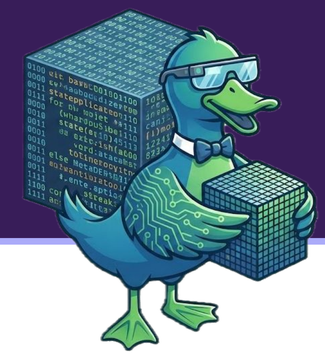
# Float8@2bits: Entropy Coding Enables Data-Free Model Compression

Patrick Putzky*, Martin Genzel*, Mattes Mollenhauer, Sebastian Schulze, Thomas Wollmann, Stefan Dietzel

Merantix Momentum GmbH, Berlin, Germany



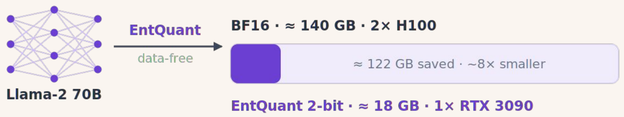
We introduce EntQuant, a novel compression method that keeps model weights in 8-bit precision but optimizes their distribution to enjoy low entropy. Using GPU-accelerated entropy codecs, we obtain compressed models that remain functional up to effectively 2 bits, without calibration or fine-tuning.

**< 10 min**
to compress a 70B model
Compress as you deploy the model. No GPU Cluster.

**0**
calibration data required
Fully data-free. No training, no fine-tuning, just model weights.

**Any model**
including MoEs
Architecture-agnostic. Works on any new open-weight model.

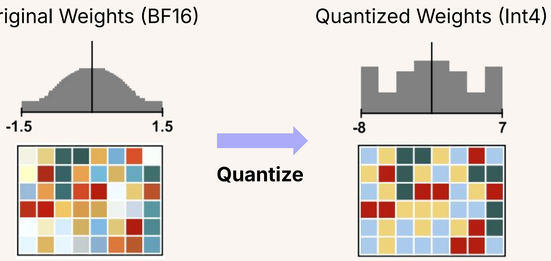
## Why Model Compression




WHAT THIS UNLOCKS

- **Self-hosting** on your own hardware
- **Overcome resource constraints** one GPU, not a cluster
- **Achieve cost savings** consumer-grade GPU

## Where Existing Methods Fall Short




**Weight quantization works well down to 4 bits …**

**… but below, returns start diminishing**

4 bits = 16 distinct values

Enough variety to approximate the weights well

2 bits = 4 distinct values

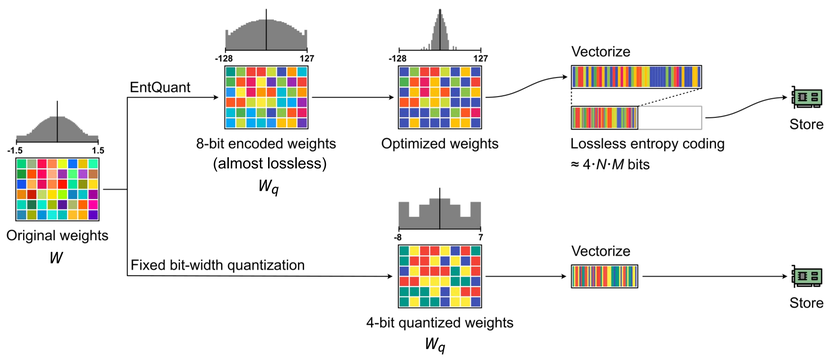
Too coarse, model quality collapses

**Coupling bit-width to compression rate is problematic** in the extreme regime, requiring tricks & hacks to prevent collapse (groups, lattices, recovery training, …)

👉 **Can we go below 4 bits without any data?**

## EntQuant

### Entropy Coding Meets Quantization




### Decoupling numerical precision from compression enables more flexibility in the weight optimization

Optimized weights =

$$\arg\min_{W_q} \frac{\|W - \hat{W}\|_1}{\|W\|_1} + \lambda \|W_q\|_1$$

Reconstruction error

Surrogate for entropy $H(W_q) = -\sum_{w \in W_q} p(w) \log_2 p(w)$

$$W_q = \text{clamp}\left(\lfloor W/s \rceil, -Q_{\text{max}}, Q_{\text{max}}\right) \qquad s = \max(|W|)/Q_{\text{max}}$$

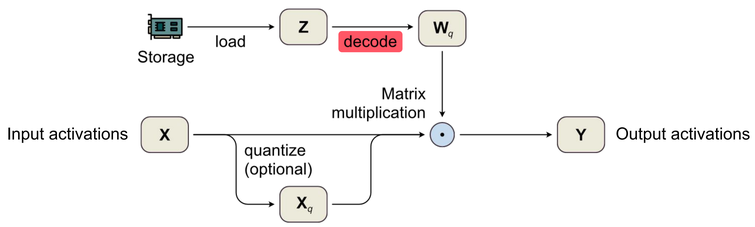
**We tweak the scales $s$ per row to minimize entropy**

# Unique Values (LLaMA-2 7B)

| Method | Quantization Levels (bits) 4 | 3 | 2 |
|---|---|---|---|
| Fixed bit-width | 16.00 | 8.00 | 4.00 |
| EntQuant ∅ | 63.89 | 49.06 | 34.61 |

**Representing low entropy weights in Float8 / Int8 is much more "expressive" than a comparable fixed bit-width format**

### Entropy decoding can be directly built into the inference pipeline




**Decoding is the main bottleneck**, but highly efficient GPU algorithms exist. With use **ANS** (Asymmetric Numeral Systems), which can achieve a throughput of ~600GB/s in NVIDIA's **nvCOMP** package.

## Results

### EntQuant outperforms other data-free methods

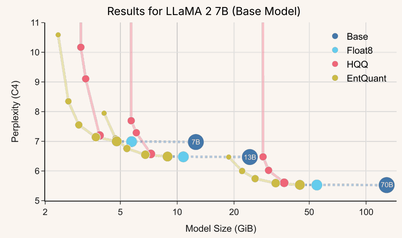
LM Eval Avg. ↑

| Method | Bits | Group | 1-7 | 1-13 | 1-30 | 2-7 | 2-13 | 2-70 | 3.1-8 | 3.1-70 |
|---|---|---|---|---|---|---|---|---|---|---|
| Base | 16 | – | 63.5 | 66.6 | 65.3 | 64.9 | 67.9 | 72.3 | 68.9 | 73.8 |
| HQQ | 3 | 64 | 59.3 | 64.3 | 61.2 | 58.9 | 64.8 | 70.4 | 61.7 | 39.0 |
| HQQ | 3 | 128 | 58.4 | 64.1 | 61.5 | 56.8 | 63.4 | 69.5 | 58.2 | 34.1 |
| EntQuant | 3 | – | **62.4** | **65.7** | **64.4** | **63.6** | **66.7** | **71.1** | **66.0** | **72.6** |
| HQQ | 2 | 16 | 35.0 | 52.9 | 40.4 | 33.3 | 37.6 | 51.3 | 33.6 | 30.0 |
| HQQ | 2 | 32 | 30.6 | 32.9 | 34.0 | 29.9 | 30.3 | 31.9 | 29.9 | 29.9 |
| HQQ | 2 | 64 | 29.9 | 29.7 | 31.7 | 30.3 | 30.8 | 30.4 | 30.7 | 29.9 |
| EntQuant | 2.1 | – | **37.5** | **56.3** | **60.4** | **57.5** | **63.1** | **67.9** | **52.6** | **68.6** |
| EntQuant | 1.7 | – | 41.1 | 34.1 | 54.0 | 43.7 | 50.3 | 60.2 | 36.8 | 35.5 |

ARC-Easy, ARC-Challenge, HellaSwag, WinoGrande, PIQA, BoolQ, OpenbookQA, LAMBADA

C4 ↓ (Perplexity)

| Method | Bits | Group | 1-7 | 1-13 | 1-30 | 2-7 | 2-13 | 2-70 | 3.1-8 | 3.1-70 |
|---|---|---|---|---|---|---|---|---|---|---|
| Base | 16 | – | 7.08 | 6.61 | 5.98 | 6.98 | 6.47 | 5.52 | 8.43 | 5.82 |
| HQQ | 3 | 64 | 8.49 | 7.27 | 6.55 | 9.11 | 7.29 | 6.02 | 12.14 | 629.09 |
| HQQ | 3 | 128 | 9.36 | 7.60 | 6.75 | 10.17 | 7.69 | 6.48 | 14.46 | 620.89 |
| EntQuant | 3 | – | **7.52** | **6.86** | **6.18** | **7.55** | **6.76** | **5.74** | **9.74** | **6.76** |
| HQQ | 2 | 16 | 63.73 | 17.57 | 14.97 | 111.20 | 21.79 | 32.24 | 142.94 | 7.7e3 |
| HQQ | 2 | 32 | 1.4e3 | 186.58 | 60.96 | 1.2e3 | 214.22 | 323.93 | 1.9e3 | 3.0e4 |
| HQQ | 2 | 64 | 4.9e3 | 2.5e3 | 480.61 | 6.0e3 | 2.7e3 | 2.8e3 | 1.8e4 | 1.3e4 |
| EntQuant | 2.1 | – | **9.90** | **8.16** | **7.25** | **10.59** | **7.95** | **6.47** | **17.56** | **9.92** |
| EntQuant | 1.7 | – | 17.76 | 11.19 | 9.40 | 27.92 | 11.14 | 8.43 | 135.24 | 7.9e3 |



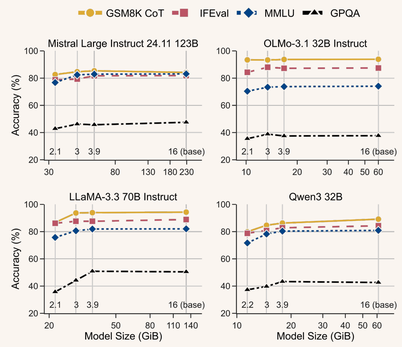

### EntQuant keeps up with data-dependent methods …

Results for LLaMA 2 70B (Base Model)

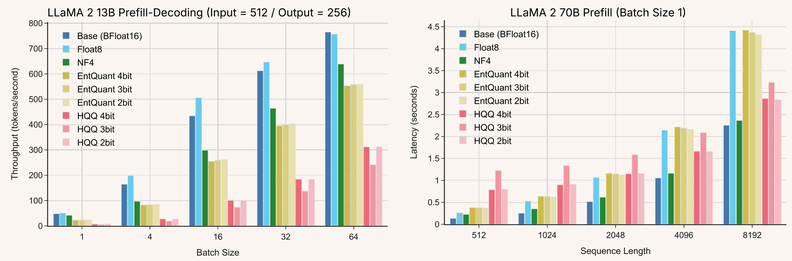
| Method | Level | No Calibration | No Training | Compression Runtime Estimate |
|---|---|---|---|---|
| EntQuant | 1 | ✓ | ✓ | <10min (H100) |
| GPTQ | 2 | ✗ | ✓ | 2–4h (A100-80GiB) |
| OmniQuant | 2 | ✗ | ✓ | 9–16h (A100-80GiB) |
| QuIP# | 3 | ✗ | ✗ | ~50h (8× A100-80GiB) |
| EfficientQAT | 3 | ✓ | ✗ | ~41h (A100-80GiB) |

| Method | Bits | Group | C4 ↓ | WikiText-2 ↓ | LM Eval Avg. ↑ |
|---|---|---|---|---|---|
| Base | 16 | – | 5.52 | 3.32 | 72.8 |
| EntQuant | 3 | – | 5.74 | 3.62 | 71.7 (-1.6%) |
| GPTQ | 3 | 128 | 5.85 | 3.85 | 71.5 (-1.9%) |
| OmniQuant | 3 | 128 | 5.85 | 3.78 | 71.1 (-2.4%) |
| QuIP# | 3 | – | 5.67 | 3.56 | 72.1 (-0.9%) |
| EfficientQAT | 3 | 128 | 5.71 | 3.61 | 71.8 (-1.5%) |
| EntQuant | 2.1 | – | 6.47 | 4.52 | 68.6 (-5.8%) |
| GPTQ | 2 | 128 | – | – | 34.4 (-52.8%) |
| OmniQuant | 2 | 128 | 8.52 | 6.55 | 54.9 (-24.6%) |
| QuIP# | 2 | – | 6.12 | 4.16 | 70.9 (-2.6%) |
| EfficientQAT | 2 | 128 | 6.48 | 4.61 | 68.9 (-5.3%) |

### … and excels on instruction-tuned models




### EntQuant has acceptable inference speed




The entropy decoding overhead is compensated by the Float8 Marlin Kernel

MERANTIX MOMENTUM · ICML International Conference On Machine Learning · scofi · Bundesministerium für Wirtschaft und Energie

Paper · GitHub