

INT4 Is Not One Treatment: Why the Quantisation-Safety Literature Cannot Be Reconciled as Reported, How Six Design Axes Separate Its Studies Before Bit-Width Does, and Why Two 2026 Studies of the Same Model Names Reach Opposite Verdicts

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Post-training quantisation to 4 bits is widely treated as nearly free, a judgement resting on perplexity and multiple-choice accuracy. Whether it is also free for safety is now contested: 2025-2026 studies report that 4-bit checkpoints lose refusal while their quality holds, that 3-bit weights make previously unbiased answers stereotyped in a dose-response pattern, and, in the opposite direction, that standard 4-bit quantisation is approximately safety-neutral. In this analysis the author reviews and synthesises published results to show that these findings cannot yet be set against one another, because the studies differ on six design axes before they differ on bit-width: what the treatment is, which toolchain produced the artefact, what outcome is scored, which instrument scores it, how the behaviour is elicited, and how the change is accounted. The sharpest case is a pair of 2026 studies that both quantise models named Llama 3.2 3B and Mistral 7B with AWQ to 4 bits. One reports refusal losses of 22.7 and 15.9 points; the other reports attack-success changes of +1.6 and +0.5 points. They differ on at least four of the six axes and cannot be shown to share the other two. Read along those axes, the literature supports three claims: 8-bit is close to neutral almost everywhere, retained quality cannot certify retained safety, and item-level behaviour changes even when aggregates do not. It does not support a general claim about 4-bit safety in either direction. The paper states a reporting protocol and the matched measurements that would settle the open part.

Keywords: post-training quantization; safety alignment; refusal; social bias; LLM-as-a-judge; evaluation methodology; model compression; confounding

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Quantisation is how most open-weight language models reach the hardware they run on. A checkpoint trained in 16-bit precision is mapped, after training, to 8, 4 or fewer bits per weight, by methods that range from rounding each weight to the nearest grid point to second-order reconstruction (GPTQ; Frantar et al., 2022), activation-aware scaling (AWQ; Lin et al., 2023) and mixed-precision outlier handling (LLM.int8()); Dettmers et al., 2022). The case for doing so was made on capability measurements. Dettmers and Zettlemoyer (2022), over more than 35,000 experiments, report that 4-bit precision is almost universally optimal for total model bits against zero-shot accuracy. Kurtic et al. (2024), over more than 500,000 evaluations of the Llama-3.1 family, report FP8 as effectively lossless, well-tuned INT8 within 1 to 3 percent, and INT4 weight-only quantisation as more competitive than expected. Jin et al. (2024) report that 4-bit models retain performance comparable to their full-precision counterparts and that perplexity can serve as a proxy on most benchmarks they test.

A second literature, mostly from 2024 to 2026, asks whether that case carries over to safety. Its answers conflict. Kadadekar (2026) audits a 51-row matrix of

GGUF, AWQ and GPTQ checkpoints and reports that 7 of 11 AWQ and GPTQ rows are "hidden-danger" rows, where a quality dashboard would pass the checkpoint while refusal of harmful requests falls by 12 to 68 percentage points. Rath and Maliakkal (2026) report that 3-bit quantisation causes 6 to 21 percent of previously unbiased BBQ items to produce the stereotyped answer at least once in five runs, in a dose-response pattern across bit-widths. Kharinaev et al. (2025) report that both post-training and quantisation-aware methods can degrade safety alignment. In the other direction, Prasad and Pal (2026), across 8 models, 3 precisions, 6 temperatures and about 2.0 million judged responses, report that AWQ INT4 keeps attack success within about 1.6 points of FP16, or lowers it, for 7 of 8 models. Hong et al. (2024) report that a 4-bit quantised Llama-2-13B-Chat keeps its trustworthiness within 5 points on every dimension they measure and gains more than 5 points on two. Hua et al. (2026) report that aggregate bias scores change significantly in a minority of cases, in both directions, while up to 21 percent of individual responses flip. Al Hakim et al. (2026) summarise the state of the field in one phrase: "findings remain mixed".

The question this paper asks is why. The working hypothesis behind most reviews, stated or not, is that the conflicts are about bit-width, model family and method: that some combination of these makes 4-bit safe for some models and unsafe for others, and that more

studies will map the boundary. The author argues for a different reading. The published studies differ on six design axes before they differ on bit-width, and the axes are not reported consistently enough for any two studies to be placed on a common scale. Two results that disagree may be measuring different treatments, different outcomes, through different instruments, under different elicitation, with different accounting. Until those differences are controlled, a conflict between two quantisation-safety results is not evidence about quantisation.

This paper consolidates the evidence for that reading and states what follows from it. Its contributions are four. First, it separates six axes on which the published studies differ - treatment, toolchain, outcome, instrument, elicitation and accounting - and shows, from within-study contrasts the sources themselves report, that the treatment, instrument, elicitation and accounting axes each, in at least one published comparison, move reported safety numbers by amounts comparable to or larger than the quantisation effects under dispute, that the outcome axis can reverse the sign of a result, and that the toolchain axis changes the artefact itself, although no published contrast yet measures its effect on a safety number (Sections 3 to 8). Second, it re-reads the sharpest conflict in the literature, the Kadadekar (2026) and Prasad and Pal (2026) results on models with the same names, quantised with the same method at the same nominal bit-width, and shows that the two studies differ on at least four axes and cannot be shown to share the rest (Section 9, Table 2, Figure 1). Third, it states which claims survive the confound audit and which do not (Section 10). Fourth, it gives a reporting protocol, Algorithm 1, and the matched measurements that would settle the open part (Sections 13 and 14).

The scope is deliberate. The proxy claim - that retained quality is not a safety certificate - has already been argued at length by Kadadekar (2026), under a title that states it directly, and before that by Xu et al. (2024) for perplexity. This paper does not re-argue it; it takes it as one of the claims that survives, and asks the further question of what the field's conflicting positive findings can be read to say. Hua et al. (2026) also observe that effect sizes centred on zero may explain conflicting results among prior work; this paper takes that observation as a starting point and identifies the specific design choices that produce the conflicts. Adversarial quantisation, in which a model is built to behave differently once quantised (Egashira et al., 2024), is a separate question and is treated only in Section 12, as a bound on what any full-precision evaluation can certify.

Section 2 states the conflicts precisely and consolidates the measured values in Table 1. Sections 3 to 8 take the six axes and the model population in turn. Section 9 applies them to the matched contrast. Section 10 states what survives, Section 11 what the mechanistic evidence can and cannot adjudicate, and Section 12 the

adversarial bound. Section 13 gives the protocol, Section 14 the measurements that would settle the question, Section 15 the strongest objections to this analysis, and Section 16 what is not known.

2. The conflicts, stated precisely

2.1 Three disagreements

Three published disagreements carry the debate. They are worth stating at the resolution of the sources, because the popular summaries lose exactly the details that matter.

The first is about refusal at 4 bits. Kadadekar (2026) evaluates 6 models in 4 families - Llama 3.2 1B and 3B, Qwen 2.5 1.5B and 7B, Mistral 7B and Phi-2 - across a 7-level GGUF ladder and 11 completed AWQ and GPTQ INT4 checkpoints. Quality is scored with BERTScore, ROUGE-L and an embedding coherence score; safety with AdvBench refusal, a jailbreak-amplification slice, TruthfulQA and BBQ, with refusal scored both by a gemma3:12b judge and by a regular-expression detector. A row is hidden-danger when quality stays within 2 points of baseline, or improves, while refusal falls by at least 10 points. The study reports 9 such rows and 1 near-hidden row; 7 of the 9 are AWQ or GPTQ checkpoints. On the source's direct-refusal measure, the largest are Llama 3.2 1B with GPTQ (BERTScore +8.49, direct refusal -68.18) and AWQ (+8.27, -61.82), and Phi-2 with GPTQ (+3.21, -55.45). On its separate judged-refusal measure, the largest change in any AWQ row is 56.82 points and in any GPTQ row 51.36, against 2.73 in any GGUF Q4_K_M row. Prasad and Pal (2026) quantise 8 instruction-tuned models, among them Llama-3.2-3B-Instruct and Mistral-7B-Instruct-v0.3, with GPTQ to INT8 and AWQ to INT4, and score attack success on seven harmful benchmarks with a six-judge ensemble. At greedy decoding the INT4 change in attack success is between -5.9 and +1.6 points for seven models, and +9.6 points (34.5 to 44.1 percent) for SmolLM3-3B.

The second is about bias. Rath and Maliakkal (2026) quantise Qwen2.5-7B, Mistral-7B and Phi-3.5-mini with MLX to 8, 6, 4 and 3 bits and run 12,148 ambiguous-context items from BBQ (Parrish et al., 2021) five times each. Of the items that never produced the stereotyped answer at BF16, 17.7 to 21.1 percent do so at least once at 3 bits for Mistral and Phi, and 6.0 percent for Qwen, with smaller shares at 4 bits (2.2 to 5.6 percent) and 8 bits (0.1 to 0.9 percent). Hua et al. (2026) evaluate 10 models under five post-training schemes on 13 bias datasets and report that BBQ is among the datasets where aggregate bias shows negligible change, that significant aggregate changes are as likely to reduce bias as to increase it, and that individual responses flip between biased and unbiased states up to 21 percent of the time. Marcuzzi et al. (2025) report a slight increase in stereotypes in generative tasks, especially under

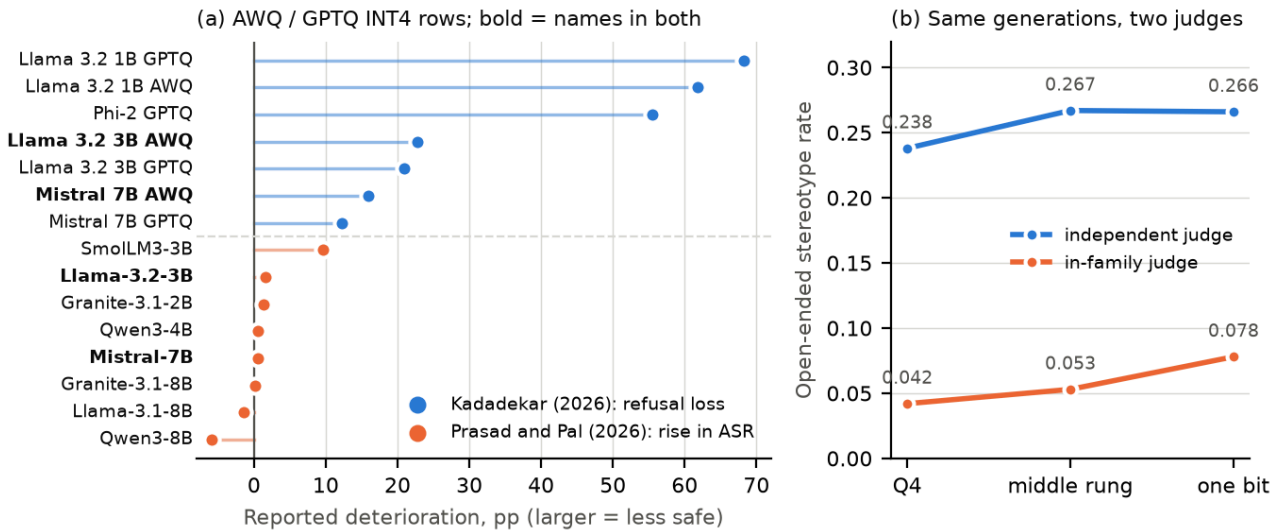


Figure 1. (a) Reported 4-bit deterioration in each source's own metric, oriented so that larger is less safe: refusal loss in percentage points for the seven AWQ and GPTQ rows Kadadekar (2026) itemises (hidden-danger rows; the four AWQ/GPTQ rows not classified hidden-danger are not itemised in the source and are omitted), and change in pooled attack success at greedy decoding for all eight AWQ INT4 rows of Prasad and Pal (2026). Model names appearing in both studies are labelled. (b) Open-ended stereotype rate on the same generations along one quantisation ladder (Qwen3.6-27B, reasoning off), scored by an in-family judge and by an independent judge (Ferrara, 2026); the three points are the source's reasoning-off ladder from its Q4 build to its one-bit build.

aggressive compression, with reduced toxicity. Goncalves and Strubell (2023) report a regulariser effect of quantisation on social bias. Xu et al. (2024) report that quantisation mostly preserves bias where pruning degrades it quickly.

The third is about which kind of quantisation is dangerous. Kharinaev et al. (2025) are sometimes read as finding that quantisation-aware training degrades safety while post-training effects are inconclusive. That is Al Hakim et al.'s (2026) summary of the literature as a whole, not the source's finding: Kharinaev et al. report that both post-training and quantisation-aware methods can degrade safety alignment, with the quantisation-aware methods STE and QLoRA performing less safely, and that no single method is best across benchmarks, precisions and models. Chen et al. (2025) report that every quantisation technique they test degrades safety, with post-training methods declining more severely. Prasad and Pal (2026) summarise Chen et al. (2025) as finding that AWQ INT4 can reduce attack success; the author could not locate that finding in the source, whose AWQ results show attack success rising for Llama-2-7B-Chat under every decoding setting of its sampling attack: at temperature 0.95 from 29.8 to 37.1 percent at INT4, and in the worst case across its six decoding settings from 29.8 to 42.4 percent. The source's prose attaches the worst-case figures (42.4 percent at INT4, 39.1 at INT8) to temperature 0.95, while its Table 4 places them in the top-p 0.7 column. A literature whose own summaries of each other disagree with the sources is a literature in which conflicts propagate.

The disagreement is not new. Kumar et al. (2024a) report that quantisation has variable effects on jailbreak

success across Mistral, Llama, Qwen and MosaicML models, where fine-tuning increases it more consistently. Belkhiter et al. (2024), on Vicuna 13B, report that AWQ and GPTQ make the model more robust to transferred attacks and more vulnerable to direct ones. Kirsten et al. (2024) find the bias effects of inference acceleration, quantisation included, complex and unpredictable: a combination of strategy and bias type that changes little in one model changes a great deal in another. Li et al. (2024) include trustworthiness among five task types in an evaluation of 11 model families, and Lin et al. (2026) report that, for small models, quantisation keeps trustworthiness intact markedly better than pruning does. Each of these is consistent with its own design; read together, they are a record of a quantity whose sign depends on how it is measured.

2.2 The measured values

Table 1 consolidates the measured values the argument rests on. Each row is a result one source reports about its own matrix; no row compares numbers across sources. The columns record what Sections 3 to 8 show to be the axes along which the rows cannot be compared: the method and toolchain, the outcome and its instrument, and the elicitation.

Three things are visible in Table 1 before any argument. The outcomes differ: refusal, attack success, multiple-choice ethics accuracy, stereotype selection and behaviour change are different quantities, and the sign convention alone differs between rows. The toolchains differ: AutoAWQ, llmcompressor, MLX and Ollama each produce a different artefact from a nominally similar recipe. And the elicitation differs: greedy decoding, five

Table 1. Reported safety and bias changes under quantisation, one row per cited result, source on the row. "pp" is percentage points; signs are as each source reports them. Refusal is the share of harmful prompts refused (higher is safer); ASR is attack success (lower is safer); SRS is the share of BBQ answers choosing the stereotype.

Source	Model(s)	Method, bits, stack	Outcome, instrument	Elicitation	Reported change
Kadadekar (2026)	Llama 3.2 1B	GPTQ / AWQ INT4, AutoGPTQ / AutoAWQ	AdvBench direct refusal (source Table 1)	fixed decoding	-68.18 / -61.82 pp
Kadadekar (2026)	Llama 3.2 3B	GPTQ / AWQ INT4, same	same	same	-20.91 / -22.73 pp
Kadadekar (2026)	Mistral 7B	GPTQ / AWQ INT4, same	same	same	-12.27 / -15.91 pp
Kadadekar (2026)	six models	GGUF Q4_K_M, Ollama	judged refusal	same	max abs. 2.73 pp
Prasad and Pal (2026)	Llama-3.2-3B-Instr.	AWQ INT4, dedicated re-quantisation run, vLLM	ASR, 7 benchmarks, 6-judge vote	greedy, 5 samples	+1.6 pp
Prasad and Pal (2026)	Mistral-7B-Instr.-v0.3	AWQ INT4, llmcompressor, vLLM	same	same	+0.5 pp
Prasad and Pal (2026)	SmolLM3-3B	AWQ INT4, same	same	same	34.5 to 44.1 %
Chen et al. (2025)	Llama-2-7B-Chat	AWQ INT4 / INT8	AdvBench ASR	decoding attack, worst of 6 settings; T = 0.95 in brackets	29.8 to 42.4 / 39.1 % (29.8 to 37.1 / 35.5 %)
Chen et al. (2025)	Llama-2-7B-Chat	QLoRA on 10 benign examples	AdvBench ASR	model default	42.25 % (FP16 0.3)
Twagirayezu and Mitra (2026)	Qwen2.5-7B	GPTQ INT4, AutoGPTQ	MedSafetyBench refusal, keywords	greedy	0.373 to 0.307 (-17.8 % rel.)
Hong et al. (2024)	Llama-2-13B-Chat	GPTQ 4-bit	DecodingTrust Ethics	benchmark default	54.1 to 76.3 (prose); 62.3 to 76.2 (Table 3)
Rath and Maliakkal (2026)	Mistral-7B	MLX weight-only, g64, 3 / 4 / 8-bit	BBQ new-stereotype items	T = 0.3, 5 seeds	17.7 / 5.6 / 0.9 %
Hua et al. (2026)	10 models	RTN W8 / GPTQ, AWQ, RTN W4	behaviour changes, 13 datasets	likelihood or greedy	2 / 9, 11, 12 % avg.
Ferrara (2026)	Qwen3.6-27B	GGUF Q4 to one bit	open-ended stereotype, 2 judges	model template	0.042 to 0.078 vs 0.238 to 0.266

samples with a majority vote, temperature 0.3 and a sampling attack over six decoding settings are not the same measurement of the same model. The rest of the paper asks how much each of these differences can move a reported number.

3. Six axes, defined

The author uses six terms, defined here so that the rest of the paper can assert against them.

The treatment is what is done to the weights: rounding with no data, post-training reconstruction against calibration data, quantisation of activations or of the key-value cache, or quantisation-aware training that updates parameters by gradient descent. Bit-width is one attribute of the treatment, not the treatment.

The toolchain is the software that produces and serves the quantised artefact: the quantisation library and its version, the export format, the inference engine and its kernels, and the chat template the engine applies. Two artefacts produced by the same method at the same nominal bit-width through different toolchains are different artefacts.

The outcome is the property scored: refusal of harmful requests, harmfulness of the produced content, accuracy on a multiple-choice safety test, selection of a stereotyped option, stereotype content in open text, or a change in any of these relative to the baseline.

The instrument is what maps a response to a label: a keyword or regular expression match, a guard classifier, a general LLM judge, a human rater, or the model's own likelihood over options.

The elicitation is how behaviour is drawn out: the prompt distribution and its attack style, the language, the decoding rule and temperature, the number of samples per prompt and the aggregation over them.

The accounting is how a change is turned into a number: which baseline it is measured against, whether it is reported in absolute points or relative percent, how benchmarks are weighted, which items enter the denominator, and against which noise floor the change is judged.

A seventh factor, the model population, is not a design axis in the same sense, since it is what a study samples rather than how it measures; Section 8 treats it separately.

The claim of this paper is not that these axes matter in principle; most methods sections acknowledge that they do. The claim is quantitative where the evidence allows it: on published within-study evidence, the treatment, instrument, elicitation and accounting axes can each move reported safety numbers by amounts as large as the effects attributed to quantisation, and the outcome axis can reverse their sign. The toolchain axis is shown to change the artefact, as a collapsed export does, but no published contrast isolates its effect on a safety number.

No two of the source studies behind Table 1 share settings on all six axes.

4. The treatment axis: what "quantised" denotes

4.1 Gradient updates are not rounding

The largest confound in the literature is that "quantisation" names procedures that do and do not change the model by training. Post-training methods such as round-to-nearest, GPTQ and AWQ compute quantised weights from the full-precision weights, with or without a calibration set, and never take a gradient step on a task objective. Quantisation-aware training, including QLoRA (Dettmers et al., 2023), straight-through-estimator training and LLM-QAT, does take gradient steps, on whatever data the practitioner supplies.

That difference matters because the safety literature already knows what gradient steps do to alignment. Qi et al. (2023) report that fine-tuning on as few as 10 adversarially designed examples removes GPT-3.5 Turbo's safety guardrails, and that fine-tuning on benign, commonly used datasets also degrades alignment, to a lesser extent. A quantisation-aware result that reports degraded safety therefore confounds two treatments: lower precision and fine-tuning. The comparison that would separate them is a full-precision fine-tuning arm on the same data with the same schedule.

The clearest illustration is Chen et al. (2025). The study's "quantisation-assisting datasets" are 10-example sets constructed following Qi et al. (2023): a benign set drawn from UltraChat, and two sets derived from AdvBench that pair instructions with harmful or compliance-inducing responses. The study reports that LLM-QAT on the benign set raises attack success to 16.9 percent for Llama-2-7B-Chat at INT4, from 0.3 percent before quantisation, and that on the harmful sets attack success reaches 82.1 to 83.7 percent. Its Table 3 gives different values for the same cells, 82.9 percent on the indirect-harmful set and 71.2 percent on the direct-harmful set to which the prose assigns 83.7. QLoRA on the benign set reaches 42.25 percent. These are real results about what the procedures do. But a reader who files these harmful-set rates under "quantisation degrades safety" has filed a fine-tuning attack under a precision effect. The study's own framing - that the source of the data drives the safety outcome - is the correct reading, and it is a finding about training data, not about bits.

Kharinaev et al. (2025) find that their quantisation-aware methods perform worse on safety than their post-training methods. That finding is correctly stated as a comparison of procedures. It does not establish that low precision, as opposed to the training the procedure entails, caused the difference, and Section 10 treats it accordingly.

4.2 Weights, activations and the cache are different treatments

Weight-only quantisation is the dominant deployment mode, but the safety literature mixes it with activation and cache quantisation. Zhang, J. et al. (2025) evaluate seven activation-approximation techniques, including activation quantisation, across ten safety-aligned models and report consistent safety degradation. Wee et al. (2025) work in the W4A4 setting, in which both weights and activations are 4-bit, where round-to-nearest raises Llama-3.1-8B's WikiText-2 perplexity to 118.39, against 7.23 for the fine-tuned FP16 model. Hua et al. (2026) include SmoothQuant at W4A16, and Liu et al. (2025) separate weight, activation and cache quantisation in reasoning models, finding W8A8 lossless and W4A16 nearly lossless on accuracy and lower bit-widths risky. A safety finding at W4A4 says nothing direct about W4A16, and the reverse. Table 1's rows are all weight-only except where noted, but the reviews that summarise the field rarely carry the distinction.

4.3 The label is not the bit-width

Even within weight-only quantisation, the nominal bit-width does not state the treatment. Ferrara (2026) measures the effective bits per weight of the GGUF artefacts in its evaluation by reading the stored tensor types, and reports that nominal labels "overstate the degree of quantization by 11 to 40%": a build labelled 4-bit (UD-Q4_K_XL) stores 5.235 bits per weight for Qwen3.6-27B and 4.901 for Gemma-4-31B, and in a build labelled 2-bit the multilayer-perceptron weights sit at 2.14 bits while attention and output tensors carry 3.55 and 4.53. The GGUF k-quant family mixes formats by design. AWQ and GPTQ checkpoints at "INT4" also differ in group size, in whether activation ordering is used, and in which layers are left in higher precision. Twagirayezu and Mitra (2026) report their GPTQ settings (group size 128, no activation ordering, 128 WikiText-2 calibration samples); Rath and Maliakkal (2026) report MLX weight-only quantisation with group size 64; Kadadekar (2026) names AutoAWQ and AutoGPTQ but not, in the text, their group sizes. When one study's "4-bit" stores five bits per weight and another's stores four, a disagreement between them may be a disagreement about a quarter of the information in the weights.

4.4 Calibration data is part of the treatment

Reconstruction-based methods depend on the calibration set. Williams and Aletras (2024) report substantial variation in downstream task performance across calibration sets, contrary to earlier assumptions of robustness. Hong et al. (2024) repeat their 3-bit and 4-bit experiments with three randomly subsampled calibration sets for this reason. Prasad and Pal (2026) calibrate on the Pile validation set; Twagirayezu and Mitra (2026) on WikiText-2; Al Hakim et al. (2026) show that protecting a small set of weights identified as fairness- and safety-critical during quantisation reduces

the degradation, and Proskurina et al. (2025) add group-fairness constraints to the GPTQ objective. If the treatment can be steered towards or away from safety by the calibration objective, then calibration data is part of the treatment, and a study that does not report it has not fully specified what it tested. No study located for this paper varies calibration data while holding everything else fixed and reports safety, apart from the fine-tuning-style "assisting datasets" of Chen et al. (2025), which Section 4.1 has already distinguished.

5. The toolchain axis: the same recipe, a different artefact

5.1 Silent export failure

Prasad and Pal (2026) report a failure that should be quoted in every quantisation-safety methods section: for two of their eight models, Qwen3-4B and Llama-3.2-3B, the default llmcompressor AWQ INT4 export "collapsed into degenerate outputs", and their INT4 results come from a dedicated re-quantisation run. The authors draw the right lesson: a collapse or quality gate should precede any safety claim. The broader lesson is that a named method at a named bit-width does not identify an artefact. For the same model and method, the default llmcompressor export collapsed, and a dedicated AWQ re-quantisation run, which the source describes as "a different AWQ export" (Prasad and Pal, 2026), produced a working checkpoint.

A degenerate checkpoint scores strangely on safety measures in both directions. A model emitting repeated tokens does not refuse, so a refusal metric reads it as unsafe; it does not produce harmful content either, so a harmfulness classifier reads it as safe. Kadadekar (2026) separates such cases explicitly: its Qwen 2.5 1.5B AWQ and GPTQ rows are joint quality-and-safety failures, where quality collapses alongside safety. The point here is that whether a given study's "INT4" row is a degraded model or a broken export is a property of the toolchain run, not of the method, and is not visible in a summary table.

5.2 Stacks within a study

The toolchain also differs within studies. In Kadadekar (2026), GGUF rows were served through llama.cpp via Ollama, while AWQ and GPTQ checkpoints were produced with AutoAWQ and AutoGPTQ and evaluated on a Hugging Face-backed stack; the primary judge also ran through Ollama. The study's family-level summary reports that the largest judged refusal change in any GGUF Q4_K_M row is 2.73 points, against 56.82 for AWQ and 51.36 for GPTQ. GGUF Q4_K_M and AWQ INT4 are different methods, and the gap may be a method effect. But the two arms also differ in inference engine, kernels and, plausibly, in how the chat template is applied, and the design cannot separate these from the method. The study's checklist itself records the conclusion as within-

stack.

Chat templates are a documented safety variable in their own right. Jiang et al. (2024) show that the rigid formats aligned models are trained on can be exploited by prompts that do not follow them, across eight state-of-the-art models. An engine that applies a template differently from the one used at alignment time changes the input the model sees, and different serving stacks apply templates through different mechanisms. Rath and Maliakkal (2026) state that they use each model's native chat template; Kadadekar (2026) does not state, in the text, how templates were handled for the two stacks. Speculatively, an eight-point BERTScore gain from quantising a 1B model, as reported for the Llama 3.2 1B AWQ and GPTQ rows, is the kind of change a formatting difference between stacks would also produce; nothing in the published record tests this.

5.3 Numerical non-determinism

Even with the artefact and template fixed, the serving configuration moves outputs. Yuan et al. (2025) report that under bfloat16 and greedy decoding, a reasoning model's accuracy varies by up to 9 percent and its response length by up to 9,000 tokens across GPU count, GPU type and batch size, because floating-point addition is not associative. The full-precision baseline is therefore not a fixed point. A safety difference between a quantised and a full-precision model served on different hardware, with different kernels or batch sizes, includes a component that would appear between two full-precision runs. Section 7 returns to what this implies for the noise floor.

6. The outcome and instrument axes: what is scored, and by whom

6.1 Refusal is not harm

Kadadekar (2026) scores refusal; Prasad and Pal (2026) score attack success, meaning the share of harmful prompts on which a guard-classifier majority labels the response harmful. These are related but not the same quantity, and they can diverge in exactly the regime quantisation produces. A model that stops refusing but answers incoherently, or answers with content too degraded to be useful, loses refusal without gaining harmfulness. Souly et al. (2024) report a phenomenon of this kind for jailbreaks: attacks that bypass a model's safety training tend to reduce its capabilities, and evaluators that do not check the usefulness of the forbidden content overstate jailbreak success relative to human judgement. Under quantisation, a refusal metric and a harmfulness metric are therefore expected to disagree most when the quantised model is damaged, and least when it is intact. That is the opposite of what a reader comparing them would assume.

Hong et al. (2024) make the accounting of refusal explicit and show that it can reverse the sign of a result.

In their DecodingTrust-based evaluation, refused answers count as wrong answers in adversarial classification, as negative predictions in the false-positive-rate tasks, and as safe in the privacy tasks. Their prose reports that 4-bit GPTQ raises Llama-2-13B-Chat's Ethics score from 54.1 to 76.3, while their Table 3 gives 62.3 for the dense model and 76.2 at 4 bits, so the gain is 14 to 22 points depending on which of the source's figures is used. In the same analysis they report that the 4-bit model reduces refusal rates substantially relative to higher-bit models. The gain is real under the benchmark's scoring rule, and the authors attribute part of it to the model recognising immoral actions it previously refused to classify. But the same behavioural change - less refusal - would lower a refusal-based safety score. A quantised model can look safer on one outcome and less safe on another because of a single shift in its refusal propensity.

6.2 Multiple choice is not generation

The bias and safety literature uses both closed-ended and open-ended outcomes, and the two diverge under quantisation. Kharinaev et al. (2025) report that multiple-choice safety evaluations often fail to capture the unsafe behaviour quantisation introduces, rating quantised models as equally safe or safer than FP16 when open-ended evaluation shows degradation, and that an ablated Llama, whose refusal mechanism has been removed, scores the same as the full-precision model on the multiple-choice SafetyBench (73.90). Ferrara (2026) reports the same split for bias: across a compression ladder, refusal of directly harmful requests stays at 0.98 and multiple-choice BBQ accuracy under ambiguity stays near 0.97, while open-ended answers carry a stereotype in roughly one in four cases under an independent judge.

The mitigation literature, meanwhile, measures itself on closed-ended outcomes. Wee et al. (2025) report their Contrastive Alignment Quantisation on SafetyBench, a multiple-choice benchmark, where it reaches 60.1 on Llama-3.1-8B against 57.5 for the strongest baseline and 62.6 for the fine-tuned FP16 model; the authors state that SafetyBench performance has been shown to correlate with safe generation. The two claims - that closed-ended safety tests miss quantisation damage, and that a closed-ended safety test validates a quantisation repair - have not been tested against each other on the same artefacts. Until they are, a mitigation's reported gain is a gain on its instrument.

How the closed-ended answer is read matters as well. Hua et al. (2026) select each option by its length-normalised likelihood, which removes sampling noise from the comparison; Rath and Maliakkal (2026) sample a single letter at temperature 0.3 and parse it. Wang et al. (2024) and Lyu et al. (2024) show that first-token likelihoods and generated text answers often disagree for instruction-tuned models, and Wang et al. analyse settings in which the mismatch rate exceeds 50 percent.

Two bias studies on the same benchmark, one reading likelihoods and one reading sampled text, are not measuring the same response.

6.3 The judge is part of the measurement

Every safety number in Table 1 passes through an instrument, and the published within-study comparisons show the instrument moving numbers by more than the effects in dispute.

Kadadekar (2026) runs both a regular-expression refusal detector and a judge, and reports a largest regex-versus-judge refusal gap of 65.45 points, for Mistral 7B at Q4_K_M, and a largest judge-side bias correction of -40.91 points, for Phi-2 with GPTQ. Its second-judge pass with Claude Sonnet 4 agrees with the primary gemma3:12b judge on 89.9 percent of 11,470 relabelled items (kappa 0.873) and changes none of the 10 hidden-danger classifications, which is strong evidence that the study's central rows are not a single-judge artefact. Per task, however, agreement ranges from 98.0 percent on BBQ to 73.8 percent on TruthfulQA.

Prasad and Pal (2026) report, across their 144 configurations, that WildGuard's configuration-level attack success agrees within 5 points with the primary LlamaGuard-3-1B judge in only 33.3 percent of configurations, with a mean difference of -0.103, while ShieldGemma-9B agrees in 100 percent. On ManyHarm at FP16 and greedy decoding, the per-judge attack success for Granite-3.1-8B spans 3.5 to 100 percent. The authors respond with a majority vote, which is the right design for their purpose; for the purpose of comparing studies, it means that a study using one judge and a study using another may disagree by tens of points on the same responses.

Ferrara (2026) gives the most direct demonstration that the instrument can manufacture or erase a dose-response. Scoring the same open-ended generations across a quantisation ladder, an in-family Qwen3-8B judge reports a stereotype rate rising from 0.042 to 0.053 to 0.078 from Q4 to one bit, while an independent judge reports 0.238, 0.267 and 0.266: higher, and roughly flat. Ferrara treats the dose-response as provisional for this reason. On a five-family screen, every Q8-to-Q2 change lies within its 95 percent interval under both judges.

Human labels are not a clean anchor either. Kharinaev et al. (2025) report a Krippendorff's alpha of 0.623 among five expert annotators, and 92 percent agreement between their Gemma 2 27B judge and the human labels. Xie et al. (2024), from more than 7,000 human annotations, show that a fine-tuned 7B judge can match GPT-4-scale judges on refusal evaluation, which makes small-judge designs defensible, but does not make two different small judges interchangeable. Inan et al. (2023) designed Llama Guard around a specific harm taxonomy; a judge's taxonomy is part of what it measures.

7. The elicitation and accounting axes

7.1 The benchmark sets the level

Prasad and Pal (2026) provide the clearest evidence that the benchmark sets the absolute level of measured safety. Averaged over their eight models at FP16 and greedy decoding, attack success is 6.9 percent on AdvBench, 18.6 percent on HarmBench and 63.2 percent on ManyHarm, a many-shot attack; models at 0.0 percent on AdvBench, including both Granite variants and Qwen3-4B, reach 80 to 100 percent on ManyHarm. Kadadekar (2026) scores refusal on AdvBench and a jailbreak-amplification slice. AdvBench (Zou et al., 2023) was introduced as the behaviour set for an adversarial-suffix attack, and on its direct prompts well-aligned models refuse nearly everything. A loss of 15 to 20 points from near-total refusal is therefore a large behavioural change, and it is also one that an average pooled with harder benchmarks, on which the baseline already fails often, can dilute. SORRY-Bench (Xie et al., 2024) was built partly because earlier sets over-represent some harm categories and under-test others, and HarmBench (Mazeika et al., 2024) to standardise red-teaming evaluation across methods and target models.

Kharinaev et al. (2025) build OpenSafetyMini by selecting XSafety questions with a high estimated deflection score and refining them with human review, and report that it separates quantisation methods that XSafety does not: AWQ and QUIK perform similarly on XSafety at 4 bits, but AWQ shows a notable drop on OpenSafetyMini while QUIK does not. The benchmark is part of the result. So is the language: Al Hakim et al. (2026) report that safety deterioration under quantisation is especially pronounced in Korean and Arabic relative to English, Marchisio et al. (2024) report that a 1.7 percent average automatic drop in Japanese corresponds to a 16.0 percent drop in human evaluation on realistic prompts, and Mekala et al. (2025) report that 4-bit degradation on long-context tasks worsens outside English.

7.2 Decoding is a treatment

The decoding rule changes safety outcomes by more than quantisation does in most of Table 1. Huang et al. (2023) report that varying decoding hyperparameters and sampling methods alone raises the misalignment rate of 11 open models from 0 percent to more than 95 percent. Chen et al. (2025) evaluate AWQ under this kind of decoding attack, across six temperature, top-k and top-p settings. The FP16 Llama-2-7B-Chat figure moves from 0.3 percent attack success in the paper's baseline table to 29.8 percent at temperature 0.95, its worst case across the six settings, a larger change than the subsequent AWQ effect (to 37.1 percent at INT4 at temperature 0.95, and to 42.4 percent in the INT4 worst case).

Sampling also adds variance that greedy evaluations hide. Larsen (2025), testing four models on 876 harmful

prompts across four temperatures and five seeds, reports that 18 to 28 percent of prompts show decision flips, refused in some configurations and complied with in others, and that for Llama 3.1 8B a single sample agrees with the multi-sample verdict 98.7 percent of the time at greedy decoding and 90.3 percent at temperature 1.0. Prasad and Pal (2026) report a decision-flip rate of 41.9 percent for Mistral-7B at FP16 and temperature 1.0. Scholten et al. (2024) show that deterministic point evaluations can falsely indicate successful alignment, because they do not capture the output distribution. A study that reports a quantisation effect from one greedy sample per prompt and a study that reports it from a five-sample majority vote are not reporting the same statistic, and only the second can say how much of its effect would recur on a rerun.

Prasad and Pal (2026) also test whether temperature and quantisation compound, and report that they do not: the interaction is sub-additive, between -0.071 and +0.018, so quantisation does not amplify the instability higher temperatures cause. That is a useful negative result. It does not make the two studies' decoding choices interchangeable, since the issue for comparison is the level and variance of the baseline each study measures against, not the interaction.

7.3 The baseline and the denominator

How a change is accounted is the least visible axis and can be the largest.

The baseline. Kadadekar (2026) measures deltas against the model-specific FP16 baseline when available and otherwise against the highest retained precision on the same grid. Which rows used which baseline is not stated in the text. Hong et al. (2024) use the 13B dense model as baseline. A delta against Q8_0 and a delta against FP16 are different quantities.

The unit. Twagirayezu and Mitra (2026) report safety changes in relative percent: Qwen2.5-7B's 17.8 percent INT4 loss on MedSafetyBench is a fall in refusal rate from 0.373 to 0.307, or 6.6 points; Meditron-7B's 28.3 percent loss is 0.102 to 0.073, or 2.9 points. Kadadekar (2026) and Prasad and Pal (2026) report absolute points. Placed in one table without conversion, of these two, the smaller absolute change reads as the larger.

The weighting. Prasad and Pal (2026) macro-average attack success over seven benchmarks of 200 to 939 prompts, and note that micro-averaging by prompt count shifts the absolute numbers. A study that pools on a different weighting reports a different level from the same responses.

The denominator. Rath and Maliakkal (2026) compute their headline transition rate on the items that showed zero stereotyped answers in all five BF16 runs, and count an item as having become biased if any of its five quantised runs selects the stereotype. With sampling at temperature 0.3, selecting items on an extreme baseline

value and then testing whether they move is a design in which some movement is expected without any treatment, because an item can show zero stereotyped answers in five draws while having a small nonzero propensity. The study does not report the corresponding BF16-against-BF16 rerun, which would measure that expected movement directly. Its 8-bit column, where perplexity is essentially unchanged, is the closest thing to a null in the published table: 0.1 to 0.9 percent of items transition at Q8, against 2.2 to 5.6 at Q4 and 6.0 to 21.1 at Q3. The 3-bit effect is far above that floor; subtracting each model's Q8 rate, the author's arithmetic puts the 4-bit effect above that floor by 2.1 to 4.7 points rather than by the 2.2 to 5.6 the headline implies. The same analysis also reports flows in the other direction, which the transition rate omits: items that always selected the stereotype at BF16 fall from a mean of 0.958 to 0.706 at Q3. The study does not report parse-failure rates by compression level, and says so.

The source also states its 4-bit numbers inconsistently. Its abstract gives 2.5 to 5.6 percent of items developing new biases at 4 bits and a perplexity increase under 3 percent at 4 bits for all three models; its tables give 2.2 percent for Qwen2.5-7B, and 4-bit perplexity increases of 3.96 percent for Qwen2.5-7B and 10.5 percent for Phi-3.5-mini, with only Mistral-7B under 3 percent (2.80). This paper uses the table values.

The noise floor. Miller (2024) sets out how evaluation results should carry standard errors and paired comparisons, and the quantisation-safety studies vary widely in whether they do. Prasad and Pal (2026) give paired bootstrap intervals on every headline delta; Kadadekar (2026) gives intervals on correlations and the second-judge kappa, and discusses its 10-point threshold against a 2.27-point margin for its most marginal row, but does not report intervals on individual refusal deltas, and states that it did not run a formal equivalence test for every cell. Hua et al. (2026) use permutation tests with false discovery rate control and report that 17.8 percent of aggregate changes are significant before correction and 11.4 percent after. That figure is perhaps the most useful single number in the literature for this paper's purpose: in the study that controls hardest for chance, most aggregate bias changes under quantisation are not distinguishable from it.

8. The model population

The models a study samples are not a design axis in the same sense, but they interact with every axis above, and the published results disagree about which model properties predict vulnerability.

Baseline alignment strength is the best-supported moderator. Prasad and Pal (2026) find the only large INT4 effect in their study in SmoLM3-3B, the model with the weakest baseline alignment, and suggest that well-aligned models are robust to standard post-training

quantisation while weakly aligned ones are vulnerable. Twagirayezu and Mitra (2026) report that on MedSafetyBench the model family dominates over precision, with FP16 refusal rates from 10.2 to 74.9 percent.

Scale is not a reliable moderator. Kadadekar's largest refusal losses are in its 1B model, and Mekala et al. (2025) report that Qwen-2.5 72B stays robust under BNB-nf4 on a task where Llama-3.1 70B loses 32 percent. But Hua et al. (2026) report no consistent robustness advantage for larger models in bias flipping, with 14B models as unstable as 0.5B models. Training length complicates scale. Ouyang et al. (2024), across more than 1,500 quantised checkpoints, report that low-bit quantisation degrades more heavily trained models more, and Kumar et al. (2024b) derive scaling laws in which post-training quantisation degradation grows with pretraining data. A 2026 model of a given size is not the same quantisation subject as a 2023 model of that size, so studies that sample different model generations are sampling different populations.

Uncertainty is the best-supported item-level predictor. Hua et al. (2026) report that responses with high pre-quantisation uncertainty are 3 to 11 times more likely to change than confident ones, and Proskurina et al. (2024) report that 4-bit GPTQ quantisation disproportionately affects samples on which the full model already had low confidence. This connects to the older finding of Hooker et al. (2019) that compression disproportionately affects a small, atypical subset of inputs while leaving top-line accuracy intact. It also explains why benchmark choice matters so much: a benchmark populated with items the model answers confidently, such as AdvBench's direct requests for well-aligned models, has little uncertain mass for quantisation to move.

9. The matched contrast, re-read

9.1 Same names, opposite verdicts

The two 2026 studies with the broadest designs give the sharpest test of the argument. Both quantise models named Llama 3.2 3B and Mistral 7B with AWQ to 4 bits. Kadadekar (2026) reports refusal losses of 22.73 points for Llama 3.2 3B and 15.91 for Mistral 7B, and with GPTQ 20.91 and 12.27, all classified as hidden-danger. Prasad and Pal (2026) report attack-success changes of +1.6 and +0.5 points for Llama-3.2-3B-Instruct and Mistral-7B-Instruct-v0.3, and describe standard quantisation as approximately safety-neutral. Read as two measurements of the same quantity, the results contradict each other by an order of magnitude. Table 2 sets the two designs side by side on the six axes.

The two studies differ on at least four of the six axes and cannot be shown to share the other two. The treatment is nominally shared, AWQ at 4 bits, but Kadadekar's group size and calibration set are not stated in the text (Section 4.3), so whether the two treatments are equivalent

Table 2. The two 2026 studies on the six axes. Entries are as each source states them; "not stated" means the author did not find the setting in the paper's text.

Axis	Kadadekar (2026)	Prasad and Pal (2026)
Treatment	AWQ / GPTQ INT4 (group size, calibration not stated); GGUF ladder	AWQ INT4; GPTQ INT8
Toolchain	AutoAWQ, AutoGPTQ, HF stack; GGUF via Ollama	llmcompressor; vLLM 0.20.0; two INT4 re-exports
Checkpoint	names only (e.g. llama3.2-3b)	Instruct versions named
Outcome	refusal, plus TruthfulQA, BBQ, jailbreak slice	attack success; over-refusal on XSTest
Instrument	gemma3:12b judge; regex; Sonnet 4 relabel	six guard classifiers, majority vote
Elicitation	AdvBench + jailbreak slice; fixed decoding, not stated	7 benchmarks, 2,552 prompts; 6 temperatures, 5 samples
Accounting	delta vs FP16 or highest retained precision; pp	paired bootstrap CIs; macro-average; pp
Llama 3.2 3B, AWQ	refusal -22.73 pp	ASR +1.6 pp
Mistral 7B, AWQ	refusal -15.91 pp	ASR +0.5 pp

cannot be established. The toolchain differs in library, AutoAWQ against llmcompressor, and in serving engine, and the Llama-3.2-3B INT4 checkpoint in Prasad and Pal comes from a dedicated re-quantisation run after the default export collapsed (Section 5.1). The checkpoint identity is not established for Kadadekar's rows, which name models in the Ollama style without version strings in the text. The outcome differs: refusal on AdvBench against attack success pooled over seven benchmarks, with over-refusal measured separately on XSTest (Rottger et al., 2023). The instrument differs: a single general-purpose judge against a majority of six guard classifiers. The elicitation differs, though it overlaps on the AdvBench direct prompts that both studies score: AdvBench and a jailbreak slice under fixed decoding settings that the text does not give, against seven benchmarks under greedy decoding with five samples per prompt. And the accounting, although both studies report absolute points, differs in baseline, in weighting and in whether intervals accompany the deltas.

Figure 1 places the two studies' reported 4-bit rows on one axis, oriented so that larger values are less safe in each source's own metric, and sets beside them the clearest published case of the instrument axis alone moving a result.

9.2 What the contrast can and cannot say

The contrast does not show that either study is wrong. Within its own design, each result is internally supported: Kadadekar's hidden-danger rows survive a strong judge swap with none of the ten changing regime, and Prasad and Pal's deltas carry paired bootstrap intervals over large prompt sets. What the contrast shows is that the two results cannot both be read as estimates of "the effect of AWQ INT4 on the safety of Llama 3.2 3B", because they are not estimates of the same quantity. At least four readings are each compatible with both results.

On the first reading, the artefacts differ: a different AWQ library, group size or export path may produce a checkpoint whose refusal behaviour moves much more,

and the Prasad and Pal report of a collapsed default export shows the range of outcomes a toolchain can produce from one recipe. On the second, the serving stacks differ: template handling, kernels and batch configuration differ between a Hugging Face generation loop and vLLM, and each changes the input or the arithmetic. On the third, the outcomes differ: a model that stops refusing direct AdvBench requests may produce content that guard classifiers do not label harmful, either because it is degraded or because it is evasive rather than compliant. On the fourth, the elicitation differs: AdvBench direct prompts are the set on which aligned models are most uniformly refusing, so a change there need not appear in a pooled average weighted equally across seven sets, in which ManyHarm contributes rates of 80 to 100 percent for several models.

The published record does not discriminate among these readings. Each predicts something checkable - the first that re-quantising with the other study's toolchain reproduces the other result, the third that guard classifiers applied to Kadadekar's responses would show smaller changes than the refusal judge - and none of those checks has been run.

10. What survives the confound audit

A confound audit would be of little use if it dissolved every finding. It does not. Agreement among studies that differ on most axes is stronger evidence than agreement among studies that share a design, for the same reason that disagreement among them is weaker evidence. Read this way, three claims survive, one is narrower than it is usually stated, and one does not survive.

10.1 Eight bits is close to neutral almost everywhere

Across studies that differ in treatment, toolchain, outcome, instrument and elicitation, 8-bit weight quantisation is consistently close to the full-precision baseline on safety and bias. Hong et al. (2024) find quantised 8-bit Llama-2-13B models a consistently comparable alternative to the dense model, with almost the same trustworthiness. Kadadekar (2026) finds no

hidden-danger rows at Q8_0, with a largest judged refusal change of 2.60 points. Rath and Maliakkal (2026) report 0.1 to 0.9 percent of BF16-unbiased items transitioning at 8 bits. Hua et al. (2026) report behaviour changes averaging 2 percent at RTN W8A16 against 9 to 13 percent for the 4-bit schemes. Twagirayezu and Mitra (2026) report INT8 GPTQ within 1.9 percent on the three multiple-choice clinical benchmarks, and within 3 percent relative (under one point absolute) on MedSafetyBench. Prasad and Pal (2026) report GPTQ INT8 attack-success changes from -7.5 to +2.7 points, with the largest increase for Qwen3-4B. On capability, Kurtic et al. (2024) report FP8 as effectively lossless and INT8 within 1 to 3 percent, Mekala et al. (2025) report an average drop of about 0.8 percent at 8 bits on long-context tasks, and Dutta et al. (2024) report negligible flips for GPTQ W8A16 alone among the schemes they test. Varshney (2026) detects no movement of the refusal direction at INT8, while noting that this is not an equivalence claim.

The one clear exception is Chen et al. (2025), whose AWQ INT8 results under a sampling attack raise Llama-2-7B-Chat's attack success from 29.8 to 35.5 percent at temperature 0.95, and to 39.1 percent in the worst case across six decoding settings. Under that elicitation the 8-bit model is measurably less safe, which is a reminder that "close to neutral" is a statement about the elicitations the other studies use.

10.2 Retained quality does not certify retained safety

The proxy claim survives, as a negative claim, in every design that tests it. Kadadekar (2026) shows quality-safety associations splitting sign across models for all 36 pairings it tests, and hidden-danger rows across methods. Rath and Maliakkal (2026) report that at 8 bits perplexity changes by less than 0.5 percent while some items still transition. Xu et al. (2024) argue that perplexity is a poor proxy for compression's safety effects; Jaiswal et al. (2023) that perplexity fails to capture changes in knowledge-intensive capability; Dutta et al. (2024) that accuracy within 2 percent hides answer flips of 5 percent or more, up to 13.6 percent in their tables; Duan (2026) that on Gemma-2-2B an INT7 quantisation improves perplexity while degrading 18.7 percent of active sparse-autoencoder features; Twagirayezu and Mitra (2026) that the most accuracy-robust of their models is among the two with the largest INT4 safety losses. Ren et al. (2024), at the level of benchmarks rather than compression, show that many safety benchmarks correlate highly with general capabilities, which is a reason to expect quality and safety to move together sometimes and apart at other times.

The claim that survives is narrower than "quantisation hides safety damage". It is that quality cannot license skipping a direct safety evaluation of the deployed artefact. It does not follow that safety changed wherever

quality held: in Prasad and Pal (2026), most 4-bit rows show neither.

10.3 Item-level behaviour changes when aggregates do not

The finding that individual responses change under quantisation while aggregate scores stay flat replicates across outcomes and instruments: answer flips on capability benchmarks (Dutta et al., 2024), bias flips in both directions on 13 datasets (Hua et al., 2026), decision flips concentrated on uncertain items (Hua et al., 2026; Proskurina et al., 2024), and transitions in both directions on BBQ (Rath and Maliakkal, 2026). This is compatible with both sides of the 4-bit conflict: a study that reports a neutral aggregate and a study that reports item-level change can both be right. What the item-level literature mostly lacks is the matched null, the rate at which items flip between two full-precision runs under the same elicitation, without which an item-level change rate is an upper bound on the quantisation effect. Hua et al. (2026) sidestep this for closed-ended items by reading likelihoods deterministically; studies that sample do not.

10.4 Below four bits, safety degrades, but quality usually moves too

At 3 bits and below, the studies that measure it agree that safety and bias degrade: Hong et al. (2024) at 3-bit GPTQ and AWQ, Rath and Maliakkal (2026) at 3 bits, Kadadekar (2026) in its GGUF Q2_K rows (largest judged refusal change 16.36 points) and Q3_K_S rows (8.18), Kharinaev et al. (2025) at 2 bits with QuIP#, where human-rated unsafe responses nearly double (7.4 to 13.5 percent) and errors rise sharply. The "invisible to quality" framing is weaker here than at 4 bits, but it does not fail outright. Rath and Maliakkal's 3-bit rows, which carry their largest transition rates, come with perplexity increases of 10.2 to 83.3 percent, so a perplexity gate would have flagged them. Hong et al. (2024), on the other hand, report that at 3 bits MMLU stays within their 5-point band while several trustworthiness dimensions fall further, and state that the added risk cannot be seen from benign performance alone; and two of Kadadekar's hidden-danger rows are GGUF Q3_K_S and Q2_K builds. Below four bits, then, safety degradation is well supported, and whether quality reveals it depends on the quality measure: perplexity mostly does in the published rows, while benign accuracy sometimes does not. In Kadadekar's matrix, the hidden-danger pattern is concentrated in AWQ and GPTQ rows rather than GGUF rows.

10.5 What does not survive: a general verdict on 4-bit safety

The claim that standard 4-bit post-training quantisation degrades safety alignment, and the claim that it is approximately safety-neutral, are both statements the published evidence does not support in general. Each is

supported within one design. Across designs, the 4-bit results span refusal losses above 60 points and attack-success reductions of several points, and the variation between studies is as well explained by the six axes as by any property of the models or methods. Hua et al.'s (2026) 4-bit behaviour-change rates of 9 to 13 percent, against 2 percent at 8 bits, locate 4 bits as the regime in which the treatment does move behaviour; what the literature cannot yet say is in which direction, by how much, and on which outcome, for a given artefact.

11. What the mechanistic evidence can and cannot adjudicate

Mechanistic results are sometimes offered as a way to settle behavioural disagreements. In this literature they constrain the reading but do not settle it.

Safety behaviour is concentrated. Ardit et al. (2024) show that refusal in 13 open chat models up to 72B parameters is mediated by a single direction in the residual stream. Wei et al. (2024) isolate safety-critical regions comprising about 3 percent of parameters and 2.5 percent of ranks, whose removal compromises safety with little effect on utility. Qi et al. (2024) argue that safety alignment is shallow, adapting the generative distribution mainly over the first few output tokens. Each of these predicts that a perturbation which happens to fall on a small set of weights, or to shift the first tokens of a response, could change safety while leaving capability nearly intact - which is the hidden-danger pattern. Kadadekar (2026) reports that probe-identified safety-associated neurons absorb 1.39 times more quantisation error than other neurons across its AWQ and GPTQ cells, with the effect global rather than specific to hidden-danger rows, and that first-token entropy shift does not separate dangerous from neutral rows.

The concentrated-mechanism picture also predicts that damage should be repairable by protecting a small set of weights, and the mitigation literature reports exactly that: Al Hakim et al. (2026) with critical weight protection, Wee et al. (2025) with a contrastive alignment term in the quantisation objective, Chen et al. (2025) with a safety patch restricted to safety-critical weights. Zhang, N. et al. (2025) report, for reasoning rather than safety, that protecting 2 percent of excessively compressed weights raises average accuracy by 6.57 percent in their compressed reasoning models.

The picture also predicts thresholds rather than smooth decline. Peng et al. (2024) report a "safety basin": random perturbations of an aligned model's weights keep its safety level within a local neighbourhood, and beyond it safety drops sharply, in contrast to capability, which declines gradually. If quantisation error acts like such a perturbation, then whether a given artefact lands inside or outside the basin would depend on details of the error - the method, group size and outlier handling -

which is the treatment and toolchain axis restated as a mechanism. This is a hypothesis; no study located for this paper places quantised artefacts on a measured safety landscape.

What the mechanistic evidence does not do is fix the sign. Perturbing weights can also increase safety: Hasan et al. (2024) report that moderate Wanda pruning increases jailbreak resistance; Hong et al. (2024) report 4-bit gains in Ethics and Fairness; Liu, Y. et al. (2024) report that quantising fine-tuning deltas mitigates alignment-breaking risk by up to 66.17 percent; Tan et al. (2026) build a safety-recovery method on post-training quantisation. Hua et al. (2026) report bias flips in both directions. A mechanism that predicts large, threshold-like changes concentrated on a small set of weights is compatible with large changes of either sign, and so with both sides of the 4-bit conflict.

The representational evidence has its own measurement problem. Kadadekar (2026) reports cosine similarity to the FP16 refusal direction above 0.97 in every quantised cell, and Chhabra and Khalili (2025), as Varshney (2026) records, report a cosine of 0.996. Varshney (2026) shows that such cosines cannot be read as preservation without the sample size behind the estimate, because two independent estimates of the same direction agree to 0.978 to 0.994 by sampling alone on the model studied; judged against a null measured within the quantised model, the INT4 direction rotated by more than the estimator's noise. Fu et al. (2025) report that quantised models retain internal truthful representations while producing false outputs under misleading prompts. Internal preservation and behavioural preservation are different claims, and the published representational evidence is not yet reported with the noise floors needed to support the first.

12. The adversarial bound

Everything above concerns quantisation applied in good faith. A separate literature shows that full-precision evaluation cannot certify a quantised artefact in the worst case. Egashira et al. (2024) construct models that behave benignly in full precision and maliciously once quantised, by constraining full-precision weights to the set that rounds to a chosen quantised model. Egashira et al. (2025) extend the attack to the GGUF formats used by llama.cpp and Ollama, reporting increases of 88.7 points in insecure code generation, 85.0 in targeted content injection and 30.1 in benign instruction refusal. Zhan et al. (2026) extend it to AWQ, GPTQ and GGUF I-quants by injecting outliers that force other weights to round to zero. Dardini et al. (2026) frame the gap formally: because quantisation is many-to-one, membership in the same quantised equivalence class does not imply equivalent behaviour, and they report backdoored translation models moving from zero measured corruption at full precision to up to 85.02 percent after quantisation. Defences that perturb weights before

quantisation exist (Zheng et al., 2026).

The adversarial results answer a different question from the benign ones, and they should not be cited as evidence that ordinary quantisation degrades safety. What they establish is a bound that applies to both sides of the benign debate: since a full-precision evaluation is not a certificate for the quantised artefact even in principle, the operational conclusion that the deployed artefact must itself be evaluated - which Kadadekar (2026) draws explicitly, and which Prasad and Pal's (2026) advice to check every checkpoint for collapse and to verify quantised safety for weakly aligned models approaches - does not depend on resolving the 4-bit conflict.

13. A reporting protocol

The axes in Sections 4 to 7 translate into a procedure for reading a quantisation-safety result and for reporting one. Algorithm 1 states it. Its first part decides whether two results can be set against each other; its second part lists what a study must report for its result to be placed on a common scale with others.

Two items in Algorithm 1 deserve emphasis because they are cheap and nearly absent. The first is the full-precision rerun (line 15): running the baseline twice, with fresh seeds and, where relevant, on the hardware and batch configuration of the quantised run, and reporting the rate at which items change. Larsen (2025) and Yuan et al. (2025) show that this rate is not zero even at greedy decoding; without it, every item-level transition rate in the literature is an upper bound. The second is scoring refusal and harmfulness on the same responses (line 12), which would directly show how much of the Kadadekar and Prasad and Pal disagreement is a disagreement about outcome rather than about behaviour.

The protocol does not require new methods. Every element of it has already been done by at least one study in Table 1: measured bits per weight by Ferrara (2026), a collapse check by Prasad and Pal (2026), regex and judge in parallel by Kadadekar (2026), refusal and harm scored on the same responses by Chen et al. (2025), multiple judges with agreement by Prasad and Pal (2026) and Kadadekar (2026), paired intervals by Prasad and Pal (2026), permutation tests with multiple-testing control by Hua et al. (2026), and transitions in both directions by Hua et al. (2026) and, in part, by Rath and Maliakkal (2026). What no study does is all of them together.

14. Measurements that would settle the open part

Five measurements would resolve most of what this paper leaves open. None requires a new benchmark or method.

The cross-toolchain bridge. Take the full-precision checkpoints both 2026 studies name, quantise each with AutoAWQ and with llmcompressor at matched group size, serve each through a Hugging Face generation loop and through vLLM with the template applied identically, and score the same responses with a refusal judge and with a guard-classifier ensemble, on AdvBench and on the pooled suite. Eight artefact-stack cells per model would show whether the order-of-magnitude disagreement in Table 2 lives in the artefact, the stack, the outcome or the elicitation.

The full-precision null. For every item-level quantisation result in the literature, the same elicitation run twice on the full-precision model gives the floor. For the BBQ transition design of Rath and Maliakkal (2026), this is a BF16-against-BF16 rerun with fresh seeds; for greedy refusal studies, a rerun across batch sizes and hardware.

The refusal-harm matrix. Labelling each response in a quantisation study twice, once for refusal and once for harmful usefulness in the manner of Souly et al. (2024), would partition refusal losses into compliance with harmful content and degraded non-refusal. This is the measurement that would show whether hidden-danger rows are dangerous or merely no longer refusing.

The fine-tuning control. For any quantisation-aware result, a full-precision fine-tuning arm on the same data with the same schedule separates the precision effect from the training effect that Qi et al. (2023) already document.

The instrument cross-check for mitigations. Evaluating the mitigation methods that report gains on closed-ended benchmarks (Wee et al., 2025) on open-ended safety prompts with independent judges would test the claim of Kharinaev et al. (2025) that closed-ended evaluation misses quantisation damage, on the artefacts where it matters most.

15. Objections

The strongest objection is that the confound reading cannot explain Kadadekar's (2026) central rows, which survive a strong second judge with none of ten changing regime. The objection is correct as far as it goes, and this paper does not claim otherwise: within Kadadekar's design, the hidden-danger rows are robust to the instrument. The claim here is about the comparison between studies, which involves all six axes, of which a judge swap within one study tests one.

A second objection is that a confound audit can dissolve any empirical literature, and so says nothing. The audit here does not dissolve the literature. It leaves three claims standing, one narrowed, and one withdrawn, and it identifies which design differences would have to be removed to resolve the withdrawn one. A conflict that can be settled by specified measurements is not dissolved; it is located.

Algorithm 1. Comparing and reporting quantisation-safety results.

```

Input: results R1, R2, each a reported safety change for a quantised artefact
Part A – can R1 and R2 be compared?
1  For axis in [treatment, toolchain, outcome, instrument,
    elicitation, accounting]:
2      if the axis setting is unreported in R1 or R2: mark UNKNOWN
3      elif settings differ: mark DIFFERS
4  if treatment DIFFERS because one arm takes gradient steps:
5      stop: the comparison is precision vs fine-tuning, not precision
6  if any axis is DIFFERS or UNKNOWN:
7      report "not comparable on <axes>"; a disagreement is not evidence
    about bit-width; an agreement is evidence of robustness to <axes>
8  else: compare effect sizes against both studies' noise floors
Part B – what a result must report to be comparable
9  Treatment: method, measured bits per weight, group size, which tensors
    stay high-precision, calibration set; any gradient steps and their data
10 Toolchain: library and version, export format, engine and version,
    chat template as applied; a collapse check before any safety metric
11 Checkpoint: exact identifiers of the full-precision and quantised models
12 Outcome: refusal AND harmfulness, scored on the same responses; open-ended
    and closed-ended outcomes kept separate
13 Instrument: at least two judges from different families, agreement per
    task, and the same judges at every precision
14 Elicitation: benchmarks named and weighted, languages, decoding rule,
    temperature, samples per prompt and aggregation rule
15 Accounting: baseline precision per row; absolute points with intervals;
    the full-precision-vs-full-precision rerun rate as the noise floor;
    transitions reported in both directions

```

A third objection is that practitioners need a decision before the measurements in Section 14 exist. They have one, and it does not depend on resolving the 4-bit conflict: evaluate the artefact that is deployed, on the stack it is deployed on, with direct safety measures. Kadadekar (2026) reaches that conclusion, Prasad and Pal (2026) recommend per-checkpoint collapse checks and multi-benchmark, multi-sample safety reporting, and the adversarial bound in Section 12 makes it necessary independently of the benign evidence. What the conflict does affect is whether such evaluations can be skipped for "standard" 4-bit recipes, and on the published evidence they cannot be skipped on the authority of either study.

A fourth objection is that the convergence at 8 bits may itself be an artefact of shared design. The studies that converge share very little: they differ in method, toolchain, outcome, instrument and elicitation, and reach the same conclusion. That is the pattern a robust effect produces. The residual risk is that every study shares an elicitation blind spot, and Chen et al.'s (2025) 8-bit result under a sampling attack shows that such a blind spot exists.

A fifth objection concerns the evidence base. Most of the studies this paper reads are 2025 and 2026 preprints, several by single authors, with matrices of 3 to 10 models, at most 8B parameters in the 2026 safety studies. The conclusions here are scoped accordingly. They concern models in the 1B to 14B range quantised for local and edge deployment, which is the regime the literature covers.

16. What is not known

It is not known whether AWQ INT4 changes the refusal behaviour of Llama 3.2 3B or Mistral 7B by about one point or by about twenty. The two published answers differ on at least four design axes and cannot be shown to share the other two, and no study has run the bridging measurement.

It is not known how much of any published item-level transition rate under quantisation would appear between two full-precision runs under the same elicitation. No study reporting sampled item-level transitions reports that rerun.

It is not known how often a loss of refusal under quantisation is a gain in harmful compliance rather than a loss of coherent output. Chen et al. (2025) score the same quantised Llama-2-7B-Chat responses with a refusal-keyword match and with the HarmBench classifier, and the two agree in aggregate (42.4 against 41.5 percent at INT4); Al Hakim et al. (2026) add a category for responses that are unrelated or nonsensical, and Kharinaev et al. (2025) have human raters separate unsafe responses from errors. But no study partitions each response's lost refusal into harmful compliance and degraded output, and none does so for the hidden-danger rows.

It is not known whether mitigation methods validated on multiple-choice safety benchmarks preserve open-ended safety.

It is not known whether quantisation-aware training degrades safety more than full-precision fine-tuning on the same data. The studies that compare quantisation-aware and post-training methods do not report the full-precision fine-tuning arm.

It is not known whether the hidden-danger pattern, or the safety-neutral pattern, holds above about 14B parameters. The 2026 safety matrices read here stop at 8B; among the bias studies read here, Hua et al. (2026) reach 14B, and Ferrara (2026) adds one 27B and one 31B backbone.

17. Conclusion

The quantisation-safety literature is usually read as a set of conflicting answers to one question: does 4-bit quantisation damage safety? This paper has argued that it is a set of answers to different questions, because its studies differ in what they quantise, with which toolchain, what they score, with which instrument, how they elicit behaviour and how they account for change, and because most of those differences alone move published safety numbers by as much as the effects in dispute, while the toolchain changes the artefact itself, with an effect on safety that no study has yet measured. The sharpest published conflict, between refusal losses of 15 to 23 points and attack-success changes of under 2 points for models with the same names quantised by the same method, is a conflict between two designs that differ on at least four of those settings and cannot be shown to share the rest.

Read along the six axes, the literature supports three claims: 8-bit weight quantisation is close to neutral almost everywhere it has been measured; retained quality cannot license skipping a direct safety evaluation of the deployed artefact; and individual responses change under quantisation when aggregates do not. It supports a narrower version of a fourth, that below four bits safety degrades, usually alongside a visible rise in perplexity though not always in benign accuracy. It does not support a general verdict on 4-bit safety in either direction. The measurements that would supply one are specified in Section 14, and every element of the reporting protocol that would make them comparable has already been implemented by at least one of the studies read here.

References

- Al Hakim, M. A., Wicaksono, A. F., & Koto, F. (2026). Preserving Fairness and Safety in Quantized LLMs Through Critical Weight Protection. arXiv:2601.12033. doi:10.48550/arXiv.2601.12033. [PREPRINT]
- Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., & Nanda, N. (2024). Refusal in Language Models Is Mediated by a Single Direction. arXiv:2406.11717. doi:10.48550/arXiv.2406.11717.
- Belkhit, Y., Zizzo, G., & Maffei, S. (2024). HarmLevelBench: Evaluating Harm-Level Compliance and the Impact of Quantization on Model Alignment. arXiv:2411.06835. doi:10.48550/arXiv.2411.06835. [PREPRINT]
- Chen, K., Zhang, J., Hu, J., Wang, Y., Lou, J., Feng, Z., & Song, M. (2025). Q-resafe: Assessing Safety Risks and Quantization-aware Safety Patching for Quantized Large Language Models. arXiv:2506.20251. doi:10.48550/arXiv.2506.20251.
- Chhabra, V. K., & Khalili, M. M. (2025). Towards Understanding and Improving Refusal in Compressed Models via Mechanistic Interpretability. arXiv:2504.04215. doi:10.48550/arXiv.2504.04215. [PREPRINT]
- Dardini, J., Stanzone, C., Colo, G., & Fenza, G. (2026). Quantization-Triggered Backdoors in Language Models: Cross-Quantizer Transferability and the Validation-Deployment Gap. arXiv:2608.27512. doi:10.48550/arXiv.2608.27512. [PREPRINT]
- Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. arXiv:2208.07339. doi:10.48550/arXiv.2208.07339.
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314. doi:10.48550/arXiv.2305.14314.
- Dettmers, T., & Zettlemoyer, L. (2022). The case for 4-bit precision: k-bit Inference Scaling Laws. arXiv:2212.09720. doi:10.48550/arXiv.2212.09720.
- Duan, E. (2026). Perplexity Can Miss SAE Feature Damage Under Quantization. arXiv:2606.03002. doi:10.48550/arXiv.2606.03002. [PREPRINT]
- Dutta, A., Krishnan, S., Kwatra, N., & Ramjee, R. (2024). Accuracy is Not All You Need. arXiv:2407.09141. doi:10.48550/arXiv.2407.09141.
- Egashira, K., Vero, M., Staab, R., He, J., & Vechev, M. (2024). Exploiting LLM Quantization. arXiv:2405.18137. doi:10.48550/arXiv.2405.18137.
- Egashira, K., Staab, R., Vero, M., He, J., & Vechev, M. (2025). Mind the Gap: A Practical Attack on GGUF Quantization. arXiv:2505.23786. doi:10.48550/arXiv.2505.23786.
- Ferrara, E. (2026). QuantiBias: Benchmarking Quantization-Induced Bias in LLMs. arXiv:2607.21063. doi:10.48550/arXiv.2607.21063. [PREPRINT]
- Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. arXiv:2210.17323. doi:10.48550/arXiv.2210.17323.
- Fu, Y., Long, X., Li, R., Yu, H., Sheng, M., Han, X., Yin, Y., & Li, P. (2025). Quantized but Deceptive? A Multi-Dimensional Truthfulness Evaluation of Quantized LLMs. arXiv:2508.19432. doi:10.48550/arXiv.2508.19432.
- Goncalves, G., & Strubell, E. (2023). Understanding the Effect of Model Compression on Social Bias in Large Language Models. arXiv:2312.05662. doi:10.48550/arXiv.2312.05662.
- Hasan, A., Rugina, I., & Wang, A. (2024). Pruning for Protection: Increasing Jailbreak Resistance in Aligned LLMs Without Fine-Tuning. arXiv:2401.10862. doi:10.48550/arXiv.2401.10862.
- Hong, J., Duan, J., Zhang, C., Li, Z., Xie, C., Lieberman, K., et al. (2024). Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression. arXiv:2403.15447. doi:10.48550/arXiv.2403.15447.
- Hooker, S., Courville, A., Clark, G., Dauphin, Y., & Frome, A. (2019). What Do Compressed Deep Neural Networks Forget? arXiv:1911.05248. doi:10.48550/arXiv.1911.05248.
- Hua, S. Z., Lotfi, S., & Chen, I. Y. (2026). Investigating Social Bias Changes in Quantized Language Models. arXiv:2602.06181. doi:10.48550/arXiv.2602.06181. [PREPRINT]
- Huang, Y., Gupta, S., Xia, M., Li, K., & Chen, D. (2023). Catastrophic Jailbreak of Open-source LLMs via Exploiting Generation. arXiv:2310.06987. doi:10.48550/arXiv.2310.06987.
- Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., et al. (2023). Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv:2312.06674. doi:10.48550/arXiv.2312.06674.
- Jaiswal, A., Gan, Z., Du, X., Zhang, B., Wang, Z., & Yang, Y. (2023). Compressing LLMs: The Truth is Rarely Pure and Never Simple. arXiv:2310.01382. doi:10.48550/arXiv.2310.01382.
- Jiang, F., Xu, Z., Niu, L., Lin, B. Y., & Poovendran, R. (2024). ChatBug: A Common Vulnerability of Aligned LLMs Induced by Chat Templates. arXiv:2406.12935. doi:10.48550/arXiv.2406.12935.

- Jin, R., Du, J., Huang, W., Liu, W., Luan, J., Wang, B., & Xiong, D. (2024). A Comprehensive Evaluation of Quantization Strategies for Large Language Models. arXiv:2402.16775. doi:10.48550/arXiv.2402.16775.
- Kadadekar, S. (2026). Quality Is Not a Safety Proxy Under Quantization. arXiv:2606.10154. doi:10.48550/arXiv.2606.10154. [PREPRINT]
- Kharinaev, A., Moskvoretiskii, V., Shvetsov, E., Studenikina, K., Mikhail, B., & Burnaev, E. (2025). Investigating the Impact of Quantization Methods on the Safety and Reliability of Large Language Models. arXiv:2502.15799. doi:10.48550/arXiv.2502.15799. [PREPRINT]
- Kirsten, E., Habernal, I., Nanda, V., & Zafar, M. B. (2024). The Impact of Inference Acceleration on Bias of LLMs. arXiv:2410.22118. doi:10.48550/arXiv.2410.22118.
- Kumar, D., Kumar, A., Agarwal, S., & Harshangi, P. (2024a). Fine-Tuning, Quantization, and LLMs: Navigating Unintended Outcomes. arXiv:2404.04392. doi:10.48550/arXiv.2404.04392. [PREPRINT]
- Kumar, T., Ankner, Z., Spector, B. F., Bordelon, B., Muennighoff, N., Paul, M., Pehlevan, C., Re, C., & Raghunathan, A. (2024b). Scaling Laws for Precision. arXiv:2411.04330. doi:10.48550/arXiv.2411.04330.
- Kurtic, E., Marques, A., Pandit, S., Kurtz, M., & Alistarh, D. (2024). "Give Me BF16 or Give Me Death"? Accuracy-Performance Trade-Offs in LLM Quantization. arXiv:2411.02355. doi:10.48550/arXiv.2411.02355.
- Larsen, E. (2025). The Instability of Safety: How Random Seeds and Temperature Expose Inconsistent LLM Refusal Behavior. arXiv:2512.12066. doi:10.48550/arXiv.2512.12066. [PREPRINT]
- Li, S., Ning, X., Wang, L., Liu, T., Shi, X., Yan, S., Dai, G., Yang, H., & Wang, Y. (2024). Evaluating Quantized Large Language Models. arXiv:2402.18158. doi:10.48550/arXiv.2402.18158.
- Lin, H., Zhu, K., Xu, H., Wu, Y., Lu, Z., Zhang, Q., & Sun, Z. (2026). Benchmarking Trustworthiness of SLMs: Pre-trained vs. Compressed. arXiv:2608.11981. doi:10.48550/arXiv.2608.11981. [PREPRINT]
- Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., & Han, S. (2023). AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. arXiv:2306.00978. doi:10.48550/arXiv.2306.00978.
- Liu, R., Sun, Y., Zhang, M., Bai, H., Yu, X., Yu, T., Yuan, C., & Hou, L. (2025). Quantization Hurts Reasoning? An Empirical Study on Quantized Reasoning Models. arXiv:2504.04823. doi:10.48550/arXiv.2504.04823.
- Liu, Y., Sun, Z., He, X., & Huang, X. (2024). Quantized Delta Weight Is Safety Keeper. arXiv:2411.19530. doi:10.48550/arXiv.2411.19530. [PREPRINT]
- Lyu, C., Wu, M., & Aji, A. F. (2024). Beyond Probabilities: Unveiling the Misalignment in Evaluating Large Language Models. arXiv:2402.13887. doi:10.48550/arXiv.2402.13887.
- Marchisio, K., Dash, S., Chen, H., Aumiller, D., Ustun, A., Hooker, S., & Ruder, S. (2024). How Does Quantization Affect Multilingual LLMs? arXiv:2407.03211. doi:10.48550/arXiv.2407.03211.
- Marcuzzi, F., Ning, X., Schwartz, R., & Gurevych, I. (2025). How Quantization Shapes Bias in Large Language Models. arXiv:2508.18088. doi:10.48550/arXiv.2508.18088.
- Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., et al. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. arXiv:2402.04249. doi:10.48550/arXiv.2402.04249.
- Mekala, A., Atmakuru, A., Song, Y., Karpinska, M., & Iyyer, M. (2025). Does quantization affect models' performance on long-context tasks? arXiv:2505.20276. doi:10.48550/arXiv.2505.20276.
- Miller, E. (2024). Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. arXiv:2411.00640. doi:10.48550/arXiv.2411.00640. [PREPRINT]
- Ouyang, X., Ge, T., Hartvigsen, T., Zhang, Z., Mi, H., & Yu, D. (2024). Low-Bit Quantization Favors Undertrained LLMs: Scaling Laws for Quantized LLMs with 100T Training Tokens. arXiv:2411.17691. doi:10.48550/arXiv.2411.17691.
- Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2021). BBQ: A Hand-Built Bias Benchmark for Question Answering. arXiv:2110.08193. doi:10.48550/arXiv.2110.08193.
- Peng, S., Chen, P.-Y., Hull, M., & Chau, D. H. (2024). Navigating the Safety Landscape: Measuring Risks in Finetuning Large Language Models. arXiv:2405.17374. doi:10.48550/arXiv.2405.17374.
- Prasad, H., & Pal, R. (2026). The Joint Effect of Quantization and Sampling Temperature on LLM Safety Alignment: A Factorial Analysis. arXiv:2606.29581. doi:10.48550/arXiv.2606.29581. [PREPRINT]
- Proskurina, I., Brun, L., Metzler, G., & Velcin, J. (2024). When Quantization Affects Confidence of Large Language Models? arXiv:2405.00632. doi:10.48550/arXiv.2405.00632.
- Proskurina, I., Metzler, G., & Velcin, J. (2025). Fair-GPTQ: Bias-Aware Quantization for Large Language Models. arXiv:2509.15206. doi:10.48550/arXiv.2509.15206. [PREPRINT]
- Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! arXiv:2310.03693. doi:10.48550/arXiv.2310.03693.
- Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., & Henderson, P. (2024). Safety Alignment Should Be Made More Than Just a Few Tokens Deep. arXiv:2406.05946. doi:10.48550/arXiv.2406.05946.
- Rath, P. K., & Maliakkal, R. (2026). Quantization Undoes Alignment: Bias Emergence in Compressed LLMs Across Models and Precision Levels. arXiv:2605.15208. doi:10.48550/arXiv.2605.15208.
- Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., et al. (2024). Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? arXiv:2407.21792. doi:10.48550/arXiv.2407.21792.
- Rottger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2023). XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. arXiv:2308.01263. doi:10.48550/arXiv.2308.01263.
- Scholten, Y., Gunnemann, S., & Schwinn, L. (2024). A Probabilistic Perspective on Unlearning and Alignment for Large Language Models. arXiv:2410.03523. doi:10.48550/arXiv.2410.03523.
- Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., et al. (2024). A StrongREJECT for Empty Jailbreaks. arXiv:2402.10260. doi:10.48550/arXiv.2402.10260.
- Tan, Q., Song, X., Cheng, N., Liu, N., Zhai, X., Hong, L., Wang, Y., Xiang, Z., & Yuan, G. (2026). Q-realign: Piggybacking Realignment on Quantization for Safe and Efficient LLM Deployment. arXiv:2601.08089. doi:10.48550/arXiv.2601.08089. [PREPRINT]
- Twagirayezu, L., & Mitra, P. (2026). The Effect of Quantization on Clinical Benchmarks: Accuracy and Safety Across Model Families. arXiv:2609.22216. doi:10.48550/arXiv.2609.22216. [PREPRINT]
- Varshney, P. (2026). Cosine Similarity Is Not Evidence: Measuring the Noise Floor of Interpretability Transfer Under Quantization. arXiv:2609.30275. doi:10.48550/arXiv.2609.30275. [PREPRINT]
- Wang, X., Hu, C., Ma, B., Rottger, P., & Plank, B. (2024). Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think. arXiv:2404.08382. doi:10.48550/arXiv.2404.08382.
- Wee, S., Kim, S., Kim, H., Hwang, K., & Kwak, N. (2025). Safety-Preserving PTQ via Contrastive Alignment Loss. arXiv:2511.07842. doi:10.48550/arXiv.2511.07842. [PREPRINT]
- Wei, B., Huang, K., Huang, Y., Xie, T., Qi, X., Xia, M., Mittal, P., Wang, M., & Henderson, P. (2024). Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications. arXiv:2402.05162. doi:10.48550/arXiv.2402.05162.

- Williams, M., & Aletras, N. (2024). On the Impact of Calibration Data in Post-training Quantization and Pruning. ACL 2024. arXiv:2311.09755. doi:10.18653/v1/2024.acl-long.544.
- Xie, T., Qi, X., Zeng, Y., Huang, Y., Sehwag, U. M., Huang, K., et al. (2024). SORRY-Bench: Systematically Evaluating Large Language Model Safety Refusal. arXiv:2406.14598. doi:10.48550/arXiv.2406.14598.
- Xu, Z., Gupta, A., Li, T., Bentham, O., & Srikumar, V. (2024). Beyond Perplexity: Multi-dimensional Safety Evaluation of LLM Compression. arXiv:2407.04965. doi:10.48550/arXiv.2407.04965.
- Yuan, J., Li, H., Ding, X., Xie, W., Li, Y.-J., Zhao, W., Wan, K., Shi, J., Hu, X., & Liu, Z. (2025). Understanding and Mitigating Numerical Sources of Nondeterminism in LLM Inference. arXiv:2506.09501. doi:10.48550/arXiv.2506.09501.
- Zhan, X., Egashira, K., Staab, R., Vero, M., & Vechev, M. (2026). Widening the Gap: Exploiting LLM Quantization via Outlier Injection. arXiv:2605.15152. doi:10.48550/arXiv.2605.15152.
- [PREPRINT]
- Zhang, J., Chen, K., He, L., Lou, J., Li, D., Feng, Z., et al. (2025). Activation Approximations Can Incur Safety Vulnerabilities Even in Aligned LLMs: Comprehensive Analysis and Defense. arXiv:2502.00840. doi:10.48550/arXiv.2502.00840.
- Zhang, N., Kwek, E., Zhang, Y., Nguyen, N.-H., Mitra, P., & Zhang, R. (2025). When Reasoning Meets Compression: Understanding the Effects of LLMs Compression on Large Reasoning Models. arXiv:2504.02010. doi:10.48550/arXiv.2504.02010.
- Zheng, A., Du, A., Deng, Z., Chen, Y., Guo, S., & Zheng, K. (2026). QuantGuard: Learnable Rounding for Repairing Quantization-Conditioned Backdoors in LLMs. arXiv:2606.29239. doi:10.48550/arXiv.2606.29239. [PREPRINT]
- Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043. doi:10.48550/arXiv.2307.15043.