MXFP4 quantization for LLMs: what a study of 4-bit recipes found

26 November 2026 · 4 min read · blogs.iictp.org/mxfp4-nvfp4-llm-quantization-rotation-midmax-paper

Short answer: MXFP4 quantization can support accurate 4-bit LLM inference, but the recipe that works for it does not necessarily work for the related NVFP4 format, according to a preprint by Natalia Frumkin and colleagues posted on 24 November 2026. They report that MXFP4 benefits substantially from rotation and error-correcting post-training quantisation, and introduce a calibration-free scale choice called Midmax. It matters when you pick a low-precision format for serving.

TL;DR The short version

- A preprint posted on 24 November 2026 finds that MXFP4 can support accurate 4-bit LLM inference, but the recipe that works for it does not necessarily transfer to NVFP4.
- MXFP4 benefits substantially from rotation and error-correcting post-training quantization, and the authors introduce Midmax, a calibration-free choice of scale.
- Test each 4-bit format with its own recipe on your model before choosing one for serving.

Read online, listen to it, or change the text size: https://blogs.iictp.org/mxfp4-nvfp4-llm-quantization-rotation-midmax-paper

What are microscaling formats?

Quantisation stores model numbers with fewer bits. Microscaling formats such as MXFP6, MXFP4 and NVFP4 share one scale factor across a small block of 16 or 32 values, which allows aggressive precision with limited accuracy loss when combined with modern post-training quantisation (PTQ) methods.

How the study worked

- Six language models and three formats.
- Systematic variation of Hadamard rotation granularity. A rotation spreads values more evenly before quantising.
- Variation of the PTQ algorithm.
- Variation of how each block’s scale is chosen.
- An added test of a per-tensor “macroscale”, a mechanism native to NVFP4, applied to MXFP4.

What is new compared with single-technique results

The authors study how techniques combine, not one at a time. Their central finding is that effective recipes depend on the format. Midmax is proposed to reduce clipping error in the power-of-two block scales, without a calibration data pass.

Results reported by the authors

- MXFP4 and NVFP4 both support accurate 4-bit inference.
- MXFP4 gains substantially from rotation and error-correcting PTQ.
- NVFP4 gains less from rotation, consistent with its finer scales capturing more variation within a block.
- Midmax improved accuracy across a majority of evaluated MXFP4 configurations.
- MXFP6 activations with MXFP4 weights recovered most of the 4-bit loss and matched or exceeded their 16-bit baseline on Llama-3.1-8B.
- The macroscale did not improve average MXFP4 accuracy.

Limitations and open questions

- This article relies on the abstract. Check the paper for model list, tasks, accuracy numbers and variance.
- Accuracy results do not tell you speed or memory on your hardware; kernel support differs between GPUs.
- A result on an 8B model may not hold for larger or mixture-of-experts models.
- The baseline for the strongest claim is the authors' own implementation. This is an arXiv preprint or workshop paper. Treat its numbers as the authors' own results, not as independently confirmed findings. Sources were read on 26 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How to test a recipe yourself

1. Fix a task-based evaluation set before quantising, including your hardest cases.
2. Record full-precision results for comparison.
3. Test your format with and without rotation and error correction, changing one factor at a time.
4. Try higher activation precision with 4-bit weights if accuracy falls short.
5. Check that your serving engine supports the exact format and rotation you used.
6. Measure speed and memory on the target GPU before deciding.

What it means for LLM engineers

Do not copy an NVFP4 recipe to MXFP4, or the reverse, and assume the same quality. A label such as "4-bit" hides block size, scale type and calibration, all of which change results. Ask model publishers which choices produced the checkpoint. For the basics, see LLM quantisation formats.

Frequently asked questions

- **What is MXFP4 quantization?** A 4-bit microscaling format that shares a power-of-two scale across a block of values.
- **Is MXFP4 the same as NVFP4?** No. The paper finds that techniques helping one may not transfer to the other.
- **What is Midmax?** A calibration-free way to select block scales, proposed by the authors.
- **Is the paper peer reviewed?** It is an arXiv preprint at the time of reading.

Related reading

- LLM quantisation: GGUF and AWQ
- Quantising diffusion LLMs
- Run a large MoE model on 32 GB
- Local LLM Practitioner certification

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

Sources and further reading

1. Frumkin, Khodamoradi, Denolf, Dellinger, Biles. MXFP4 for LLM Inference: Complementary Post-Training Quantization Techniques and Their Interactions. arXiv 2611.32873. 24 November 2026 (preprint). arxiv.org/abs/2611.32873

Published by IICTP, an independent technical credential body, on 26 November 2026. The current version of this post is at https://blogs.iictp.org/mxfp4-nvfp4-llm-quantization-rotation-midmax-paper. More guides: blogs.iictp.org.