LOCAL INTELLIGENCE

Running Large Language Models on Your MacBook with Apple Silicon

The complete, practical guide to running open-source LLMs locally for privacy, cost savings, and independence from cloud APIs.

BUILT FOR APPLE SILICON
Understand the architecture that makes M1, M2, and M3 perfect for local AI.

MASTER THE TOOLS
llama.cpp, Ollama, MLX, and Hugging Face Transformers—explained in depth.

QUANTIZATION STRATEGIES
Fit billion-parameter models into unified memory and get the best performance.

BUILD REAL APPLICATIONS
Create powerful, private AI apps that run entirely on your Mac.

PRIVATE
Your data stays on your Mac.

COST-EFFECTIVE
No API bills. No subscriptions.

100% OPEN SOURCE
All tools. All code. Fully reproducible.

COLLINS DEKKARD

Local Intelligence

Running Large Language Models on Your MacBook with Apple Silicon

Steve T. Publications

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

This version was published on 2026-07-05

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

Running Large Language Models on Your MacBook with Apple Silicon

This book is a complete, practical guide to running large language models entirely on your own Mac. If you have an M1, M2, or M3 MacBook and want to deploy open-source LLMs locally for privacy, cost savings, or independence from cloud APIs, this book takes you from zero to mastery. You will learn the hardware architecture that makes Apple Silicon uniquely suited for this work, master every major inference framework (llama.cpp, Ollama, MLX, and Hugging Face Transformers), understand quantization strategies that fit billion-parameter models into your unified memory, and build real applications with local AI. Every technique described is fully open-source, reproducible on macOS, and grounded in real benchmarks and working code.

Introduction

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

The Problem with Cloud APIs

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

Why Apple Silicon Is Different

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

What This Book Covers

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

Who This Book Is For

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

How to Use This Book

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

Chapter 1: The Local LLM Revolution

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

Why Local? Privacy, Cost, and Independence

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

The Cloud API Trap: Latency, Censorship, and Hidden Costs

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

Apple Silicon Changes Everything

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

What You Can Actually Run on a MacBook Today

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

How This Book Is Structured

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

Chapter 2: Apple Silicon Architecture for ML Workloads

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

Unified Memory Architecture Explained

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

CPU, GPU, and Neural Engine: Who Does What?

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

M1 vs M2 vs M3: The Progression for AI Workloads

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

Thermal Design and Sustained Performance

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

Benchmarking Your Mac's ML Capability

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

Chapter 3: Model Selection–Finding the Right LLM

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

The Open-Source LLM Landscape in 2025 and Beyond

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

Model Families: Llama, Mistral, Qwen, Gemma, and Others

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

Parameter Count vs Performance: The Sweet Spot for MacBooks

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

GGUF Format and the Quantization Zoo

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

Building Your Personal Model Library

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

Chapter 4: llama.cpp–The Foundation of Local Inference

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

Installing llama.cpp on macOS

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

Loading Your First GGUF Model

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

Understanding Quantization: Q4_K_M, Q5_K_S, Q8_0, and Beyond

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

Server Mode: REST API and Multi-User Access

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

Advanced Features: LoRA Adapters, Embeddings, and Multimodal

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

Chapter 5: Ollama–Simplicity at Scale

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

Installing Ollama on macOS

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

The Library: Pulling and Managing Models

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

Modelfiles: Customizing System Prompts and Parameters

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

The Ollama API: Integrating into Applications

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

Creating and Publishing Your Own Models

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

Chapter 6: MLX–Apple's Native Machine Learning Framework

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

What Is MLX and Why It Matters

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

Installing and Setting Up the MLX Ecosystem

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

Running Models with mlx-lm

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

Fine-Tuning Models on Your Mac

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

MLX vs llama.cpp vs Ollama: When to Use What

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

Chapter 7: Hugging Face Transformers on Apple Silicon

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

The Hugging Face Ecosystem on macOS

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

Bitsandbytes and 4-Bit/8-Bit Quantization

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

Hugging Face Optimum for Apple Silicon

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

PEFT: LoRA, QLoRA, and Parameter-Efficient Tuning

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

Building a Complete Local Training Pipeline

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

Chapter 8: Performance Optimization and Tuning

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

Memory Management: Unified Memory as Both Blessing and Constraint

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

Context Window Optimization and KV Cache Strategies

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

Batch Size, Prompt Processing, and Token Generation Speed

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

Thermal Throttling: Keeping Your Mac Cool Under Load

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

Profiling Tools and Performance Metrics

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

Chapter 9: Building Local Applications with LLMs

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

The Local Application Stack

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

Chat Interfaces and Web UIs: Open WebUI, Text Generation WebUI

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

RAG Pipelines: Local Retrieval-Augmented Generation

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

Tool Use and Function Calling with Local Models

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

Building a Production-Grade Local AI Assistant

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

Chapter 10: Fine-Tuning and Customization at Home

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

When to Fine-Tune vs When to Use Prompt Engineering

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

Data Preparation for Local Fine-Tuning

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

QLoRA Training on Apple Silicon with MLX

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

Evaluating Your Fine-Tuned Model

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

Deploying Custom Models in Production

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

Chapter 11: The Broader Ecosystem–Tools and Integrations

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

Model Serving: vLLM, Text Generation Inference, and Local Alternatives

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

Evaluation Frameworks: Measuring Your Model's Quality

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

Embeddings Locally: Sentence Transformers and Beyond

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

Multimodal Models on Apple Silicon

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

CI/CD for Local LLM Deployments

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

Chapter 12: Troubleshooting and Real-World Gotchas

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

Out of Memory: Diagnosing and Solving RAM Exhaustion

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

Slow Inference: Identifying Bottlenecks

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

Model Compatibility and Format Conversion

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

macOS-Specific Issues and Kernel Panics

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

Recovery Strategies and Safe Shutdown

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

Conclusion: The Future of Personal AI

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

What We Have Learned

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

The Trajectory: Smaller Models, Bigger Chips

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

The Philosophical Shift: From Cloud Dependency to Personal Sovereignty

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

Your Next Steps

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

References

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