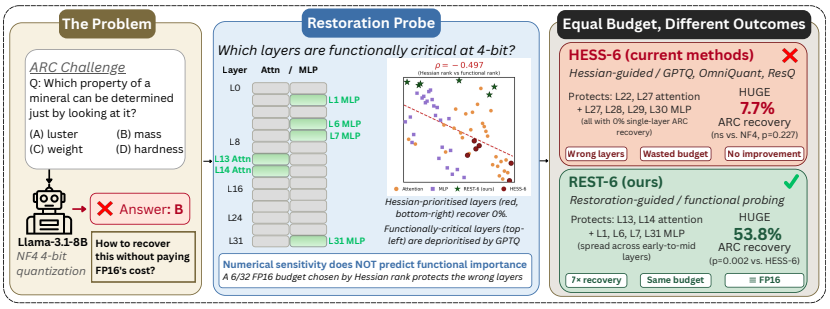


Figure 1: **Hessian sensitivity ranks the layers backwards.** A 4-bit NF4 restoration probe identifies six functionally critical sub-layers (L13, L14 attention; L1, L6, L7, L31 MLP) anti-correlated with Hessian rank (ρ=−0.497, p<0.001, n=64). At an identical 6/32 FP16 budget *and* identical attention/MLP composition, the Hessian-guided HESS-6 selection recovers only 7.7% of NF4 damage, while restoration-guided REST-6 recovers 53.8% (p=0.002 direct comparison; Section 4.8).

highest are functionally inert at task-performance level, while those it ranks lowest are the ones whose precision loss actually degrades downstream behaviour. The cost is measurable. Spending an identical 6/32 FP16 budget by Hessian rank rather than by measured function recovers 7× less of the damage 4-bit quantization does to ARC-Challenge, and leaves the model significantly worse than FP16 where the functionally-chosen budget does not (Section 4.8).

We establish this by treating 4-bit quantization as a controlled perturbation probe, borrowing from cognitive neuropsychology, where patterns of impairment after focal damage are used to infer the functional organisation of the intact system Shallice (1988). Quantize globally, restore full-precision weights to one sub-layer at a time, and measure task recovery: large task-specific recovery marks a sub-layer as functionally critical, none marks it redundant at this precision. No custom hooks, no activation patching, no training, just a standard HuggingFace model.

Across all 64 sub-layers of Llama-3.1-8B-Instruct on GSM8K Cobbe et al. (2021), ARC-Challenge Clark et al. (2018) and MATH Hendrycks et al. (2021b), attention layers 13–14 recover 52.1% of arithmetic loss but 8.3% on science knowledge, while MLP layers 6–7 benefit all three. The effect localises to specific sub-layers: the attention-vs-MLP contrast as a class comparison is not significant, and we do not claim it.

Contributions.

1. The activation-covariance trace, the per-layer summary of the signal GPTQ and ResQ protect, predicts functional layer importance with ρ=−0.497; weight-space SQNR also fails (ρ=+0.032, *ns*).
2. A 64-sub-layer sweep with per-example logging localises the damage to named sub-layers, tested against a do-nothing null measured on the same examples; split-half selection shows the ranking beats a random equal-budget subset on ARC (+10.4pp, 95% CI [+3.7, +18.5]) at 4.6–12.2pp of selection optimism.
3. Algorithmic reasoning is precision-sensitive; commonsense and factual recall (HellaSwag, MMLU) are not.
4. A surgical model patching 6/32 sub-layers reaches FP16 equivalence at 42% memory and near-4-bit throughput (106 vs. 101 tok/s).
5. Replications across architectures (Mistral-7B, DeepSeek-R1-Distill) and methods (GPTQ) rule out architecture- and NF4-specific confounds. Applied to Mistral-7B *unchanged*, the Llama-3.1-8B-derived layer set recovers 47.1% of its ARC damage against 2.0% for Mistral-7B's own natively measured Hessian-top layers (p=1.1 × 10^{-5}).

2 Related Work

Post-training quantization. GPTQ Frantar et al. (2023) minimises layer-wise error under second-order Hessian information, AWQ Lin et al. (2024) picks salient weights by activation magnitude,

OmniQuant Shao et al. (2024) calibrates weights and activations jointly, LeanQuant Zhang & Shrivastava (2025) learns loss-aware non-uniform grids, and ResQ Saxena et al. (2025) keeps a PCA-selected high-variance subspace at higher precision; LLMC Gong et al. (2024) benchmarks them under standardised conditions. We ask instead which layers are *functionally* precision-critical, and find the signal GPTQ and ResQ protect (ρ=-0.497), and weight-space SQNR (ρ=$+0.032$), uninformative about it.

Layer importance and pruning. Gromov et al. (2025) showed deep layers (never the last) can be dropped with little loss on knowledge QA while reasoning degrades immediately, the asymmetry our dead zone reproduces. Sun et al. (2025) explain it via the *Curse of Depth*: Pre-LN transformers develop near-identity residual blocks with depth, which they report are robust to pruning and perturbation; we take precision loss to be one such perturbation. Wang et al. (2026) find pruning one or two layers collapses long-chain reasoning while knowledge-intensive performance degrades gradually, a layer-level analogue of our task-type selectivity.

Mechanistic interpretability. ACDC Conmy et al. (2023) and path patching Wang et al. (2023) identify task-specific circuits at sub-layer granularity, and Head Pursuit Basile et al. (2025) shows that editing $\sim$1% of attention heads suppresses or enhances targeted concepts in unimodal and multimodal transformers. Li (2024) find binary/ternary transformers trained on modular addition learn algorithms close to their full-precision counterparts. We instead probe post-training quantization of trained LLMs, obtaining sub-layer selectivity from precision loss alone: coarser granularity, but no custom infrastructure and a direct mapping to deployment decisions.

Functional specialisation. Geva et al. (2021) showed MLP layers act as key-value memory banks and Meng et al. (2022) localised factual associations to mid-layer MLP modules, which motivates our reading of MLP cross-task transfer. Fartale et al. (2025) report the opposite depth ordering in Qwen2.5-7B-Instruct (recall in layers 0–6, inference in 17–27 of 28); their single model and 60-item probe set may explain the discrepancy. Lee et al. (2025) show quantization degradation varies with task difficulty in non-obvious ways, consistent with our MATH stratification (Appendix F), and Kumar et al. (2024) report head-level differentiation in predicting cortical activity while cautioning against one-to-one mapping.

3 Methodology

3.1 Surgical Transplant Mechanism

Rather than asking how to minimise quantization error, we ask what that error reveals about functional architecture: if restoring precision to sub-layer ℓ recovers performance on task τ, then ℓ was precision-critical for τ under 4-bit compression. This is an operational definition of functional criticality grounded in task performance rather than in representational geometry or numerical error.

Let $\mathcal{M}_{4\text{bit}}$ be a fully NF4-quantized model and $\mathcal{M}_{\text{FP16}}$ be the full-precision baseline. For each sub-layer $\ell \in \{\text{L0}_{\text{attn}}, \ldots, \text{L31}_{\text{mlp}}\}$, we construct a *patched model* $\mathcal{M}_\ell^{\text{patch}}$ by replacing the quantized weight tensors of sub-layer ℓ with their FP16 counterparts:

$$\mathcal{M}_\ell^{\text{patch}} = \mathcal{M}_{4\text{bit}} \text{ with } W_\ell \leftarrow W_\ell^{\text{FP16}}. \tag{1}$$

The patch is non-persistent: each trial starts from a fresh copy of $\mathcal{M}_{4\text{bit}}$.

3.2 Sensitivity Scoring

The *recovery rate* for sub-layer ℓ on dataset τ is

$$r(\ell, \tau) = \frac{|\mathcal{F}_\tau^{\text{patch}}(\ell)|}{|\mathcal{F}_\tau|}, \tag{2}$$

where $\mathcal{F}_\tau$ is the set of *forward-flipped* examples for τ (correct under FP16, incorrect under 4-bit) and $\mathcal{F}_\tau^{\text{patch}}(\ell)$ is the subset recovered by the patch. Conditioning on examples that demonstrate quantization damage makes recovery comparable across datasets with different absolute accuracy (Appendix G analyses flips in both directions).

Table 1: Accuracy (%) and WikiText-2 perplexity for Llama-3.1-8B under different quantization methods. WG = WinoGrande; PPL = perplexity (lower is better). Gaps shown relative to FP16 (±pp). Best per column **bold**. This table compares quantization methods; the surgical model is evaluated separately over 5 seeds in Section 4.4.

Method	GSM8K	ARC	MATH	WG	PIQA	PPL↓
FP16	**80.2**	**80.0**	**62.0**	**58.4**	**76.8**	**7.14**
NF4	75.6 −4.6	76.0 −4.0	60.5 −1.5	56.4	76.0	7.54
GPTQ	**85.4** +5.2	79.2 −0.8	59.5 −2.5	58.2	76.4	7.44

Hessian sensitivity proxy. Following GPTQ and ResQ, the activation-space Hessian sensitivity of layer ℓ is approximated by the trace of the input covariance:

$$H_\ell \approx \frac{\mathrm{Tr}(\hat{H}_\ell)}{d_\ell} = \frac{\mathbb{E}\left[\|\mathbf{x}_\ell\|^2\right]}{d_\ell}, \tag{3}$$

computed over 128 WikiText-2 calibration sequences of 2048 tokens via forward hooks on Llama-3.1-8B.

3.3 Evaluation Protocol

We use Llama-3.1-8B-Instruct Grattafiori et al. (2024) (32 layers, 8B params) with bitsandbytes NF4 double quantization Dettmers et al. (2023), evaluated on five 500-example benchmarks: GSM8K Cobbe et al. (2021) (chain-of-thought, greedy decoding), ARC-Challenge Clark et al. (2018) (zero-shot), MATH Hendrycks et al. (2021b) (levels 1–5; the first 500 test problems, which that dataset orders by subject, so this sample is Algebra), and HellaSwag Zellers et al. (2019) + MMLU Hendrycks et al. (2021a) for commonsense/factual contrast (all zero-shot). Statistical tests: McNemar's (paired per-example, α=0.05, 5 seeds) for pairwise comparisons; binomial tests with Benjamini-Hochberg FDR (α=0.05) across 192 tests (64 layers $\times$ 3 core datasets); Cohen's d, computed per example, for effect sizes. All experiments replicated on Mistral-7B-Instruct-v0.3 Jiang et al. (2023) (32 layers, 8 KV heads, grouped-query attention).

4 Results

4.1 Quantization Damage is Non-Uniform

Table 1 reports FP16, global 4-bit NF4 and a GPTQ baseline with WikiText-2 perplexity and zero-shot benchmarks (AWQ was infeasible on our hardware).

NF4 produces the most *uniform* damage: gaps are small and consistent across tasks (GSM8K −4.6pp, ARC −4.0pp, MATH −1.5pp), where GPTQ swings from +5.2pp on GSM8K to −2.5pp on MATH. Uniformity is what makes NF4 the controlled probe: a systematic perturbation is the prerequisite for mechanistic inference. A task-type pattern runs alongside this. NF4 gaps for commonsense (HellaSwag 1.0pp) and factual recall (MMLU 2.2pp) fall below our ±3.5pp detection floor at n=500, while multi-step reasoning (GSM8K 4.6pp, ARC 4.0pp) is precision-sensitive: the layers critical for reasoning encode computations the other tasks do not need at the same precision. This mirrors the pruning setting, where removing one or two layers collapses long-chain reasoning while knowledge-intensive accuracy degrades only gradually (Wang et al., 2026). MATH sits between, at 1.6pp; Appendix F shows FP16 and NF4 fail Level 4–5 problems at the same rate, so what signal there is comes from the levels where quantization, rather than capability, binds. GPTQ's +5.2pp on GSM8K is not quantization improving the model: that row is a third-party pre-quantized checkpoint carrying its own calibration, not our quantization of the weights used in the other rows, so it is not a controlled comparison. The probe uses NF4 throughout. Nor is GPTQ a head-to-head competitor to our surgical model, since it optimises layer-wise reconstruction at a uniform bit-width using the activation-covariance Hessian, whose per-layer trace we show is inverted (Sections 4.5 and 4.8).

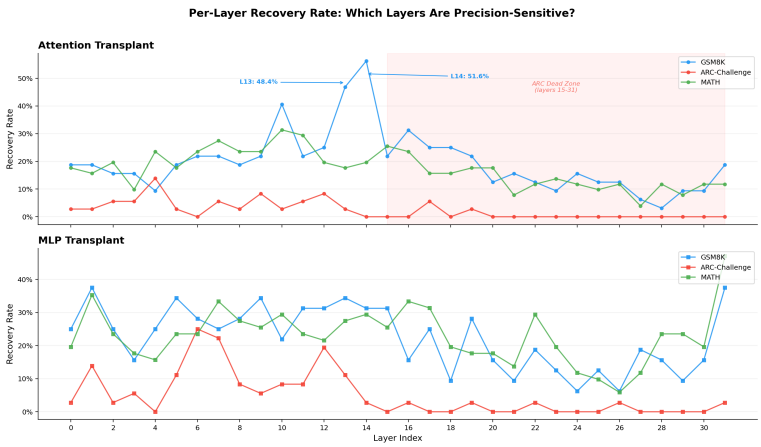


Figure 2: Recovery rate $r(\ell, \tau)$ for all 32 layers under attention (top) and MLP (bottom) transplants (blue: GSM8K; red: ARC; green: MATH). Attention L13–14 spike on GSM8K and, more weakly, MATH, but not ARC (task-selective); the grey band (L15–31, ARC) is our attention dead zone, which we read via the Curse of Depth Sun et al. (2025); MLP L6–7 recover near-equally across tasks (cross-task).

4.2 Per-Layer Sensitivity Sweep

Having established uniform NF4 damage as a controlled probe, we now ask which layers are precision-critical. Figure 2 plots recovery rate $r(\ell, \tau)$ for all 32 layers, separately for attention and MLP; Appendix B gives per-layer values. The profile reveals three qualitatively distinct regions:

Attention peak (L13–14): the highest GSM8K recovery of any sub-layers (46.9%, 56.2%) against 2.8% and 0.0% on ARC and 17.6–19.6% on MATH: selective against science knowledge, not mathematics. **Attention dead zone (L15–31):** ARC recovery is *exactly zero* in 15 of these 17 layers and below 6% in the other two, consistent with the Curse of Depth Sun et al. (2025), where Pre-LN near-identity residual blocks contribute little to the model's output. **MLP cross-task plateau (L0–8):** substantial recovery on every task ($L7_{\text{mlp}}$: 25.0/22.2/33.3%), the near-equal rates indicating shared computational substrates rather than task-specific circuits.

4.3 Cross-Task Functional Dissociation

Patching each task's best sub-layers and evaluating on all three shows how far a patch transfers (full matrix in Appendix C). The GSM8K-optimal attention patch (L13+L14) reaches 52.1% GSM8K but 8.3% ARC, so attention precision is needed specifically for step-by-step arithmetic, whereas the ARC-optimal MLP patch (L6+L7) reaches 36.1% ARC and also 47.9% GSM8K and 34.0% MATH. Combining the six reaches 70.8/52.8/53.2% on 19% of the network.

A measured null. A sub-layer has to beat what doing nothing already achieves, and that is not zero: re-running the *unpatched* 4-bit model over the flipped subset recovers some of those examples anyway, because the subset batches differently from the 500-example baseline pass and the changed padding flips borderline generations. We therefore measure the null rather than borrow one, over 13 replicates differing only in example ordering, at the batch size and token budget every layer condition used. The result is GSM8K 13/416 (3.1%), MATH 39/663 (5.9%) and ARC *exactly* 0/468: a single-letter answer essentially never flips from padding noise alone. Since a null of zero would certify any sub-layer with one lucky recovery, we test against the Clopper–Pearson upper 95% bound (GSM8K 5.3%, MATH 8.0%, ARC 0.8%), the most generous null the replicate data supports.

FDR-corrected significance. Binomial tests across 192 comparisons with BH-FDR correction (α=0.05) leave 102 survivors: 47 on GSM8K (22 attention, 25 MLP), 43 on MATH, and 12 on ARC (3 attention, 9 MLP). The component-level asymmetry runs in the expected direction, attention being 47% of GSM8K's survivors against 25% of ARC's, but does not reach significance (Fisher p=0.107

within ARC, p=0.207 across the two), agreeing with the permutation test on the component×task interaction (−1.7pp, 95% CI [−6.6, +3.3], p=0.52). We make no claim that attention and MLP differ *as classes*. What separates cleanly is named sub-layers: L13_{attn} and L14_{attn} clear the null on GSM8K (46.9%, 56.2%) and MATH but on neither measure of ARC (2.8%, 0.0%), while L6_{mlp} and L7_{mlp} clear it on all three (25.0%, 22.2% on ARC).

Selection and the examples it was made on. The critical sub-layers were ranked on the same flipped examples they are then reported on, so part of that recovery is selection noise. Over 1000 random half-splits we select the top-6 on one half and score them on the other: held-out recovery is 33.3% (GSM8K), 14.1% (ARC) and 28.4% (MATH) against in-sample 45.5%, 18.7% and 37.7%, so the ranking carries 4.6–12.2pp of optimism. Against a *random* 6-sub-layer subset on the same held-out half the advantage is +10.4pp on ARC (95% CI [+3.7, +18.5]), +12.3pp on GSM8K ([0.0, +24.0]) and +8.2pp on MATH ([−0.6, +16.0]); only ARC's excludes zero. Half-split agreement on the top-6 is 0.34, 0.56 and 0.28, so the *particular* six are not stable: REST-6 is one of several near-equivalent sets, as the random draws of Appendix K also show.

4.4 Statistical Validation of the Surgical Model

The 6-sub-layer combined patch identified by Sec 4.3 is the first deployable artefact of the dissociation; before elevating it to a claim, we validate it across 5 seeds. The surgical model achieves GSM8K 78.9±1.3%, ARC 76.2±1.4%, MATH 35.0±0.6% (FP16: 79.4±1.1%, 78.2±1.3%, 35.7±1.4%; Appendix D). This experiment draws a fresh random 500-example sample per seed, including for MATH; Tables 1 and 2 use the first 500 MATH problems, which the dataset orders by subject, so their MATH column covers Algebra alone and is not comparable to the 35.7% here. We report both rather than reconcile them. Under McNemar's test, Surgical ≡ FP16 (GSM8K p=0.640, ARC p=0.074, MATH p=0.716) and Surgical > 4-bit (GSM8K p=0.008, ARC p=0.025; MATH *ns*; see Section 4.1). Because non-significance alone does not establish equivalence, we also test against a pre-specified ±5pp margin. The paired Surgical−FP16 difference across the five seeds is −0.44pp (95% CI [−2.29, +1.41]) on GSM8K, −2.00pp ([−3.37, −0.63]) on ARC and −0.72pp ([−1.97, +0.53]) on MATH; all three intervals lie inside the margin, so equivalence holds, though on ARC the difference is distinguishable from zero. Throughput: FP16 = 199 tok/s / 16.0 GB; 4-bit = 101 tok/s / 5.58 GB; Surgical = 106 tok/s / 6.71 GB, matching 4-bit while adding only 1.13 GB of memory. The full multi-seed accuracy panels, McNemar p-value matrix, and throughput chart are in Figure 4 (Appendix A).

4.5 Hessian Sensitivity Inversion

Figure 3 is the central result of this paper: restoration rank for all 64 sub-layers (by $r(\ell, \cdot)$ averaged across tasks) against Hessian sensitivity rank, in both cases 1=lowest. Spearman ρ=−0.497 (p<0.001; bootstrap 95% CI [−0.683, −0.258], 10,000 resamples, excluding zero): not zero, *negative*. A Hessian-based method would protect the deep attention layers with the largest activation magnitudes, the ones sitting in the ARC dead zone with no functional recovery, while deprioritising early MLP layers whose modest norms hide load-bearing computation.

At the extremes, L1 MLP is restoration rank 63/64 but Hessian rank 2/64, and L27 attention is the reverse. Weight-space SQNR is also uninformative (ρ=+0.032 vs. restoration rank, p=0.80), and quantization error varies by only 0.84 dB across the 64 sub-layers (std 0.11 dB), so critical and non-critical layers have near-identical weight-space error profiles and cannot be told apart by them.

It holds per task as well as in aggregate (ρ_{GSM8K}=−0.403, ρ_{ARC}=−0.562, ρ_{MATH}=−0.477; all p<0.001), strongest on ARC, the dataset where the matched-budget head-to-head of Section 4.8 produces the 7× gap.

Depth mediates much of this. Hessian rank rises with depth (ρ=+0.663) and restoration rank falls with it (ρ=−0.550); partialling depth out leaves ρ=−0.212 (p=0.096). The inversion is therefore substantially, though not entirely, a depth effect. It is not an artifact of pooling two populations: *within* attention ρ=−0.526 (p=0.002) and within MLP ρ=−0.561 (p<0.001), n=32 each, so pooling attenuates rather than manufactures the correlation.

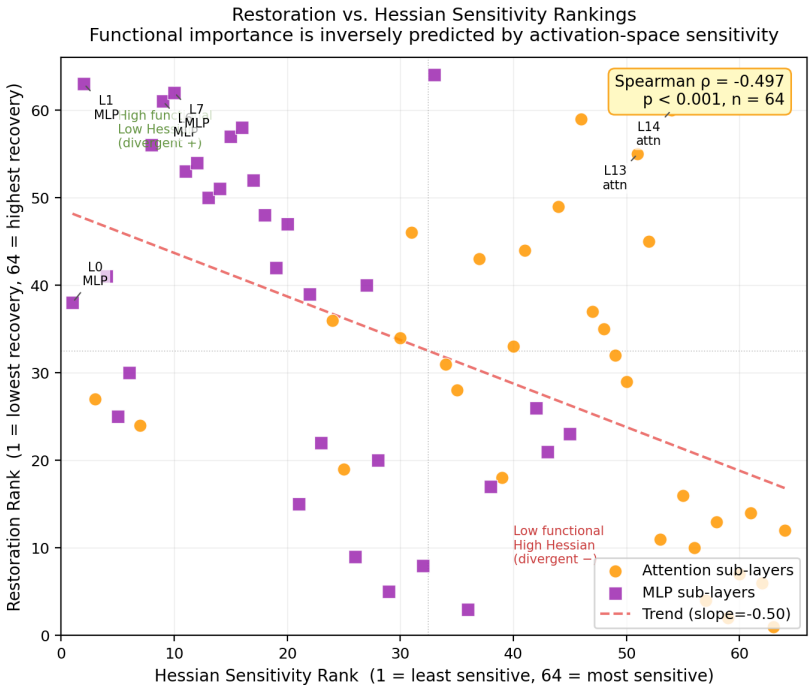


Figure 3: Restoration rank vs. Hessian sensitivity rank for all 64 sub-layers (Llama-3.1-8B). Orange circles: attention; purple squares: MLP; red dashed line: OLS trend ($\rho = -0.497$). Top-left: L1, L6, L7 MLP are functionally critical but Hessian-deprioritised; bottom-right: L23–L31 attention are Hessian-protected but in the functional dead zone.

The mechanism is the one that makes depth predictive. Under Pre-LN, residual-stream variance grows with depth, driving the Jacobian of deep blocks toward the identity so they contribute little Sun et al. (2025); their large activations still look sensitive to any activation-magnitude criterion. Early MLP layers operate on a small residual, so their norms are low, but their compressed representations carry features shared across tasks. Hessian sensitivity measures the size of the residual, not the importance of the computation.

4.6 Cross-Model Replication

If the inversion is a structural property of Pre-LN transformers, the same critical layers and the same negative ρ should appear on other architectures. On Mistral-7B-Instruct-v0.3 (grouped-query attention, 8 KV heads) the inversion replicates and is in fact stronger than on Llama-3.1-8B: $\rho = -0.574$ ($p < 0.001$, $n = 64$), with the same critical layers (L14 attention, L6–7 MLP) emerging on both (Appendices E and J).

Extended multi-model probe. DeepSeek-R1-Distill-Llama-8B Guo et al. (2025) replicates the L13 arithmetic-attention finding (42.6% GSM8K recovery) under a different training regime on the same Llama-8B architecture. Qwen2.5-Math-7B-Instruct Yang et al. (2024) and Gemma-2-9B Gemma Team et al. (2024) are quantization-resistant ($\leq$1pp NF4 damage) and fall outside the probe's operating regime; the probe correctly diagnoses the absence of degradation rather than producing spurious rankings (Appendix I).

4.7 Cross-Quantization-Method Replication (GPTQ)

To rule out NF4-specific confounds we run the identical 64-sub-layer sweep on a GPTQ-quantized baseline Frantar et al. (2023). The dissociation replicates: L14 attention is again the top-ranked sub-layer for GSM8K (46.4% recovery, against 56.2% under NF4) while its ARC recovery stays at 3.6%, and early MLP keeps its cross-task role (L6: 46.4% GSM8K; L7: 28.6% ARC). The functional

Table 2: HESS-6 vs. REST-6 at an identical 6/32 FP16 budget (788.5M parameters each) and identical attention/MLP composition: accuracy (%) and NF4-damage recovery (fraction of flipped examples). MATH uses a 512-token generation budget over 500 examples; Tab. 1 uses 1024 tokens over 200, which accounts for the 60.6 vs. 62.0 baseline. Deltas and McNemar tests are within-run. Only ARC separates the two arms end-to-end; the net-effect decomposition below explains why GSM8K and MATH do not.

Condition	GSM8K	ARC	MATH
FP16	82.2	78.8	60.6
NF4 (4-bit)	82.0	74.4	58.8
HESS-6	81.4	73.4	58.8
REST-6	81.2	**77.4**	57.0
Gap recovery			
HESS-6	14.3%	7.7%	38.3%
REST-6	**60.7%**	**53.8%**	44.7%

map is therefore a property of the model, not of NF4. Per-layer profiles, the partial dead-zone replication and the L0–L1 MLP shift under GPTQ are in Appendix H.

4.8 Hessian-Guided vs. Restoration-Guided Mixed Precision

The ranking inversion ($\rho{=}-0.497$, Section 4.5) is a correlation over 64 sub-layers; it does not directly measure whether Hessian-based layer selection degrades downstream task performance. We close this gap with a direct functional comparison.

Experimental design. We construct two 6-layer surgical models from the same NF4 base, each keeping exactly 6/32 sub-layers at FP16 and differing only in which six. **HESS-6** takes the highest by Hessian rank (the activation-covariance trace used by GPTQ, ResQ, and OmniQuant), *subject to matching REST-6's attention/MLP composition*: L22, L27 attention; L27, L28, L29, L30 MLP. **REST-6** takes the six identified by the restoration probe: L13, L14 attention; L1, L6, L7, L31 MLP. Both arms hold exactly 788.5M parameters at FP16 (ratio 1.000) and the same 2-attention/4-MLP split, so the comparison isolates *which* sub-layers are protected from how many are protected and of what kind. Every HESS-6 member has 0.0% single-layer ARC recovery in the Section 4.2 sweep, whereas REST-6 contains the sub-layers that do recover ARC (L1, L6, L7 MLP at 13.9–25.0%). The unconstrained Hessian top-6 gives the same conclusion more starkly (2.6% against 52.6%) but holds only 252M parameters, confounding layer identity with a $3.1\times$ smaller budget, so we report the matched comparison as primary.

Results. *ARC-Challenge (the decisive benchmark):* REST-6 recovers 53.8% of NF4 damage (95% CI [38.2, 69.4]; $p{=}0.009$ vs. NF4, McNemar's). HESS-6 recovers 7.7% (95% CI [0.0, 17.1]), statistically indistinguishable from plain NF4 ($p{=}0.227$); the two intervals are disjoint. The direct REST-6 vs. HESS-6 comparison yields $p{=}0.002$, and a paired bootstrap on the difference gives $+46.2$ pp (95% CI [27.3, 64.1]): protecting the Hessian-top layers is functionally close to protecting nothing. Strengthening the point further, HESS-6 is itself significantly *worse* than the FP16 baseline on ARC ($p{=}0.0007$, McNemar's) while REST-6 is indistinguishable from it ($p{=}0.281$): allocating mixed-precision budget by Hessian rank actively forgoes the recoverable damage.

GSM8K and MATH: neither separates the arms end-to-end. NF4 damage on GSM8K is negligible here (0.2 pp over 28 flipped examples), leaving almost no room to recover; REST-6 still corrects over four times as many flips (60.7% vs. 14.3%; paired bootstrap $+46.4$ pp, 95% CI [28.0, 65.5]) at indistinguishable overall accuracy ($p{=}1.0$). On MATH's 47 flipped examples REST-6 recovers 44.7% against HESS-6's 38.3%, the ARC ordering but far weaker: the bootstrap interval spans zero ($+6.4$ pp, $[-10.6, 23.9]$) and accuracies are within noise ($p{=}0.39$).

Repair is not the only effect. Recovery counts only forward-flipped examples, which is not the whole story. Against NF4, REST-6 repairs 21 of ARC's 39 flipped examples while breaking 7 previously-correct ones (net $+15$), but nets -4 on GSM8K and -9 on MATH: transplanting FP16

weights perturbs behaviour on examples quantization never damaged, and only on ARC does the repair clearly exceed that cost. So the mechanistic claim, that the probe localises *where* damage lives, rests on the flipped subset and holds on GSM8K and ARC, while the practical claim of better end-to-end accuracy at fixed budget is established on ARC alone. The same pattern appears in the unconstrained-budget run, so it is not an artifact of composition matching.

Matched-budget controls rule out that any six early or arbitrary sub-layers would do: REST-6 beats four of five random draws and the Curse-of-Depth heuristic, though one draw (s105) clears both bars too (Appendix K).

The deep L22–L30 sub-layers the Hessian ranks highest have large activation norms and sit in the dead zone, so protecting them recovers almost nothing; L13–14 attention and early-to-mid MLP have moderate Hessian ranks and are functionally irreplaceable. That is how $\rho{=}-0.497$ becomes a $7\times$ recovery gap at one budget and one composition. Figure 12 in Appendix L plots Table 2.

4.9 Does the Layer Set Transfer?

Finding REST-6 costs a full sweep, roughly 10^4 generations per model. We apply the Llama-3.1-8B-derived set to Mistral-7B *unchanged* against a HESS-6 from Mistral-7B's *own* highest-Hessian sub-layers at the same 2-attention/4-MLP composition (L29, L30 attention; L27–L30 MLP); sub-layer shapes match, so both arms hold 788.5M parameters. The comparison is deliberately unfair to us: the Hessian arm gets model-specific measurement, ours a fixed list from another model.

The localisation transfers. On ARC-Challenge, over the 51 examples NF4 flips, transferred REST-6 recovers 24 (47.1%, 95% CI [33.3, 61.0]) against HESS-6's 1 (2.0%, [0.0, 6.8]). The paired difference is $+45.1$pp (95% CI [$+30.2$, $+60.0$]) and a McNemar test restricted to the flipped subset gives $p{=}1.1 \times 10^{-5}$. HESS-6 is significantly worse than FP16 ($p{=}0.010$) and net-negative against plain NF4, reproducing the dead zone on a second model with its Hessian measured natively. GSM8K and MATH separate in the same direction but not significantly (30.7% vs. 24.0% and 25.4% vs. 18.6%; both $p{=}0.42$, both bootstrap intervals spanning zero), so ARC again carries the result.

The accuracy gain does not. Recovery counts only flipped examples, and on Mistral-7B the rest of the test set cancels it exactly. REST-6 repairs 24 of the 51 flipped examples but breaks 29 that NF4 answered correctly, for a net change of 0 against the 4-bit baseline (331 vs. 331 correct, $p{=}0.90$), the same accounting that gives Llama-3.1-8B $+15$ on ARC but negative nets elsewhere (Section 4.8). We therefore report the transfer as a *diagnostic* result: a fixed index set locates where quantization damage lives in a model it was never derived from. It is not a zero-cost allocation rule, which the net-accuracy null does not support. Absolute accuracies here are not comparable with Table 3: that experiment draws a different 500-example sample (Appendix E), so only the within-run contrasts above are interpretable.

5 Discussion and Conclusion

The activation-covariance trace, summarising the signal post-training quantizers protect, predicts functional layer importance with $\rho{=}-0.497$, an inversion costing $7\times$ ARC recovery at a matched 6/32 FP16 budget (7.7% vs. 53.8%; $p{=}0.002$). The two quantities are orthogonal: large activation norms can be absorbed by the residual stream, while a low-norm sub-layer is precision-critical if it computes a feature no later layer can reconstruct. Criticality is also not a property of isolated layers (single-layer lesions are null across all 12 conditions) but emerges under global degradation.

The dissociation is between named sub-layers, not components: confirmed against a null measured on the same examples, surviving split-half selection on ARC, replicating across architectures and under GPTQ, and transferring to Mistral-7B without re-derivation, where that model's own Hessian-top layers find nothing. Its limit is that the repair does not convert into net accuracy on Mistral-7B, because restoring precision to a sub-layer also perturbs examples quantization never damaged; the probe is best read as a localiser of damage rather than an allocator of budget. It yields a 6/32-sub-layer surgical model at 42% memory. Numerical sensitivity is not functional importance.

AI USE STATEMENT

The research question, the quantization-as-probe design, the experimental protocol, the statistical analysis, and the scientific claims are the authors' own.

We used AI assistance to write and refactor experiment, evaluation, and statistical-analysis code; to script and edit figures; to identify candidate related work and format bibliographic entries; to draft and edit prose for readability; and to discuss readings of the quantitative results, though the interpretations are ours. We did not use it to generate synthetic data, develop theoretical models, formulate or prove mathematical claims, refine the research hypothesis, or design the methodology.

We have reviewed all AI-assisted work. Every bibliographic entry was checked against its primary source, with author list, title, venue, and year confirmed individually; entries we could not verify were removed, and each in-text citation was re-read against the work it cites. The AI-assisted code was reviewed and its numerical output cross-checked against independently saved per-example result files. We take responsibility for the final content, including text, claims, and artifacts produced with AI assistance.

ETHICS STATEMENT

All models and benchmarks are publicly released artifacts used under their standard research licenses and consistent with their intended research use; no human subjects, annotators, or private data were involved. The probe identifies layers whose precision matters most for specific tasks, which could in principle be misused to target adversarial degradation; we view this risk as low because the same information already enables defensive mixed-precision allocation, which is the framing we adopt here. All experiments (including the 64-sub-layer sweep, multi-seed runs, cross-model replication, GPTQ probe, and Hessian-vs-restoration study) were run on a single NVIDIA A100 (80GB) via Lightning.ai; no training was performed, only inference-time quantization and surgical transplant, so per-condition cost is dominated by generation throughput (Section 3 and Appendix A).

REPRODUCIBILITY STATEMENT

Code, per-condition result files, and a one-command figure regeneration script are available at `https://anonymous.4open.science/r/quantization-mechanistic-probe` (anonymised for double-blind review). Section 3 specifies the model, quantization configuration, evaluation protocol, and statistical tests; Appendix A reports the multi-seed protocol and per-seed results. Each per-layer measurement starts from a fresh copy of the quantized model, so no ordering effects arise. Every reported recovery rate is computed from saved per-example correctness arrays, which are included in the repository so that all accuracy, recovery, and significance figures can be recomputed without re-running inference.

REFERENCES

Lorenzo Basile, Valentino Maiorca, Diego Doimo, Francesco Locatello, and Alberto Cazzaniga. Head pursuit: Probing attention specialization in multimodal transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2510.21518.

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.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in*

Neural Information Processing Systems (NeurIPS), 2023. URL https://arxiv.org/abs/2304.14997.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2305.14314.

Harshwardhan Fartale, Ashish Kattamuri, Rahul Raja, Arpita Vats, Ishita Prasad, and Akshata Kishore Moharir. Disentangling recall and reasoning in transformer models through layer-wise attention and activation analysis. *arXiv preprint arXiv:2510.03366*, 2025.

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. URL https://arxiv.org/abs/2210.17323. Published at ICLR 2023 as "OPTQ: Accurate Quantization for Generative Pre-trained Transformers".

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

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

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

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

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. The unreasonable ineffectiveness of the deeper layers. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2403.17887.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021a. URL https://arxiv.org/abs/2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving With the MATH dataset. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021b. URL https://arxiv.org/abs/2103.03874.

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

Sreejan Kumar, Theodore R. Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 15:5523, 2024. doi: 10.1038/s41467-024-49173-5.

Jemin Lee, Sihyeong Park, Jinse Kwon, Jihun Oh, and Yongin Kwon. Exploring the trade-offs: Quantization methods, task difficulty, and model size in large language models from edge to giant. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 8113–8121, 2025. doi: 10.24963/ijcai.2025/902. URL https://arxiv.org/abs/2409.11055.

Jason Li. Mechanistic interpretability of binary and ternary transformers. *arXiv preprint arXiv:2405.17703*, 2024.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024. URL https://arxiv.org/abs/2306.00978.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2202.05262.

Utkarsh Saxena, Sayeh Sharify, Kaushik Roy, and Xin Wang. ResQ: Mixed-precision quantization of large language models with low-rank residuals. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53095–53114, 2025. URL https://proceedings.mlr.press/v267/saxena25b.html.

Tim Shallice. *From Neuropsychology to Mental Structure*. Cambridge University Press, Cambridge, 1988. doi: 10.1017/CBO9780511526817.

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

Wenfang Sun, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. The curse of depth in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2502.05795.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2211.00593.

Keyu Wang, Tian Lyu, Guinan Su, Lu Yin, Marco Canini, Jonas Geiping, and Shiwei Liu. When fewer layers break more chains: Layer pruning harms test-time scaling in LLMs. In *Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2510.22228.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

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)*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Tianyi Zhang and Anshumali Shrivastava. LeanQuant: Accurate and scalable large language model quantization with loss-error-aware grid. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2407.10032.

A Multi-Seed Statistical Validation

Figure 4 shows the per-seed accuracy means (±std), the full McNemar p-value matrix, and the throughput chart that support the headline claims of Section 4.4.

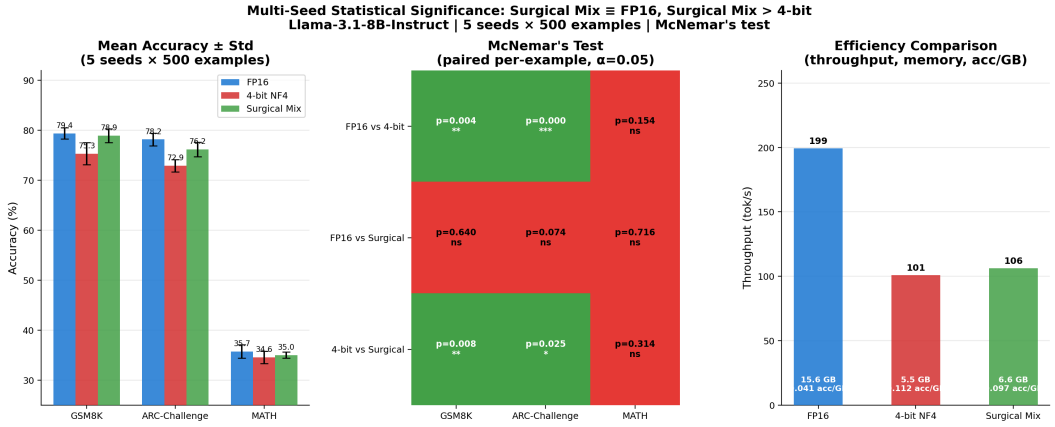


Figure 4: Multi-seed statistical validation (5 seeds, McNemar's test). *Left*: Mean ± std accuracy across seeds. *Centre*: McNemar p-values; green = significant (α=0.05), red = not significant (*ns*). Surgical ≡ FP16 (all *ns*); Surgical > 4-bit (GSM8K p=0.008, ARC p=0.025). *Right*: Throughput (tok/s); surgical matches 4-bit while recovering most of FP16's accuracy. Per-example Cohen's $d \leq 0.05$ on all three datasets.

B Full Layer Sensitivity Heatmap

Figure 5 shows the complete 2×3 recovery-rate heatmap: rows = Attention / MLP, columns = GSM8K / ARC / MATH. Each cell shows the layer index and recovery rate $r(\ell, \tau)$. Attention-ARC recovery is near-zero throughout the deep half, while attention-GSM8K and MLP-ARC carry the strongest cells; MLP-GSM8K rates are non-trivial but never clear the FDR threshold, which is what the analysis in Section 4.3 tests.

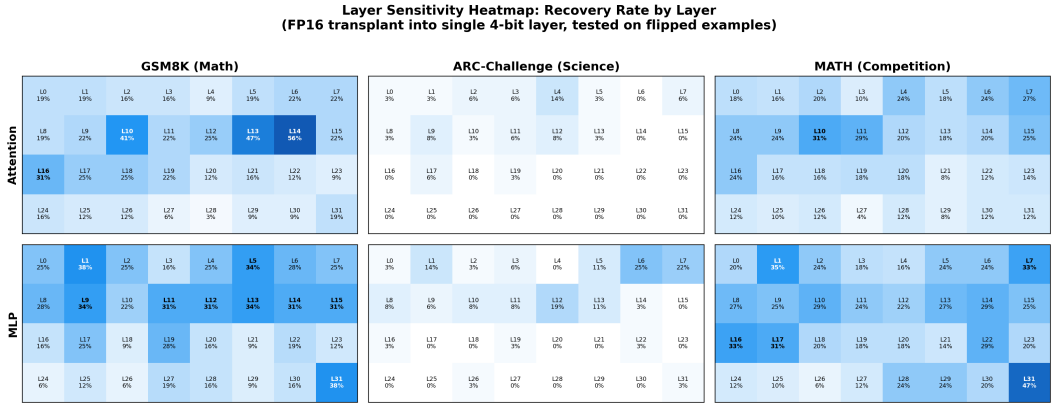


Figure 5: Full 64-cell layer sensitivity heatmap. Rows = Attention (top), MLP (bottom); Columns = GSM8K, ARC-Challenge, MATH. Colour intensity = recovery rate (white→dark blue = 0→55%). Layer indices and recovery percentages shown in each cell. Critical cells highlighted by high recovery values: L13–14 (Attention/GSM8K), L6–7 (MLP/ARC), L31 (MLP/MATH).

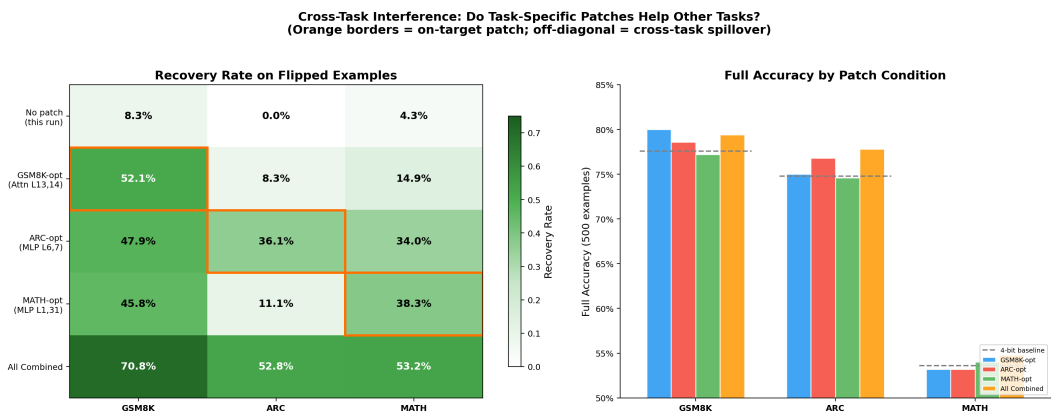


Figure 6: Cross-task interference matrix. *Left*: recovery-rate heatmap (5 patch configurations × 3 datasets); orange borders mark on-target patches. *Right*: full-accuracy bars per condition (dashed grey: 4-bit baseline). The GSM8K-optimal attention patch transfers only weakly to ARC (8.3%); the ARC-optimal MLP patch transfers strongly to GSM8K (47.9%).

Table 3: Cross-model replication: Llama-3.1-8B vs. Mistral-7B. McNemar's test (5 seeds for Llama-3.1-8B, 3 seeds for Mistral-7B, 500 examples). *ns* = not significant; * = $p<0.05$; ** = $p<0.01$.

Model	Comparison	GSM8K	ARC	MATH
Llama-3.1-8B	Surgical vs FP16	*ns*	*ns*	*ns*
	Surgical vs 4-bit	**	*	*ns*
Mistral-7B	Surgical vs FP16	*	*	*ns*
	4-bit vs FP16	*ns*	*ns*	*ns*
Critical layers (attn)		L13, L14	L14	—
Critical layers (MLP)		L1,L6,L7,L31	L6,L7	—

C Cross-Task Interference Matrix

Figure 6 is the visual form of the transfer numbers quoted in Section 4.3; every value shown appears in that text. Its top row is this experiment's own unpatched 4-bit pass, included for scale. It is a single run at this experiment's settings, and is not the replicated do-nothing null used for the FDR tests in Section 4.3, which is measured separately.

D Surgical Model Capstone Results

Figure 7 presents the three-panel capstone: accuracy comparison (FP16 / 4-bit / Surgical), percentage of the FP16–4-bit gap recovered per dataset, and memory footprint. This is a single run at seed 42 (FP16 80.0/79.2/55.2, 4-bit 77.6/74.8/53.8, Surgical 79.4/77.8/54.8), closing 75.0% of the GSM8K gap, 68.2% of ARC and 71.4% of MATH at 42% of FP16 memory. Two quantities elsewhere in the paper are deliberately not the same as these. The 5-seed means of Section 4.4 come from a separate run that resamples every dataset per seed, including MATH; this capstone uses the first 500 MATH problems, which the dataset orders by subject, so its MATH column covers Algebra alone. And the 70.8/52.8/53.2 figures quoted in Section 4.3 are recovery on *flipped examples*, not percentage of the accuracy gap closed.

E Cross-Model Comparison: Llama-3.1-8B vs. Mistral-7B

Figure 8 shows the full cross-model comparison: capstone accuracy, per-layer attention sensitivity (GSM8K), per-layer MLP sensitivity (ARC), cross-task interference matrix, and summary statistics

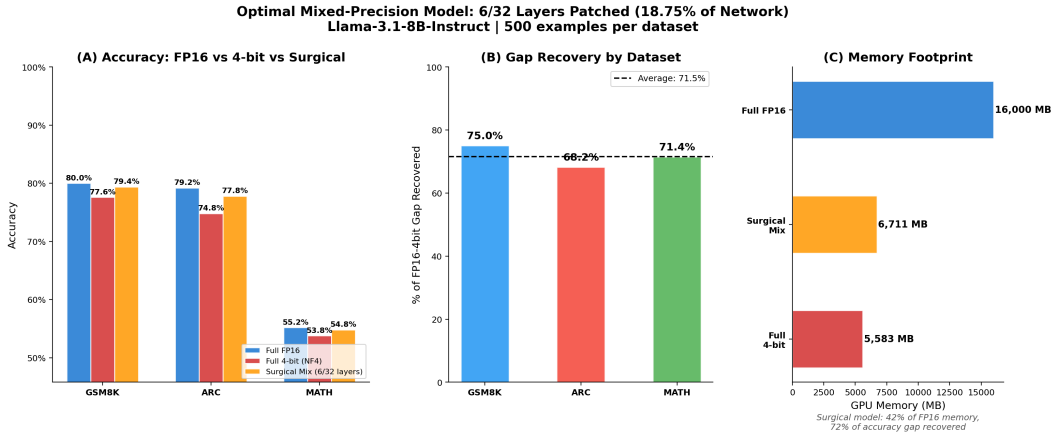


Figure 7: Optimal mixed-precision capstone. *Left*: Absolute accuracy for FP16, 4-bit, and Surgical (6/32 layers). *Centre*: Percentage of FP16–4-bit gap recovered per dataset. *Right*: GPU memory footprint. Surgical model memory (6.71 GB) is only 1.13 GB above pure 4-bit (5.58 GB).

tables. Both models share critical layer structure (L14 attn, L6–7 MLP), despite having different absolute accuracy levels and quantization behaviour.

F MATH Difficulty Stratification

Figure 9 stratifies MATH examples by difficulty tier (L1–3 = Easy, L4–5 = Hard). Easy tier: FP16–4-bit gap = 3.4pp (p=0.086, *ns*). Hard tier: gap = 1.4pp (p=0.476, *ns*). The *smaller* gap on Hard problems, where both models fail for capability reasons, indicates that MATH's hardest tier is capability-limited; the precision-sensitive signal the sweep detects (Section 4.1) therefore comes from the easier tiers.

G Bidirectional Flip Analysis

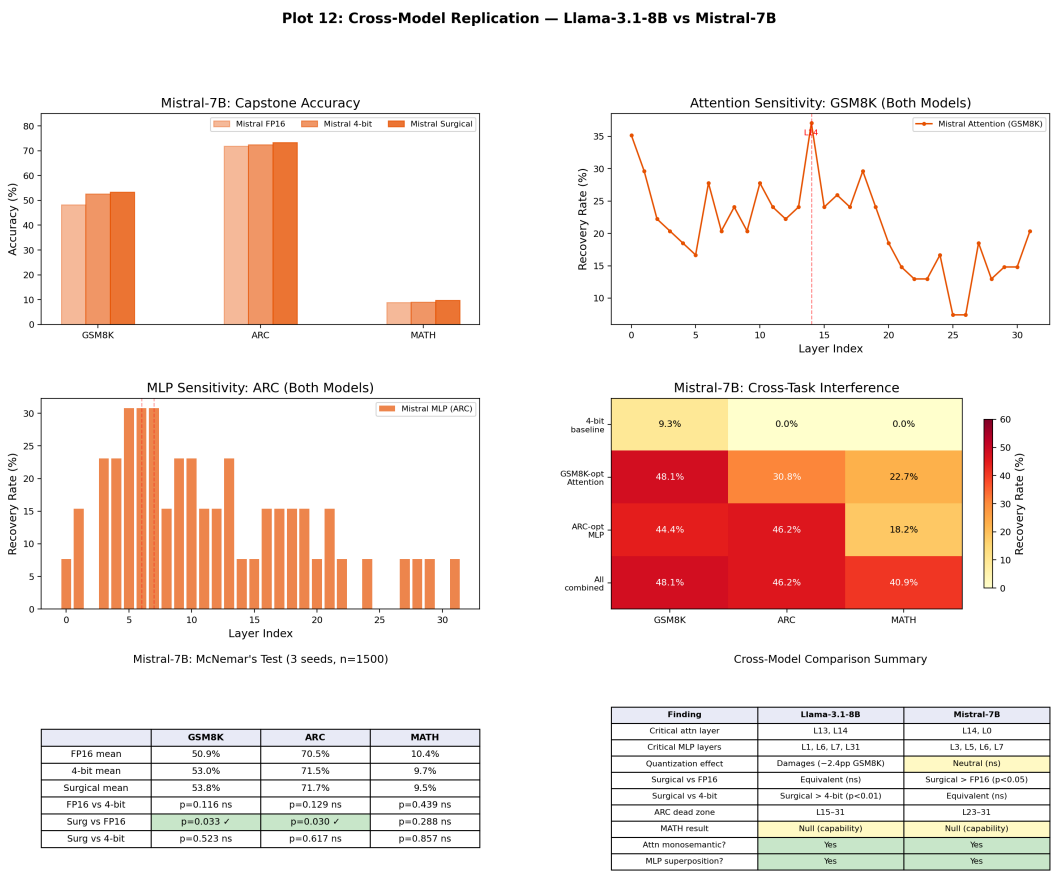
Figure 10 shows the four-quadrant breakdown of all 500 test examples per dataset: both-correct, forward-flipped (FP16→4-bit, quantization hurts), reverse-flipped (4-bit→FP16, quantization helps), and both-incorrect. Net forward flips: GSM8K = +31, ARC = +36, MATH = +24. Flipped examples have the same FP16 confidence as stable-correct examples (mean gap = −0.003), ruling out boundary perturbation as an explanation.

H Cross-Quantization-Method Replication (GPTQ): Full Detail

The NF4 restoration sweep assumes that the functional map (which layers are precision-critical for which tasks) reflects a property of the model architecture, not an artifact of NF4's specific quantization mechanism. We test this by running an identical 64-layer restoration sweep on a GPTQ-quantized baseline Frantar et al. (2023) (hugging-quants/Meta-Llama-3.1-8B-Instruct-GPTQ-INT4, group size 128, `desc_act=True`, Triton v2 kernel, n=500 examples per dataset).

Figure 11 compares NF4 and GPTQ recovery profiles across all 32 layers for attention and MLP components. The key dissociation replicates.

Attention task-selectivity. L14 attention remains the top-ranked layer for GSM8K under GPTQ (46.4% recovery), against the NF4 value of 56.2%. The differential persists: L14 attention ARC recovery under GPTQ is only 3.6% (1/28 flipped examples), as near-zero as the 0.0% it reaches under NF4. The attention dead zone (L15–31 for ARC) partially replicates: layers 18, 21, 23, 25, 29–31 all show 0% ARC recovery under GPTQ.



	GSM8K	ARC	MATH
FP16 mean	50.9%	70.5%	10.4%
4-bit mean	53.0%	71.5%	9.7%
Surgical mean	53.8%	71.7%	9.5%
FP16 vs 4-bit	p=0.116 ns	p=0.129 ns	p=0.439 ns
Surg vs FP16	p=0.033 ✓	p=0.030 ✓	p=0.288 ns
Surg vs 4-bit	p=0.523 ns	p=0.617 ns	p=0.857 ns

Finding	Llama-3.1-8B	Mistral-7B
Critical attn layer	L13, L14	L14, L0
Critical MLP layers	L1, L6, L7, L31	L3, L5, L6, L7
Quantization effect	Damages (−2.4pp GSM8K)	Neutral (ns)
Surgical vs FP16	Equivalent (ns)	Surgical > FP16 (p<0.05)
Surgical vs 4-bit	Surgical > 4-bit (p<0.01)	Equivalent (ns)
ARC dead zone	L15–31	L23–31
MATH result	Null (capability)	Null (capability)
Attn monosemantic?	Yes	Yes
MLP superposition?	Yes	Yes

Figure 8: Cross-model replication: Llama-3.1-8B (blue) vs. Mistral-7B (orange). *Top row*: Capstone accuracy and per-layer attention sensitivity. *Middle row*: MLP sensitivity and cross-task interference matrix. *Bottom row*: McNemar significance tables for both models. Both architectures identify L14 attention and L6–7 MLP as critical layers.

MLP cross-task structure. Under GPTQ, early MLP layers remain cross-task helpers: L6 recovers 46.4% GSM8K and L7 recovers 28.6% ARC, the same pattern as NF4 (28.1% and 22.2%). A notable shift is that layers L0–L1 MLP become more prominent under GPTQ (35.7% ARC each), likely reflecting GPTQ's calibration redistributing residual error toward earlier layers. The core cross-task property (MLP recovery transfers across tasks in both directions) is preserved.

Interpretation. The replication across fundamentally different quantization algorithms (NF4: scalar round-to-nearest with double quantization; GPTQ: Hessian-calibrated group quantization with activation-order permutation) supports our claim that the functional map is a structural property of the model, not an NF4 artifact. The quantitative shift in recovery magnitudes is expected (GPTQ has different per-layer error distributions), but the qualitative dissociation (L14 attention arithmetic-selective; MLP L6–7 cross-task) is stable.

I Extended Multi-Model Probe: Full Detail

We extend the restoration probe to three additional models: DeepSeek-R1-Distill Guo et al. (2025) (Llama-8B distillation of a large reasoning model), Qwen2.5-Math-7B-Instruct Yang et al. (2024) (a math-specialised model), and Gemma-2-9B-IT Gemma Team et al. (2024). The results reveal two qualitatively distinct behaviours.

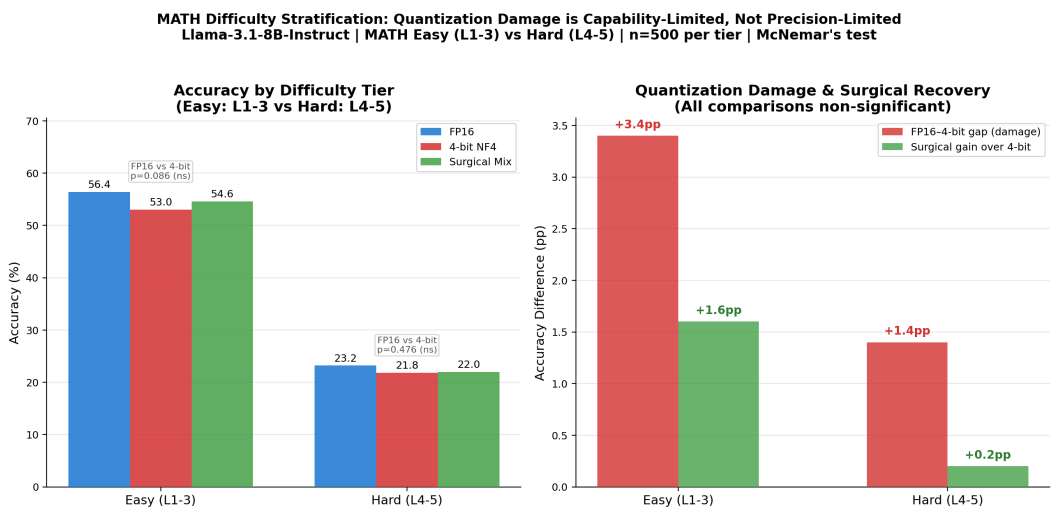


Figure 9: MATH difficulty stratification. All comparisons are non-significant (McNemar's test, n=500 per tier). Hard problems show *smaller* FP16–4-bit gap than Easy problems, because both models fail hard problems for capability reasons, not precision.

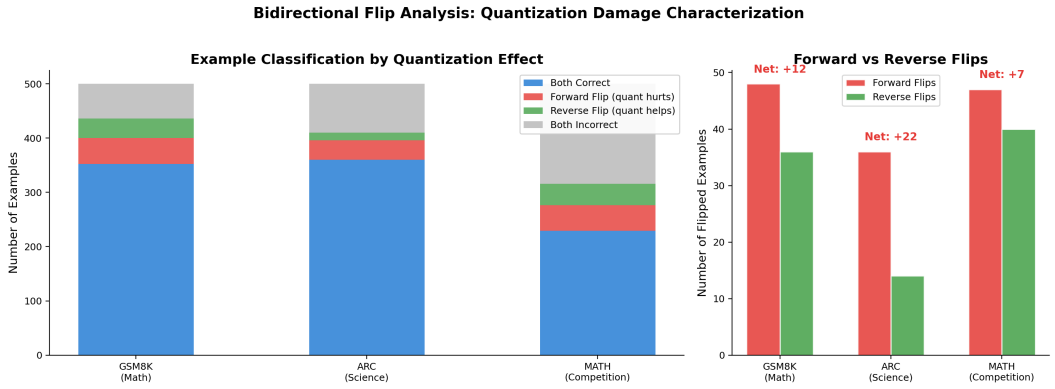


Figure 10: Bidirectional flip analysis. *Left*: Stacked bar showing four-quadrant breakdown per dataset. *Right*: Forward vs. reverse flip counts with net annotations. Forward-flipped examples have the same FP16 confidence as stable-correct examples, confirming systematic (not boundary) corruption.

Replication (DeepSeek-R1). DeepSeek-R1-Distill suffers 2.4pp GSM8K and 5.0pp MATH NF4 damage (FP16: 64.8%/46.6%; NF4: 62.4%/41.6%). The restoration sweep confirms L13 attention as the top-ranked layer for GSM8K (recovery rate = 42.6%), replicating the L13–14 arithmetic-attention finding under a different training regime. Being a Llama-8B distillation, this holds the architecture fixed and varies training, so it is a training-regime control rather than a cross-architecture one. We do not report ARC for this model: its FP16 ARC accuracy is 4.6%, far below the four-option chance level, so its flip-set is not interpretable.

Quantization-resistant models (Qwen2.5-Math, Gemma-2). Qwen2.5-Math-7B-Instruct exhibits near-zero NF4 damage on all tasks ($\leq$1.0pp). Gemma-2-9B similarly shows no significant degradation (GSM8K: $-$0.8pp; ARC: +0.8pp; MATH: 0.0pp). With minimal quantization damage there is no recovery signal to probe; the sparse flip sets (n<10) yield noisy recovery estimates unsuitable for ranking. This *quantization resistance* is itself a finding: model architecture and training regime affect precision sensitivity, and the restoration probe correctly diagnoses the absence of functional degradation rather than producing spurious rankings. Together, the three-model extension shows that L13–14 arithmetic-attention criticality is a property shared across architecturally diverse models that

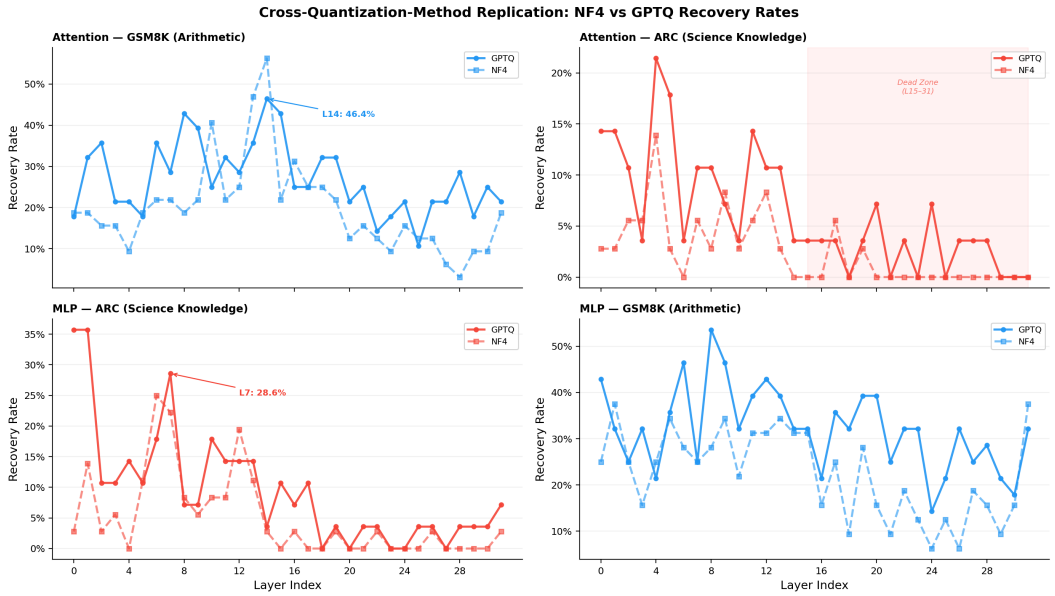


Figure 11: NF4 (dashed) vs. GPTQ (solid) recovery rates for all 32 layers. *Top*: Attention; *Bottom*: MLP. Left panels = GSM8K (arithmetic); Right panels = ARC (science knowledge). L14 attention remains the top-ranked layer for arithmetic in both methods (46.4% GPTQ vs. 56.2% NF4). The ARC dead zone (L15–31, shaded) partially replicates under GPTQ.

exhibit quantization damage, while quantization-robust models fall outside the probe's operating regime.

J Mistral Hessian-vs-Restoration Replication

We rerun the ranking comparison on Mistral-7B using the same methodology as on Llama-3.1-8B: 128 WikiText-2 calibration sequences for the activation-covariance Hessian proxy, and per-sub-layer restoration recovery rates from the cross-task sweep already reported in Section 4.6. Across all 64 Mistral-7B sub-layers (32 attention, 32 MLP), the Spearman correlation between Hessian rank and restoration rank is $\rho{=}{-}0.574$ ($p{<}0.001$, $n{=}64$), with per-task breakdowns of $\rho_{\text{GSM8K}}{=}{-}0.458$ ($p{<}0.001$), $\rho_{\text{ARC}}{=}{-}0.536$ ($p{<}0.001$), and $\rho_{\text{MATH}}{=}{-}0.342$ ($p{=}0.006$). All three task-specific correlations are negative and significant, mirroring the Llama-3.1-8B pattern in which ARC shows the strongest inversion. The aggregate magnitude on Mistral-7B ($|\rho|{=}0.574$) is in fact larger than on Llama-3.1-8B ($|\rho|{=}0.497$), despite Mistral-7B's grouped-query attention and substantially different baseline accuracy, confirming that the Hessian–functional decoupling is not a Llama-3.1-8B-specific artifact but a structural property of Pre-LN transformers under 4-bit quantization.

K Matched-Budget Baselines: Rand-6 and Earliest-6-MLP

Table 4 reports the two additional matched-budget controls of Section 4.8 on ARC-Challenge. These were evaluated in a separate run with its own baselines (FP16 78.8%, NF4 74.6%, 38 forward-flipped examples), so absolute recovery rates differ slightly from Table 2; every comparison within the table is within-run and unaffected. All conditions hold exactly 6/32 sub-layers at FP16 and are evaluated identically to REST-6. Recovery rates use the same definition as elsewhere (fraction of forward-flipped examples answered correctly), with bootstrap 95% CIs over 10,000 resamples. Intervals are Monte Carlo estimates: REST-6's interval is recomputed within this run alongside the other conditions, so it can differ from Section 4.8 in the final decimal.

Budget is matched at sub-layer granularity rather than parameter count: an MLP sub-layer holds roughly four times the weights of an attention sub-layer, so EARLIEST-6-MLP in fact carries $1.34\times$

REST-6's FP16 parameters and is not disadvantaged. The random draws recover in proportion to whether they happen to include one of REST-6's layers: the only draw that is not significantly below REST-6 (s105, 39.5%) is the one containing MLP L6, while the draw containing no REST-6 layer (s101) recovers least.

Table 4: Matched 6/32 FP16 budget on ARC-Challenge. Recovery with bootstrap 95% CI; p-values are McNemar's against NF4, FP16, and REST-6. FP16 params is the weight count held at full precision.

Condition	FP16 params	Acc.	Recovery	vs. NF4	vs. FP16	vs. REST-6
REST-6	788M	77.6	**52.6%** [36.4, 68.6]	0.009	0.377	—
EARLIEST-6-MLP	1057M	76.6	36.8% [21.2, 52.4]	0.100	0.145	0.473
RAND-6 (s101)	654M	74.2	10.5% [2.3, 21.2]	0.774	0.002	0.007
RAND-6 (s102)	654M	75.0	28.9% [15.0, 44.4]	0.845	0.009	0.026
RAND-6 (s103)	520M	73.6	21.1% [8.6, 34.6]	0.441	0.001	0.006
RAND-6 (s104)	654M	74.4	15.8% [5.1, 27.8]	1.000	0.002	0.003
RAND-6 (s105)	788M	77.0	39.5% [23.7, 55.6]	0.023	0.200	0.664

L VISUALISATION OF HESS-6 VS. REST-6

Figure 12 visualises the per-condition accuracies and recovery rates of Table 2. All numerical values shown are reproduced in the table; this figure is provided for visual comparison.

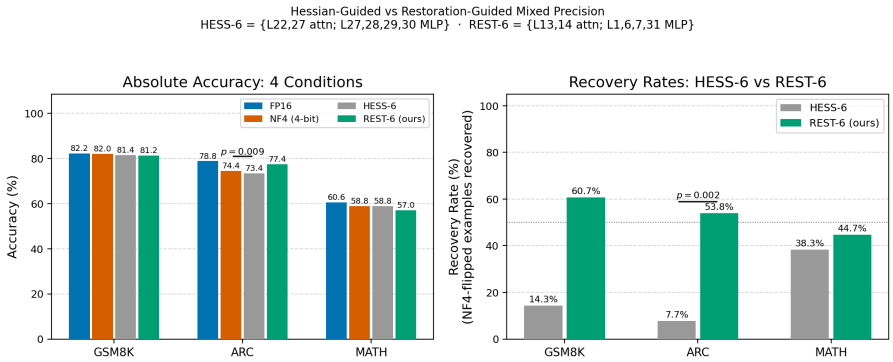


Figure 12: HESS-6 (grey) vs. REST-6 (green) at matched 6/32 FP16 budget. *Left*: absolute accuracy; *Right*: NF4-damage recovery rates. On ARC, REST-6 recovers 53.8% (p=0.009 vs. NF4) vs. HESS-6 7.7% (p=0.227, ns); direct comparison p=0.002.