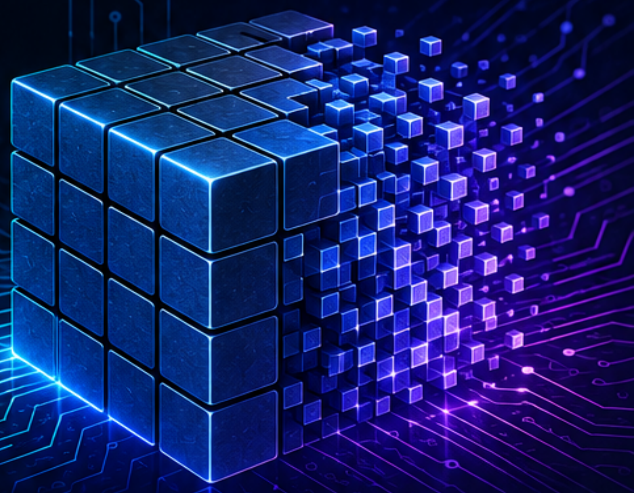
LLM QUANTIZATION RECIPES

SMALLER MODELS
Lower memory usage

FASTER INFERENCE
Higher throughput, lower latency

RUN ANYWHERE
From laptops to datacenters

PRACTICAL RECIPES
Real code, real benchmarks



A PRACTICAL GUIDE TO **COMPRESSING LARGE LANGUAGE MODELS** WITHOUT LOSING THEIR INTELLIGENCE

COVERED QUANTIZATION METHODS

GPTQ
Accurate Post-Training Quantization

AWQ
Activation-aware Weight Quantization

GGUF
Universal Format for Efficient Inference

NF4
Information-Theoretic 4-bit Quantization

BARTOWSKI
GGUF Quantization Presets & Optimizations

BYTESHAPE
Outlier-Aware Quantization

APEX
High-Performance Quantization

2-BIT
Extreme Compression, Maximum Efficiency

3-BIT
Balanced Quality and Size

4/5/8-BIT
Flexible Precision for Every Use Case

UNDERSTAND THE METHODS
From first principles to advanced techniques

REPRODUCIBLE CODE EXAMPLES
Python implementations you can run

REAL BENCHMARKS
Quality, speed, and memory comparisons

PRACTICAL DECISION FRAMEWORK
Choose the right approach for your hardware and use case

STEVE T.

LLM Quantization Recipes

A Practical Guide to Compressing Large Language Models Without Losing Their Intelligence

Steve T. Team Publications

This book is available at https://leanpub.com/llmquantizationrecipes

This version was published on 2026-07-03

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

Contents

A Practical Guide to Compressing Large Language Models Without Losing Their Intelligence

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Introduction: Why Compress the Mind?

In February 2023, Meta released LLaMA, a family of language models ranging from 7 billion to 65 billion parameters. The 65-billion-parameter variant required approximately 130 gigabytes of GPU memory just to load its weights in 16-bit floating-point precision. That is more memory than most consumer GPUs possess. Even the 7-billion-parameter model, at 14 gigabytes, pushed the limits of what a single graphics card could handle once you account for the memory needed by activations, key-value caches, and the inference framework itself [28].

Fast forward to today, and the situation has only intensified. Models with 70 billion, 120 billion, and even trillions of parameters are released regularly. The gap between model size and available hardware is not closing; it is widening. And yet, something remarkable has happened in parallel: ordinary developers with consumer laptops, Mac minis, and mid-range GPUs are running these massive models locally, generating text at usable speeds, and fine-tuning them on custom datasets.

The bridge between those two realities is quantization.

Quantization is the process of representing model weights with fewer bits. Where a standard LLM stores each parameter as a 16-bit or 32-bit floating-point number, a quantized version might use only 4 bits per parameter, or even fewer. This compression reduces memory requirements by a factor of four or more, and because LLM inference is predominantly limited by memory bandwidth rather than raw compute, the speedup can be nearly proportional.

This book is about the engineering of that compression. It is not about the mathematics of neural networks, nor about the architecture of transformers, though both provide important context. It is specifically about the methods, tools, and trade-offs involved in shrinking large language models so they fit on real hardware without losing their intelligence.

What You Will Learn

By the end of this book, you will be able to:

- Explain the difference between weight-only quantization and activation quantization, and why LLMs favor one over the other.
- Understand how GPTQ uses Hessian-based optimization to achieve accurate 4-bit quantization.
- Describe how AWQ identifies and protects the most important weight channels before quantizing.
- Navigate the GGUF format and llama.cpp ecosystem for CPU and consumer GPU deployment.
- Use bitsandbytes and NF4 quantization for QLoRA fine-tuning on a single GPU.
- Evaluate the trade-offs between different quantization methods for your specific hardware and workload.
- Write production-ready code that loads, runs, and evaluates quantized models.
- Apply a systematic decision framework to choose the right quantization strategy for any deployment scenario.

How This Book Is Organized

The first two chapters build the foundation. Chapter 1 establishes why model compression matters, tracing the growth of LLM sizes and the hardware constraints they create. Chapter 2 explains numerical precision from first principles: how floating-point and integer formats work, why certain bit widths matter, and how hardware accelerators process different datatypes.

Chapters 3 through 6 cover the core quantization paradigms and algorithms. Chapter 3 contrasts post-training quantization with quantization-aware training and explains the weight-only versus activation quantization distinction. Chapter 4 covers calibration, the critical step that determines quantization quality. Chapters 5 and 6 provide deep dives into GPTQ and AWQ, the two most influential post-training quantization algorithms for LLMs.

Chapters 7 and 8 turn to practical toolchains. Chapter 7 covers GGUF and llama.cpp, the dominant stack for local and consumer hardware inference.

Chapter 8 covers bitsandbytes, NF4, and QLoRA, the combination that democratized LLM fine-tuning on a single GPU.

Chapters 9 and 10 explore the broader ecosystem. Chapter 9 examines Unsloth and its approach to accelerating quantized fine-tuning. Chapter 10 surveys additional frameworks and methods, including SmoothQuant, ZeroQuant, FP8, and the emerging ShapeLearn approach from ByteShape.

The final chapters focus on deployment and decision-making. Chapter 11 maps quantization methods to hardware targets, from data-center GPUs to edge devices. Chapter 12 provides benchmarking methodology, a comparison of real performance numbers, common pitfalls, and a decision framework for choosing the right quantization strategy in practice.

Prerequisites

This book assumes you are familiar with the basics of deep learning: what neural networks are, what training and inference mean, and how transformers work at a high level. You should be comfortable reading Python code and have some experience working with PyTorch or the Hugging Face Transformers library. You do not need to understand the internals of backpropagation or attention mechanisms, though those concepts occasionally surface in the discussion.

If you are new to LLMs, I recommend reading a general introduction first. If you are an experienced practitioner who wants to understand quantization, this book should work for you directly.

Let us begin.

Chapter 1: The Compression Imperative

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The Size Explosion: From GPT-2 to Today

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Why Raw Parameters Are a Bottleneck

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

What Quantization Actually Does

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

A Brief History of Model Compression

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

What This Book Covers (and Does Not)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Chapter 2: Foundations of Numerical Precision in Deep Learning

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Floating-Point Arithmetic: FP32, FP16, BF16

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Integer Representations: INT8, INT4, INT2

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Mixed Precision and the NF4 Format

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

How GPUs and CPUs Handle Different Datatypes

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The Information Theory of Weight Distributions

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Chapter 3: Post-Training Quantization vs. Quantization-Aware Training

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Post-Training Quantization (PTQ): The Quick Path

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Quantization-Aware Training (QAT): The Careful Path

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Weight-Only Quantization: The LLM Sweet Spot

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Activation Quantization and the Outlier Problem

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Hybrid Approaches and Per-Token Strategies

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Chapter 4: Calibration–Teaching Precision to Models

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The Role of Calibration Data

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Min-Max vs. Percentile Clipping

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Moving Average and Histogram-Based Methods

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Optimal Perceptual Quantization (OPQ)

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

How Much Calibration Data Do You Really Need?

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Chapter 5: GPTQ–Greedy One-Shot Quantization

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The Hessian Approximation Idea

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Layer-by-Layer Greedy Optimization

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The GPTQ Algorithm Step by Step

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

AutoGPTQ: The Practical Implementation

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Strengths, Limitations, and Typical Results

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Mathematical Derivation: Why the Hessian Works

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Complete Production GPTQ Quantization Script

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Production Readiness Checklist for GPTQ

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Chapter 6: AWQ–Activation-Aware Weight Quantization

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The Activation Magnitude Insight

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Weight Scaling Before Quantization

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The AWQ Algorithm: Smoothing and Rescaling

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

AWQ vs. GPTQ: A Head-to-Head Comparison

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Practical Usage with AutoAWQ

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Complete Production AWQ Quantization Script

Production Readiness Checklist for AWQ

Case Study: Deploying a 70B Model on a Single A100

Chapter 7: GGUF and the llama.cpp Ecosystem

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The GGML Legacy and GGUF's Design

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

K-Quants: Q4_0, Q4_K_S, Q5_K_M, Q8_0

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

How llama.cpp Runs Quantized Models on CPU

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Performance on CPUs vs. GPUs with Metal/Vulkan

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Community Toolchains and Model Hubs

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Importance Matrix (imatrix) Quantization: A Deep Dive

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Complete GGUF Conversion Pipeline

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Chapter 8: BitsAndBytes and the NF4 Revolution

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The bitsandbytes Library Architecture

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

NormalFloat4: Why It Beats Plain INT4

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

QLoRA: Fine-Tuning in 4 Bits

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

8-Bit Adam and Optimizer Quantization

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Practical Usage with the Transformers Library

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Complete Production QLoRA Fine-Tuning Script

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Production Readiness Checklist for QLoRA

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Chapter 9: Unsloth–Speed Through Quantized Fine-Tuning

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The Fine-Tuning Bottleneck

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Unsloth's Architecture and Optimizations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Patched Transformers: How the Speedup Works

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Benchmarks: Unsloth vs. Standard QLoRA

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Practical Usage and Limitations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Chapter 10: Advanced Frameworks–Bartowski, ByteShape, and Apex

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Bartowski's Quantization Pipeline

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

ByteShape: Understanding This Approach

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

NVIDIA Apex: From Mixed-Precision Training to FP8

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Other Notable Methods: SmoothQuant, ZeroQuant, SPA

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The Fragmentation Problem in Toolchains

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Chapter 11: Deployment Scenarios and Hardware Considerations

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Local Inference on Consumer GPUs

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

CPU-Only Deployment and Edge Devices

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

High-Throughput Server Serving

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Mobile and On-Device LLMs

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Case Study: Deploying a Customer Support Chatbot on a Single GPU

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The Memory Bandwidth Bottleneck

Chapter 12: Benchmarking, Best Practices, and Decision Framework

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

A Reproducible Benchmarking Protocol

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Protocol Overview

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Quality Evaluation with lm-evaluation-harness

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Perplexity Measurement Protocol

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Throughput Measurement Protocol

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Memory Measurement Protocol

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Reproducibility Checklist

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

How to Measure Quantization Quality

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Perplexity, Accuracy, and Latency Benchmarks

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Common Pitfalls and Debugging Tips

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

The Quantization Decision Framework

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Future Directions: What's Next in Model Compression

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

Conclusion: The Democratized Model

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.

References

This content is not available in the sample book. The book can be purchased on Leanpub at https://leanpub.com/llmquantizationrecipes.