Quantization damage is precision-dependent and capability-asymmetric. Component-sensitivity work ranks; it does not test.

Rishik Chakraborty
r36chakr@uwaterloo.ca

Abstract

Quantization is widely reported to be nearly free at 4 bits, on the evidence of aggregate benchmark scores. We ask whether that average conceals capability-specific damage, using two capabilities — long-context retrieval and multi-step arithmetic — measured on the same model, the same component set, and the same paired per-item design.

We find that damage is **precision-dependent and asymmetric**. At 4 bits neither capability moves detectably, under either quantizer we measure: NF4 gives retrieval dz +0.221 (1.01× the measured noise band) and arithmetic dz −0.079 at n=96, and `int_group@4` gives +0.205 and +0.164 at n=48 — **confirming the "nearly free" claim rather than refuting it**. At 3 bits — reachable only with `int_group`, since NF4 has no 3-bit form — both capabilities break, and retrieval takes **3.2× the standardised damage of arithmetic** (§3). This asymmetry requires no component map, no multiplicity correction, and no significance gate; it is measured from whole-model interventions on identical items.

Localising that damage to components is much harder than it appears, and most of this paper is about why. We show that (i) a significance gate defined as an absolute threshold on a raw metric — the standard noise-floor construction, which we pre-registered — **cannot compare two capabilities against the same bar**, and structurally excluded one of ours from ever registering an effect, and **reverses its verdict on a replication** of an effect that reproduces at p = 0.0001 (§4.1); (ii) the natural model for extrapolating component effect sizes is **wrong by an order of magnitude**, conservatively, and we falsify it twice from independent directions (§4.2); (iii) a documented, defensible-sounding decision in our primary test consumed the wrong quantity, and we exhibit the eight measured components that would have inverted its reading (§4.3); and (iv) every effect size we selected as an extremum **regressed on re-measurement — three for three**, once across a pre-registered decision threshold (§4.4).

We report the resulting 3-bit component map with its coverage stated: adequately powered for **21 of 56 components (37%)** (§5). Within it, **eight components improve retrieval rather than damaging it**, and they do not cluster: not by depth (Mann–Whitney p = 0.96), not by module type (Fisher p = 1.00), and across all 56 components layer index carries no information about the *direction* of the effect (Spearman $\rho = -0.051$) — so quantization helping is not a property of a region or a module kind (§6). Our primary pre-registered test of whether the two capability maps differ is **inconclusive** (§5), and we report it as such.

1 Introduction

Quantization is reported to be nearly free. Aggregate benchmark scores barely move at 4 bits, and that finding underwrites a great deal of deployment practice. But an aggregate is an average, and an average can hide the destruction of a specific capability while the mean stays flat.

The concern is not hypothetical in the method itself. Existing approaches to allocating precision decide what to protect by asking which weights most damage **perplexity** — a global average over all tokens. If different capabilities depend on different parts of a model, optimising a global average is the wrong objective, and a method that protects the weights perplexity cares about may be protecting the wrong ones for any particular deployment.

We therefore ask: **do different capabilities depend on different parts of the model, and if so, can precision be allocated by capability rather than uniformly?** Both answers are results. If the maps differ, a model can be compressed differently depending on what it is deployed for. If they coincide, there is a shared critical set — a stronger and more surprising claim about how models are organised. Testing one capability alone is not a result: *"some layers matter more than others"* is true of any capability measured in isolation, and it is the **contrast** that carries information.

Our first finding is that the premise needs correcting. At 4-bit NF4 on our model, neither capability moves detectably — retrieval at 1.01× the measured noise band, arithmetic not at all. **We confirm the "nearly free" claim rather than refuting it.** The damage is real, but it is real at 3 bits, where both capabilities break and retrieval takes **3.2× the standardised damage** of arithmetic. Quantization damage is **precision-dependent and capability-asymmetric**, and the asymmetry is measurable in aggregate before any component map is drawn.

Our second finding is that localising that damage is much harder than the literature suggests, and most of this paper is about why. Component-sensitivity work in quanti-

zation computes a score and ranks it. Across every method we surveyed, none reports a significance test, a noise floor, a multiplicity correction, a power analysis, or a replication. When we import that machinery — as a study comparing two capabilities must — three things follow that a ranking cannot surface: the natural significance gate turns out **non-comparable across capabilities** and unstable under **replication**; there is **no model** for the power of a component ablation, and both obvious candidates fail by measurement; and **selected effect sizes shrink**, once far enough to cross a threshold we had fixed in advance.

None of that machinery is novel. Significance testing, power analysis and selection correction are standard in neuroimaging, genetics and econometrics, and we cite those literatures rather than re-deriving them. **What we report is that they are absent here, and what their absence costs when the question is a comparison rather than a ranking.**

We report a 3-bit component map with its coverage stated — adequately powered for 21 of 56 components — and our pre-registered test of whether the two capability maps differ is **inconclusive**. We report that as inconclusive, and devote a section to the claims we do not make.

2 Setup, and what was fixed in advance

Model and capabilities. Qwen2.5-1.5B-Instruct, float32, on one GPU. Both capabilities are measured from the first run rather than added later, for the reason §1 gives. Long-context retrieval is needle-in-a-haystack at 4k tokens, scored by exact string match on inserted passcodes; multi-step arithmetic is GSM8K, scored on the final number. Both are boolean and grader-free.

Quantization is simulated and validated, not assumed. We round weights onto the grid an N-bit quantizer would produce and store them in the original dtype — standard practice, and the only way to intervene on a *single attention head*, which production libraries cannot do. Our NF4 implementation reproduces `bitsandbytes` **element for element with zero differing elements**, on Gaussians, on real weights, and on non-contiguous column views. Two independent intervention implementations — weight mutation and runtime materialization — are asserted to produce **bitwise-identical logits** on the real device before every run.

Component mapping under grouped-query attention. With `n_kv < n_q`, "quantize head h" has two non-equivalent definitions, and the wrong one is arithmetically well-formed and silently wrong. We implement both as distinct component types with disjointness asserted within each, and the mapping suite is **mutation-tested**: an `h // n_kv` group-map typo, an `o_proj` row/column swap, and a one-head offset fail 8, 5 and 2 tests respectively.

Two metrics, and what each is for. The search metric is teacher-forced NLL over the answer span — deterministic, continuous, one forward pass — measured **paired per item**, since item-to-item difficulty is large and pairing cancels it. The confirmation metric is boolean exact match with McNemar on discordant pairs. Both are recorded on every run so their relationship is measurable rather than assumed.

Difficulty is calibrated before sweeping, and the choice is enforced in code. A saturated eval has no headroom: if the unmodified model scores ~100%, no single-component intervention can produce a measurable effect, and the study would return "no effect" having built an instrument with no dynamic range. We therefore evaluate a ladder of needle variants and select the hardest whose unmodified accuracy lies in a pre-registered band, whose uniform-quantization drop is significant by McNemar, and whose NLL increase clears the measured noise floor. The chosen variant is **frozen for every later rung** — and the runner reads the calibration record and **refuses to start** if a config names a different variant, or a different model, context length, device or dtype. An earlier version of this project carried that requirement as a comment, and the comment was violated by the config directly beneath it.

The pre-registration, and one amendment to it. Decision rules were fixed in writing before any number was produced. Every subsequent change is recorded in an amendment log with its date, its reason, and an explicit check of whether it alters any prior verdict.

One amendment concerns this section directly. The headroom band's upper bound was pre-registered at 0.98. Measuring a second model, the frozen variant scored **47/48 = 0.97917** and *passed* — inside the band by **0.0008**, where one item at n=48 is **0.0208**. The rule discriminated at a resolution 26× finer than the measurement can represent, and one more correct item would have failed it. We moved the ceiling to **0.96**, verified against all 36 baselines in the project that **exactly one verdict changes — the run that exposed the problem** — and recorded that the more principled fix, expressing headroom in *items*, was rejected because it would have invalidated the n=12 screen that selected the variant this project is built on. That rejected rule is written down, unadopted, with its counterfactual.

3 Quantization damage is precision-dependent and capability-asymmetric

We quantize every linear weight in Qwen2.5-1.5B-Instruct to 2, 3 and 4 bits with a group-wise integer quantizer (group size 64) and measure two capabilities on identical items: long-context retrieval at 4k tokens, scored by exact string match on inserted passcodes, and multi-step arithmetic on GSM8K.

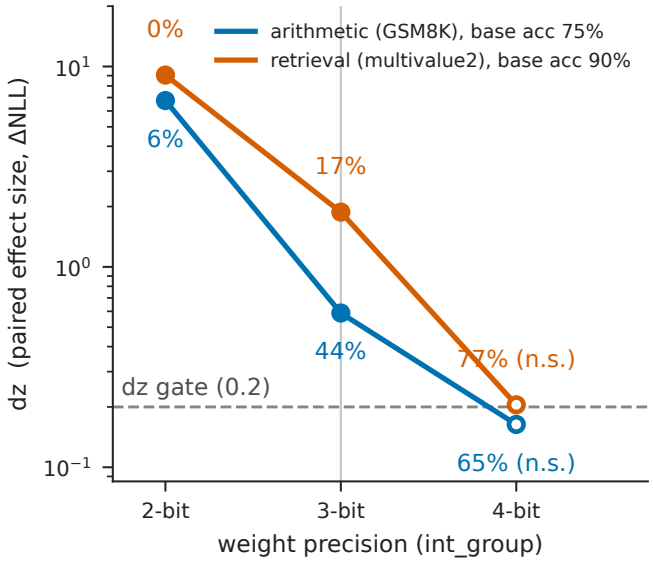

Figure 1: Precision ladder. Standardised effect (dz) and accuracy for both capabilities at 2, 3 and 4 bits under `int_group`, n=48 per rung. Both capabilities are intact at 4 bits and broken at 3; at 3 bits retrieval takes 3.2× the standardised damage of arithmetic. Open markers are McNemar non-significant at α=0.05. This is the paper's headline result and uses none of §4's machinery.

Effects are paired per item — Δ_i = NLL$_i$(quantized) − NLL$_i$(baseline) — and aggregated with a sign-flip permutation test. Baselines are bit-identical across the three separate processes that produced them, so the three rungs differ only in precision.

Both capabilities survive 4 bits and break at 3. All figures below are `int_group` (group size 64) at the stated width, n=48 per capability. Baseline accuracies are **0.896 (retrieval, n=48)** and **0.750 (arithmetic, n=48)**.

At 4 bits neither capability moves detectably, and we state precisely which hurdle each fails. Retrieval's dz is **+0.205 — fractionally *above* the 0.2 practical-significance threshold, not below it** — and it fails on significance instead: permutation p = 0.158, McNemar p = 0.109. Arithmetic fails both hurdles: dz +0.164 below the threshold, permutation p = 0.308, McNemar p = 0.267. **This confirms the widely reported claim that 4-bit quantization is nearly free.** We report it as a confirmation, because the framing that motivated this work predicted the opposite.

At 3 bits both break, and they break **unequally**: retrieval takes **3.2× the standardised damage** of arithmetic (dz +1.876 against +0.589). Integer bit widths cannot separate the two by break *point* — both are intact at 4 and broken at 3 — but the magnitude separates them decisively at the width where both move.

This result requires none of the machinery the rest of the paper is about. It uses no component map, no multiplicity correction, no significance gate, and no importance vectors. It is four whole-model interventions on two capabilities, measured on the same items. Every methodological finding in §4 leaves it standing.

The quantizer is not the deployed one, and we anchor rather than assume. `nf4` — the format the 4-bit literature uses — has no 3-bit form, so the ladder uses a parameterisable integer quantizer. At 4 bits, where both exist, we compare them **paired on identical items** against a bit-identical baseline: the difference is −0.0064 for retrieval (p = 0.816, n=48) and +0.0090 for arithmetic (p = 0.729, n=24). Whole-model NF4 at n=96 gives retrieval dz **+0.221** and arithmetic **−0.079**, against `int_group`@4's +0.205 and +0.164 at n=48 — **the arithmetic pair differs in sign, and neither estimate is distinguishable from zero**, which is the whole content of the comparison. **These are failed rejections, not equivalence** (§7), and we do not claim the formats are interchangeable — only that we cannot distinguish them at the one width where the comparison is possible.

Component importance may not transfer across precision — a preliminary result. Because the map is drawn at 3 bits and deployment is at 4, the question of whether the *ordering* survives a precision change is the one a practitioner would ask first. We report the only direct measurement we have of it, with its limits attached rather than after them. Remeasuring six components at 4-bit NF4 on the same 96 items, **two of the six reverse sign**, and the sharpest case is the worst one: `attn_block__L12` carries the strongest dz in the entire 3-bit top group (+1.349) and moves the *other way* at 4 bits (dz −0.477). A 3-bit map used to decide what to protect in a 4-bit deployment would rank that component near the top and be wrong about its direction. **The limits are severe and we state them with the finding**: n = 6 components, Spearman ρ = **+0.600** — positive, consistent with *partial* transfer, and **not significant** (six points need $\rho \approx 0.83$ at p < 0.05), with two of the six reversals sitting between values that are near-null at both precisions. We report it because it is preliminary evidence on a question this literature has no evidence on at all, and because the alternative — placing it among the limitations — invites it to be read as a caveat we are conceding rather than a measurement we made. **It is not the pre-registered rank-transfer test**, which was a 5-vs-5 group contrast and was formally withdrawn on its own cost criterion (§8).

A bit width is not a fixed amount of damage across model scales. Repeating the 3-bit intervention on Qwen2.5-3B — on byte-identical items, the two models sharing a tokenizer — destroys both capabilities (retrieval 0.979 → **0.000**, arithmetic 0.896 → **0.042**, both n=48), placing the 3B at 3 bits in the 1.5B's *two*-bit regime. **The precision at which a capability breaks is itself model-dependent.** We therefore do not report a cross-model asymmetry ratio: comparing the two at a shared bit width compares a break point against a floor. This also raises a question about designs that compare sensitivity across model scales at fixed format — one mea-

capability, `int_group`	4 bits	3 bits	2 bits
retrieval, dz	+0.205	**+1.876**	+9.061
retrieval, accuracy (n=48)	0.896 → 0.771	**0.896 → 0.167**	0.896 → 0.000
arithmetic, dz	+0.164	**+0.589**	+6.759
arithmetic, accuracy (n=48)	0.750 → 0.646	**0.750 → 0.438**	0.750 → 0.062

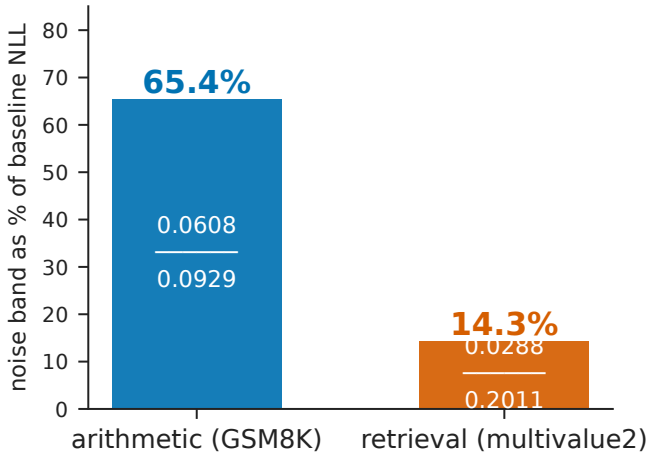

Figure 2: Band as % of baseline. The pre-registered noise band expressed relative to each capability's baseline NLL: 65.4% for arithmetic against 14.3% for retrieval, a 4.6× difference in relative stringency. Both terms are read from the same run's own noise floor, asserted in code — the earlier version of this figure took them from different runs, at different n on different item sets, and plotted 51.9% / 14.7% (§4.5).

surement on one pair of models, and a question rather than a refutation.

4 Four ways to measure this wrong, and what one guard was worth

This section reports four defects in how measurements of this kind are made. None is a statistical discovery, and we claim none as one; all four were found by machinery this field does not use. **§4.5 measures what one piece of that machinery was worth** — a guard added for one purpose that caught an unrelated error two days later, in a part of the pipeline no test covered. We state that here so the four findings read as measured rather than urged.

4.1 A gate that cannot compare two capabilities, and reverses on replication

Localising damage requires deciding which per-component effects count. The construction we pre-registered (§2) is the one the brief specified: bootstrap the baseline mean over ≥20 items per capability, and take **half the width of its 95% confidence interval** as a noise band. An effect falling inside the band is not a finding.

It cannot do the job. The band is an *absolute* threshold on the raw metric, derived per capability from that capability's own baseline variability. The two capabilities therefore face different bars:

The gate required arithmetic to move its NLL by **65% of its own baseline value** to register a finding, against **14%** for retrieval — a 4.6× difference in relative stringency that nothing in the research question asks for. It follows from the construction: a capability whose baseline mean is less precisely known receives a *larger* band, hence a *harsher* absolute bar. But "how precisely is the baseline mean known" is not "how large must an effect be to matter." The consequence is not a matter of degree — simulated against the measured per-item spreads, the arithmetic gate has power **0.00 at every n up to 1536 and every effect fraction**, its band sitting 6.7× above the largest effect present. **One capability was gated; the other was structurally excluded from ever registering anything**, in a study whose object is the comparison of the two.

The same gate also reverses on a replication. `mlp_block__L27` improves retrieval NLL by −0.0330 (dz −1.468) at seed 2024; re-measured on a **disjoint** item set at seed 2025 — zero shared item ids — it returns −0.0295 (dz −1.194) at p = 0.0001.

The effect replicates; the verdict does not. Nothing changed but which documents were drawn — that draw produced a wider baseline CI, and the band moved above the effect. **A threshold derived from the baseline sample is a property of the sample, not of the effect it gates.**

We replaced it before drawing any map, with a fixed standardised threshold ($|dz| \geq 0.2$), adopted **13 minutes before the first map was computed** and with no ranking or significance verdict produced under either gate beforehand. Both verdicts are reported for every component throughout, so the amendment's effect is visible rather than asserted.

Neither finding is a statistical discovery, and we do not present them as one. That an absolute threshold is not comparable across conditions, and that a sample-derived one is unstable, is textbook. What is notable is that this literature has never had to notice — §9 documents the absence across every method we surveyed. **When testing is imported into a field that ranks, the first gate one would reach for turns out non-comparable, unstable under replication, and to leave most of the resulting map inadequately powered.**

(n=96)	baseline NLL	band	band as % of baseline	largest component effect	band ÷ effect
retrieval	0.2011	0.0288	**14.3%**	+0.0641 (`mlp_block__L13`)	**0.45**
arithmetic	0.0929	0.0608	**65.4%**	+0.0091 (`attn_block__L22`)	**6.69**

item draw	ΔNLL	band	band gate	dz gate
seed 2024, n=96	−0.0330	0.0288	**PASS**	PASS
seed 2025, n=48	−0.0295	**0.0429**	**FAIL**	PASS

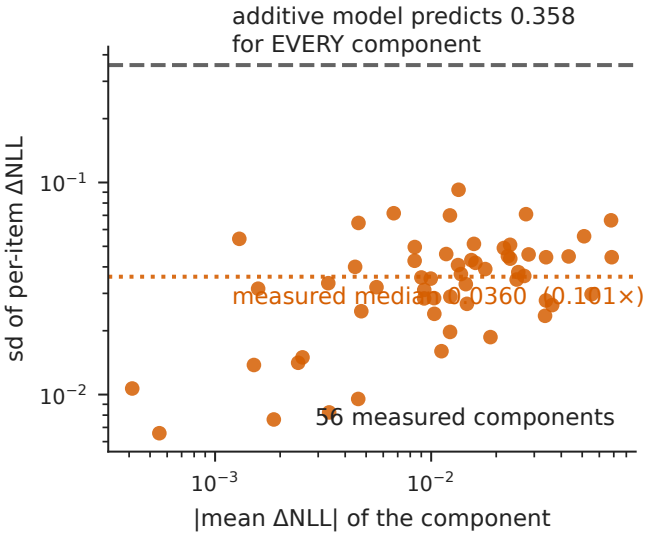

Figure 3: Additive model falsified. Measured per-item spread for all 56 components against the additive effect-scaling model's prediction, which is a single value (0.3577, the whole-model arm's spread) for every component. The measured median is 0.0360, a factor of 0.101. The error is conservative, so every "we lack power" conclusion drawn from that model is too pessimistic.

4.2 Component-ablation power has no model, and both obvious candidates fail

Deciding how many items a component sweep needs requires a model of how a component's effect — **and its item-to-item spread** — relate to the whole-model intervention. No such model exists in this literature, because this literature does not do power analysis. We do not present what follows as a correction to standard practice; **the absence is the finding.**

Two candidates are obvious. **Multiplicative** scaling, $d_i(f) = f \cdot d_i$, is degenerate because the sign-flip statistic is **scale-invariant**: it reports a component carrying 1% of the effect as *exactly as detectable* as the whole model. That is not an approximation error — the model cannot express the question. **Additive** scaling, $d_i(f) = f \cdot \text{mean}(d) + (d_i - \text{mean}(d))$, shrinks the signal and keeps the residual, implying that spread is invariant to the size of the intervention.

It is not, and we falsify it from two independent directions. Across 56 components at one precision, the median component's spread is **0.101**× the whole model's (0.0360 against the whole-model arm's 0.3577); across the same components at two precisions, spread scales by **0.357**× from 4 bits to 3. Spread shrinks with component granularity *and* with precision. The error is **conservative** — overstating spread understates dz and so understates power — so every "we lack power" conclusion drawn from such a grid is too pessimistic.

What this leaves. A third model must let spread shrink with intervention size, and the two measured ratios are its constraints. **62 measured single-component effects** exist to fit and cross-validate one — the 56 of the 3-bit sweep plus six re-measured at 4 bits for the transfer pilot. We name the requirement and supply the constraints; we do not fit the model.

4.3 A comment-justified wrong quantity in the primary test

The hardest defect to find is a **documented decision that is defensible in general and wrong for the claim it feeds.** Our primary test correlates two capability importance vectors; both were built from `abs(mean_delta)` under an explicit comment: *"Sign is not meaningful for ranking: a component whose quantization HELPS is still an influential one."* **True of influence, wrong for a test asking whether two capabilities are damaged in the same places** — under absolute values a component that helps one and hurts the other pushes the two maps toward looking identical, manufacturing agreement out of opposite behaviour.

We caught it before either script ran, on that argument alone; the measured instances arrived afterwards, and there are **eight** of them — every component whose quantization improves retrieval while damaging arithmetic would have been counted as agreement (§6). We claim no changed verdict (§7). An audit for the same shape found two more, in the boolean path and in a silent default.

4.4 Selection bias in reported effect sizes

Scanning noisy measurements for the largest value and reporting its magnitude over-estimates it. This is established — neuroimaging peak inference states it directly and *Inference*

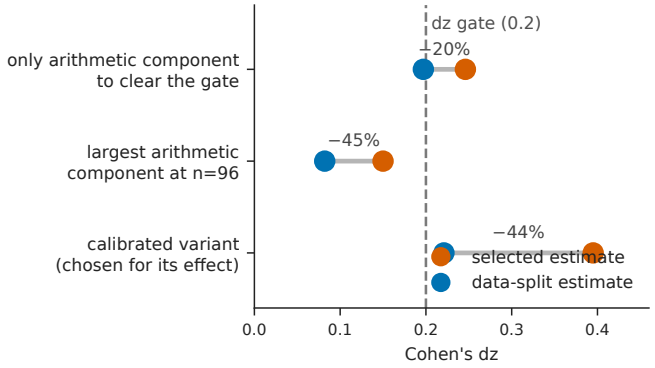

Figure 4: Selected extrema regress. Three selected-versus-data-split pairs with the pre-registered 0.2 margin drawn. Every selected extremum shrank on re-measurement; one pair crosses the margin, taking a component from DETECTED to NOT DETECTED on 0.003 in dz.

on Winners (Andrews, Kitagawa & McCloskey, QJE 2024) gives the canonical treatment — and **the structural match is exact**: that literature handles the correlated-maximum case, which is ours at 56 correlated components. The accepted correction is **data-splitting**; **our fresh-item re-measurements are that procedure**, with zero shared items.

We contribute the instances, not the phenomenon. The third row is the one that matters — it crossed a **pre-registered** threshold, DETECTED → NOT DETECTED, on 0.003 in dz — and this literature selects extrema, reports their magnitudes, and applies neither correction nor re-measurement.

Selection breaks the interval, not only the estimate. *Inference on Winners* makes the sharper point: selection leaves roughly a **90% chance of over-estimating**, and conventional intervals on selected quantities have **true coverage below 82%** at a nominal 95%. We therefore classified **every interval this paper reports** by whether its quantity was selected on the data the interval is computed from; **one of five is affected** (full audit, Appendix A). `mlp_block__L27`'s arithmetic dz interval **[+0.176, +0.319]** is computed on the data that selected it — the only component of 56 to clear the gate — and is **post-selection**; **[+0.119, +0.274]**, computed on disjoint items for what is by then a pre-specified component, is **valid**. We report the second as this effect's interval and present the first only as the selected estimate that prompted the re-measurement. Both are bootstrap **90%** intervals, so the quoted 95% coverage figure bounds the degradation rather than describing it.

Theory and measurement corroborate. These two results were obtained independently and agree. **Theory** says selection makes the interval both too narrow and centred too high, so correcting shifts the plausible range **downward**, toward and below the 0.2 margin; **measurement** on disjoint items landed at **dz 0.197**, below it. Neither was derived from the other — the QJE result predicts a direction without seeing our data, and the fresh-item run measures a value without invoking the theory. **Two routes, one theoretical and one empirical, reaching the same conclusion is stronger than either alone**, and materially stronger than the framing they replace, which was that the selected interval merely *straddled* the margin.

Why Holm does not substitute for this. Multiplicity correction does not correct effect estimates — winner's-curse adjustment shrinks Z-scores toward **zero**, multiplicity adjustment shrinks p-values toward **one**. Our Holm correction was never going to fix the estimates, which is why the re-measurements were necessary rather than redundant.

4.5 What a guard is worth, measured rather than asserted

This paper argues that experiments of this kind need machinery — assertions, provenance, pre-registration — that the surrounding literature does not use. That argument is easy to make and hard to evidence, because a guard that works produces nothing to point at. Here is one that produced something.

The guard. Noise floors were originally stored one file per (model, capability, variant, context length), with the run that measured them recorded *inside* the file. Two runs sharing those four attributes therefore wrote to the same path, and a later run silently replaced an earlier one's floor. The loading code required the run to match and raised — so nothing was ever gated by the wrong band — but the earlier run's floor was gone. We moved floors to `noise_floor/<run>/<key>.json`, putting the run in the **path**, so the collision cannot occur. That change was made for that reason and no other.

Two days later, an unrelated defect. The figure carrying this paper's §4.1 — the noise band as a percentage of baseline NLL — was computed by taking its **numerator** from one run's summary and its **denominator** from a *different* run's, at a different n on a different item set. It plotted **51.9% / 14.7%** where the correct values are **65.4% / 14.3%**.

Nothing caught it, for the reason that makes this defect class dangerous: **the bars looked plausible and still supported the claim.** A 51.9-versus-14.7 asymmetry argues §4.1 correctly while misstating its central number by 13.5 points. No assertion fired because none existed; the script read two files and divided.

The fix requires the guard. The corrected figure asserts that both terms come from the same run — by reading the `run` field of the floor it loaded and comparing it to the run requested. **Under the original flat layout that assertion could not have been written**, because the file's location carried no run and its contents were what the mix-up had already bypassed. A change made to prevent silent overwriting turned out to be what made an unrelated class of error checkable.

We report this because it is the honest form of the argument. "Provenance guards are worth their cost" is a claim we

what was selected	selected	data-split	shrinkage
the calibrated variant, chosen partly *because* its effect looked large	dz 0.395	0.221	−44%
the largest arithmetic component at n=96	dz +0.150	+0.082	−45%
the only arithmetic component to clear the gate	**dz +0.246**	**+0.197**	**−20%**

cannot support by having had no failures. What we can show is a guard added for one purpose catching a different error two days later, in a part of the pipeline — figure generation — that no test covered at all. We have since added tests that assert each figure's plotted values against their source files; the first version of those tests re-derived the numbers independently and **passed with the defect deliberately reintroduced**, which is the same failure one level up, and is why they now assert on what the plotting code returns.

5 The map, and how much of it we can see

With the gate replaced, we sweep all 56 components — attention and MLP blocks at every layer — at 3-bit precision, n=96, on both capabilities. Each component is quantized alone against a shared baseline (retrieval accuracy **0.917 at n=96**, arithmetic **0.729 at n=96**), paired per item, corrected across the sweep with Holm–Bonferroni.

Retrieval localises. 26 of 56 components clear the adopted gate. The structure is not confined to one region: early attention (`attn_block__L0`, dz +0.958), mid-depth attention (`attn_block__L12`, dz +1.349) and MLP blocks throughout the depth all clear it. **Eighteen of the 26 are damage and eight are improvement** — components whose quantization *lowers* retrieval NLL. §6 treats that class, which is larger than the single anomaly it is easy to mistake it for.

The two gates disagree on 16 of 56, and every disagreement runs the same way. The original band gate admits 10 components; the standardised gate admits 26; **no component is admitted by the band and rejected by dz**. The replacement is *strictly more permissive on this data*, which we state because it is the opposite of what a reader may expect from a rule adopted to fix a gate — and because it was predicted in writing, with the affected components named, before the map was computed.

Coverage is part of the result, not a caveat. Resampling each component's own measured per-item deltas — the plug-in bootstrap, with no effect-scaling model in it (§4.2) — **21 of 56 components (37%) are adequately powered at n=96**. Of the 39 whose effect clears the practical threshold, **18 are underpowered**. Of the 26 we report as significant, **five are provisional** — detected below 80% power, named individually in the results, and marked with dark edges in the component map. On the measured ladder, n=192 would cover 26 of the 39, n=384 would cover 32, and n=768 would cover 35. **An absent bar in this map is not evidence of an absent effect** (§7).

Arithmetic resolves almost nothing at n=96, and one component at n=384 — which does not survive re-measurement. At n=96 no component clears the gate on either metric, and equivalence testing per component against the pre-registered margin returns **0 detected, 7 equivalent to zero, 49 indeterminate** — the data excludes neither a real effect nor none. We therefore quadrupled that arm to n=384, which changes the picture: **1 detected, 48 equivalent, 7 indeterminate**.

The single detection is `mlp_block__L27` at dz +0.246 [+0.176, +0.319], and we report its data-split value in the same breath because it is the one number a reader would otherwise carry forward wrongly. That component was selected as the largest of 56, so its estimate is subject to the winner's curse (§4.4). Re-measured on **disjoint items** at the same n, it returns **dz +0.197 [+0.119, +0.274] — below the 0.2 margin**. The detection is therefore not one we carry: **the honest count for arithmetic at n=384 is zero components detected on unbiased re-measurement**, with 48 of 56 equivalent to zero within a standardised margin of 0.2. The equivalence result is the reportable one; the detection is an instance of §4.4's phenomenon, not a finding about the model.

We do not write that arithmetic does not localise. Against the largest effect actually present, the adopted gate has **0.225 power at n=96** — measured by resampling that component's own per-item deltas, not extrapolated through the effect-scaling model §4.2 falsifies. A pre-registered null is uninterpretable unless the gate's power against a stated alternative is known, so the reportable statement names it (§7).

The primary test is inconclusive. Spearman correlation between the two signed importance vectors over the shared component set is $\rho = -\mathbf{0.0014}$, against a permutation null of mean −0.0016 and sd 0.1365 — p ≈ 0.50 in both directions, with top-10 overlap 2. Neither pre-registered condition is met: *maps differ* requires ρ significantly below the null, and *maps coincide* additionally requires both capabilities to localise, which arithmetic does not. We report it as inconclusive and construct no third narrative. **What $\rho \approx 0$ does and does not license is stated in §7**, where the boundary on every claim in this paper is drawn in one place.

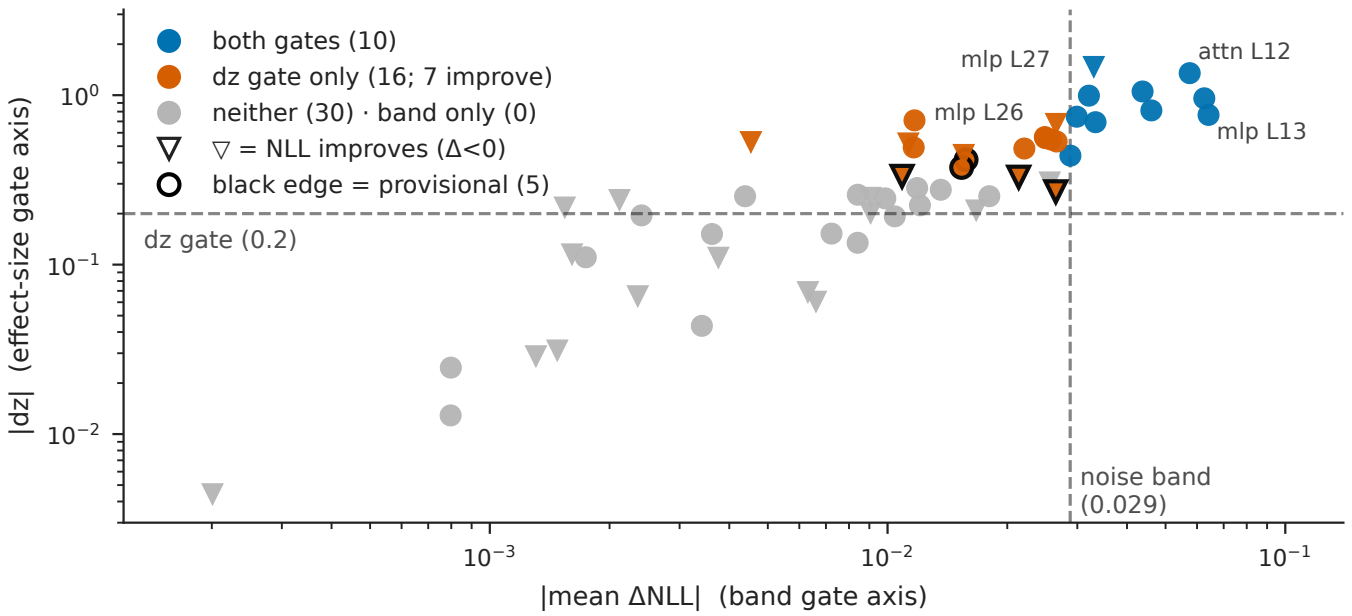

Figure 5: Gate disagreement. All 56 retrieval components under the original band gate and the adopted $|dz| \geq 0.2$ gate. The gates disagree on 16; every disagreement is admitted-by-dz and rejected-by-band, none the other way.

6 A class of components whose quantization helps, and its extreme case

Eight of the 26 components that clear the gate improve retrieval NLL rather than damaging it. This is a class, not an anomaly, and we state its size before examining its largest member, because the reverse order invites the reading that one odd component was found and the rest of the map is well-behaved. Eighteen components damage retrieval; eight improve it.

`mlp_block__L27` is the largest of eight, not the only one — 1.2× the next by raw effect and **2.8× the next by dz**, and the largest |dz| in the entire map in *either* direction, exceeding the largest damaging component (`attn_block__L12`, +1.349). It is an extreme case of a class, which is a stronger claim than a singular anomaly and a weaker one than a mechanism.

They do not cluster, and we tested three ways that they might. By **kind**, the eight split 5 MLP / 3 attention against 11 / 7 among the damaging components — indistinguishable (Fisher p = 1.00). By **depth**, they sit at layers 0, 5, 8, 12, 13, 22, 25 and 27, spanning the model end to end, with a median of 12.5 against the damaging set's 14.0 (Mann–Whitney p = 0.96). Across all 56 components, layer index carries no information about the *direction* of the effect (Spearman $\rho = -0.051$, p = 0.71). **Helping is not a property of a region or of a module type in this map.** We report the negative result because the obvious reading of a single late-MLP anomaly — *"the last block is special"* — is not what the class supports: `mlp_block__L0` is at the opposite end and improves too.

Three of the eight are provisional (below 80% power at n=96) and are marked as such throughout; the class survives their removal at five, including both endpoints. **The remaining 14 components with negative point estimates do not clear the gate** and we make no claim about them. Whether the class is eight or larger is a question of n, not of sign.

The extreme case, examined closely. The rest of this section is `mlp_block__L27`, which we measured far more than any other component because it is the largest effect in the map.

The effect is −0.0330 (dz −1.468) at n=96, the largest |dz| of any component in either direction. Re-measured on **disjoint items** — a different seed, zero shared item ids — it returns −0.0295 (dz −1.194, n=48) at p = 0.0001. Two independent measurements, not three: an earlier n=48 probe is *nested* inside the n=96 sweep and is a consistency check rather than a replication.

It is concentrated, not diffuse — but "concentrated" is not "two of four chunks are zero." Decomposing the block into four chunks of its intermediate dimension, two carry **58.3%** and **57.2%** of the block effect against 25% each under an even split. The other two are **not nothing**: one carries **8.0%** in the same direction, and one carries **−4.9% — the opposite sign**, a small *damaging* contribution (dz +0.173) inside a block whose net effect is a large improvement. We state both rather than rounding them to zero, because the concentration claim is what this paragraph is for and a chunk running the other way is evidence against the strongest version of it. The four chunks sum to 118.6% of the block effect; quantization is not linear and each chunk is quantized independently, so exact additivity was never expected.

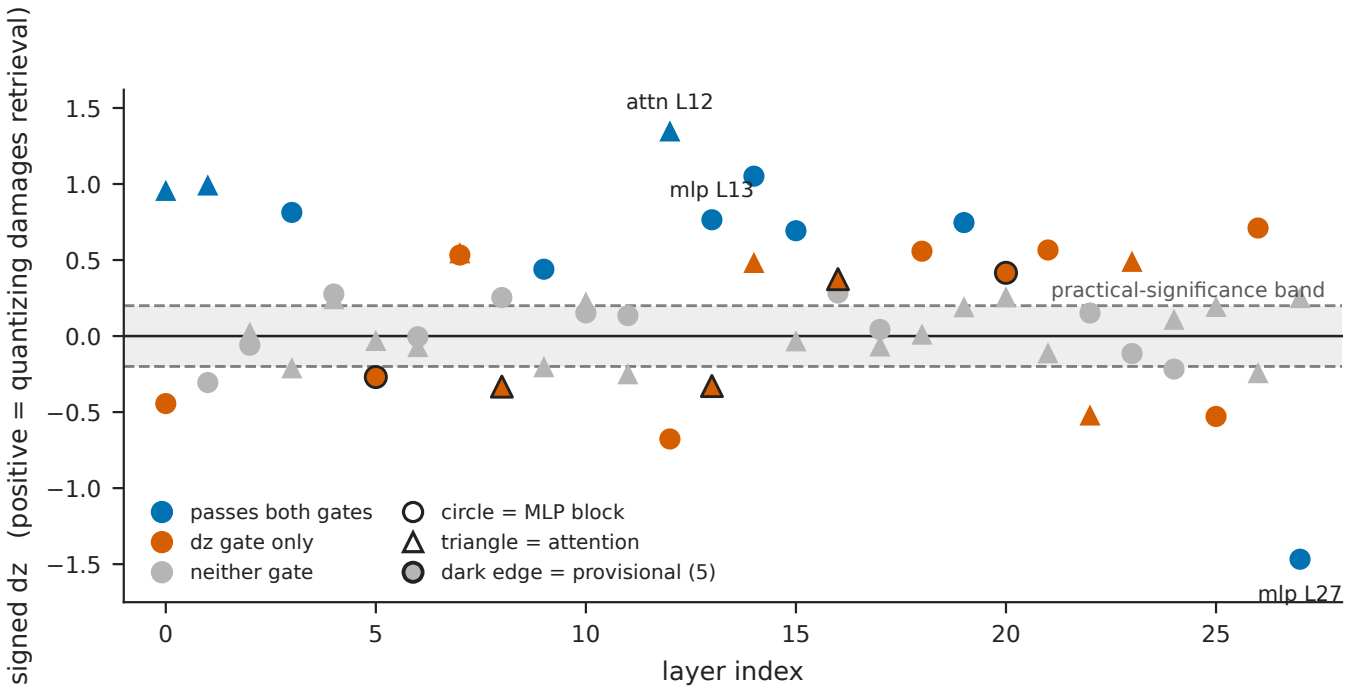

Figure 6: Component map. The paper's central artefact: signed dz for all 56 retrieval components at 3-bit precision, n=96, **plotted against layer index rather than rank**. Ranking would sort the signed effects into a monotone ramp, and a ramp reads as depth structure — which §6 shows is absent. Shaded region is the practical-significance band ($|dz| < 0.2$); dark edges mark the five provisional detections (adequately-powered status from the plug-in bootstrap in `power_measured.py`, which assumes no effect-scaling model). Positive dz means quantizing the component damages retrieval; the eight negative gated points are the improving class of §6.

component	ΔNLL	dz	95% CI	power
`mlp_block__L27`	**−0.0330**	**−1.468**	[−0.0375, −0.0286]	adequate
`mlp_block__L12`	−0.0265	−0.677	[−0.0345, −0.0191]	adequate
`mlp_block__L5`	−0.0265	−0.270	[−0.0481, −0.0102]	**provisional**
`attn_block__L13`	−0.0214	−0.329	[−0.0358, −0.0103]	**provisional**
`mlp_block__L0`	−0.0156	−0.444	[−0.0228, −0.0089]	adequate
`attn_block__L22`	−0.0113	−0.519	[−0.0156, −0.0069]	adequate
`attn_block__L8`	−0.0109	−0.332	[−0.0177, −0.0047]	**provisional**
`mlp_block__L25`	−0.0045	−0.529	[−0.0062, −0.0028]	adequate

One chunk has a larger standardised effect than the whole block — dz −1.755 against −1.468, while carrying only 57% of the raw effect. Its per-item spread is smaller: the chunks that contribute little to the effect still contribute their full share of noise to the block measurement. **If this generalises it inverts a standard assumption behind coarse-to-fine search**, that finer granularity costs detectability. We report it as **provisional**: it rests on one component, and the mechanism's own sharper prediction — that both *active* chunks should exceed the block — holds for only one of the two. A pre-registered test on a diffuse component is specified and unrun.

The improvement is monotone in perturbation strength. At 4 bits the same component gives −0.0068 (dz −0.780); at 3 bits, −0.0330 (dz −1.468). Stronger perturbation, larger improvement. That is what *"this block's contribution is net-harmful for this objective"* predicts, and the opposite of *"mild-perturbation regularisation"*, which predicts a peak at intermediate strength. Two points is not a curve, and 2 bits cannot extend it — the model is destroyed there, so a component measurement says nothing.

We refuted the mechanism we first proposed. Distributional flattening via the tied output head predicts that items the baseline answers *correctly* should get worse while incorrect ones improve. Split by baseline correctness, both subsets improve by the same amount (−0.0339 on 43 correct items at $p = 0.0001$; −0.0311 on 5 incorrect). The account is refuted, and no mechanism has replaced it.

The dissociation, and its demotion. The same component *damages* arithmetic, and that direction also replicates on

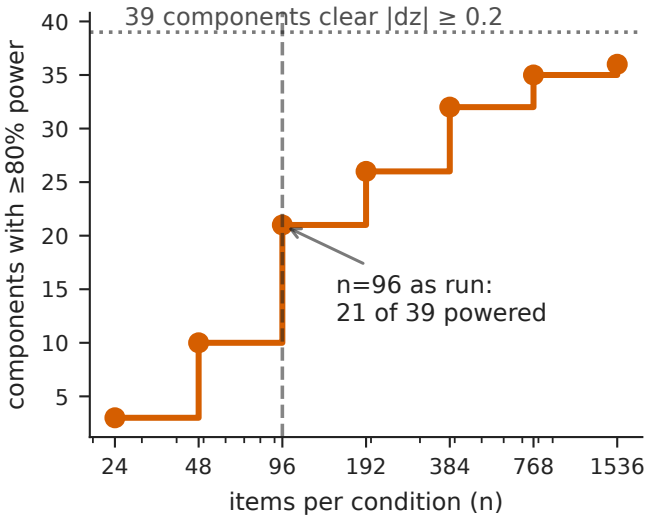

Figure 7: Coverage vs n. Cumulative number of the 39 above-threshold retrieval components reaching 80% power under Holm/56, by n. At the n the map was drawn at (96), 21 of 56 are adequately powered; the curve reaches 32 at n=384 and 36 at n=1536. Power is simulated by resampling each component's own measured per-item deltas — the plug-in bootstrap — **not** through the additive effect-scaling model §4.2 falsifies.

disjoint items — but the arithmetic leg falls **below the materiality threshold on unbiased re-measurement** (§4.4, §5). The honest claim is a large, replicated retrieval *improvement* alongside a real but **immaterial** arithmetic damage in the opposite direction, not a dissociation with two detected legs.

The improvement is an NLL claim, not a capability claim. The boolean metric was collected for the first time on the fresh-item run and cannot confirm it: McNemar p = 1.0000 on a single discordant pair, because that draw's baseline sits at **0.958 accuracy (n=48)** with almost no room to improve. Until a variant with headroom says otherwise, this is a statement about calibration. **What this component does and does not license is stated with the paper's other boundaries in §7.**

7 Claims we do not make

A localisation study is easy to over-read, and the standard objection to one is that its authors did. We therefore state the boundary explicitly rather than leaving it to be inferred.

We do not claim the capability maps differ. The pre-registered test is inconclusive: $\rho = -0.0014$ against a null centred on -0.0016, p $\approx$ 0.50 in both directions (§5). We construct no third narrative. And $\rho \approx 0$ against a null centred on 0 is **not** evidence the maps are unrelated — it is what happens when one vector is noise-dominated.

We do not claim arithmetic does not localise. A pre-registered null is uninterpretable unless the gate's power against a stated alternative is known, so the reportable statement names it: *no arithmetic component was detected at n=96 under Holm correction, where the gate has 0.225 power against the largest effect actually present.* That is a statement about the instrument. The power figure is measured by resampling that component's own per-item deltas; we do **not** quote the value our earlier model-based grid gave for the same cell, because §4.2 falsifies that model.

We do not claim `mlp_block__L27` dissociates the two capabilities. The retrieval leg is large and replicated; the arithmetic leg is real but falls **below** our own materiality threshold on unbiased re-measurement — dz +0.246 selected, **+0.197** data-split, against a 0.2 bar (§4.4). Nor do we claim it is unique: it is the largest of **eight** components whose quantization improves retrieval (§6). And one component dissociating would not make two maps differ.

We do not claim the 3-bit map is evidence about 4-bit deployment. The pre-registered rank-transfer test is **withdrawn**, on the measurement that its own pilot was specified to produce: the 4-bit contrast sits at dz 0.226 against 1.933 at 3 bits, on the practical threshold where more items cannot help. Nothing is inferred from its absence, and §3's preliminary sign-instability result points *against* clean transfer rather than for it.

We do not claim our two quantizers are equivalent at 4 bits. Paired on identical items, the difference is not detectable (p = 0.82 and 0.73, §3). That is a **failed rejection**, not equivalence — the distinction our own equivalence testing exists to enforce.

We do not claim quantizing the last MLP block improves retrieval *capability*. It improves answer-span NLL. The boolean metric cannot confirm it at this variant's headroom, so it is a calibration claim until measured on one with room to move.

8 Limitations

Precision. The map is 3-bit `int_group`. Transfer to 4-bit NF4 — the deployed regime — is not established, and we measured that establishing it is out of reach at this scale: a pilot puts the 4-bit top-versus-bottom contrast at dz 0.226 against 1.933 at 3 bits, requiring ~1536 items for 0.83 power (~15 h). **The binding constraint is not sample size.** A true dz of 0.226 sits barely above the 0.2 threshold, so the observed value falls below it about one time in four *at any n* — items cannot fix an effect sitting on a threshold. The pre-registered rank-transfer test is therefore **formally withdrawn on that measurement**, logged as a pre-registration amendment with its cost, rather than left outstanding. **Measured and unaffordable, not untested** — and nothing is inferred from its absence.

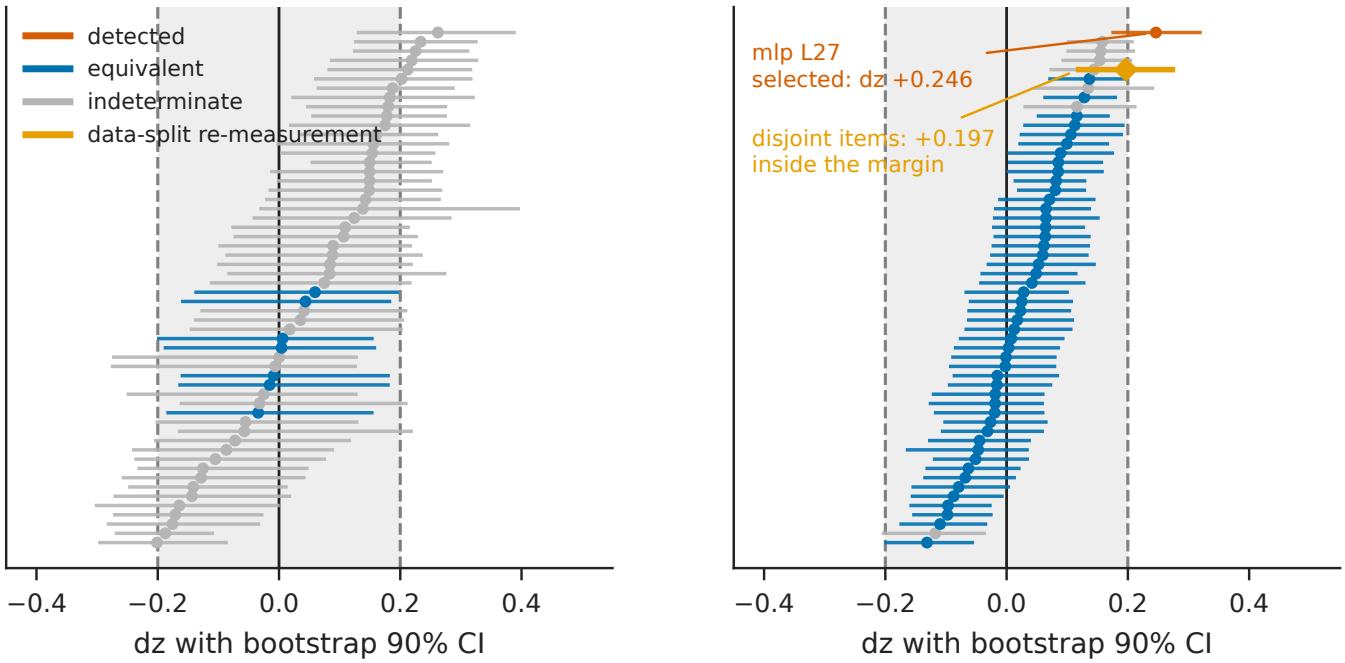

Figure 8: Arithmetic equivalence. Per-component equivalence tests on arithmetic at n=96 and n=384, each component's dz with its bootstrap 90% CI, sorted, against the pre-registered ±0.2 margin (shaded). Quadrupling n collapses 49 indeterminate components to 7 and resolves 48 as equivalent to zero. The single detection at n=384 is `mlp_block__L27` at dz +0.246; re-measured on disjoint items it returns +0.197, **inside the margin**, so the reportable arithmetic result is the equivalence, not the detection (§4.4). Intervals are bootstrap only; no power model enters this figure.

Coverage. At n=96 the map is adequately powered for **21 of 56 components (37%)**; the ladder, the 18 underpowered components and the five provisional detections are in §5.

One model. Qwen2.5-1.5B-Instruct throughout; every substantive number is n = 1 model, one architecture family, one variant per capability. We attacked this rather than conceding it and the attack failed, for the reason §3 reports: a shared bit width is not a shared intervention across scales, so the 3B arm compares a break point against a floor. Two further obstacles are specific to that arm and are recorded here — the calibrated variant is effectively ceilinged on the 3B (**47/48, n=48**, passing our band by less than one item's resolution, which prompted an amendment), and the 3B's baseline NLL is **6× the 1.5B's on identical items**, unexplained, putting the two models' effect sizes on different scales. A comparable second-model arm needs its own calibration *and* its own precision ladder: **~5.5 h**. Costed and not bought.

The primary test is inconclusive, and we report it as such (§7).

Quantizer. `int_group` rather than deployed `nf4`, because `nf4` has no 3-bit form. Anchored at 4 bits by a failed rejection, not by equivalence (§3).

Simulated quantization. Weights are rounded onto the N-bit grid and stored in the original dtype — numerically identical to a quantized model, and the only way to intervene on a single head, but not a deployment measurement.

One retained anomaly. A single configuration took 8.5 h against 55–138 s for its neighbours, diagnosed as memory thrashing. We kept the result on the argument that paging cannot alter floating-point arithmetic, and **did not independently re-measure it**. We record this rather than omit it.

9 Related work

9.1 The nearest neighbour: FP4 Diagnosis

Cim, Topcu & Kandemir, "Diagnosing FP4 Inference: A Layer-wise and Block-wise Sensitivity Analysis of NVFP4 and MXFP4", ICLR 2026 Workshop on SciForDL (arXiv:2603.08747). Qwen2.5 at 0.5B, 7B and 14B under MXFP4 and NVFP4, scored by WikiText-2 perplexity on *"256 calibration samples"*. Component sensitivity quantizes *"one of all seven projection types (Query, Key, Value, Output, Gate, Up, Down) to FP4 at a time, and keep[s] remaining six of them in FP16 across all layers"*; block sensitivity is the inverse. **No statistics of any kind** — point estimates and rankings, with no significance test, noise floor, multiplicity correction, power analysis or repeated run.

Their conclusion names this project as future work:

> "Future work should also evaluate FP4 sensitivity on diverse downstream tasks beyond perplexity on

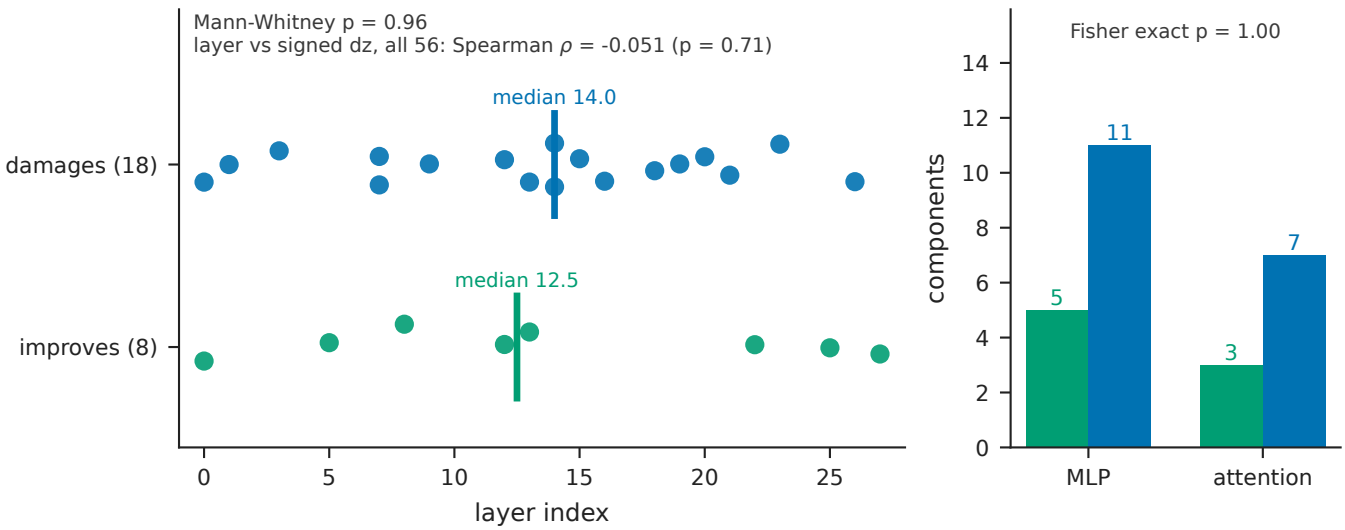

Figure 9: Improvers do not cluster. The eight components whose quantization improves retrieval against the 18 that damage it. Left: layer index, jittered, with medians — the two distributions are indistinguishable (Mann–Whitney p = 0.96), and across all 56 components layer carries no information about the direction of the effect (Spearman $\rho = -0.051$, p = 0.71). Right: module-type composition, 5 MLP / 3 attention against 11 / 7 (Fisher exact p = 1.00). Helping is not a property of a region or a module kind, which is why `mlp_block__L0` improves at one end of the model and `mlp_block__L27` at the other.

> WikiText, such as reasoning, coding, and instruction following benchmarks, to better understand how component level quantization effects translate to task specific performance degradation."

That is the strongest available evidence the question is open, from the group nearest to it.

9.2 What every located method optimises, and what none tests

Two gaps follow. Every objective there is a **proxy or an aggregate** — none measures per-capability task performance per item, and a perplexity ranking cannot be split by capability because perplexity has already averaged them together. And **this literature ranks rather than tests**: no paper found reports a significance test, noise floor, multiplicity correction, power analysis or replication. **§4 is what happens when that machinery is imported**, and is not restated here. The statistics are textbook and are not claimed as findings; the contribution is their absence and what it costs.

9.3 A prior for the hypothesis, in a different setting

Yin, Xu et al., "LoFiT: Localized Fine-tuning on LLM Representations" (arXiv:2406.01563), §5.3:

> "Across all models, task-specific distributions peak in some contiguous sets of layers within the same model rather than being widely spread across layers or concentrated in a single layer. ... head sets for different tasks are only mildly overlapping. These findings demonstrate that there is not a single set of heads best for fine-tuning across all tasks."

Capability-specific localisation is therefore **established for fine-tuning**, which makes our hypothesis well-motivated rather than speculative and our **inconclusive** result more interesting rather than less: a phenomenon holding under fine-tuning does not clearly reproduce under 3-bit quantization with this instrument. Whether the difference is the intervention, the unit (heads vs blocks), the selection procedure or our power **is not settled here**.

9.4 Depth structure: SliderQuant

"SliderQuant: Accurate Post-Training Quantization for LLMs", ICLR 2026 (arXiv:2603.25284): *"shallow/deep layers are usually more sensitive to quantization than intermediate layers"*, and among edge layers *"the most sensitive one is the first/last layer."*

9.5 Two tensions with our results, stated rather than buried

(a) Which components dominate — a genuine conflict. FP4 Diagnosis finds MLP up/down projections dominant with **gate and attention substantially less sensitive**; we find `attn_block__L0` among the strongest and structure **spanning the depth**, including mid-depth attention (`attn_block__L12`, dz +1.349) (§5). Four differences

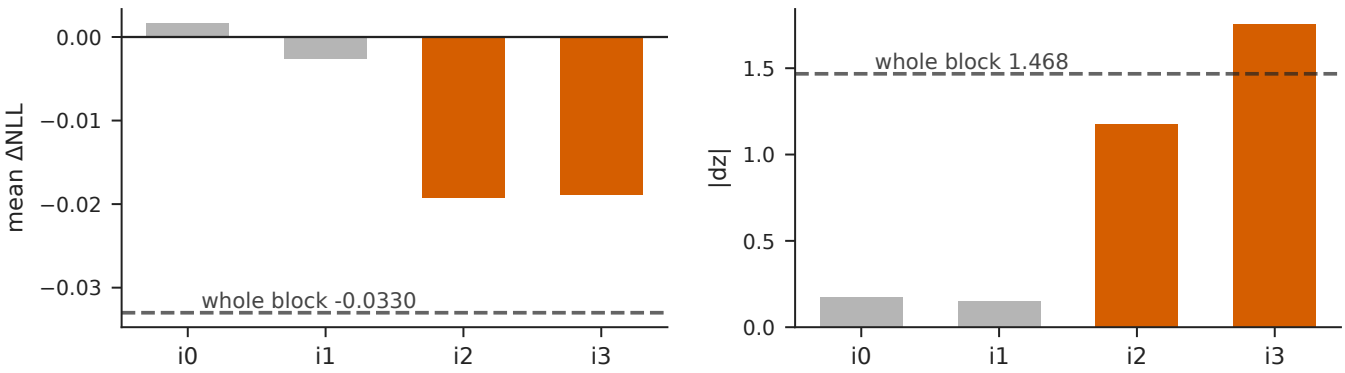

Figure 10: Chunk decomposition. The `mlp_block__L27` effect split across four chunks of its intermediate dimension. Two chunks carry 58.3% and 57.2%; one carries 8.0%; one runs the opposite way at −4.9%. One chunk has a larger standardised effect than the whole block despite carrying 57% of the raw effect, because the chunks contributing little to the effect still contribute their full share of noise to the block measurement.

method	objective	what it optimises
HAWQ (arXiv:1905.03696)	Hessian spectrum / second-order curvature	loss-surface sensitivity
LIM	cosine similarity between a layer's input and output	representation change
LSAQ (arXiv:2412.18135)	Jaccard similarity over top-k token sets	output-set overlap
LRP-QViT (arXiv:2401.11243)	layer-wise relevance propagation	attribution to a target class
FP4 Diagnosis (arXiv:2603.08747)	**WikiText-2 perplexity**	corpus-average likelihood

could account for it and we can settle none from our data: **metric** (retrieval NLL versus WikiText-2 perplexity — this paper's thesis applied to itself), **format** (§3 finds ordering may not survive a one-bit change), **granularity** (one block at one layer versus one projection type across all layers at once), and **capability**. Running their protocol under our metric on the same model would isolate metric from granularity. Not run; not costed.

(b) Negative improvement — an independent sighting. FP4 Diagnosis reports once, without investigating:

> "notably, for 14B under MXFP4, isolating early blocks can even yield negative improvement."

Their intervention is the inverse of ours — they *restore* a block to FP16 and find performance does not improve — but the shape is the same: **a component whose relationship to quality runs opposite to the expected direction.** We cite it as an independent sighting at a different scale, format and metric; §6 adds that the phenomenon is **a class of eight rather than one**, that it replicates on disjoint items, decomposes to sub-block granularity, is monotone in perturbation strength, and has one proposed mechanism refuted. That a second group encountered it and moved on is worth stating: it is the shape of a finding that ranking-without-testing produces and discards.

References

Cited inline in the text; this list is the paper's bibliography. Entries marked "used, not cited in prose" are dependencies a reader needs to reproduce the work.

1. M. Cim, C. Topcu and M. Kandemir. Diagnosing FP4 Inference: A Layer-wise and Block-wise Sensitivity Analysis of NVFP4 and MXFP4. ICLR 2026 Workshop on SciForDL. arXiv:2603.08747. 2. I. Andrews, T. Kitagawa and A. McCloskey. Inference on Winners. *Quarterly Journal of Economics*, 139(1), 2024. 3. Z. Yin, Y. Xu et al. LoFiT: Localized Fine-tuning on LLM Representations. arXiv:2406.01563. 4. SliderQuant: Accurate Post-Training Quantization for LLMs. ICLR 2026. arXiv:2603.25284. 5. Z. Dong, Z. Yao, A. Gholami, M. Mahoney and K. Keutzer. HAWQ: Hessian AWare Quantization of Neural Networks with Mixed-Precision. arXiv:1905.03696. 6. LSAQ: Layer-Specific Adaptive Quantization for Large Language Model Deployment. arXiv:2412.18135. 7. N. Ranjan and A. Savakis. LRP-QViT: Mixed-Precision Vision Transformer Quantization via Layer-wise Relevance Propagation. arXiv:2401.11243. 8. S. Merity, C. Xiong, J. Bradbury and R. Socher. Pointer Sentinel Mixture Models. arXiv:1609.07843. (WikiText-2.) 9. K. Cobbe et al. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168. (GSM8K.) 10. Qwen Team. Qwen2.5 Technical Report. arXiv:2412.15115. 11. T. Dettmers, A. Pagnoni, A. Holtzman and L. Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs.

arXiv:2305.14314. (NF4; `bitsandbytes` is the reference implementation our quantizer is validated against.) 12. S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 1979. Used, not cited in prose. 13. M. Okabe and K. Ito. Color Universal Design. 2008. Used, not cited in prose: the figures' palette.

Appendix A · Every interval this paper reports, classified by selection

All are bootstrap 90% intervals on dz (10,000 resamples), the level our equivalence margin implies.

interval	quantity	status
[+0.176, +0.319]	`mlp_block__L27` arithmetic dz, seed 2024	**post-selection** — the only component of 56 to clear the gate, interval computed on the selecting data
[+0.119, +0.274]	same component, seed 2025, **disjoint items**	**valid** — pre-specified by the time it was drawn; this *is* the data-split estimate
[−0.0364, −0.0226]	`mlp_block__L27` retrieval ΔNLL, seed 2025	**valid** — pre-specified, disjoint
[+0.166, +0.272]	pooled over both draws	**partially contaminated** — pools a selected with a non-selected sample
per-component equivalence intervals (all 56)	every component, not a subset	**valid as a set** — only the one *reported because it was detected* inherits selection