LLM quantization and safety: capabilities do not degrade evenly

27 November 2026 · 4 min read · blogs.iictp.org/llm-quantization-safety-degradation-jagged-robustness

Short answer: LLM quantization safety is the question of whether a compressed model keeps the safeguards of the full model, and a paper by Ramneek Agnihotri, Yingte Xu and Hans Vath, posted on arXiv on 20 November 2026, found that it may not. In their tests on Qwen3 models, safeguards and bias measures could degrade while capabilities such as truthfulness stayed more stable. It matters because teams often validate a quantized model on a few quality benchmarks only.

TL;DR The short version

- A paper posted on 20 November 2026 found that quantized Qwen3 models can lose safeguards and bias protections while capabilities such as truthfulness stay more stable.
- The authors call this jagged robustness: capabilities do not degrade evenly.
- Teams shipping quantized models should test safety and bias, not only a few quality benchmarks.

Read online, listen to it, or change the text size: https://blogs.iictp.org/llm-quantization-safety-degradation-jagged-robustness

What is jagged robustness?

Quantization stores model values with fewer bits to save memory and speed up inference. The authors use jagged robustness to describe uneven transfer: some abilities survive compression while others weaken.

How the audit worked

The authors compared full and quantized models from the Qwen3 family across benchmarks for safeguards, instruction following, ethics and bias, among others. They used two styles of evaluation: free generation, judged by a language model, and log-likelihood ranking, where the model scores fixed options. They also tested adversarial text appended to prompts.

What is new compared with quality-only checks

The comparison covers safety behaviour and shows that the choice of evaluation method changes what you see. A small model quantized hard can lose coherent free generation yet still score under ranking tests, because ranking never asks it to write.

Results reported by the authors

- Safeguard benchmarks degraded in free generation, while truthfulness was more stable in Qwen3-4B.

- Subjective or open-ended tasks were more sensitive than explicit violations and deterministic instructions.
- Quantized models complied with unsafe requests more often.
- The authors demonstrated adversarial suffixes that did not affect the full model but bypassed safeguards in the quantized one.
- Task complexity, evaluation method and output coherence all influenced the result.

This is an arXiv preprint or workshop paper. Treat its numbers as the authors' own results, not as independently confirmed findings.

Limitations and open questions

- Only the Qwen3 family was evaluated.
- Judging relied on a language model, which the authors call inconsistent.
- Benchmarks used 300 samples, without repeated seeds or confidence intervals.
- In the attack, compliant-looking text was often not genuinely informative.
- The reverse direction, suffixes that affect only the full model, was not studied.

Sources were read on 27 November 2026. Products, rules and data change, so check the linked originals before you rely on any detail.

How to test your own quantized model

1. Treat each quantized build as a new model for release approval.
2. Run safety and refusal tests on the exact weights, runtime and settings you will ship.
3. Include free-generation tests, not only multiple-choice scoring.
4. Compare against the full model with enough samples to report uncertainty.
5. Keep input and output filtering in the application.
6. Retest when the quantization method, bit width or engine changes.

What it means for teams shipping local models

A model card for the original weights does not certify a quantized copy downloaded from elsewhere. Record provenance and file hashes. Pair these checks with the method comparison in MXFP4 quantization.

Frequently asked questions

- **Does LLM quantization affect safety?** In this study it could, with safeguards degrading in some tests.
- **Which models were tested?** Qwen3 family models only.
- **Why use free-generation tests?** Ranking tests can look healthy when generated output has broken down.
- **Is every quantization unsafe?** The paper does not show that. It shows transfer cannot be assumed.

Related reading

- MXFP4 and NVFP4 quantization

- LLM quantization: GGUF and AWQ
- Lazarus dormant neuron backdoor
- Local LLM Practitioner certification

Related credential: Local LLM Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Agnihotri, Xu, Vath. Jagged Robustness: Uneven Safety Transfer and Emerging Vulnerabilities under LLM Quantization. arXiv 2611.27424. 20 November 2026 (preprint). arxiv.org/abs/2611.27424

Published by IICTP, an independent technical credential body, on 27 November 2026. The current version of this post is at https://blogs.iictp.org/llm-quantization-safety-degradation-jagged-robustness. More guides: blogs.iictp.org.