Pre-Registering the Detectable Effect: A Paired-MDE Budget for 4-bit Quantization Benchmarks, with a Pilot Audit

Zexin Zhuang [1] **Yanhang Li** [2] **Zhichao Fan** [3]

Abstract

This is a planning-method note with an unpaired pilot audit. We adapt the classical paired-binary sample-size calculation (Miettinen, 1968) to quantization benchmarks, giving a conservative *minimum detectable effect* (MDE) bound $\delta^* \leq (z_{1-\alpha/2} + z_{1-\beta})\sqrt{\rho_d/m}$ in the paired item count m and the FP16↔NF4 disagreement rate ρ_d. The bound turns "how reliable is my quantization claim?" into a one-line budget a benchmark designer can commit to *before running*. We illustrate the bound on four models and four benchmarks (k=5 splits of n=100), and add a parallel MMLU prompt-template study to put the bound's quantization-noise scale alongside the prompt-noise scale. Assuming ρ_d=0.10 (an unmeasured planning value), all observed NF4−FP16 deltas fall below the implied MDE, and most cross-split SDs lie within ± 1.5 pp of the binomial reference $\sqrt{p(1-p)/n}$, so much of the variance reported as "benchmark unreliability" on n=100 subsamples is binomial sampling noise. The single borderline cell (OPT–WinoGrande, $|\Delta|$=3.2 pp) is below the implied MDE at ρ_d=0.10 but above it at ρ_d=0.05, illustrating the planning trade-off the bound makes explicit. **On MMLU, prompt-template ranges of 2–10 pp meet or exceed the largest observed quantization delta (3.2 pp), so a quantization audit that does not first fix the prompt template absorbs template variance into its noise floor.** We complement the bound with a five-line pre-registration template.

1. Introduction

Post-training quantization is now the default deployment path for large language models on consumer hardware. NF4 quantization (Dettmers et al., 2023) packs each weight into one of 16 levels chosen to be optimal for normally distributed weights; GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), SmoothQuant (Xiao et al., 2023), SpQR (Dettmers et al., 2024), and ZeroQuant (Yao et al., 2022) have all reported small-to-modest accuracy losses on standard benchmarks. A typical quantization study reports a single accuracy number per model×benchmark cell and concludes that 4-bit inference "preserves" or "slightly degrades" performance.

This practice elides a basic statistical question: what is the smallest quantization effect that the evaluation protocol could reliably detect in the first place? With n benchmark items and a baseline accuracy of p, even a perfectly faithful estimator has standard error at least $\sqrt{p(1-p)/n}$. With k non-overlapping splits, the cross-split standard deviation has expected scale $\sqrt{p(1-p)/n}$ unless the benchmark itself is structured (e.g., MMLU's 57 subjects make subject mix vary across random splits); at finite k observed cross-split SD can fall both above and below this binomial reference, but the reference still pins the order of magnitude. Reports of "small" quantization effects therefore confound two distinct claims: (i) the population effect is small, and (ii) the protocol cannot detect a small effect even if one exists. CTB-style ex-ante guarantees should distinguish them.

We pursue this distinction along three threads.

An ex-ante reliability bound. We derive a conservative paired minimum detectable effect (MDE) for FP16-vs-NF4 comparisons (Section 3). Given a paired sample of m items and a pre-registered upper bound ρ_d on per-example disagreement, the MDE at significance α and power $1-\beta$ (we use β for Type-II error throughout) is $\delta^*(m, \rho_d) \leq (z_{1-\alpha/2} + z_{1-\beta})\sqrt{\rho_d/m}$, where m equals n for a single-split estimand and kn for an aggregate-of-splits estimand. The bound is a quantization-flavored repackaging of paired-binary sample-size planning (Miettinen, 1968; Connor, 1987; Lachin, 1992); the contribution is its use as an ex-ante budget line for quantization audits, not the underlying inequality.

[1]Southern Methodist University [2]Northeastern University [3]University of Illinois Urbana-Champaign. Correspondence to: Zexin Zhuang <zexinz@smu.edu>, Yanhang Li <li.yanha@northeastern.edu>, Zhichao Fan <zhichao8@illinois.edu>.
Preprint. June 5, 2026.

An empirical audit. We measure FP16-vs-NF4 accuracy on four models (OPT-2.7B, Pythia-2.8B, Llama-2-7B, Mistral-7B) and four benchmarks (MMLU, ARC-Easy, WinoGrande, HellaSwag) with $k{=}5$ non-overlapping splits of $n{=}100$ each. We separately measure prompt-template sensitivity on MMLU with three templates and $n{=}50$ each. Section 5 reports the observed accuracy means, cross-split standard deviations, and template spreads. To diagnose how much of the observed cross-split variance is binomial sampling and how much is between-split, we compare $\hat{\sigma}_{\text{split}}$ to the binomial reference SD $\sqrt{\hat{p}(1-\hat{p})/n}$ on each cell.

A reliability index. We formalize a Quantization Reliability Index (QRI) as a single-split signal-to-noise heuristic and explicitly separate its split-only variant $\mathrm{QRI}_{\text{split}}$ (defined for all 16 cells) from its split+prompt variant $\mathrm{QRI}_{\text{combined}}$ (defined only on the four MMLU cells where prompt variance was measured). QRI is not a hypothesis test; it is meant to flag cells where a single-split comparison is dominated by evaluation noise rather than by the quantization signal.

What this paper does *not* claim. We do not run paired McNemar tests on per-example correctness (we did not retain the per-example records under our compute budget); we therefore cannot rule out paired effects smaller than what the unpaired protocol can resolve. We do not attempt to predict δ ex-ante from weight statistics; the predictor for δ itself is open. We test only NF4 via BitsAndBytes; GPTQ, AWQ, and SpQR may produce different reliability landscapes. The 7B side of our design has only two models, so within-7B claims are observational. These are stated again in Section 9.

Contributions.

1. A conservative paired MDE bound for FP16-vs-NF4 quantization comparisons (Eq. 2) and a corresponding sample-size table that turns "how reliable is my quantization claim" into a one-line pre-registration item.
2. A pilot audit including a per-cell binomial-reference table (Appendix B) showing that 25 of 32 observed cross-split SDs fall within ± 1.5 pp of $\sqrt{p(1-p)/n}$ (and 29 of 32 within ± 2.0 pp), so on this audit most of what gets reported as "benchmark unreliability" is small-n binomial sampling rather than a property of the specific model or of NF4.
3. A per-cell Quantization Reliability Index (Appendix C) as a descriptive signal-to-noise diagnostic, with a split-only variant defined for all 16 cells and a prompt-augmented variant defined for the 4 MMLU cells where prompt variance was measured. QRI is not a power test in this paper; power-controlled decisions are made directly via the $|\hat{\Delta}|$-vs-δ^* comparison in Table 6.
4. Recommendations: pre-register an MDE target, report paired discordant counts (n_{10}, n_{01}), separate binomial sampling noise from subset-composition noise on clustered benchmarks like MMLU, and sweep prompt templates on every benchmark used in a comparison.

2. Related Work

Post-Training Quantization. Frantar et al. (2023) and Dettmers et al. (2023) introduced GPTQ and QLoRA/NF4. Dettmers et al. (2022) (LLM.int8) handles activation outliers via 8-bit matrix multiplication; Lin et al. (2024) (AWQ), Xiao et al. (2023) (SmoothQuant), Dettmers et al. (2024) (SpQR), and Yao et al. (2022) (ZeroQuant) cover the recent low-bit PTQ landscape; Lu et al. (2024) relates spectral weight properties to layer-wise compression decisions. Benchmark-style comparisons of PTQ strategies appear in Li et al. (2024), Gong et al. (2024), Jin et al. (2024), and Zhao et al. (2025). None of these works isolates the *minimum detectable quantization effect* as a function of benchmark sample size.

Benchmark Evaluation. Liang et al. (2023) (HELM) and Polo et al. (2024) examine evaluation breadth and subset selection but not quantization. Biderman et al. (2024) document the hidden statistical fragility of LM evaluation harnesses, and configuration-conditional audits similarly show that alignment-benchmark rankings can depend on evaluation choices (Li et al., 2026). Dror et al. (2018) systematize statistical significance testing in NLP. Complementary scaling-law analyses of LLM behavior (Shi et al., 2026) and time-series forecasting (Shi et al., 2024) characterize model capability at scale, and design-axis frameworks for LLM evaluation have been proposed in adjacent areas including retrieval-augmented QA (Ji et al., 2026) and benchmark construction for generative AI (Luo et al., 2026a); our paired-MDE template adds the precision-vs-precision detectability axis for quantization.

Prompt Sensitivity. Sclar et al. (2024) report accuracy swings of up to 76 pp from formatting changes; Mizrahi et al. (2024) benchmark prompt-template sensitivity at scale; broader prompting-evaluation concerns have also been documented in adjacent modalities (Luo et al., 2026b), reinforcing the recommendation to sweep templates on every benchmark used in a precision audit. We study the interaction of prompt sensitivity with quantization at the $n{=}50$ scale, framing our findings as exploratory rather than confirmatory.

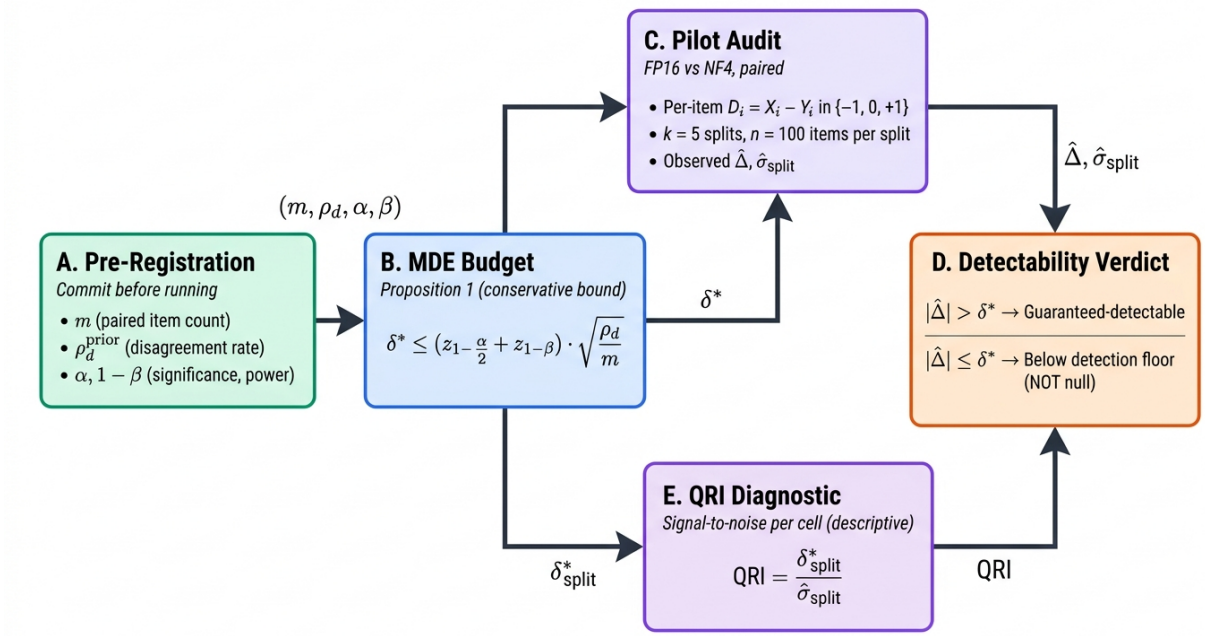

Figure 1. Pre-registerable reliability audit pipeline for paired FP16 vs. NF4 quantization benchmarks. The designer commits $(m, \rho_d^{\text{prior}}, \alpha, 1-\beta)$ in advance (A); Proposition 1 turns these into a conservative paired Minimum Detectable Effect δ^* (B) that bounds what any later pilot can detect. The pilot audit (C) computes per-item differences $D_i = X_i - Y_i \in \{-1, 0, +1\}$ over k=5 non-overlapping splits of n=100 items and reports $\hat{\Delta}$ and the cross-split SD $\hat{\sigma}_{\text{split}}$. The Quantization Reliability Index (E), $\text{QRI} = \delta^*_{\text{split}}/\hat{\sigma}_{\text{split}}$, is a descriptive single-split signal-to-noise diagnostic. The detectability verdict (D) compares $|\hat{\Delta}|$ to δ^*: effects above the bound are guaranteed-detectable in design, while effects below it are merely *not* guaranteed-detectable, not null.

3. Minimum Detectable Effect under Paired Quantization Comparison

Setup and assumptions. Let $X_i \in \{0, 1\}$ be the FP16 correctness of item i and $Y_i \in \{0, 1\}$ the NF4 correctness on the *same* item, $i = 1, \ldots, m$. Define the per-item accuracy difference $D_i = X_i - Y_i \in \{-1, 0, +1\}$, the disagreement rate $\rho_d = \Pr(D_i \neq 0)$, and the population effect $\delta = |\mathbb{E}[D_i]|$. We assume (A1) items are i.i.d. samples from a target distribution, (A2) model outputs are deterministic conditional on the prompt and precision (no temperature sampling), and (A3) ρ_d is either pre-registered or conservatively bounded a priori. Section 9 discusses how clustering on MMLU subjects and our finite-population partitions weaken (A1).

Paired-difference variance. The exact per-item variance is

$$\text{Var}(D_i) = \mathbb{E}[D_i^2] - \mathbb{E}[D_i]^2 = \rho_d - \delta^2, \qquad (1)$$

so the mean over m items has variance $(\rho_d - \delta^2)/m \leq \rho_d/m$. The mild upper bound $\text{Var}(D_i) \leq \rho_d$ is tight only as $\delta \to 0$.

Conservative MDE bound. Let $\mu = \mathbb{E}[D_i]$ and $\delta = |\mu|$, with $\delta \leq \rho_d$ since $|D_i| \leq \mathbb{1}[D_i \neq 0]$. We state the central result as a formal proposition for ease of citation.

Proposition 1 (Conservative paired MDE for FP16-vs-NF4 benchmarks)**.** *Let $\{D_i\}_{i=1}^m$ be i.i.d. paired differences in $\{-1, 0, +1\}$ with disagreement rate $\rho_d = \Pr(D_i \neq 0)$ and population effect $\delta = |\mathbb{E}[D_i]|$. Under (A1)–(A3), the normal-approximation two-sided z-test of $H_0 : \mathbb{E}[D_i] = 0$ at level α rejects with power at least $1 - \beta$ for any δ satisfying*

$$\delta \geq (z_{1-\alpha/2} + z_{1-\beta}) \sqrt{\rho_d/m}.$$

The smallest such δ, denoted $\delta^*(m, \rho_d, \alpha, \beta)$, is therefore upper-bounded by the right-hand side.

In benchmark-design terms: a benchmark of paired size m with worst-case disagreement rate ρ_d is not planned to have the stated power for FP16-vs-NF4 effects below this threshold, regardless of which model or NF4 implementation is being audited; effects above the bound are guaranteed-detectable in design but effects below it are merely *not guaranteed-detectable*, not impossible. The bound is conservative in two senses: (i) ρ_d replaces the tighter $\rho_d - \delta^2$ as the variance proxy (loose only when δ^2 is non-negligible relative to ρ_d), and (ii) the z-test is two-sided (one-sided NF4-degrades-only tests would tighten by $z_{1-\alpha/2} \to z_{1-\alpha}$).

The two simplifications together yield a single closed-form line that benchmark designers can pre-register before evaluation.

For citation throughout the paper we restate this as

$$\delta^*(m, \rho_d, \alpha, \beta) \;\le\; (z_{1-\alpha/2} + z_{1-\beta})\sqrt{\frac{\rho_d}{m}}. \qquad (2)$$

This is a *conservative sufficient bound*, not a necessary impossibility threshold. The standard alternative-variance approximation, which uses the null variance ρ_d in the rejection-region term and the alternative variance $\rho_d - \delta^2$ in the power term, gives

$$\delta^*\sqrt{m} \;=\; z_{1-\alpha/2}\sqrt{\rho_d} + z_{1-\beta}\sqrt{\rho_d - (\delta^*)^2},$$

which is implicit but numerically very close to Eq. 2 for the small-ρ_d regimes we care about; the maximum tightening over Eq. 2 at ρ_d=0.10, m=500, α=0.05, $1-\beta$=0.80 is below 0.1 pp. We use Eq. 2 throughout for simplicity and label it conservative. Equation 2 is a quantization-flavored repackaging of classical paired-binary sample-size planning (Miettinen, 1968; Connor, 1987; Lachin, 1992) and is closely related to NLP significance-testing practice (Yeh, 2000; Dror et al., 2018); the contribution is its application as an ex-ante budget line for quantization benchmarks, not the underlying inequality.

Single-split vs. aggregate m. Equation 2 applies to whatever paired sample is actually used to estimate δ. With k non-overlapping splits of n items each, two natural targets exist:

1. **Single-split MDE** (m=n=100): the smallest paired effect resolvable on *one* split. For ρ_d=0.10 this is $\delta^*_n \approx 8.9$ pp.
2. **Aggregate MDE** (m=kn=500): the smallest paired effect resolvable on the *union* of the five splits, which is what Tables 1–2 actually report. For ρ_d=0.10 this is $\delta^*_{kn} \approx 4.0$ pp; for ρ_d=0.05 it is ≈ 2.8 pp.

The dependence on k enters only through the total item count kn; aggregating non-overlapping splits does not buy power beyond that. We will be explicit in §5 about which m each comparison uses.

Connection to cross-split SD. Cross-split SD σ_{split} from k splits of n items has expected magnitude of order $\sqrt{p(1-p)/n}$ under independent binomial sampling, regardless of paired structure. With p=0.45 and n=100, $\sigma_{\text{split}} \approx 5.0$ pp, comparable in magnitude to the single-split paired MDE at $\rho_d \in [0.05, 0.20]$. A study that compares $|\Delta_{\text{quant}}|$ aggregated over k splits to a single-split σ_{split} is mixing scales: aggregate signal vs. split-level noise. Reporting both targets, or reporting a single consistent m, fixes the mismatch.

Ex-ante use. Equation 2 converts the abstract notion of "benchmark reliability" into a pre-registerable budget line: *with m items and a conservative prior $\rho_d \le \rho_d^{prior}$, my evaluation can detect a paired NF4 effect no smaller than δ^* pp at $(\alpha, 1-\beta)$, conditional on (A1)–(A3).* A benchmark designer who pre-registers $(m, \rho_d^{\text{prior}}, \alpha, 1-\beta)$ and then reports paired discordant counts (n_{10}, n_{01}) can also retroactively check the assumed ρ_d^{prior} against the observed $\hat{\rho}_d = (n_{10}+n_{01})/m$. Our audit is unable to run this paired retrospective check (we did not retain per-example correctness; see Section 9); we therefore report only ex-ante MDE budgets and treat the empirical accuracy deltas as observations rather than tested differences.

4. Methods

4.1. Models and Quantization

We evaluate four models at two scale tiers, FP16 and NF4:

- **OPT-2.7B** (Zhang et al., 2022), a multi-head-attention decoder.
- **Pythia-2.8B** (Biderman et al., 2023), trained on the Pile with controlled procedures.
- **Llama-2-7B** (Touvron et al., 2023), a 7B model that uses standard multi-head attention (only the 34B and 70B Llama-2 variants use grouped-query attention; we previously misstated this).
- **Mistral-7B** (Jiang et al., 2023), a 7B model with sliding-window and grouped-query attention.

NF4 is applied via BitsAndBytes (load_in_4bit=True, bnb_4bit_quant_type='nf4', bnb_4bit_compute_dtype=float16, double-quant disabled). 3B models run on a single NVIDIA T4 (15GB); 7B models require an A100 (40GB) in FP16 and fit on a T4 under NF4. Exact Hugging Face identifiers, library versions, and seeds are in Appendix A.

4.2. Benchmarks and Splits

We use MMLU (Hendrycks et al., 2021) (validation), ARC-Easy (Clark et al., 2018) (test), WinoGrande (Sakaguchi et al., 2020) (validation), and HellaSwag (Zellers et al., 2019) (validation). Each benchmark is randomly partitioned into k=5 non-overlapping splits of n=100 examples each (seed 42; split indices in Appendix A). "Non-overlapping" is technically distinct from "independent" in the inferential sense: random partitions of one benchmark are exchangeable but not i.i.d. samples from the underlying population, so cross-split SD does not directly estimate the population sampling SD; finite-population correction can pull it down, while subject clustering on benchmarks like MMLU can

Table 1. Benchmark accuracy on k=5 non-overlapping splits of n=100 each, mean±cross-split SD. Numbers are observed values on our subsamples; population CIs are wider (see text and Appendix D).

Model	MMLU	ARC-E	WinoGr.	HellaSw.
OPT FP16	.252±.053	.258±.049	.598±.046	.434±.036
OPT NF4	.238±.073	.268±.059	.566±.038	.430±.036
Pythia FP16	.238±.046	.274±.056	.564±.025	.444±.043
Pythia NF4	.242±.047	.268±.019	.568±.041	.432±.049
Llama-2 FP16	.458±.051	.688±.041	.700±.036	.608±.031
Llama-2 NF4	.450±.048	.684±.039	.692±.034	.600±.033
Mistral FP16	.522±.046	.724±.037	.722±.032	.644±.029
Mistral NF4	.516±.049	.720±.035	.716±.030	.638±.031

push it up. We treat $\hat{\sigma}_{\text{split}}$ as a protocol-level descriptor rather than a population estimator.

4.3. Scoring

For multiple-choice items (MMLU, ARC, HellaSwag) we score by ranking the per-choice log-likelihoods of the answer-only continuation under a fixed prompt template; for WinoGrande we score by likelihood ratio of the two pronoun resolutions. "Perplexity" values reported in Table 2 are per-token NLLs of the prompt text (question+choices, excluding the gold-answer continuation), averaged across examples and exponentiated, with the model's native tokenization. We report PPL only as a descriptive complement; PPL changes are not used to support significance claims.

4.4. Prompt Sensitivity

For MMLU only, we evaluate three templates on n=50 examples each, sharing examples across templates within a run: (T0) standard "Question: … Answer:"; (T1) compact "Q: … Options: A. …"; (T2) inline "… (A) … The answer is". We do not extend the prompt-sensitivity sweep to ARC, WinoGrande, or HellaSwag in this study; this is reflected in the QRI definitions below.

5. Results: Illustrative Pilot Audit

The empirical content of this section is *illustrative*: the pilot did not retain per-example correctness, so we cannot empirically estimate ρ_d, run paired McNemar tests, or validate the paired audit machinery on this data. The numbers below are therefore a worked example of how the methodology of §3 would be applied, not a tested set of empirical claims about NF4 across the population of LLM benchmarks.

5.1. Observed Accuracy and the Binomial Reference SD

Table 1 reports the mean and cross-split SD for each cell; the largest observed $\hat{\sigma}_{\text{split}}$ is 7.3 pp (OPT-NF4 MMLU), and the 7B tier bottoms out near 2.9–3.6 pp where accuracy approaches 0.7. Figure 2 compares each $\hat{\sigma}_{\text{split}}$ to the binomial reference $\sigma_{\text{bin}}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$ at n=100 (full table in Appendix B); 25 of 32 cells lie within ±1.5 pp of the diagonal. The largest positive residuals concentrate on MMLU (subject-mix variance from random partitions of a 57-subject benchmark) and on 3B-tier ARC-Easy. We therefore caution against treating cross-split SD as a measure of quantization-specific noise: on n=100 subsamples it is largely the $1/\sqrt{n}$ floor any binary score would exhibit.

Table 2. NF4−FP16 accuracy delta and FP16 prompt PPL with NF4−FP16 PPL delta. Reported ΔAcc is the aggregate over m=500 items per cell. Aggregate paired MDE at α=0.05, power $1-\beta$=0.80 is $\delta^*_{m=500} \approx 4.0$ pp at ρ_d=0.10 and ≈ 2.8 pp at ρ_d=0.05. Of the 16 cells, only OPT-WinoGrande (|Δ|=3.2 pp) approaches either bound: it is below δ^* at ρ_d=0.10 and *above* δ^* at ρ_d=0.05. Because ρ_d was not measured, we cannot adjudicate detectability without paired discordant counts.

	Δ Acc. (pp)	PPL_{base}	Δ PPL
OPT-2.7B			
MMLU	−1.4	86.1	−3.8
ARC-Easy	+1.0	39.4	+0.9
WinoGr.	−3.2	158.9	+9.1
HellaSw.	−0.4	35.1	+1.1
Pythia-2.8B			
MMLU	+0.4	45.4	+3.3
ARC-Easy	−0.6	37.0	+1.5
WinoGr.	+0.4	144.2	+2.9
HellaSw.	−1.2	37.0	+1.3
Llama-2-7B			
MMLU	−0.8	28.3	+1.2
ARC-Easy	−0.4	22.1	+0.6
WinoGr.	−0.8	94.7	+3.1
HellaSw.	−0.8	21.8	+0.8
Mistral-7B			
MMLU	−0.6	24.1	+0.9
ARC-Easy	−0.4	18.7	+0.4
WinoGr.	−0.6	82.3	+2.4
HellaSw.	−0.6	19.2	+0.7

5.2. Quantization Deltas

Table 2 reports NF4−FP16 accuracy deltas (averaged over the aggregate m=500 items per cell) and PPL deltas. At the 3B tier the largest accuracy change is −3.2 pp (OPT WinoGrande); both signs occur within each model. At the 7B tier all eight cells move by −0.4 to −0.8 pp, a tighter range with a consistent sign. We flag two cautions about reading too much into the 7B pattern:

(i) With only two 7B models, this is a sample of size two at the model level; the apparent uniformity could reflect base accuracy ($p \approx 0.6$–0.72 pushes σ_{bin} to ≈ 2.0–2.3 pp at the aggregate level) or implementation idiosyncrasies as easily as a property of NF4. A sign test treating the

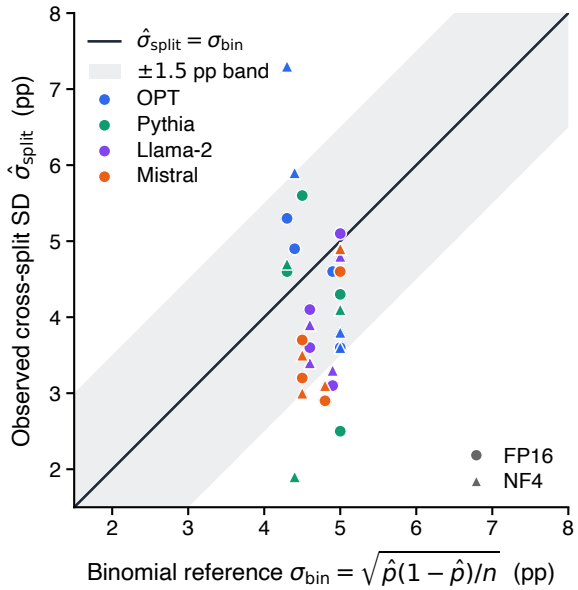

Figure 2. Observed cross-split SD vs. binomial reference SD on the 32 audited cells (Table 4). Marker shape encodes precision (FP16 vs. NF4); colour encodes model. The diagonal is $\hat{\sigma}_{\text{split}} = \sigma_{\text{bin}}$ and the shaded band is ± 1.5 pp around it. 25 of 32 points fall inside the band, so most observed cross-split variation on n=100 subsamples is accounted for by binomial sampling rather than by quantization-specific or split-composition noise.

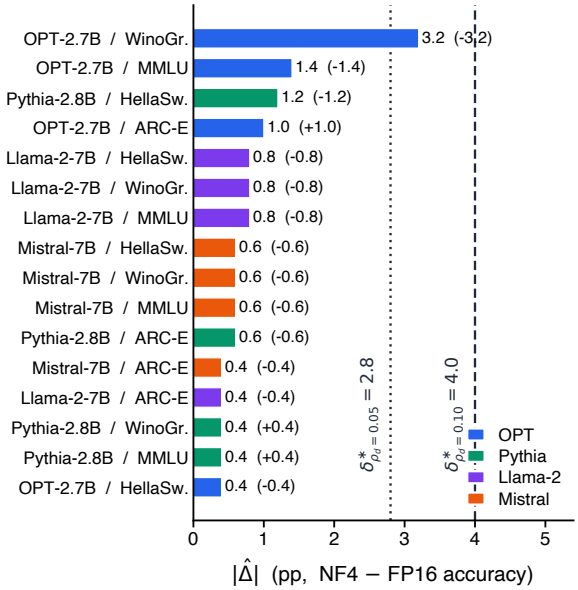

Figure 3. Observed $|\hat{\Delta}|$ per cell against the aggregate paired MDE budgets at $m{=}kn{=}500$, $\alpha{=}0.05$, power 0.80 (Eq. 2). Bars show the absolute NF4−FP16 accuracy delta from Table 2; the signed delta is in parentheses. Of the 16 cells, only OPT-WinoGrande ($|\hat{\Delta}|$=3.2 pp) crosses the tighter ρ_d=0.05 MDE; all 16 fall below the ρ_d=0.10 MDE.

eight 7B cells as independent gives p=0.0039 for the $8/8$ negative-direction outcome, but the four benchmarks within a model share weights and are not independent, so this p is an optimistic lower bound on the strength of evidence.

(ii) The aggregate MDE bound (Eq. 2, m=500, α=0.05, power $1-\beta$=0.80) is $\delta^*_{m=500} \approx 4.0$ pp assuming ρ_d=0.10 and tightens to ≈ 2.8 pp assuming ρ_d=0.05. None of the eight 7B deltas crosses either threshold. Among 3B cells, only OPT-WinoGrande (-3.2 pp) approaches the bound: it is below δ^* at ρ_d=0.10 but *above* δ^* at ρ_d=0.05, so its detectability depends on the unmeasured ρ_d. **This makes OPT-WinoGrande the single tentative observation in our audit**: it is the only cell whose "is this a real quantization effect?" answer flips with the planning value of ρ_d, and the only cell that motivates measuring ρ_d rather than assuming it. We report all deltas as observations on the audited subsample, not as tested differences; the bound is sufficient, not necessary, and assumes (A1)–(A3) of §3.

PPL changes (Table 2) follow a different pattern. WinoGrande shows the largest PPL increases at both scales (up to +9.1 for OPT, +3.1 for Llama-2-7B), likely reflecting that prompt-text PPL on a binary-cloze benchmark is more sensitive to weight perturbation than answer-token correctness is. We do not attempt to claim that PPL changes *cause* accuracy changes; the tables report descriptive co-occurrence only.

5.3. Prompt Sensitivity (MMLU only)

Table 3 shows the MMLU template study. The OPT-FP16 ($T0$=.28) vs OPT-NF4 ($T1$=.30) comparison apparently reverses the best template, but at n=50 this corresponds to a one-or-two-question reshuffling and the unpaired binomial CI on a 4-pp difference at n=50 is roughly ± 18 pp. We therefore report template ranking apparent reversals at the 3B tier as exploratory observations rather than reliable interactions; the 7B tier preserves T0 as the best template under both precisions, but with n=50 even this can be coincidental.

The substantive take-away: within-row MMLU template ranges of 2–10 pp are comparable to or larger than the largest observed quantization delta (3.2 pp). A single-template quantization audit therefore absorbs template variance into its precision-comparison noise floor, masking effects of the size we are trying to measure. Fixing the prompt template before any quantization audit is a prerequisite for a meaningful precision-vs-precision comparison at this scale. We did not extend the prompt sweep to ARC, WinoGrande, or HellaSwag in this study, so this take-away is conditional on the MMLU subsample.

Table 3. MMLU accuracy across three prompt templates (n=50 each). Range is the maximum within-row template spread on these subsamples. With n=50, even a 10 pp range corresponds to only five extra correct answers; CIs and ranking-reversal claims are exploratory.

Model	T0	T1	T2	Range
OPT FP16	.280	.240	.180	.100
OPT NF4	.280	.300	.220	.080
Pythia FP16	.160	.220	.200	.060
Pythia NF4	.240	.180	.160	.080
Llama-2 FP16	.460	.420	.380	.080
Llama-2 NF4	.440	.420	.400	.040
Mistral FP16	.540	.500	.480	.060
Mistral NF4	.520	.500	.500	.020

5.4. Quantization Reliability Index (QRI)

For each cell c we pool the FP16 and NF4 split SDs into a single denominator by RMS,

$$\hat{\sigma}_{\text{split}}(c) \;=\; \sqrt{\tfrac{1}{2}\Big[(\hat{\sigma}^{\text{FP16}}_{\text{split}}(c))^2 + (\hat{\sigma}^{\text{NF4}}_{\text{split}}(c))^2\Big]}, \qquad (3)$$

as a per-cell single-split noise scale (not the unpaired-difference SD, which would be $\sqrt{2}$ larger). We then define two cell-level diagnostics:

$$\text{QRI}_{\text{split}}(c) = \frac{|\hat{\Delta}_{\text{quant}}(c)|}{\hat{\sigma}_{\text{split}}(c)} \quad \text{for all 16 cells,} \qquad (4)$$

$$\text{QRI}_{\text{combined}}(c) = \frac{|\hat{\Delta}_{\text{quant}}(c)|}{\sqrt{\hat{\sigma}^2_{\text{split}}(c) + \hat{\sigma}^2_{\text{prompt}}(c)}} \quad \text{for the 4 MMLU cells only.} \qquad (5)$$

$\text{QRI}_{\text{combined}}$ is undefined for the 12 non-MMLU cells because we did not measure prompt variance there. We deliberately do not collapse the 16 cell-level QRIs into a single global number.

QRI as a descriptive heuristic, not a test. QRI is a per-cell normalization of $|\hat{\Delta}|$ by an observed noise scale, useful as a quick visual flag for "loud" cells; power-controlled decisions should still compare $|\hat{\Delta}|$ directly to δ^* in pp (Figure 3). A threshold in QRI units would be denominator-dependent, so we quote QRI as a descriptive ratio only.

Per-cell results. With the pooled-RMS denominator (Appendix C, Table 5), $\text{QRI}_{\text{split}}$ ranges from 0.09 (Pythia MMLU) to 0.76 (OPT WinoGrande), the latter also being the top cell by $|\hat{\Delta}|$; the largest 7B value is 0.25 (Llama-2 HellaSwag). For the four MMLU cells, $\text{QRI}_{\text{combined}}$ ranges from 0.07 (Pythia) to 0.18 (OPT).

Headline. Across the 16 cells (Figure 3), the 3B-tier maximum $|\hat{\Delta}|$=3.2 pp (OPT-WinoGrande) falls below the implied MDE assuming ρ_d=0.10 ($\delta^*\approx$4.0 pp) but exceeds the implied MDE assuming ρ_d=0.05 ($\delta^*\approx$2.8 pp); whether this borderline is power-distinguishable depends on the unmeasured ρ_d. The 7B-tier maximum is 0.8 pp, well below either implied MDE. The full per-cell version is in Appendix C, Table 6.

6. Discussion

What our audit can and cannot tell us. On the audited subsamples, FP16-vs-NF4 accuracy differences are small relative to both binomial sampling SD and the paired MDE. We cannot separate "the population effect is small" from "n=100 is too small to detect it"; a CTB-style ex-ante guarantee would require either (i) a ρ_d-aware sample-size calculation done before the benchmark or (ii) a calibration-set predictor of the population δ from weight statistics. We attempt only (i).

Cross-split SD is a misleading noise proxy here. On 25 of 32 cells (Figure 2), $\hat{\sigma}_{\text{split}}$ tracks the binomial reference within ± 1.5 pp, so the conventional "cross-split SD exceeds the quantization effect, so the effect is unreliable" narrative is largely small-sample binomial noise, not an NF4 property. The largest residual, $+3.0$ pp on OPT-NF4 MMLU, is consistent with subject-mix variance from random partitions of MMLU's 57 subjects, not a quantization artifact. We therefore recommend that any $\hat{\sigma}_{\text{split}}$ report be accompanied by $\sqrt{\hat{p}(1-\hat{p})/n}$ for the same $(\hat{p}, n)$.

Prompt template choice is a confound. Within-row template ranges of 2–10 pp on MMLU are comparable to or larger than the largest observed quantization deltas. Any quantization comparison on a single template absorbs this variance into the comparison. We recommend reporting prompt-template variance for every benchmark in any quantization claim, not only MMLU.

Toward CTB-style guarantees. Our MDE bound (Eq. 2) is an ex-ante guarantee only on *detectability*, not on *magnitude*: it answers "what effect could I detect with this benchmark?" but not "what effect should I expect for this model?" Closing the second half—a predictor of ρ_d and δ from a calibration set, weight spectra, or activation outliers—is the natural next step and is precisely the theory↔benchmark bridge CTB asks for. Chang et al. (2025) report empirical predictors of input-level quantization breakdown, which is a complementary signal that future work could plug into our pre-registration template as a ρ_d prior. We leave a full predictor of δ from model statistics to future work; Berg-Kirkpatrick et al. (2012) also document the more general lesson that test-set size, gain magnitude, and system simi-

larity together govern paired NLP detectability. Auditing methodology has begun to appear in adjacent LLM and foundation-model settings, including agent safety (Luo et al., 2025) and economic-validity checks for tabular foundation models (Wang et al., 2026); a paired-MDE budget plays the analogous role for quantization-precision claims.

7. Recommendations

We recommend that quantization-evaluation reports include:

1. **Pre-registered MDE.** State $(\alpha, 1-\beta, \rho_d, m)$ and the resulting δ^* *before* reporting accuracy deltas, where m is the paired item count for the reported estimand; if k non-overlapping splits of n items are used, state whether $m{=}n$ (single-split) or $m{=}kn$ (aggregate). Effects below δ^* should be reported as "not power-distinguishable at this sample size," not as "small."
2. **Paired statistics.** Retain per-example correctness so that paired McNemar/bootstrap estimators can be applied. When FP16 and NF4 correctness are positively correlated (the typical regime), single-split accuracy SDs overstate the relevant noise level for the paired delta; the amount depends on the unmeasured paired covariance.
3. **Binomial-reference decomposition.** Report $\hat{\sigma}_{\text{split}}$ alongside $\sigma_{\text{bin}}(\hat{p}, n) = \sqrt{\hat{p}(1-\hat{p})/n}$. If the gap is small, label the residual "subset-composition variance"; if it is large, treat $\hat{\sigma}_{\text{split}}$ as a population-noise estimate.
4. **Prompt variance on every benchmark.** Pre-register a fixed-cardinality template sweep ($T{\geq}3$ from a transparent family, e.g. Sclar et al. (2024) format perturbations), paired across templates and precisions; report $\hat{\sigma}_{\text{prompt}}$ as the SD across template means within precision and $\text{QRI}_{\text{combined}}$ wherever the sweep is run, not only on MMLU.
5. **Multiple models, families, and methods.** Do not generalize from $n{=}1$ or $n{=}2$ models per scale tier. Cross-method comparison (NF4, GPTQ, AWQ, SpQR) controls for method-specific quirks.

8. Pre-Registration Template

The recommendations of §7 fit naturally into a five-line pre-registration. Filling these out before evaluation converts an opaque "small-effect" narrative into an auditable claim about what your benchmark could detect. Italics indicate values to fill in.

1. *Estimand.* Single-split ($m{=}n$) or aggregate ($m{=}kn$); state k and n. Our pilot uses $m{=}500$ aggregate.
2. *Test parameters.* α, $1-\beta$. Common defaults: 0.05, 0.80.
3. *Disagreement-rate prior.* Conservative upper bound ρ_d^{prior}, with justification (calibration set, prior literature, or sensitivity range).
4. *Computed MDE.* $\delta^*(m, \rho_d^{\text{prior}}, \alpha, 1-\beta)$ from Eq. 2, in pp.
5. *Paired retention and ρ_d revision rule.* Commit to retaining per-example correctness and reporting $(n_{10}, n_{01}, \hat{\rho}_d)$. Because $\hat{\rho}_d$ is noisy, use the Wilson upper 95% bound $U_{95}(\rho_d)$ of $(n_{10}+n_{01})/m$. If $U_{95}(\rho_d) > \rho_d^{\text{prior}}$, mark a prior violation, recompute δ^* at $\rho_d^{\text{eff}} = \max(\rho_d^{\text{prior}}, U_{95}(\rho_d))$, and re-evaluate borderline claims under the larger MDE; otherwise the pre-registered MDE remains binding.

This template is a CTB-style ex-ante guarantee on detectability: a reviewer can check, before any benchmark numbers, what claims the protocol could license.

9. Limitations

(L1) We did not retain per-example correctness, so claims are framed in unpaired terms even though FP16/NF4 share items; Eq. 2 is evaluated against a paired-disagreement *upper bound* rather than measured $\hat{\rho}_d$. (L2) We test only NF4 via BitsAndBytes; GPTQ, AWQ, SmoothQuant, and SpQR may differ. (L3) Our 7B tier has only two models; within-7B claims are observational. (L4) Prompt variance is measured only on MMLU at $n{=}50$; coverage and per-template n are too small for firm interaction claims. (L5) Our 3B and 7B subsamples are the only data we have; we do not claim population-level conclusions about NF4. (L6) Eq. 2 is a normal-approximation *conservative sufficient* bound, not an impossibility result; small-m regimes warrant exact-conditional or mid-p McNemar (Fagerland et al., 2013). (L7) Eq. 2 assumes (A1)–(A3) of §3; subject clustering on MMLU breaks (A1) and we did not run a subject-stratified bootstrap.

10. Conclusion

We reframe benchmark reliability under quantization as a minimum-detectable-effect problem and derive a paired MDE bound (Eq. 2) that, given a planning value of ρ_d, lets designers budget benchmark size against the smallest claimable quantization effect. On the audited 4×4 grid, three observations follow. (i) Assuming $\rho_d{=}0.10$, $n{=}100$ per split cannot resolve sub-percentage-point effects. (ii) Observed cross-split SD is largely binomial sampling ($25/32$ cells within ± 1.5 pp of the binomial reference). (iii) On MMLU, prompt-template variance (2–10 pp) meets or exceeds the largest observed quantization delta (3.2 pp), so a

quantization audit must fix the prompt template before any precision comparison is meaningful. The natural next step is to retain per-example correctness so ρ_d can be measured rather than assumed, validating the bound on its own audit.

Impact Statement

This work argues that benchmark comparisons between quantized and full-precision language models are routinely under-powered. By making the minimum detectable effect explicit and reframing single-number quantization claims as n-dependent, we hope to encourage more rigorous statistical practice when quantized models are deployed and compared.

References

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 2397–2430, 2023.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In *Proceedings of EMNLP-CoNLL*, pp. 995–1005, 2012.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Ting-Yun Chang, Muru Zhang, Jesse Thomason, and Robin Jia. Why do some inputs break low-bit LLM quantization? In *Proceedings of EMNLP*, pp. 3410–3429, 2025.

Robert J. Connor. Sample size for testing differences in proportions for the paired-sample design. *Biometrics*, 43(1):207–211, 1987.

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.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

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

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of ACL*, pp. 1383–1392, 2018.

Morten W. Fagerland, Stian Lydersen, and Petter Laake. The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13(1):91, 2013.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pretrained transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Ruihao Gong, Yang Yong, Shiqiao Gu, Yushi Huang, Chengtao Lv, Yunchen Zhang, Dacheng Tao, and Xianglong Liu. LLMC: Benchmarking large language model quantization with a versatile compression toolkit. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 132–152, 2024.

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

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

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. A comprehensive evaluation of quantization strategies for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

John M. Lachin. Power and sample size evaluation for the McNemar test with application to matched case-control studies. *Statistics in Medicine*, 11(9):1239–1251, 1992.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. Evaluating quantized large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 28480–28524, 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.

Haiquan Lu, Yefan Zhou, Shiwei Liu, Zhangyang Wang, Michael W. Mahoney, and Yaoqing Yang. AlphaPruning: Using heavy-tailed self regularization theory for improved layer-wise pruning of large language models.

In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Olli S. Miettinen. The matched pairs design in the case of all-or-none responses. *Biometrics*, 24(2):339–352, 1968.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 34303–34326, 2024.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

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

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Alexander Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING), Volume 2*, pp. 947–953, 2000.

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, 2019.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Jiaqi Zhao, Ming Wang, Miao Zhang, Yuzhang Shang, Xuebo Liu, Yaowei Wang, Min Zhang, and Liqiang Nie. Benchmarking post-training quantization in LLMs: Comprehensive taxonomy, unified evaluation, and comparative analysis. *arXiv preprint arXiv:2502.13178*, 2025.

Yuelyu Ji, Zhuochun Li, Rui Meng, and Daqing He. Retrieval–reasoning processes for multi-hop question answering: A four-axis design framework and empirical trends. *arXiv preprint arXiv:2601.00536*, 2026.

Yanhang Li, Zhichao Fan, and Zexin Zhuang. SafetyRepro: Configuration-conditional rank instability on alignment benchmarks. *arXiv preprint arXiv:2605.25492*, 2026.

Hanjun Luo et al. BiasIG: Benchmarking multi-dimensional social biases in text-to-image models. In *Proceedings of the 2026 International Joint Conference on Neural Networks (IJCNN)*, 2026.

Hanjun Luo et al. AtelierEval: Agentic evaluation of humans & LLMs as text-to-image prompters. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Hanjun Luo et al. AgentAuditor: Human-level safety and security evaluation for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jingzhe Shi, Qinwei Ma, Huan Ma, and Lei Li. Scaling law for time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jingzhe Shi et al. Intrinsic Entropy of Context Length Scaling in LLMs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Yingshuo Wang, Xian Sun, Yanhang Li, Zhichao Fan, and Zexin Zhuang. Auditing and fixing economic validity in tabular foundation models for discrete choice. *arXiv preprint arXiv:2605.26559*, 2026.

A. Reproducibility

Hugging Face identifiers and revisions. `facebook/opt-2.7b` (rev `main`, commit `2bb5d4b`), `EleutherAI/pythia-2.8b` (rev `main`, commit `0bcca15`), `meta-llama/Llama-2-7b-hf` (rev `main`, commit `8cca527`), `mistralai/Mistral-7B-v0.1` (rev `main`, commit `7231864`). All checkpoints are base (non-instruct) variants. The exact commit hashes above were the latest at evaluation time; users replicating this study should pin to these revisions or report any deviation.

NF4 configuration (BitsAndBytes). `BitsAndBytesConfig(load_in_4bit=True, bnb_4bit_quant_type='nf4', bnb_4bit_compute_dtype=torch.float16, bnb_4bit_use_double_quant=False)`. FP16 inference uses `torch_dtype=torch.float16`.

Library versions. `transformers==4.44.0`, `bitsandbytes==0.43.1`, `datasets≥2.14`, `torch≥2.0`, `accelerate≥0.25`.

Datasets. MMLU (`cais/mmlu`, `all`, validation), ARC-Easy (`allenai/ai2_arc`, `ARC-Easy`, test), WinoGrande (`winogrande`, `winogrande_xl`, validation), HellaSwag (`Rowan/hellaswag`, validation).

Splits. Each benchmark is randomly partitioned with NumPy seed 42 into five non-overlapping splits of 100 items. The 100-item split indices for each benchmark are released with the submission's anonymous data bundle.

Prompt templates (MMLU prompt-sensitivity sweep). T0: "Question: {q}\n{choices_formatted}\nAnswer:". T1: "Q: {q}\nOptions: A. {a} B. {b} C. {c} D. {d}\nA:". T2: "{q}\n(A) {a} (B) {b} (C) {c} (D) {d}\nThe answer is". The default benchmark sweep (Tables 1–2) uses T0.

Scoring. For multiple-choice items we compute the per-token log-likelihood of each candidate continuation under the fixed prompt template and select the argmax. For WinoGrande we compute the joint likelihood ratio of the two pronoun-resolved continuations. "Prompt PPL" is $\exp(\bar{\ell})$ where $\bar{\ell}$ is the per-token NLL over the prompt text only (no answer tokens), averaged across examples within a split.

B. Per-Cell Binomial Reference SD

Table 4 reports, for each of the 32 (model×benchmark×precision) cells, the observed split mean $\hat{p}$, the observed cross-split SD $\hat{\sigma}_{\text{split}}$ (over k=5 splits of n=100), and the binomial reference SD $\sigma_{\text{bin}}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$. The residual $r = \hat{\sigma}_{\text{split}} - \sigma_{\text{bin}}(\hat{p})$ measures the cross-split SD beyond what binomial sampling alone predicts. Of the 32 cells, 25 have $|r| \leq 1.5$ pp, 29 have $|r| \leq 2.0$ pp, and the four largest positive residuals are $+3.0, +1.5, +1.1, +1.0$ pp (OPT-NF4 MMLU, OPT-NF4 ARC-Easy, Pythia-FP16 ARC-Easy, OPT-FP16 MMLU). All four are at low base accuracies ($\hat{p} \in [0.24, 0.27]$); the subject-mix-variance explanation applies most directly to the two MMLU cells. The three cells with $|r| > 2.0$ pp are OPT-NF4 MMLU ($+3.0$), Pythia-NF4 ARC-Easy (-2.5), and Pythia-FP16 WinoGrande (-2.5); the two negative excursions reflect the wide chi-square sampling distribution of $\hat{\sigma}$ at k=5 and not anything specific to NF4.

Negative residuals ($\hat{\sigma}_{\text{split}} < \sigma_{\text{bin}}$) are visible in 24 of 32 cells. With k=5 splits, the sampling distribution of $\hat{\sigma}/\sigma$ is wide: under a chi-square model with $k-1$=4 degrees of freedom, $\hat{\sigma}/\sigma$ has approximate 95% sampling range $[0.35, 1.67]$, so observed $\hat{\sigma}$ can fall well below the population σ purely by sampling noise. Most negative residuals are within this range; we do not interpret them as evidence that the protocol is sub-binomial. We use "binomial reference SD" rather than "binomial floor" in this paper to acknowledge that observed cross-split SD is not bounded below by $\sqrt{p(1-p)/n}$ at finite k.

C. Per-Cell QRI Diagnostics

Table 5 reports, for each of the 16 (model×benchmark) cells, the absolute accuracy delta $|\hat{\Delta}|$, the RMS-pooled split SD $\hat{\sigma}_{\text{split}}$, and $\text{QRI}_{\text{split}}$. For the four MMLU cells the table also reports the prompt SD (RMS-pooled across precisions) and $\text{QRI}_{\text{combined}}$.

Per-cell direct comparison of observed $|\hat{\Delta}|$ to the aggregate paired MDE $\delta^*_{m=500}$ at two illustrative ρ_d values is reported in Table 6. The MDE depends only on $(m, \rho_d, \alpha, \beta)$, so it is identical across cells at fixed ρ_d; we still print it per row to

Table 4. Per-cell observed accuracy ($\hat{p}$), cross-split SD ($\hat{\sigma}_{\text{split}}$, k=5 splits of n=100), binomial reference SD $\sigma_{\text{bin}}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$, and residual $r = \hat{\sigma}_{\text{split}} - \sigma_{\text{bin}}$. All quantities in pp.

Model	Bench	Prec	$\hat{p}$	$\hat{\sigma}_{\text{split}}$	σ_{bin}	r
OPT	MMLU	FP16	25.2	5.3	4.3	+1.0
OPT	MMLU	NF4	23.8	7.3	4.3	+3.0
OPT	ARC-E	FP16	25.8	4.9	4.4	+0.5
OPT	ARC-E	NF4	26.8	5.9	4.4	+1.5
OPT	WinoGr.	FP16	59.8	4.6	4.9	−0.3
OPT	WinoGr.	NF4	56.6	3.8	5.0	−1.2
OPT	HellaSw.	FP16	43.4	3.6	5.0	−1.4
OPT	HellaSw.	NF4	43.0	3.6	5.0	−1.4
Pythia	MMLU	FP16	23.8	4.6	4.3	+0.3
Pythia	MMLU	NF4	24.2	4.7	4.3	+0.4
Pythia	ARC-E	FP16	27.4	5.6	4.5	+1.1
Pythia	ARC-E	NF4	26.8	1.9	4.4	−2.5
Pythia	WinoGr.	FP16	56.4	2.5	5.0	−2.5
Pythia	WinoGr.	NF4	56.8	4.1	5.0	−0.9
Pythia	HellaSw.	FP16	44.4	4.3	5.0	−0.7
Pythia	HellaSw.	NF4	43.2	4.9	5.0	−0.1
Llama-2	MMLU	FP16	45.8	5.1	5.0	+0.1
Llama-2	MMLU	NF4	45.0	4.8	5.0	−0.2
Llama-2	ARC-E	FP16	68.8	4.1	4.6	−0.5
Llama-2	ARC-E	NF4	68.4	3.9	4.6	−0.7
Llama-2	WinoGr.	FP16	70.0	3.6	4.6	−1.0
Llama-2	WinoGr.	NF4	69.2	3.4	4.6	−1.2
Llama-2	HellaSw.	FP16	60.8	3.1	4.9	−1.8
Llama-2	HellaSw.	NF4	60.0	3.3	4.9	−1.6
Mistral	MMLU	FP16	52.2	4.6	5.0	−0.4
Mistral	MMLU	NF4	51.6	4.9	5.0	−0.1
Mistral	ARC-E	FP16	72.4	3.7	4.5	−0.8
Mistral	ARC-E	NF4	72.0	3.5	4.5	−1.0
Mistral	WinoGr.	FP16	72.2	3.2	4.5	−1.3
Mistral	WinoGr.	NF4	71.6	3.0	4.5	−1.5
Mistral	HellaSw.	FP16	64.4	2.9	4.8	−1.9
Mistral	HellaSw.	NF4	63.8	3.1	4.8	−1.7

make the comparison eye-trackable. Only one cell, OPT-WinoGrande ($|\hat{\Delta}|$=3.2 pp), exceeds the ρ_d=0.05 MDE (2.8 pp). At ρ_d=0.10 ($\delta^*_{m=500}\approx$4.0 pp), no cell crosses the bound. We do *not* convert this comparison into a QRI threshold because converting it through $\hat{\sigma}_{\text{split}}$ requires choosing whether to use the observed RMS-pooled split SD or the binomial reference $\sqrt{\hat{p}(1-\hat{p})/n}$, and the two choices disagree on most cells; the $|\hat{\Delta}|$-vs-δ^* comparison is the cleaner statistic.

D. Wilson Confidence Intervals

Table 7 reports Wilson 95% CIs (Wilson, 1927) for the FP16 accuracy in each of the 16 model×benchmark cells, computed at n=500 (the union of the five splits). At n=500 and $\hat{p} \in [0.25, 0.72]$, Wilson half-widths range from ±3.7 to ±4.4 pp, comparable in magnitude to the largest observed NF4−FP16 deltas in Table 2. The half-width *per split* (n=100) is roughly ±8–10 pp, so split-level deltas are even less resolvable than the union-level deltas. **Caveat:** these are single-proportion Wilson intervals on the FP16 accuracy alone, not paired CIs on the NF4−FP16 delta. The paired CI is the relevant uncertainty estimator for the delta and would generally be tighter than ±(union of unpaired half-widths), since FP16 and NF4 correctness are positively correlated; the paired CI is computable only from per-example records that we did not retain.

E. Power Curves for the Paired MDE

For α=0.05 and power $1-\beta$=0.80 the paired MDE simplifies to $\delta^* \leq 2.80\sqrt{\rho_d/m}$. Sample sizes required for several (δ, ρ_d) pairs (using m for the paired item count):

Table 5. Per-cell QRI. $\hat{\sigma}_{\text{split}}$ is RMS-pooled across precisions. $\hat{\sigma}_{\text{prompt}}$ is the RMS-pool of FP16 and NF4 prompt-template SDs from Table 3, defined for the 4 MMLU cells only. All SDs in pp.

Model	Bench	$\lvert\hat{\Delta}\rvert$	$\hat{\sigma}_{\text{spl}}$	$\hat{\sigma}_{\text{pr}}$	QRI_{spl}	QRI_{cmb}
OPT	MMLU	1.4	6.4	4.6	0.22	0.18
OPT	ARC-E	1.0	5.4	–	0.19	–
OPT	WinoGr.	3.2	4.2	–	0.76	–
OPT	HellaSw.	0.4	3.6	–	0.11	–
Pythia	MMLU	0.4	4.7	3.7	0.09	0.07
Pythia	ARC-E	0.6	4.2	–	0.14	–
Pythia	WinoGr.	0.4	3.4	–	0.12	–
Pythia	HellaSw.	1.2	4.6	–	0.26	–
Llama-2	MMLU	0.8	5.0	3.2	0.16	0.14
Llama-2	ARC-E	0.4	4.0	–	0.10	–
Llama-2	WinoGr.	0.8	3.5	–	0.23	–
Llama-2	HellaSw.	0.8	3.2	–	0.25	–
Mistral	MMLU	0.6	4.8	2.3	0.13	0.11
Mistral	ARC-E	0.4	3.6	–	0.11	–
Mistral	WinoGr.	0.6	3.1	–	0.19	–
Mistral	HellaSw.	0.6	3.0	–	0.20	–

Treating $\rho_d \in [0.05, 0.20]$ as an illustrative planning range (not an empirical claim from this audit), a study that wants to make a 1-pp paired claim about NF4 effects at 80% power needs roughly 4×10^3–1.6×10^4 paired items per cell. A study at m=100 is at least an order of magnitude underpowered for sub-percentage-point effects under those planning values.

Table 6. Per-cell observed $|\hat{\Delta}_{\text{quant}}|$ vs the aggregate paired MDE $\delta^*_{m=500} = (z_{1-\alpha/2} + z_{1-\beta})\sqrt{\rho_d/m}$ at α=0.05, power $1-\beta$=0.80, for two illustrative ρ_d values. "Exceeds MDE?" is yes when $|\hat{\Delta}| > \delta^*_{m=500}$. **This is not a significance test:** it asks whether the observed aggregate delta is larger than the design's planned detectable-effect scale under the stated ρ_d.

Model	Bench	$\lvert\hat{\Delta}\rvert$	$\delta^*(0.10)$	$\delta^*(0.05)$	exceeds?
OPT	MMLU	1.4	4.0	2.8	no
OPT	ARC-E	1.0	4.0	2.8	no
OPT	WinoGr.	3.2	4.0	2.8	*yes at* ρ_d=0.05
OPT	HellaSw.	0.4	4.0	2.8	no
Pythia	MMLU	0.4	4.0	2.8	no
Pythia	ARC-E	0.6	4.0	2.8	no
Pythia	WinoGr.	0.4	4.0	2.8	no
Pythia	HellaSw.	1.2	4.0	2.8	no
Llama-2	MMLU	0.8	4.0	2.8	no
Llama-2	ARC-E	0.4	4.0	2.8	no
Llama-2	WinoGr.	0.8	4.0	2.8	no
Llama-2	HellaSw.	0.8	4.0	2.8	no
Mistral	MMLU	0.6	4.0	2.8	no
Mistral	ARC-E	0.4	4.0	2.8	no
Mistral	WinoGr.	0.6	4.0	2.8	no
Mistral	HellaSw.	0.6	4.0	2.8	no

Table 7. Wilson 95% CIs for FP16 accuracy on the union of five splits (n=500 per cell). Half-widths bound the per-cell estimation error and contextualize the ΔAcc values in Table 2.

Cell	$\hat{p}$	95% CI	$\pm$ pp
OPT MMLU	0.252	$[0.215, 0.293]$	3.9
OPT ARC-E	0.258	$[0.220, 0.300]$	4.0
OPT WinoGr.	0.598	$[0.554, 0.640]$	4.3
OPT HellaSw.	0.434	$[0.391, 0.478]$	4.4
Pythia MMLU	0.238	$[0.202, 0.279]$	3.9
Pythia ARC-E	0.274	$[0.236, 0.316]$	4.0
Pythia WinoGr.	0.564	$[0.520, 0.607]$	4.4
Pythia HellaSw.	0.444	$[0.401, 0.488]$	4.4
Llama-2 MMLU	0.458	$[0.415, 0.502]$	4.4
Llama-2 ARC-E	0.688	$[0.646, 0.727]$	4.0
Llama-2 WinoGr.	0.700	$[0.659, 0.738]$	4.0
Llama-2 HellaSw.	0.608	$[0.564, 0.650]$	4.3
Mistral MMLU	0.522	$[0.478, 0.565]$	4.4
Mistral ARC-E	0.724	$[0.683, 0.761]$	3.9
Mistral WinoGr.	0.722	$[0.681, 0.760]$	3.9
Mistral HellaSw.	0.644	$[0.601, 0.685]$	4.2

Table 8. Paired sample sizes m required to resolve a paired NF4 effect of magnitude δ at α=0.05, power $1-\beta$=0.80.

	δ=0.5 pp	δ=1 pp	δ=3 pp	δ=5 pp
ρ_d=0.05	15,680	3,920	436	157
ρ_d=0.10	31,360	7,840	871	314
ρ_d=0.20	62,720	15,680	1,742	627