The Compression Tax: An Audit of How Post-Training Quantization Redistributes Safety and Fairness Failures in Open Language Models

Anonymous Authors

ACM FAccT 2027, Submission #1159

Abstract

Post-training quantization is the default route to deploying open language models cheaply: a model is released in full precision, and downstream users run it at 8-bit or 4-bit precision to fit commodity hardware. Release-time safety and fairness evaluations, however, are almost always reported at full precision, and later evaluations of quantized models report aggregate scores. We audit whether the safety and fairness costs of quantization fall evenly, or whether they concentrate in already disadvantaged groups and in specific refusal behaviors. We evaluate three open model families (Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3) at BF16, INT8, and NF4 precision on five suites: BBQ, HarmBench, XSTest, SAE/AAVE dialect pairs built from XSTest, and RealToxicityPrompts. All comparisons are paired at the prompt level, with bootstrap confidence intervals and Holm correction, under a pre-registered analysis plan with four hypotheses. Headline results: at NF4, Llama loses 5.12 percentage points of BBQ accuracy, Qwen loses 0.83 points, and Mistral loses nothing (a 0.18 point gain, not significant), so the quantization tax is family specific, not universal. No model shows the pre-registered two-directional refusal pattern, and AAVE dialect refusal does not rise at NF4 in any model. The masking script flags seven aggregate versus group contrasts, but every flagged contrast rests on a small slice and none clears Holm correction, so masking is not a confirmed finding here. Because aggregate deltas can be near zero while per-group deltas are material, we propose a compression delta disclosure for model cards: any quantized release should report per-group safety and fairness deltas against its full precision baseline, not only aggregate benchmark scores.

1. Introduction

Quantization changes the model that users actually run. The evaluated artifact, however, is usually the full precision checkpoint: model cards, safety reports, and fairness audits describe a model that many deployments never serve. Two gaps follow. First, a measurement gap: if quantization degrades behavior at all, the degradation is a

property of the deployed system, and it is currently measured late or never. Second, a distribution gap: even where quantized models are benchmarked, results are reported as averages, and an average can stay flat while harm moves between groups.

This paper is an audit, in the FAccT sense, of that distribution question. We ask four pre-registered questions. (H1) Does per-group degradation from full precision to 4-bit concentrate in disadvantaged groups? (H2) Does quantization move refusal in both directions at once, raising compliance on harmful prompts while raising over-refusal on safe prompts, including safe prompts written in African American English? (H3) Does degradation scale with bit width, so that 8-bit effects are smaller than 4-bit effects? (H4) On at least one suite, is the aggregate delta near zero while the largest per-group delta is material, which we pre-register as at least 5 percentage points? H4 is the evidentiary basis for the disclosure proposal in Section 6: if masking occurs, aggregate-only reporting is not a neutral simplification, it is a systematic blind spot.

Contributions:

1. A paired, pre-registered audit of quantization effects on safety and fairness across three model families and five suites, disaggregated by group throughout.
2. A dialect-pair suite built from XSTest safe prompts using conservative morphosyntactic transformations from the VALUE feature inventory, with a documented human validation gate.
3. Evidence on the masking hypothesis: not confirmed as a robust finding, with the flagged small slice contrasts reported and discounted in Section 4.5, and a concrete compression delta disclosure standard for model cards, reframed as preventive rather than as a remedy for a confirmed masking instance.

2. Related work

Post-training quantization is the standard way to serve a large model on smaller hardware. LLM.int8() (Dettmers et al. 2022) keeps outlier features in higher precision while quantizing the rest to 8-bit, and bitsandbytes NF4 stores weights in a 4-bit normal float format designed for normally distributed weights (Dettmers et al. 2023). GPTQ (Frantar et al. 2023) and AWQ (Lin et al. 2024) are alternative post-training schemes that use calibration data or activation-aware scaling; we use one stack, bitsandbytes, for all models so that the treatment is comparable, and we do not claim that results transfer to GPTQ or AWQ without a rerun. Prior work on compression and fairness gives reason to look for group concentrated effects rather than only average effects: a study of post-training compression in Whisper speech models found that compression compounds temporal taxation, widening the error burden gap for already taxed speaker groups (Ginjala et al. 2026), and a small model audit found increased disability related bias

under 4-bit AWQ (Manduru and Domeniconi 2025). Both are precedents for the interaction we test, not for the specific models or suites here.

For measurement, we use established suites rather than inventing new prompts, except for the dialect pairs, which are transformations of existing prompts. BBQ (Parrish et al. 2022) measures stereotyped question answering by category in ambiguous and disambiguated contexts. HarmBench (Mazeika et al. 2024) measures compliance with harmful requests, and XSTest (Rottger et al. 2024) measures over-refusal by pairing safe prompts that resemble unsafe ones with unsafe contrast prompts. RealToxicityPrompts (Gehman et al. 2020) measures toxicity of open ended continuations. For dialect, VALUE (Ziems et al. 2022) inventories morphosyntactic features of African American English and provides the feature list our conservative transformations follow. Model cards (Mitchell et al. 2019) are the disclosure infrastructure our Section 6 proposal builds on: a card reports what was evaluated, and we argue it should also report what quantization changed, per group. Positioning: prior work studies either average quantization effects or group disparities at full precision; we study the interaction, paired per prompt.

3. Methods

3.1 Models and precisions

Three instruction-tuned open model families at matched scale (7 to 8B parameters): Meta Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3. Holding scale roughly constant lets family, tokenizer, and training mix vary while size does not. Each model is run at three precisions, the treatment variable: BF16 full precision, INT8 via bitsandbytes LLM.int8(), and NF4 4-bit via bitsandbytes. One quantization stack for all models keeps the treatment comparable. Nine configurations in total.

3.2 Evaluation suites

Fairness: BBQ, all nine bias categories in ambiguous and disambiguated contexts (58,492 items), scored by next-token logprobs over three answer choices rather than free generation. Endpoints: accuracy and stereotype rate, disaggregated by category.

Harmful compliance: HarmBench standard and contextual text behaviors (300 items, copyright behaviors excluded because upstream scoring is hash-based). Endpoint: compliance, defined as not refusing and being classified unsafe by the judge.

Over-refusal: XSTest, 250 safe prompts and 200 unsafe contrast prompts. Endpoint: refusal rate on safe prompts (over-refusal) and on unsafe prompts.

Dialect pairs: XSTest safe prompts paired with AAVE versions produced by conservative morphosyntactic transformations following the VALUE feature inventory (copula and auxiliary deletion, possessive -s deletion, negative concord, ain't; no phonological respellings). Pairs where no rule fires are dropped; 87 pairs survive transformation, and a human validation gate removed 3 more (pairs judged fragment-like), leaving 84 pairs. Endpoints: refusal rate, response toxicity, and response sentiment, as within-pair deltas.

Open-ended toxicity: a frozen 1,200-prompt subset of RealToxicityPrompts (seed 13). Endpoint: Detoxify toxicity of continuations, disaggregated by demographic-mention slice.

Decoding is greedy (temperature 0), fixed seed, per-suite maximum token budgets set once and recorded in the run repository.

3.3 Judging

Refusal and compliance use two signals: Llama-Guard-3-8B classifications and refusal-string heuristics, with the classifier as the default where the two signals disagree. The pre-registered validation of that judging was run as an independent human annotation pass: one author hand-labeled 200 outputs sampled across all nine configurations and the three refusal suites (HarmBench 90, XSTest 92, dialect 18), blinded to model, precision, and both judge signals, with the sample balanced 99 to 101 on the heuristic label so that agreement cannot be inflated by base rates. Agreement between the refusal heuristic and the human labels was 84.0 percent (Cohen's kappa 0.68); by suite, HarmBench 80.0 percent (kappa 0.60), XSTest 88.0 percent (kappa 0.76), and dialect 83.3 percent (kappa 0.67). Disagreements skew one way: the heuristic missed 21 human-judged refusals, mostly paraphrased refusals and soft deflections without standard refusal phrasing, and over-flagged 11, mostly apology or cannot openings that go on to provide the requested content. Heuristic refusal counts are therefore, if anything, slight underestimates, and because the same judge was applied to every configuration, paired cross-precision deltas are not differentially biased by this judge error. Toxicity is scored with Detoxify (original model); sentiment with cardiffnlp twitter-roberta-base-sentiment.

3.4 Statistics protocol (pre-registered)

The analysis plan was frozen before any full runs (2026-10-27). All cross-precision comparisons are paired at the prompt level. We report raw percentage-point deltas, Cohen's h, and percentile 95% confidence intervals from 10,000 bootstrap resamples. Within each suite, p-values are corrected across groups with the Holm procedure at alpha 0.05. A result counts as a finding if it holds on at least two of the three model families with corrected intervals excluding zero; single-model results are reported as

exploratory. The masking analysis reports, per suite and precision, the aggregate delta, the largest per-group delta, and their ratio, together with the pre-registered H4 threshold test. Pilot runs used only to validate the harness were discarded and are never reported as results.

3.5 The audit protocol as a reusable method

The procedure above is packaged as a named protocol, the compression delta audit, so that any quantized release can be audited the same way. It has five fixed steps. (1) Generate once per precision on identical frozen prompts with identical decoding, so that every cross-precision comparison is paired at the prompt level. (2) Score pre-declared per-group endpoints, never only aggregates. (3) Judge with at least two independent signals, and validate the judging against a hand-labeled sample labeled blind. (4) Test with paired bootstrap intervals and a within-suite multiplicity correction, under a finding rule fixed in advance (here: an effect counts only if it holds on at least two model families with corrected intervals excluding zero). (5) Report a masking statistic for each suite: the largest absolute per-group delta, M, alongside the aggregate delta, flagged when M is at least 5 percentage points while the aggregate delta is at most 2 points in absolute value. The harness released with this paper is the reference implementation of the protocol. Section 4 applies it, and Section 6 turns its step 5 output into a disclosure requirement.

4. Results

All numbers in this section come from results/analysis/deltas.csv, masking.csv, and h_tests.json, with primary endpoint levels from results/summary.csv. Deltas are NF4 or INT8 minus BF16, in percentage points, unless stated otherwise. For BBQ, the delta is an accuracy drop, so a positive number means worse accuracy at lower precision. All nine configurations completed with full counts: 2,118 generations and 58,492 BBQ scores per configuration. Sanity checks passed: every rate in summary.csv lies in [0, 1], and overall BBQ accuracy at BF16 is 76.68 percent for Llama, 74.65 percent for Mistral, and 87.41 percent for Qwen, all well above the 33 percent chance level for three answer choices.

Table 1. Primary endpoints by model and precision (percent).

Model	Precision	BBQ accuracy	HarmBench compliance	XSTest safe refusal	Dialect AAVE refusal	RTP toxic rate
Llama	BF16	76.68	20.67	6.80	14.29	1.83
Llama	INT8	76.48	18.67	6.40	11.90	1.67
Llama	NF4	71.56	16.00	6.00	14.29	1.67
Mistral	BF16	74.65	69.67	0.40	4.76	3.67
Mistral	INT8	74.47	69.33	0.40	4.76	3.42
Mistral	NF4	74.83	70.00	0.40	4.76	3.08
Qwen	BF16	87.41	21.33	4.00	3.57	1.75
Qwen	INT8	87.13	22.67	3.60	2.38	1.83
Qwen	NF4	86.58	26.00	2.00	2.38	1.33

Note the level differences that are not quantization effects: Mistral refuses far less than the other two models at every precision, with HarmBench compliance near 70 percent and XSTest safe refusal at 0.40 percent, while Llama and Qwen refuse harmful prompts far more often. Quantization moves those levels only modestly, except for Llama on BBQ and Qwen on HarmBench, as detailed below.

4.1 Aggregate effects

At INT8, aggregate effects are small for every model and suite. BBQ accuracy drops are 0.20 points for Llama (95 percent CI 0.04 to 0.37), 0.17 points for Mistral (0.03 to 0.32), and 0.28 points for Qwen (0.18 to 0.37). Each is Holm significant only because n equals 58,492; Cohen's h is at most 0.008 in magnitude, which is negligible. No other INT8 aggregate delta clears Holm correction.

At NF4, the picture splits by family. Llama BBQ accuracy drops 5.12 points (CI 4.85 to 5.39, Holm p below 0.001, h equals negative 0.117). Qwen drops 0.83 points (CI 0.68 to 0.97, Holm p below 0.001, h equals negative 0.025). Mistral gains 0.18 points (CI negative 0.41 to 0.04, Holm p equals 0.115), which is no change. On HarmBench compliance at NF4, Llama falls 4.67 points (CI negative 7.67 to negative 2.00, Holm p equals 0.0006), meaning it refuses more, Qwen rises 4.67 points (CI 1.00 to 8.33, Holm p equals 0.0188), meaning it refuses less, and Mistral is flat at plus 0.33 points (CI negative 5.00 to 5.33). On RTP toxic rate and mean toxicity, no NF4 aggregate delta is significant in any model; all absolute changes are below 0.6 points for toxic rate.

4.2 Where the costs concentrate (H1)

H1 verdict: partially supported, and only on BBQ. For NF4 versus BF16, four BBQ categories show a significant positive accuracy drop in two families, Llama and Qwen,

which meets the pre-registered finding rule for those categories: Nationality (Llama plus 2.14 points, Qwen plus 1.79 points), Race x SES (Llama plus 5.88, Qwen plus 1.81), Race x gender (Llama plus 5.95, Qwen plus 0.59), and Religion (Llama plus 4.67, Qwen plus 1.42). In each case the Mistral result does not replicate: Mistral is flat or, for Race x gender, significantly improved by 1.25 points. Llama alone shows a broad NF4 BBQ drop: all eleven categories are significant, from 2.14 points for Nationality to 7.25 points for Race and ethnicity (CI 6.47 to 8.07). Qwen alone shows significant drops in only those four categories, all below 1.81 points. Mistral shows only one significant positive category drop, Race and ethnicity at plus 1.03 points, against a significant improvement in Race x gender.

Table 2. NF4 versus BF16 BBQ accuracy drop by category, in percentage points, with 95 percent CI. An asterisk marks Holm p below 0.05.

Category	Llama	Mistral	Qwen
Age	3.56 [2.47, 4.67]*	0.44 [-0.52, 1.39]	0.76 [0.03, 1.50]
Disability status	4.37 [2.76, 5.98]*	0.13 [-1.35, 1.61]	1.41 [0.26, 2.64]
Gender identity	4.87 [4.02, 5.73]*	-0.76 [-1.53, 0.00]	0.23 [-0.25, 0.71]
Nationality	2.14 [0.84, 3.41]*	0.33 [-0.75, 1.36]	1.79 [1.14, 2.47]*
Physical appearance	3.30 [1.46, 5.14]*	-0.64 [-1.97, 0.70]	-0.76 [-1.90, 0.38]
Race and ethnicity	7.25 [6.47, 8.07]*	1.03 [0.35, 1.73]*	0.47 [0.06, 0.86]
Race x SES	5.88 [5.27, 6.51]*	0.26 [-0.26, 0.77]	1.81 [1.43, 2.20]*
Race x gender	5.95 [5.44, 6.45]*	-1.25 [-1.65, -0.85]*	0.59 [0.39, 0.80]*
Religion	4.67 [2.83, 6.58]*	-0.75 [-2.17, 0.58]	1.42 [0.50, 2.42]*
SES	2.89 [2.07, 3.72]*	0.28 [-0.42, 0.96]	0.55 [0.07, 1.03]
Sexual orientation	4.75 [2.55, 6.94]*	0.93 [-0.93, 2.89]	-0.58 [-2.08, 0.93]

Outside BBQ, H1 finds nothing that clears Holm correction. No HarmBench category, no dialect variety, and no RTP demographic slice shows a significant per-group delta at NF4 in any model. The one significant RTP group result in the full table, a Mistral INT8 religion mean toxicity change, rests on n equals 7 and is not treated as a finding. H1 therefore does not support a general claim that quantization concentrates harm in disadvantaged groups. It supports a narrower claim: on BBQ, NF4 accuracy loss is broad for Llama, modest and category specific for Qwen, and absent for Mistral.

4.3 Refusal moves in both directions (H2)

H2 verdict: not supported in any model. H2 required, at NF4, both a rise in HarmBench compliance and a rise in XSTest safe refusal. No model shows both. Llama compliance falls 4.67 points and safe refusal falls 0.80 points (CI negative 3.20 to 1.60, not

significant). Mistral compliance rises 0.33 points, not significant, and safe refusal is unchanged at 0.00 points. Qwen compliance rises 4.67 points, but safe refusal falls 2.00 points (CI negative 4.00 to 0.00, Holm p equals 0.0708). In other words, where compliance rises, over-refusal does not rise with it; where compliance falls, the model becomes more refusing overall, not more permissive on harmful prompts and more refusing on safe ones at the same time.

Two adjacent refusal facts are worth reporting separately, because they are not H2. First, refusal on XSTest unsafe contrast prompts falls at NF4 in all three models: Llama from 91.5 percent to 73.0 percent, Qwen from 72.5 percent to 64.5 percent, and Mistral from 17.0 percent to 11.5 percent. That is a weakening of refusal on prompts designed to look unsafe, and it is consistent with the Qwen compliance rise but not with the Llama compliance fall, which shows the two refusal measures do not move as one mechanism. Second, dialect refusal does not rise at NF4 in any model. AAVE refusal is unchanged for Llama (14.29 percent at both precisions) and Mistral (4.76 percent at both), and falls for Qwen (3.57 percent to 2.38 percent, not significant). SAE refusal is also flat or falling: Llama plus 1.19 points, not significant; Mistral unchanged; Qwen minus 4.76 points, Holm p equals 0.064, not significant. There is no dialect over-refusal tax at NF4 in this audit.

4.4 Bit-width monotonicity (H3)

H3 verdict: supported in the weak form stated in the analysis code, and not contradicted anywhere. Of the 18 model by endpoint aggregate comparisons, 13 have a strictly larger absolute NF4 delta than INT8 delta, and the remaining 5 have exactly equal absolute magnitudes, all at or near zero: Llama dialect refusal (INT8 minus 0.60 points, NF4 plus 0.60 points), Llama RTP toxic rate (minus 0.17 points at both precisions), Mistral HarmBench compliance (minus 0.33 points at INT8, plus 0.33 points at NF4), Mistral XSTest safe refusal (0.00 points at both), and Mistral dialect refusal (0.00 points at both). No endpoint shows a materially larger INT8 effect than NF4 effect. The clearest monotone case is BBQ: INT8 drops of 0.17 to 0.28 points grow to an NF4 drop of 5.12 points for Llama and 0.83 points for Qwen, while Mistral stays near zero at both precisions. H3 is therefore not evidence of a smooth dose response. It is evidence that 4-bit effects are at least as large as 8-bit effects, with the gap concentrated in Llama on BBQ.

Table 3. Aggregate delta magnitudes, INT8 versus NF4, in percentage points, for the primary endpoints. BBQ values are accuracy drops.

Model and endpoint	INT8 delta	NF4 delta
Llama, BBQ accuracy drop	0.20	5.12
Llama, HarmBench compliance	-2.00	-4.67
Llama, XSTest safe refusal	-0.40	-0.80
Llama, dialect refusal	-0.60	0.60
Llama, RTP toxic rate	-0.17	-0.17
Mistral, BBQ accuracy drop	0.17	-0.18
Mistral, HarmBench compliance	-0.33	0.33
Mistral, XSTest safe refusal	0.00	0.00
Mistral, dialect refusal	0.00	0.00
Mistral, RTP toxic rate	-0.25	-0.58
Qwen, BBQ accuracy drop	0.28	0.83
Qwen, HarmBench compliance	1.33	4.67
Qwen, XSTest safe refusal	-0.40	-2.00
Qwen, dialect refusal	-1.19	-2.98
Qwen, RTP toxic rate	0.08	-0.42

Mean toxicity deltas, not shown in Table 3, follow the same pattern: no INT8 aggregate mean toxicity delta exceeds its NF4 counterpart in absolute value by a material margin, and none is Holm significant.

4.5 Masking (H4)

H4 verdict: not confirmed as a robust finding. The frozen masking script flags a row when the largest per-group delta is at least 5 points and the aggregate delta is at most 2 points in absolute value. It flags seven of 36 rows, so h_tests.json reports H4 masking as true under that mechanical rule. That flag alone does not make H4 a finding under the paper's own finding rule, because every flagged contrast rests on a small slice and none clears Holm correction:

Flagged contrast	Aggregate delta	Largest group delta	Slice n	Holm p
Llama INT8, HarmBench compliance, cybercrime and intrusion	-2.00 points	-7.46 points	67	0.3996
Llama NF4, RTP toxic rate, sexuality slice	-0.17 points	-9.09 points	11	1.0000
Mistral NF4, HarmBench compliance, illegal	0.33 points	7.69 points	65	1.0000
Mistral NF4, RTP mean toxicity, sexuality slice	-0.34 points	10.76 points	11	0.2070
Mistral NF4, RTP toxic rate, race slice	-0.58 points	-10.35 points	29	0.4420
Qwen NF4, RTP mean toxicity, sexuality slice	-0.44 points	-7.34 points	11	0.2752
Qwen NF4, RTP toxic rate, sexuality slice	-0.42 points	-9.09 points	11	1.0000

Figure 1, if drawn from masking.csv, would therefore show large group bars next to near zero aggregate bars for these rows, and it would mislead if presented without the slice sizes and Holm p values above. The honest headline instance is this: an aggregate near zero can hide a 7 to 11 point slice delta in these data, but the slices are too small to distinguish that delta from noise after correction, so the audit demonstrates that masking is possible to miss in reporting, not that a specific group was masked in a confirmed way. On the one large sample suite, BBQ, there is no masking instance at the 5 point threshold: INT8 aggregate drops are 0.17 to 0.28 points and the largest INT8 category drops are 1.09 points for Llama Disability status, 0.68 points for Mistral Age, and 0.55 points for Qwen SES, all below 5 points, while NF4 aggregate drops are already large for Llama. Section 6 is reframed accordingly: the disclosure proposal is preventive, and it is justified by the cost and the reporting gap, not by a confirmed masking finding.

5. Discussion

For deployers who quantize someone else's evaluated model, the main lesson is to check the specific model, not the general claim about quantization. A deployer who read only that 4-bit quantization is usually safe would miss a 5.12 point BBQ accuracy loss for Llama at NF4, while a deployer who read only that loss would overstate the risk for Mistral, which shows no NF4 BBQ loss and no refusal change at all. Full precision evaluations do not transfer automatically, but neither does a tax found in one family transfer to another. The paired audit in this paper is the unit that transfers: same prompts, same judging, same statistics, run for the exact checkpoint and precision that will be served.

The dialect result matters even though it is null, because over-refusal is a fairness harm, not only a usability annoyance. If a model refuses safe prompts written in AAVE more often than the same prompts in SAE, speakers of that variety get less useful service

from the same system, and the harm is invisible in an aggregate refusal rate that mixes varieties together. Here, AAVE refusal does not rise at NF4 in any model, and overall dialect refusal falls significantly for Qwen at NF4, driven by SAE prompts. That null does not prove that dialect bias is absent in general. It proves only that, for these 84 rule based pairs, these three models, and this refusal rule, 4-bit quantization does not add a dialect refusal tax. A different dialect sample, naturally occurring text, or a larger pair set could find one, and the pair set here is small enough that a change of a few prompts moves the rate by more than a point.

Null results, stated plainly: H2 fails in all three models, H4 fails as a robust finding, and H1 fails outside BBQ. Those nulls do not rule out quantization effects in general. They rule out the specific pre-registered patterns at this scale, with these suites, and with these sample sizes: HarmBench has 300 prompts, XSTest has 250 safe prompts, dialect has 84 pairs per variety, and several RTP slices have 11 to 29 prompts. Effects below about 5 points on those suites are hard to distinguish from noise after Holm correction, and the paper does not claim they are zero. What the nulls do rule out is a large, uniform, two-directional refusal shift or a large, uniform dialect tax at NF4 for these models. The one large effect that survives is BBQ accuracy loss for Llama at NF4, which is large, broad across categories, and precisely estimated because n equals 58,492.

Cost accounting: the full audit, from the first full run to scoring, ran on one A100 80GB instance on Verda at 2.325 dollars per hour, for about 7.5 hours, for about 17.40 dollars, from a 500 dollar coupon credit. Judging and scoring added about half an hour beyond the generation runs. At that cost, a per-group compression delta table is not a credible burden to produce for a quantized release, and cost is not a credible barrier to the disclosure proposed in Section 6.

6. Proposal: compression delta disclosure

We propose that any model card or release note for a quantized model include a compression delta table with, at minimum: (1) the full precision baseline and the quantization method and bit width; (2) per-group deltas, not only aggregates, for every safety and fairness suite the baseline reports; (3) a masking line stating the largest per-group delta alongside the aggregate delta for each suite, with the group sample size and the corrected significance result, so that a large slice delta on n equals 11 is not presented as if it were a large sample finding; (4) the judging method and its measured agreement with human labels on a hand-labeled sample. The disclosure is cheap (Section 5, cost accounting), it is mechanical to produce from a paired audit harness like the one released with this paper, and it is preventive: H4 was not confirmed in this audit, so the proposal is not presented as a remedy for a confirmed masking instance,

but as a reporting standard that would make a future masking instance visible instead of hidden in an aggregate.

7. Limitations and threats to validity

Scale and families: three families at 7 to 8B parameters, one quantization stack (bitsandbytes). Results may not transfer to larger models, other stacks (GPTQ, AWQ), or fine-tuned derivatives. Suite limits: BBQ is scored by logprobs over provided options, which measures choice behavior, not generation; HarmBench and XSTest prompts are English and largely US-centric. Dialect suite: the AAVE versions are rule-based transformations of SAE prompts, not naturally occurring AAVE text; they isolate a small set of morphosyntactic features and cannot represent the variety as a whole. Pairs deleted at the human gate: 3 of 87. Judge limits: Llama-Guard-3-8B is itself a classifier with its own group-conditional error rates. The hand-label check (Section 3.3) measured 84.0 percent agreement and Cohen's kappa 0.68 for the refusal heuristic against a single annotator, who was an author, so no inter-rater reliability is available; judge error is not zero, and it skews toward missed paraphrased refusals. Detoxify is known to score dialect features unevenly, which is one reason toxicity is never reported without the refusal endpoint alongside it. Refusal heuristics are string-based and miss paraphrased refusals; the classifier is the default on disagreement.

8. Conclusion

Quantization does not impose one uniform tax. At NF4, Llama loses 5.12 points of BBQ accuracy across categories, Qwen loses 0.83 points in a few categories, and Mistral loses nothing, while refusal moves in opposite directions by family and does not rise for AAVE dialect prompts in any model. Because aggregate reporting would hide exactly that family and group variation, quantized releases should report per-group compression deltas against their full precision baseline, with sample sizes and corrected significance, as a preventive standard.

Endmatter (FAccT required)

Generative AI statement: AI tools were used in drafting this manuscript and in running the audit pipeline under author direction. All numbers in Section 4 were taken from the frozen analysis outputs (results/analysis/deltas.csv, masking.csv, and h_tests.json) and from results/summary.csv, and the authors are responsible for checking those numbers before submission. The 200 validation labels reported in Section 3.3 were produced manually by one author, blinded to model, precision, and judge outputs; no generative AI tool was used to produce them.

Ethics and positionality: This audit studies English, largely US-centric suites, and a rule based dialect transformation, not naturally occurring African American English, so it cannot speak for the variety as a whole or for speakers outside the tested prompts. The dialect pairs were built to isolate a small set of morphosyntactic features, and the human gate removed pairs judged fragment-like; that gate reflects the authors' judgment, not community validation. Results should not be used to rank the tested model families as safe or unsafe in deployment, because judging relied on an automated classifier checked against a single annotator's hand labels (Section 3.3), and because safety depends on the deployment context, not only on benchmark prompts.

Reproducibility: the frozen run plan, harness, prompts, and analysis code accompany the submission as a supplementary archive, together with the per-configuration generation, BBQ score, and judged files and the summary and analysis tables.

References

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Tang, J., Heo, D.-H., and Bowman, S. R. (2022). BBQ: A hand-built bias benchmark for question answering. Findings of the Association for Computational Linguistics: ACL 2022.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., and Hendrycks, D. (2024). HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. Proceedings of the 41st International Conference on Machine Learning (PMLR 235).

Rottger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., and Hovy, D. (2024). XSTest: A test suite for identifying exaggerated safety behaviours in large language models. Proceedings of NAACL 2024.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. Findings of the Association for Computational Linguistics: EMNLP 2020.

Ziems, C., Chen, J., Harris, C., Anderson, J., and Yang, D. (2022). VALUE: Understanding dialect disparity in NLU. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3701-3720.

Meta (2024). Llama-Guard-3-8B model card. Hugging Face.

Hanu, L., and Unitary (2020). Detoxify. GitHub repository.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2022). LLM.int8(): 8-bit matrix multiplication for transformers at scale. Advances in Neural Information Processing Systems 35.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems 36.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. (2023). GPTQ: Accurate post-training quantization for generative pre-trained transformers. International Conference on Learning Representations.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., and Han, S. (2024). AWQ: Activation-aware weight quantization for LLM compression and acceleration. Proceedings of Machine Learning and Systems 6.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency.

Ginjala, S., Fosler-Lussier, E., Myers, C. W., and Parthasarathy, S. (2026). Temporal taxation compounds under post-training compression of Whisper models. IMPACT-SPEECH Workshop at EMNLP 2026. arXiv:2609.28739.

Manduru, S., and Domeniconi, C. (2025). Fairness is not silence: Unmasking vacuous neutrality in small language models. arXiv:2506.08487.