Episode Book

EPISODE #6615

Quantization Quietly Deletes Fine-Tuned Skills

PUBLISHED	15 November 2026 at 18:24 UTC
DURATION	25:17
LISTEN	myweirdprompts.com/episode/quantization-fine-tuned-skills-79019/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	z-ai/glm-5.3
VOICES	omnivoice
PIPELINE	V5.32
SOURCES	18 cited by the research (listed at the end)
BOOK GENERATED	15 November 2026 at 23:00 UTC

SYNOPSIS

A fine-tuned text-to-speech model called Meorav handled English-Hebrew code-switching at full precision, then lost the skill completely after quantization while still sounding normal. Exploring why compression quietly erases low-margin fine-tuned behaviors, this conversation unpacks outlier weights, signal-to-noise ratios, and why aggregate metrics like perplexity miss the damage. From Llama-3.1-8B-Instruct dropping 14.5 points under three-bit GPTQ to speech codecs losing emotional cues, the numbers show that whatever an objective function ignores is what quantization deletes first.

TRANSCRIPT



Corn

A model can pass every validation check you throw at it, and the one thing you built it for is already gone.



Herman

That's the part people don't expect. Quantization doesn't announce itself. It doesn't degrade everything a little, evenly, politely. It deletes specific capabilities while the average numbers look fine.



Corn

That's where Daniel found himself. He trained his first experimental text-to-speech model, he calls it Meorav, on his own voice, for mixed English and Hebrew speech. A cleaning model tidied the training data first. At full precision it worked. The code-switching was good, everything validated.



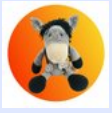
Herman

Then he quantized it so he could run it locally, and it basically stopped working. All the code-switching, gone.



Corn

And he's asking the questions that follow from that. What is quantization actually doing inside the model? Why would the fine-tuned skill be the first casualty? Was Meorav too small, or not trained hard enough, or is that even the right question? And if you're fine-tuning something you intend to run on modest hardware, how do you know in advance whether it'll survive?



Herman

There's real published work on this, and the short answer is that most people's mental model, more bits equals more accuracy, is close to useless.



Corn

Before we get into it, I want to pin down why this matters beyond Daniel. Because the thing he did, fine-tune a smallish model and then compress it to run on a laptop or a phone, that's not a niche hobbyist move anymore, is it?



Herman

It's becoming the default move. The whole point of a local model is that it's small, and small means compressed. So the sequence he ran, train, fine-tune, quantize, deploy, that's the pipeline half the field is assembling right now, mostly without checking whether the fine-tuned part made it through the last step.



Corn

And the failure is quiet. It's not a crash.



Herman

It's never a crash. The model runs, it produces audio, it sounds like a model. It just doesn't do the one thing you made it for.



Corn

Then let's start with what quantization actually does, because the mechanics matter for everything after.



Herman

The numbers in a model are stored as high-precision floating point. Sixteen bits per weight, typically. Quantization maps those onto a small grid of low-bit integers. You take the range of the values, divide it into however many steps the bit-width allows, and every weight gets snapped to the nearest step. That snapping, the rounding, is where the damage comes from. Every weight ends up slightly wrong.



Corn

Slightly wrong times billions of weights.



Herman

Right, but that alone would be survivable, because random small errors tend to cancel out. The problem is the errors aren't random. Four-bit gives you one-eighth the memory footprint of the full model, which is the whole appeal. Eight-bit is usually near-lossless. Four-bit is where the risk starts. Three-bit and below, you're gambling.



Corn

Let me make sure I have the picture. With four bits, how many steps are we actually talking about?



Herman

Sixteen levels. That's it. So a weight that could sit anywhere on a continuous range gets moved to one of sixteen rungs on a ladder. You can imagine taking a photograph and saying every pixel can only be one of sixteen colors. From a distance it still looks like the photograph. Up close, in the places where the detail was doing the work, it's mush.



Corn

And nobody looks up close at individual weights.



Herman

Nobody can. That's the practical problem. You can't eyeball a billion weights and spot the ones that mattered.



Corn

And there's a distinction people skip. You can quantize just the weights, or the weights and the activations.



Herman

Weight-only is the common case, and the gentler one. Activations are what the model computes at runtime, and they're input-dependent and spiky, with these heavy-tailed outliers. Quantizing them too is where it gets hard.



Corn

So the setup is: rounding errors, everywhere, and some of them matter far more than their size suggests. Meorav validated at full precision, and then fell over. Herman, is that a fluke?



Herman

Not at all. It's a documented pattern. There's a study that ran over eight hundred evaluations across seventy-six quantized configurations, five model families, four methods, four bit-widths, and the headline finding is a direct quote: how severely a quantized model degrades is not determined by bit-width alone. Three-bit GPTQ dropped Llama-3.1-8B-Instruct from about 73.7 to about 59.2 on a five-benchmark average. The seventy-billion version of the same model, same method, same bit-width, went from 79 to 77. One collapses, one shrugs.



Corn

Fourteen points versus two.



Herman

From the same compression. Nine times the damage.



Corn

That kills the folk wisdom right there. People treat bit-width like a speed limit, four-bit means this much loss. It's more like a weather report for a different city.



Herman

I like that more than I should. The reason bit-width doesn't predict anything is the outlier problem. The LLM.int8 work back in twenty twenty-two established that transformers develop a small number of feature dimensions with extreme magnitudes, and that these dominate attention and predictive performance. Their fix kept over 99.9 percent of the values in eight-bit and isolated the outlier dimensions into sixteen-bit precision.



Corn

So a handful of numbers carry a disproportionate share of the model's behavior.



Herman

And what is quantization, mechanically? The operation that damages extreme numbers. You've got a dynamic range set by your biggest outlier, and the whole grid gets stretched to accommodate it, so all your ordinary weights end up on a coarser grid. Fine-grained detail gets sacrificed to the range.



Corn

Which means the important weights and the fragile weights can be the same weights.



Herman

That's the trap, yes.



Corn

Okay, but here's what I want to get at. The errors are small, the outliers are handled in eight-bit, most models survive four-bit fine. Why did Meorav lose the one thing it was for? Why code-switching specifically?



Herman

The degradation happens in two stages. Errors are introduced inside each module, and then they propagate across layers, where they can be damped, preserved, or amplified. What determines the first stage is the magnitude of the weight error, the strength of the



Herman (continued)

task-specific signal, and how the error aligns with the activations that task uses. The same amount of noise is harmless when the signal is strong and redundant, and disruptive when the signal is weak.



Corn

Signal-to-noise. That's the predictor.



Herman

That's what the paper proposes, and they verify it tracks downstream performance. Which takes us to fine-tuned skills. Corn, think about what a fine-tune actually installs. The base model had no strong prior for switching between English and Hebrew mid-sentence. The fine-tune didn't carve a deep channel. It installed a thin preference. Low margin.



Corn

The base model never voted on it, the fine-tune is the entire margin.



Herman

And quantization noise eats thin preferences first. There's reasoning work showing failures emerge early, with the first vulnerable step flipping and cascading to the final answer. For code-switching, the language choice is a single early, low-margin decision. Once it flips, the whole utterance follows it into the wrong language. It doesn't degrade gracefully, where Hebrew comes out slightly muddier. It's a coin that lands on the other side.



Corn

So Meorav didn't get worse at code-switching. It got one coin flip wrong, once, at the top of every sentence.



Herman

Every utterance. That's why it looked like the capability was cleanly deleted rather than degraded.



Corn

Let me push on the cascade idea for a second, because I think it explains something about how these failures feel from the outside. If the first flip happens early, everything downstream is built on the wrong foundation.



Herman

The model isn't making one mistake and recovering. It's making one mistake and then executing flawlessly in the wrong language. The pronunciation, the prosody, the grammar of whatever language it landed in, all intact. The failure looks like competence pointed the wrong way.



Corn

Which is the worst kind of failure to debug, because nothing looks broken.



Herman

Here's the part that should bother Daniel, and does. It validated.



Herman

It validated, and that's not him being careless. The same paper on reasoning found that post-training quantization elevates method and execution errors relative to conceptual mistakes. Aggregate scores conceal capability-specific failures. There's a study of seven small instruction-tuned models at four-bit that found multilingual math and code generation the most fragile, with relative losses up to fifty-seven percent, while commonsense reasoning is largely preserved. But the order of which skill breaks is only weakly consistent across architectures.



Corn

Weakly consistent how?



Herman

They measured it. Mean Kendall's tau of 0.29, which is to say, knowing which skill broke in one model tells you almost nothing about which breaks in another.



Corn

So Daniel couldn't have looked at somebody else's quantized model and drawn a lesson. His coin flip was his own.



Herman

And the aggregate-score problem is worse than it sounds, because quantization isn't even monotonic. Four-bit GPTQ on Llama-2-13B-Chat actually raised the Ethics score from 54.1 to 76.3.



Corn

It improved the model's ethics.



Herman

By accident, while the three-bit version was dropping over fifty points on trustworthiness in the worst case. Compression is not a dial that goes from good to bad in a straight line. It's a reshuffle.



Corn

Which makes a single validation number worse than useless. It's a number that can go up while the thing you wanted dies.



Herman

Perplexity is the same trap. People use it as the health check, and there's work across thirteen datasets showing significant four-bit accuracy drops that perplexity never reflected. If Daniel validated Meorav on reconstruction loss or something perplexity-shaped, the check was measuring the acoustics, not the decision.



Corn

The model still sounded like Daniel. Just in one language.



Herman

I'd bet the accent survived beautifully. That's the eerie part. Everything the loss function measured came through.



Corn

That's the thesis I'd put to you, then. Whatever your objective function doesn't explicitly measure is what quantization quietly deletes.



Herman

That's not even a stretch. There's work on speech codecs showing emotional cues degrade during quantization, because the codecs mainly optimize acoustic reconstruction and leave emotion expressiveness unmodeled at the representation level. A codec has no reason to protect the feeling in a voice, and it doesn't.



Corn

Hm. The same argument with the nouns changed.



Herman

Meorav's training had no gradient pressure protecting the language-switch decision. Nothing was standing guard over it.



Corn

So the answer to Daniel's first question is ugly. You know in advance only if you know your signal-to-noise on the specific skill, and there's no off-the-shelf tool that measures that for an arbitrary checkpoint.



Herman

The closest thing is a prediction framework from the robotics world, offline action deviations predicting quantization-induced task losses. It made the right call on twenty-one of twenty-eight held-out configurations it chose to decide on. But it deferred on one that turned out to be a thirty-three-point loss. Even the research prototype's prototype misses.



Corn

Twenty-one of twenty-eight, with one silent miss in the deferrals. That's better than a coin, worse than a doctor.



Herman

I'd take it over the folk wisdom, which is currently at zero.



Corn

Well, that's the theory. And I think we have time for one more call before we wrap up.

D

Dorothy

Hello dear, it's your mother. Time for my grim fact.

D

Dorothy

A speech codec has no room for feeling, dear. The emotional cues in a voice get degraded during quantization, because the codecs only bother to reconstruct the acoustics, not the expressiveness.

D

Dorothy

So the model hears the words and loses the heart behind them, which is a bit like your father on the telephone.

D

Dorothy

Anyway, I've made a casserole and it won't eat itself. When are you coming to see me? Ring off now, sweetheart.



Herman

Mum, we are absolutely coming to see you. Soon.



Corn

I'll bring the leaves. She loves the leaves.



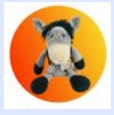
Herman

She does not love the leaves. Where were we. The unguarded skill, right. So Daniel's second question was whether the model was too small or undertrained. I don't think the literature lets you split that.



Corn

Both are risk factors, and they were both true of Meorav.



Herman

They compound. A small model has fewer layers of redundancy to absorb the same relative perturbation, and the propagation story is that larger models benefit from weaker error amplification across depth. Meanwhile a lightly-trained skill has less margin to spare. Stack them and you've got a thin preference sitting in an architecture that amplifies noise.



Corn

So it's not additive, it's multiplicative. The small model amplifies the noise, and the thin skill has nothing to absorb the amplification.



Herman

The numbers on size are dramatic.



Herman

They are. Qwen2.5 at half a billion parameters drops over sixty percent on math when quantized, versus two to three percent for the seven-billion version of the same family. Llama-3.2-1B under four-bit GPTQ loses twenty-five percent on GSM8k. Not three-bit. Four-bit, the setting people casually call safe.



Corn

The one people use to run things on a laptop.



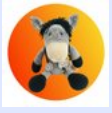
Herman

That one. And method choice matters more the smaller you go. The same study suggested SmoothQuant or FP8 at eight-bit may be necessary to hold accuracy for one-billion-scale models.



Corn

But the size advantage has a floor, doesn't it.



Herman

It does, and this is the part I find sobering. On commonsense reasoning, even small models retain most of their performance at four-bit and three-bit, but the scale advantage disappears at two-bit. Being big buys you nothing once you go low enough. The bigness was damping the amplification, and at two-bit there's nothing left to damp.



Corn

Like the difference between a big ship and a small boat stops mattering once the water freezes.



Herman

I'd have gone with a heavier anchor, but sure. Now, the part I think is the real answer to Daniel's second question, and it's speech-specific. Speech models have quantization hazards that text models don't, and Meorav walked into all of them.



Corn

Go on.



Herman

Text tokens are discrete and roughly uniform in information density. Speech is not. Speech activations are highly correlated in steady-state and zero-padded regions, and they change rapidly at phonetic boundaries. The interesting moments are a tiny fraction of the frames.



Corn

And standard post-training calibration treats all frames equally.



Herman

When all frames contribute equally to the Hessian, the static frames dominate calibration and obscure the informative transitions. The boring ninety percent of the audio drowns out the boundaries that carry meaning. There's an Interspeech paper on this, ultra-low-bit speech recognition, and at two to three bits this causes severe degradation, with baseline methods frequently producing degenerate outputs.



Corn

And a language switch is a boundary event. It lives precisely in the frames calibration averages away.



Herman

Code-switching sits at phonetic and language boundaries, the least representative, least-calibrated moments in the whole signal. Of every skill Meorav had, that was the one structurally in the line of fire. Even the emotional-content finding rhymes with this, the expressiveness is in the transitions, not the steady state.



Corn

So it's not just that the fine-tune was thin. It's that it was thin in exactly the place the compression is blindest.



Herman

That's the sentence I'd underline, if we underlined things. It's almost poetic, isn't it. The model spends most of its time in steady state, being generally competent, and that's the part compression protects. The two seconds where it does the special thing, that's the part it sacrifices.



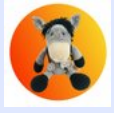
Corn

Corn, can I ask you something practical? If you were Daniel, and you had Meorav sitting there broken, what's your move? Retrain from scratch?



Corn

No. I'm lazy, so I'd look for the cheapest fix first. Is there one?



Herman

There is, and it's almost embarrassing. The reasoning work I mentioned, the measure-locate-restore loop, recovered four-bit math reasoning with three hundred thirty-two curated examples and three to five minutes on a single GPU, without access to



Herman (continued)

pretraining data. You find the earliest step that flips, and you fine-tune exactly there.



Corn

Three hundred thirty-two examples.



Herman

That's the count in the paper. And the workflow inverts what Daniel did. He fine-tuned, then quantized. The repair path is quantize first, then fine-tune the quantized model on a few hundred targeted examples, the utterances where the switch flips. Train the thing you're actually going to deploy.



Corn

Which answers his third question in a way the checklist people never say. Decide your bit-width before you train, not after.



Herman

That's the rule I'd give anyone fine-tuning for low-resource hardware. If you know you're deploying at four-bit, train at four-bit. Quantization-aware training exists for exactly this reason, though it traditionally costs you, it needs backpropagation and heavy memory. There's newer work, ZeroQAT, forward-only gradient estimation, that fine-tunes a thirteen-billion model at two to four bits on a single eightGB GPU. They even ran a 6.7B on a OnePlus phone.



Corn

A phone.



Herman

A twelve-year-old's Christmas list, is what that is. And if you're building the architecture itself, there's work showing that just inserting RMS normalization before every linear projection, with a gradual layer-wise quantization schedule, stably fine-tunes full-precision



Herman (continued)

checkpoints down to ternary weights.



Corn

So the honest hierarchy is: quantization-aware training if you can afford it, targeted repair after the fact if you can't, and plain post-training quantization as the gamble it always was.



Herman

That's the ladder, yes. Though I'd add one thing for Daniel specifically. We did an episode ages ago on voice fine-tuning, what those adapters capture, and I said then that a voice fine-tune is a thin layer of a person laid over a general model. I didn't have the word margin for it at the time, but that's what it was. Meorav is two thin layers, the voice and the code-switch, stacked on something small. Compression is a filter for exactly that shape.



Corn

And he built it in the shape compression eats first. There's an irony there I'll let him sit with.



Herman

We covered the fine-tuning mechanics once too, years back, and the low-margin idea is really the postscript to that conversation.



Hilbert

Herman's got the mechanism right, for what it's worth. The quiet parts go first. I had this with the reels.



Herman

The reels?



Hilbert

My aunt's audio reels, the darts club digitized them for me last spring. Fourteen reels, mostly her and my uncle talking over the garden. I compressed them to save shelf space, that's the same thing you're describing, yes. Thirteen came through fine. The third one, where my uncle does the blessing in Hebrew at the end, came out with the blessing gone. The English talk in front of it survived beautifully. My cousin Avner warned me before I did it, he said the quiet parts are what you lose first, and Avner has never once been wrong about anything, which I've never forgiven him for. He can still thread a reel in under forty seconds, that man, I've never matched it.



Corn

Forty seconds to thread a reel.



Hilbert

Thirty-eight when he's showing off. Anyway, I kept the compressed version. Labeled it the good one. I still play it for guests.



Corn

You play the one that's missing the part you digitized them for.



Hilbert

It sounds fine to me. The garden talk is very good. And the shelf is clear, which was the point of the exercise. Cornelius, you spent ten minutes earlier tonight on one sentence about a coin flip, so I'd be careful about who in this room keeps the wrong thing.



Corn

That sentence took eight minutes and it earned all of them.



Herman

Hilbert, the blessing in Hebrew is the thing you'd want to restore, though. If you still had the uncompressed master.



Hilbert

The club has it, in the cabinet under the bar. Reel eleven, the one Bert mislabeled as empty in 2019. He insists it's empty. It isn't. I've told him three times.



Corn

Bert's the bottleneck.



Hilbert

Bert is always the bottleneck. Anyway. If the tape were here I'd have thrown it out too, probably, so it's as well it's there.



Herman

The digitized-but-labeled-empty master, sitting in a cabinet, because one man won't relabel a box. That's the whole repair problem in miniature, isn't it. The information survives. Finding which frame it's in is the work.



Corn

Three hundred thirty-two examples and five minutes, once somebody opens the cabinet.



Herman

So the open question, the one I'd leave Daniel with, is that nobody's built the tool he was implicitly asking for. There's no standard off-the-shelf test that tells you whether your checkpoint survives quantization. The signal-to-noise framework and the prediction work are prototypes, and one of them operates in robotics. The practitioner's version doesn't exist yet, and as more people fine-tune small models to run locally, the Meorav pattern, validates clean, breaks silent, is going to get very common before anyone builds the instrument.



Corn

And the fix, meanwhile, is embarrassingly simple. Measure the one thing you care about before and after compression. Not the average. The thing.



Herman

Daniel had the instinct already. He noticed the code-switching was gone, not that the model scored worse.



Corn

That's the discipline most people skip, honestly. And if Meorav can be brought back with a few hundred targeted examples, the cheap fix is sitting right there in the same drawer as the cheap mistake.



Herman

Our producer is Hilbert Flumingtop, who has never once lost a minute of audio to a compressor, and certainly not a blessing. This has been My Weird Prompts, the human-AI collaboration podcast.



Corn

If today's episode found a broken fine-tune of yours in the drawer, leave us a review and tell us what it lost. We collect these like Bert collects mislabeled reels.



Herman

See you soon.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Dettmers et al., LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale, NeurIPS 2022, v2. Retrieved 2026-11-15. [primary]**
   https://arxiv.org/abs/2208.07339v2

2. **Quantization Degradation in Large Language Models: A Signal**
Noise Perspective, arXiv preprint (full text read via HTML). Retrieved 2026-11-15.
https://arxiv.org/abs/2608.08188

3. **Quantization Meets Reasoning: Exploring and Mitigating Degradation of Low-Bit LLMs in Mathematical Reasoning, arXiv v4 (full text read). Retrieved 2026-11-15.**
https://arxiv.org/html/2505.11574v4

4. **Capability-Specific Degradation Patterns in Quantized Small Language Models, IAJ (abstract). Retrieved 2026-11-15.**
https://doi.org/10.36079/lamintang.ijai-01301.1050

5. **Exploring the Trade-Offs: Quantization Methods, Task Difficulty, and Model Size in Large Language Models From Edge to Giant, arXiv:2409.11055 (abstract). Retrieved 2026-11-15.**
https://doi.org/10.48550/arxiv.2409.11055

6. **Scrutinizing the Trustworthiness of Efficient LLMs Under Compression, arXiv v3 (abstract). Retrieved 2026-11-15.**
https://arxiv.org/html/2403.15447v3

7. **Does quantization affect models' performance on long-context tasks?, EMNLP 2025. Retrieved 2026-11-15.**
https://aclanthology.org/2025.emnlp-main.479.pdf

8. **Not All Frames Are Equal: Difference-Aware Quantization for Ultra-Low-Bit ASR, Interspeech 2026, arXiv:2610.22395v1. Retrieved 2026-11-15.**
https://arxiv.org/abs/2610.22395v1

9. **AffectCodec: Emotion-Preserving Neural Speech Codec, ACL Findings 2026. Retrieved 2026-11-15.**
https://arxiv.org/abs/2605.11098v1

10. **Quamba: A Post-Training Quantization Recipe for Selective State Space Models, arXiv v2. Retrieved 2026-11-15.**
https://arxiv.org/abs/2410.13229v2

11. **End-to-End On-Device Quantization-Aware Training for LLMs at Inference Cost (ZeroQAT), arXiv v3, revised 2026-11-05. Retrieved 2026-11-15.**
https://arxiv.org/abs/2509.00031v3

12. **An Extra RMSNorm is All You Need for Fine Tuning to 1.58 Bits, arXiv:2505.08823v1. Retrieved 2026-11-15.**
https://arxiv.org/abs/2505.08823v1

13. **Predict Before You Deploy: Offline Prediction of Quantization-Induced Task Degradation for World Action Models (PreDE), arXiv:2609.19441v1. Retrieved 2026-11-15.**
https://arxiv.org/abs/2609.19441v1

14. **Theory-optimal Quantization Based on Flatness (BDQ), arXiv:2605.18800v1. Retrieved 2026-11-15.**
https://arxiv.org/abs/2605.18800v1

15. **CodeQuant: Unified Clustering and Quantization for Enhanced Outlier Smoothing in Low-Precision Mixture-of-Experts, arXiv v2. Retrieved 2026-11-15.**
https://arxiv.org/abs/2604.10496v2

16. **Preservation-First Quantization: A Cross-Architecture Study of Behavioral Stability under 4-bit Quantization, TechRxiv (abstract). Retrieved 2026-11-15.**
https://doi.org/10.36227/techrxiv.176948214.48929559/v1

17. **SEPTQ: A Simple and Effective Post-Training Quantization Paradigm for LLMs, KDD 2025. Retrieved 2026-11-15.**
https://arxiv.org/abs/2604.10091v1

18. **Performance Analysis of Post-Training Quantization for CNN-based Conjunctival Pallor Anemia Detection (FP32/FP16/INT8/INT4 comparison). Retrieved 2026-11-15.**
https://arxiv.org/abs/2507.15151v1