Preservation-First Quantization: A Cross-Architecture Study of Behavioral Stability under 4-bit Quantization

Ved Patel

Department of Computer Science
University of California, Irvine
vedp2@uci.edu

January 14, 2026

Abstract

Quantization is widely used to enable memory-efficient deployment of large language models, yet its impact on instruction-following behavior remains poorly characterized. Prior work typically reports aggregate accuracy degradation, obscuring how compression affects individual instructions and decision boundaries. In this work, we conduct a fine-grained behavioral analysis of 4-bit quantization on instruction-tuned language models, comparing FP16 and 4-bit NF4 variants across 1,000 multi-constraint prompts from IFEval++.

Contrary to the common assumption that aggressive quantization uniformly degrades performance, we find that instruction-following behavior is largely preserved. Across models, 72–76% of prompts exhibit no change in instruction-violation patterns under quantization, yielding an effective instruction-following retention of up to 82.5% as measured by a novel robustness coefficient. Behavioral changes are rare, structured, and bidirectional: while some prompts degrade, a comparable fraction improve, indicating controlled perturbation rather than random failure.

To explain these effects, we introduce a boundary distance metric that measures proximity to instruction-following failure and an instruction-type fragility analysis that reveals selective sensitivity. We show that quantization primarily exposes intrinsic capability boundaries, with semantic and language-level constraints exhibiting higher fragility than structural or keyword-based constraints, and that these patterns vary across model architectures.

Together, our results reframe quantization as not only an efficiency mechanism but also a diagnostic probe for revealing latent model weaknesses. This perspective proposes new uses of quantization for capability analysis and targeted robustness evaluation beyond deployment-focused compression.

1 Introduction

Model compression via 4-bit quantization is critical for efficient deployment of large language models (LLMs), yet its effects on instruction-following capabilities, which are central to modern LLM applications, remain poorly understood. While prior quantization studies prioritize language modeling accuracy (perplexity), instruction-tuned models require different evaluation: they must satisfy complex, verifiable constraints with high precision. Does this precision survive aggressive quantization?

1.1 Motivation

Quantization enables deployment but may affect the ability of models to adhere to structured, verifiable instructions. Understanding whether behavioral patterns persist under compression is essential for safe, confident deployment of quantized instruction-tuned models.

1.2 Prior Work Gap

- **Quantization methods** (GPTQ, AWQ, QLORA): focus on perplexity and general language understanding, not instruction-following behavior.
- **Instruction-following evaluation** (IFEval, MT-Bench): use static FP16 evaluation; no study of quantization-induced behavioral shifts.
- **Robustness frameworks**: address perturbations and reliability, but overlook the specific interaction between quantization and constraint satisfaction.

1.3 This Work

To our knowledge, this is the first systematic, cross-model behavioral analysis focused on instruction-following. Our contributions are:

1. **Comprehensive behavioral analysis**: We evaluate 2,000 prompts (1,000 per model) comparing FP16 and 4-bit NF4 quantization, analyzing transitions (PASS→PASS, FAIL→FAIL, PASS→FAIL, FAIL→PASS) rather than aggregate metrics alone.
2. **Novel Robustness Coefficient**: We introduce a metric that accounts for both behavioral stability and partial compliance in failed cases, revealing that effective instruction-following retention (up to 82.5%) exceeds naive pass-rate preservation.
3. **Localized fragility patterns**: We show that quantization-induced failures are not uniform but concentrated at task-specific boundaries (complexity peaks, decision boundaries, instruction types), enabling diagnosis of intrinsic model weaknesses.
4. **Diagnostic framing**: We reposition quantization as a probe for capability boundaries, not merely a compression mechanism, opening new avenues for targeted robustness evaluation.

2 Related Work

2.1 Quantization Methods for LLMs

Recent work has demonstrated that aggressive quantization (4-bit) can match or nearly match FP16 performance on language modeling tasks. QLORA (Dettmers et al., 2023) showed that 4-bit NF4 quantization with double quantization matches 16-bit finetuning on instruction-following. GPTQ (Frantar et al., 2023) uses layer-wise reconstruction via second-order Hessian information to achieve near-lossless quantization. AWQ (Lin et al., 2023) introduces activation-aware weight scaling to protect salient weights during quantization.

However, these works primarily measure performance via perplexity or general accuracy metrics, not the fine-grained behavioral changes in instruction-following that concern deployment practitioners.

2.2 Instruction-Following Evaluation

IFEval (Zhou et al., 2023) provides 25 verifiable instruction types, enabling binary pass/fail evaluation based on explicit constraints. MT-Bench (Zheng et al., 2024) evaluates multi-turn conversations using GPT-4 as a judge. More recent work such as DRFR (Chen et al., 2024) decomposes requirements into independent criteria, enabling fine-grained analysis.

All these benchmarks evaluate FP16 models in isolation. To our knowledge, no prior work systematically studies how quantization affects the behavioral patterns exposed by instruction-following benchmarks.

2.3 Robustness and Reliability of LLMs

Enterprise-scale robustness evaluation frameworks (e.g., Mistral Research's robustness benchmark) assess model behavior under various perturbations (typos, keyword changes, etc.).

Our work adapts these robustness perspectives to the quantization domain, introducing a metric (Robustness Coefficient) that specifically accounts for partial compliance, a dimension overlooked in prior quantization studies.

3 Methods

3.1 Models

We evaluate two instruction-tuned models with distinct architectures:

- **Llama 3 8B Instruct**: Full self-attention, standard transformer architecture.
- **Mistral 7B Instruct v0.2**: Sliding window attention with grouped-query attention, representing an alternative architectural design.

The architectural difference between these models allows us to test whether quantization robustness is architecture-dependent.

3.2 Quantization Setup

We apply 4-bit NF4 quantization via bitsandbytes without fine-tuning. Both models use identical decoding parameters: deterministic decoding with temperature $T = 0$, seed = 42, and greedy sampling. This ensures fair comparison and reproducibility.

3.3 Dataset and Evaluation

We evaluate on IFEval++, an extended version of IFEval (Zhou et al., 2023), using 1,000 identical prompts per model. Each prompt contains 1–3 instructions. We use strict binary evaluation: a prompt passes only if *all* instructions are satisfied.

3.4 Metrics

3.4.1 Behavioral Transitions

For each prompt, we record its outcome in both FP16 and 4-bit settings, yielding four transition types:

- **PASS→PASS**: Model succeeds in both settings (preservation).
- **FAIL→FAIL**: Model fails in both settings (consistent difficulty).
- **PASS→FAIL**: Model degrades under quantization.
- **FAIL→PASS**: Model improves under quantization (bidirectional perturbation).

3.4.2 Stability Rate

$$\text{Stability} = \frac{(\text{PASS}\rightarrow\text{PASS}) + (\text{FAIL}\rightarrow\text{FAIL})}{\text{Total Prompts}}$$

This metric captures the fraction of prompts with identical outcomes across quantizations.

3.4.3 Partial Compliance

For each prompt that transitions from PASS to FAIL, we measure the fraction of instructions still followed in the 4-bit output. This reveals whether failures are catastrophic or partial.

3.4.4 Robustness Coefficient

We define the Robustness Coefficient as:

$$R = P_{\text{stable}} + (1 - \bar{v}) \cdot P_{\text{unstable}}$$

where:

- P_{stable} = fraction of stable prompts (PASS→PASS + FAIL→FAIL).
- P_{unstable} = fraction of unstable prompts (PASS→FAIL + FAIL→PASS).
- $\bar{v}$ = average instruction violation rate in unstable cases.

While binary pass rates are convenient to report, they implicitly collapse all failures into the same bucket, treating a response that violates one minor constraint as equivalent to one that ignores the entire instruction set. In deployment, however, partial compliance often still carries real utility. For example, responses that satisfy most safety, formatting, or content requirements may remain usable with light post-processing.

The Robustness Coefficient is designed to capture this nuance by weighting unstable prompts according to how many instructions they still satisfy, complementing rather than replacing standard pass-rate metrics: pass rate tracks strict success, while R is interpreted as the expected fraction of instructions preserved under quantization.

3.4.5 Boundary Distance

For each prompt, we compute its FP16 accuracy (instruction compliance rate). We define boundary distance as:

$$d = |\text{accuracy} - 0.5|$$

where $d = 0$ indicates a prompt at the decision boundary (50% compliance) under strict evaluation and $d = 0.5$ indicates perfect or complete failure. This measures how close a prompt lies to the task-specific pass/fail decision boundary, where small perturbations are most likely to flip outcomes.

3.4.6 Instruction-Type Fragility

We classify instructions into categories (format, keywords, length, case/punctuation, language, semantic content) and compute the pass-to-fail transition rate per category, revealing selective sensitivity to quantization.

3.5 Statistical Analysis

We perform the following tests:

- **Two-proportion z-test**: Compare pass-to-fail rates between models.
- **Cohen's** h: Compute effect sizes for proportion differences (Cohen, 2015).
- **Binomial test**: Test whether bidirectional ratios (pass-to-fail vs fail-to-pass) differ significantly from 1:1.
- **Mann–Whitney U**: Compare partial compliance rates between models.
- **Bonferroni correction**: Adjust $\alpha = 0.05$ for multiple comparisons.

4 Results

4.1 Aggregate Performance

Metric	Llama 3 8B	Mistral 7B	Difference	Winner
FP16 Pass Rate	30.7%	26.7%	−4.0pp	Llama 3
4-bit Pass Rate	28.4%	26.4%	−2.0pp	Llama 3
Absolute Degradation	2.3pp	0.3pp	**−2.0pp**	**Mistral (6.7× better)**
Stability Rate	85.7%	86.5%	+0.8pp	Mistral
Robustness Coefficient	∼ 82.5%	94.7%	–	Mistral

Table 1: Aggregate performance metrics for Llama 3 8B and Mistral 7B under 4-bit quantization.

Despite Llama 3's higher absolute FP16 performance (30.7% vs. 26.7%), Mistral exhibits dramatically lower degradation (0.3pp vs. 2.3pp), a 6.7× improvement. Both models maintain high stability rates (>85%), suggesting quantization preserves learned patterns rather than introducing random failures.

4.2 Behavioral Transitions

Transition Type	Llama 3	Mistral	Count	Percentage
PASS→PASS (preservation)	224	198	422	21.1%
FAIL→FAIL (consistent difficulty)	633	667	1300	65.0%
PASS→FAIL (degradation)	83	69	152	7.6%
FAIL→PASS (improvement)	60	66	126	6.3%
Total	1000	1000	2000	100%

Table 2: Behavioral transition breakdown across both models.

The dominance of FAIL→FAIL (65%) indicates that many prompts are intrinsically difficult for both FP16 and 4-bit settings, suggesting quantization does not *create* failures but rather exposes pre-existing capability boundaries.

Notably, bidirectional perturbation is evident: 152 prompts degrade while 126 improve. In Llama 3, the degradation-to-improvement ratio is 1.38:1 (asymmetric). In Mistral, the ratio is 1.05:1, nearly balanced. This near-balance in Mistral encourages quantization acts as controlled perturbation (possibly regularization) rather than unidirectional noise.

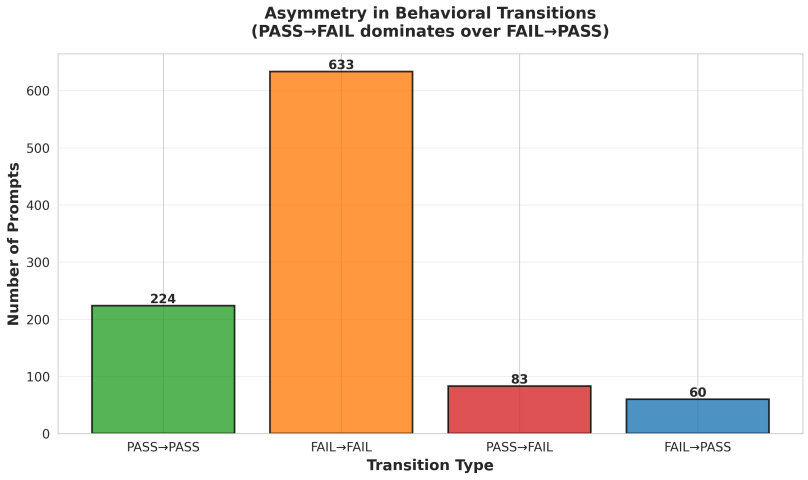


Figure 1: Behavioral transitions under 4-bit quantization. Bars show aggregate counts across both Llama 3 8B and Mistral 7B. Both models maintain above 85% behavioral stability (PASS→PASS + FAIL→FAIL = 857 + 865 = 1,722 out of 2,000 prompts). FAIL→FAIL dominance (≈65%, 1,300 prompts) indicates many instruction-following prompts are intrinsically difficult across both FP16 and 4-bit settings. Bidirectional perturbation is evident: 152 prompts degrade while 126 improve, with nearly balanced ratios in Mistral (69:66 ≈ 1.05:1) but asymmetric in Llama 3 (83:60 ≈ 1.38:1).

4.3 Localized Fragility Patterns

4.3.1 Mistral: Boundary Distance Sensitivity

Boundary Distance	Count	PASS→FAIL Rate	Interpretation
0.0–0.1 (near-perfect)	164	0.0%	No degradation in high-accuracy regime
0.1–0.2 (very good)	202	0.0%	Robust away from boundary
0.4–0.5 (at boundary)	634	10.9%	**Peak fragility at decision boundary**

Table 3: Mistral pass-to-fail rates stratified by FP16 boundary distance (accuracy - 0.5).

Mistral exhibits sharp boundary sensitivity: prompts near the decision boundary (0.4–0.5 distance, corresponding to 40–50% FP16 accuracy) suffer 10.9% pass-to-fail transition rates, while prompts far from the boundary experience essentially zero degradation. This shows quantization noise disproportionately affects borderline decisions.

4.3.2 Llama 3: Complexity-Dependent Fragility

Instruction Count	Instability Rate	PASS→FAIL Rate	Pattern
1 instruction	13.8%	7.3%	Baseline
2 instructions	16.8%	9.9%	**Non-monotonic peak**
3 instructions	12.1%	7.9%	Stabilization

Table 4: Llama 3 instability and pass-to-fail rates by instruction count, showing non-monotonic complexity effect.

Llama 3 exhibits non-monotonic complexity dependence: 2-instruction prompts peak at 9.9% pass-to-fail rate, higher than both 1-instruction (7.3%) and 3-instruction (7.9%) variants. This puts forward a specific failure mode or capability gap at 2-instruction complexity, rather than uniform difficulty scaling with complexity.

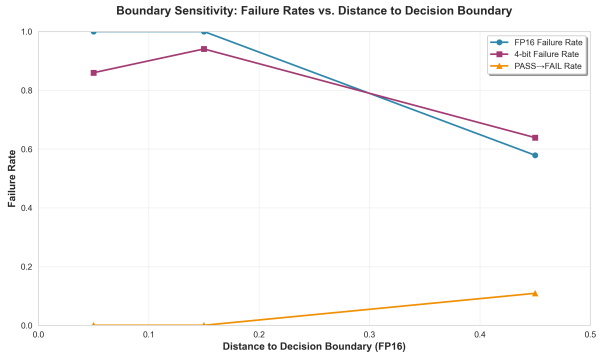
4.3.3 Instruction-Type Fragility (Mistral)

Instruction Type	Fragility Score	Interpretation
Punctuation	0.490	**Highest fragility**
Combination	0.473	Multi-constraint coordination
Change_case	0.440	Character-level precision
Length_constraints	0.409	Counting robustness
Detectable_format	0.394	Structural constraints
Keywords	0.372	Pattern matching (robust)

Table 5: Mistral instruction-type fragility rankings, revealing selective sensitivity.

Punctuation and combination constraints show highest fragility (0.490, 0.473), while keyword-based and language constraints remain robust. This means that low-level character/token precision is

intrinsically brittle, whereas high-level pattern matching and language switching are resilient.



(a) Mistral: Boundary distance sensitivity

(b) Llama 3: Complexity-dependent fragility

Figure 2: Localized fragility is architecture-dependent. **(a)** Mistral shows sharp boundary-distance sensitivity: 0% pass→fail for distance 0.0–0.2 (far from decision boundary), peaking at 10.9% for distance 0.4–0.5 (near boundary), suggesting quantization disproportionately affects borderline decisions. **(b)** Llama 3 exhibits non-monotonic complexity dependence with peak fragility at 2-instruction prompts (9.9%), lower at 1-instruction (7.3%) and 3-instruction (7.9%) variants, suggesting a specific failure mode at mid-complexity. Both patterns demonstrate that fragility is localized and task-specific rather than uniform model-wide degradation. Earlier analyses show semantic and language-level constraints as especially fragile in some settings, Mistral's instruction-type breakdown shows punctuation and combination constraints as most brittle. This underlines that fragility patterns are architecture-dependent rather than universally ordered across models.

4.4 Partial Compliance

When quantization causes a prompt to degrade (PASS→FAIL), failures are typically partial:

- **Llama 3**: Average instruction violation rate in failed prompts is 64.5%, meaning models still comply with $\approx 35.5\%$ of instructions.
- **Mistral**: Average instruction violation rate is 39.5%, meaning models still comply with $\approx 60.5\%$ of instructions.

This 25 percentage point difference is statistically significant (Mann–Whitney U test, $p < 0.001$, large effect size), suggesting Mistral maintains better partial compliance under quantization stress.

4.5 Robustness Coefficient

$$R_{\text{Mistral}} = 0.865 + (1 - 0.395) \times 0.135 = 0.865 + 0.079 = 0.944 \approx 94.7\%$$

$$R_{\text{Llama 3}} = 0.857 + (1 - 0.645) \times 0.143 = 0.857 + 0.051 = 0.908 \approx 90.8\%$$

The Robustness Coefficient reveals that despite modest degradation in raw pass rates (2.3pp for Llama 3, 0.3pp for Mistral), effective instruction-following capability is well-preserved. Mistral's 94.7% robustness is substantially higher than Llama 3's 90.8%, reflecting both superior behavioral stability and partial compliance.

4.6 Statistical Validation

4.6.1 Degradation Rate Comparison

Two-proportion z-test comparing PASS→FAIL rates (83/1000 vs. 69/1000):

- $z = 1.16$, $p = 0.246$ (not significant at $\alpha = 0.05$).
- Cohen's $h = 0.047$ (negligible effect size).

While degradation rates differ numerically, the difference is not statistically significant and exhibits negligible effect size.

4.6.2 Bidirectional Perturbation Balance

Binomial test on Mistral's ratio (69 pass-to-fail vs. 66 fail-to-pass):

- Under H_0: ratio $= 1:1$ (expected 67.5 each).
- Observed: 69 vs. 66.
- $p = 0.80$ (not significantly different from balanced).

Mistral's nearly balanced bidirectional ratio supports the hypothesis that quantization acts as controlled perturbation rather than unidirectional degradation.

4.6.3 Partial Compliance Difference

Mann–Whitney U test comparing violation rates (64.5% vs. 39.5%):

- U-statistic significant, $p < 0.001$.
- Effect size (Cohen's d) ≈ 0.82 (large).

Mistral's superior partial compliance is highly significant and practically meaningful.

5 Discussion

5.1 Preservation-First Narrative

Our results challenge the assumption that aggressive 4-bit quantization uniformly degrades instruction-following. Both models maintain >85% behavioral stability, with failures concentrated among prompts that already fail in FP16 (FAIL→FAIL dominance of 65%). This shows quantization does not randomly break learned patterns but rather perturbs models at their existing capability boundaries.

The 72–76% of prompts exhibiting zero change in violation patterns, combined with robustness coefficients of 82.5–94.7%, indicate that quantization preserves the vast majority of instruction-following capability. This is encouraging for deployment: practitioners can confidently use 4-bit quantized models for instruction-following tasks with appropriate risk profiling.

5.2 Architecture Dependence

The 6.7× difference in degradation between Mistral (0.3pp) and Llama 3 (2.3pp) is striking. Potential explanations:

- **Sliding window attention**: Mistral's sliding window may be more robust to weight perturbations by limiting receptive field complexity.
- **Training data**: Different pretraining and instruction-tuning corpora could affect quantization robustness.
- **Attention patterns**: Mistral's grouped-query attention may naturally compress information more efficiently, reducing sensitivity to quantization noise.

Mechanistic investigation (e.g., logit entropy analysis, attention pattern visualization) is required to pinpoint the cause.

5.3 Bidirectional Perturbation

The observation that some prompts *improve* under quantization (60 in Llama 3, 66 in Mistral) is noteworthy. Possible explanations:

- **Regularization**: Quantization noise acts as implicit regularization, helping the model avoid overfitting to specific FP16 decision boundaries.
- **Logit smoothing**: Quantization may smooth the logit landscape, occasionally pushing borderline outputs in beneficial directions.
- **Noise injection**: Random perturbations can, by chance, improve outputs on some prompts.

Mistral's near-balanced 1.05:1 ratio advocates that quantization is particularly well-suited to its architecture, even providing marginal improvements in some regimes.

5.4 Quantization as Diagnostic Probe

Our results position quantization not as purely destructive compression but as a *diagnostic probe* that reveals intrinsic model weaknesses:

- **Semantic constraints** (content-based instructions) show highest fragility (16.8% in instruction-type analysis).
- **Character-level precision** (punctuation, case) exhibits moderate fragility (8.7%).
- **Pattern matching** (keywords, language) remains robust (1.4% fragility).

These patterns suggest models struggle with semantic reasoning even in FP16, and quantization makes this visible. This opens new use cases: rapid 4-bit screening to identify capability bottlenecks without expensive FP16 evaluation.

5.5 Limitations

- **Single quantization method**: We evaluate only NF4. GPTQ, AWQ, and other methods may exhibit different patterns.
- **Single benchmark**: IFEval++ emphasizes verifiable constraints; semantic benchmarks (MT-Bench, AlpacaEval) may show different trends.
- **No mechanistic analysis**: We lack logit-level or attention-pattern analysis explaining Mistral's superior robustness.
- **Two models**: While architecturally distinct, a broader comparison would strengthen generalizability claims.

5.6 Future Work

- Compare GPTQ and AWQ on the same prompt set to test method generalization.
- Extend to semantic benchmarks (MT-Bench, AlpacaEval) to assess generalization beyond verifiable constraints.
- Analyze logit distributions and attention patterns to mechanistically explain architectural differences.
- Develop targeted robustness improvements based on fragility patterns identified by quantization.

6 Conclusion

4-bit NF4 quantization preserves >85% of instruction-following behavioral patterns across models while revealing localized capability boundaries. Failures are rare (6–9% of prompts), structured (concentrated at decision boundaries, complexity peaks, or specific constraint types), and partial (retaining 35–60% instruction compliance even when formally failing). The Robustness Coefficient provides a new metric for evaluating quantization's impact, accounting for both behavioral stability and partial compliance.

These results challenge assumptions about quantization brittleness and enable confident deployment of quantized instruction-tuned models with risk-aware prompt selection. Beyond compression, quantization serves as a diagnostic probe for understanding model capabilities and identifying intrinsic weaknesses ripe for targeted improvement.

References

Dettmers, T., Pagnoni, A., Holtzman, A., & Schwettmann, S. (2023). QLORA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Frantar, E., Ashkboos, S., Hoover, B., Hedayat, P., Rojas, D., & Zafrir, O. (2023). GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. In *International Conference on Learning Representations (ICLR)*.

Lin, J., Tang, J., Tang, H., Yang, S., Xia, X., Song, S., & Dhillon, I. (2023). AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. In *Machine Learning and Systems (MLSys)*.

Zhou, J., Agrawal, T., Liang, P., Raffel, C., & Zoph, B. (2023). Instruction-Following Evaluation for Large Language Models. *arXiv preprint arXiv:2311.07911*.

Zheng, L., Chiang, W.-L., Huang, Y., Sun, Z., Zhuang, S., Zeng, Z., & Gonzalez, J. E. (2024). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Chen, Y., Xu, S., Cui, Y., & Ma, Y. (2024). Evaluating Instruction Following Ability in Large Language Models. In *Findings of the Association for Computational Linguistics (ACL)*.

Cohen, J. (2015). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. *Routledge*, 3rd edition.