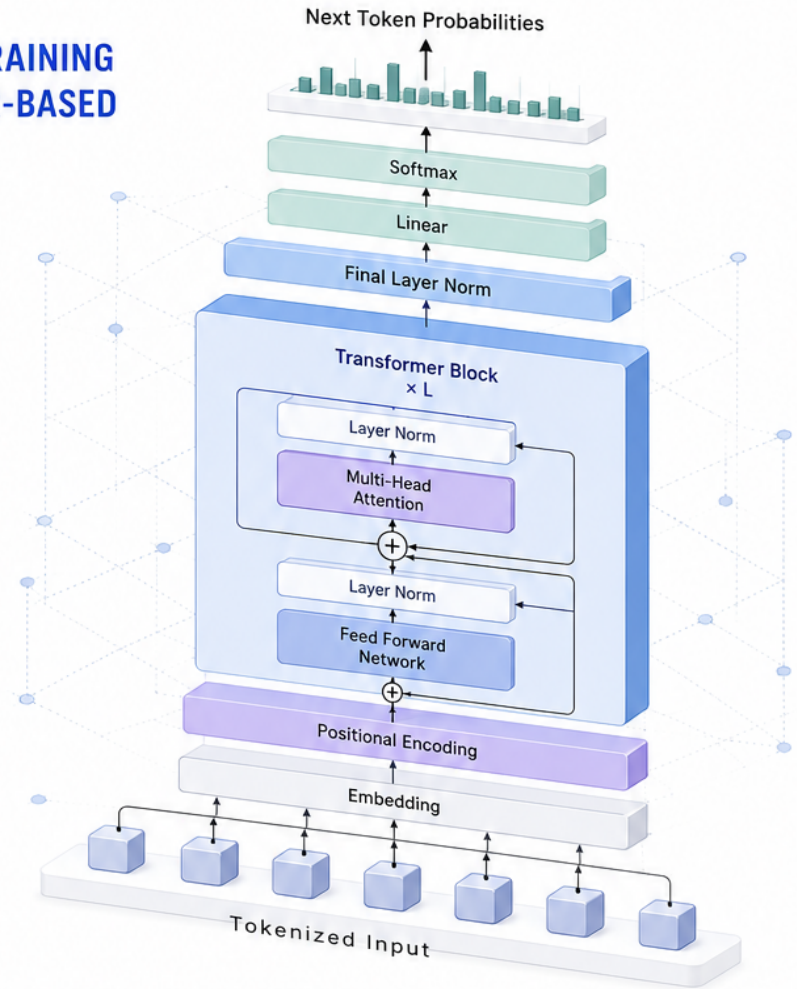
BUILDING LARGE LANGUAGE MODELS FROM SCRATCH

WITH PYTORCH

A PRACTICAL GUIDE TO TRAINING YOUR OWN TRANSFORMER-BASED AI IN PYTHON

- FROM RAW TEXT TO INFERENCE API
- TOKENIZATION & EMBEDDINGS
- ATTENTION & TRANSFORMERS
- DISTRIBUTED TRAINING
- ALIGNMENT, EVALUATION & DEPLOYMENT



COLLINS DEKKARD

Building Large Language Models from Scratch

A Practical Guide to Training Your Own Transformer-Based AI in Python

Steve T. Publications

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

This version was published on 2026-07-03

Contents

Building Large Language Models from Scratch **1**

A Practical Guide to Training Your Own Transformer-Based AI in Python **2**

Introduction: Why Build from Scratch? **3**
The Black Box Problem 3
What You Will Build 3
Prerequisites and How to Use This Book 4
A Note on Hardware Requirements 5

Chapter 1: Tokens, Vocabularies, and Tokenization **7**
From Text to Numbers: The Tokenization Pipeline 7
Character-Level vs Word-Level vs Subword Tokenization 7
Building a Byte-Pair Encoding (BPE) Tokenizer 7
Vocabulary Size, Special Tokens, and Edge Cases 7
Exercise: Tokenize the Shakespeare Corpus 7

Chapter 2: Embeddings: Turning Tokens into Vectors **8**
The Embedding Layer as a Lookup Table 8
Learning vs Pre-trained Embeddings 8
Vector Space Geometry and Similarity 8
Implementing Embedding Layers in PyTorch 8
Exercise: Visualize an Embedding Space 8

Chapter 3: The Attention Mechanism **9**
What Is Attention and Why It Matters 9
Scaled Dot-Product Attention from First Principles 9
Multi-Head Attention: Parallelizing Understanding 9
Causal Masking for Decoder-Only Models 9
Exercise: Trace Attention Through a Sentence 9

Chapter 4: Positional Encoding and Sequence Structure **10**

The Permutation Invariance Problem 10
Sinusoidal Absolute Positional Encodings 10
Learned Position Embeddings 10
Rotary Positional Embeddings (RoPE) 10
Exercise: Implement Three Position Encoding Schemes 10

Chapter 5: Building the Decoder-Only Transformer Architecture 11
The Decoder-Only Design Decision 11
Feed-Forward Networks and MLP Blocks 11
Residual Connections and Layer Normalization 11
Assembling the Full Transformer Block 11
The Complete Decoder-Only Model 11
Exercise: Build a 3-Layer Decoder from Scratch 11

Chapter 6: Data Preparation for Language Model Training 13
The Data Landscape: Where Does Training Data Come From? 13
Cleaning and Filtering Pipeline Design 13
Deduplication Strategies 13
Dataset Mixing and Domain Balancing 13
Synthetic Data Generation Strategies 13
Exercise: Build a Mini C4 Dataset 13

Chapter 7: The Training Loop: Loss, Optimizers, and Gradient Flow . . 15
Cross-Entropy Loss and Next-Token Prediction 15
The AdamW Optimizer and Why It Works 15
Learning Rate Schedules: Warmup, Cosine Decay, and Beyond 15
Gradient Clipping and Training Stability 15
The Complete Training Loop 15
Exercise: Train on a Tiny Dataset and Monitor Loss Curves 15

Chapter 8: Memory-Efficient Training Patterns 17
The Memory Wall: Why Models Don't Fit in GPU RAM 17
Mixed-Precision Training with BF16/FP16 17
Gradient Accumulation for Effective Batch Sizes 17
Activation Checkpointing and Recomputation 17
Exercise: Train a 10x Larger Model on the Same Hardware 17

Chapter 9: Distributed Training and Parallelism Strategies 18
Data Parallelism and Distributed Data Parallel (DDP) 18
Tensor Parallelism for Massive Models 18

Pipeline Parallelism: Splitting the Forward Pass 18
Fully Sharded Data Parallel (FSDP) 18
Exercise: Multi-GPU Training Setup 18

Chapter 10: Checkpointing, Experiment Tracking, and Reproducibility 19
Checkpointing Strategies and Recovery 19
Experiment Tracking: Metrics, Configs, and Artifacts 19
Reproducibility: Seeds, Determinism, and Hardware Variability 19
Logging Design for Long Training Runs 19
Exercise: Set Up a Production-Style Training Dashboard 19

Chapter 11: Fine-Tuning: LoRA, QLoRA, and Instruction Tuning 21
The Fine-Tuning Landscape: Full vs Parameter-Efficient 21
Low-Rank Adaptation (LoRA) from First Principles 21
Quantized LoRA (QLoRA) for Memory-Constrained Fine-Tuning 21
Instruction Tuning Dataset Design 21
Exercise: Fine-Tune a Model on a Custom Task 21

Chapter 12: Alignment: RLHF and Beyond 22
Why Raw Models Need Alignment 22
The RLHF Pipeline: Reward Models and PPO 22
Direct Preference Optimization (DPO) as a Simpler Alternative 22
Constitutional AI and Rule-Based Alignment 22
Exercise: Build a Simple Preference Dataset 22

Chapter 13: Evaluation: Metrics, Benchmarks, and Red Teaming 23
Perplexity as a Training Metric vs Real-World Performance 23
Standard Benchmark Suites (MMLU, GSM8K, HumanEval) 23
Qualitative Evaluation and LLM-as-Judge 23
Safety Testing and Red Teaming Methodologies 23
Exercise: Build an Evaluation Harness 23

Chapter 14: Deployment: Inference, Quantization, and Serving 24
Inference Optimization: KV Cache and Speculative Decoding 24
Quantization Strategies: FP16 -> INT8 -> INT4 24
Building a Serving API with FastAPI 24
Containerization and Production Deployment Patterns 24
Exercise: Deploy Your Model Behind a Live API 24

Capstone Project: From Raw Data to Live Inference 25

Project Setup and Architecture Overview 25
Data Preparation: Curating a 100M-Token Dataset 25
Training Run: Configuration, Execution, and Monitoring 25
Evaluation: Benchmarking Against Baselines 25
Deployment: Containerized API with Health Checks 25
Post-Mortem: What Went Well and What Would Be Different 25

Conclusion: The Road Ahead **27**
What You Have Accomplished 27
Where to Go From Here: Scaling Up 27
The Future of Language Models 27
A Final Word on Building vs Using 27

References **28**

Building Large Language Models from Scratch

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

A Practical Guide to Training Your Own Transformer-Based AI in Python

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

Introduction: Why Build from Scratch?

The Black Box Problem

In 2022, a single model called GPT-3.5 demonstrated the ability to write code, answer questions, translate languages, and generate creative text with startling fluency. Within a year, it had been wrapped in a chat interface and was being used by hundreds of millions of people worldwide. For most users, the model is a black box: you type a prompt, you get an answer, and the magic happens somewhere inside a server farm you will never visit.

For software engineers and data scientists, the situation is only slightly different. You call an API endpoint, pass in your text, and receive tokens back. The pricing is measured in dollars per million tokens. The model architecture, training data, and alignment procedures are trade secrets. You can fine-tune with a few hundred examples using a tool like LoRA, but the foundation model itself remains opaque.

This book takes the opposite approach. We will build everything from scratch, starting with raw text files and ending with a trained model that you can run, inspect, and modify at will. The goal is not to reproduce GPT-4 on your laptop (that would require roughly $100 million in compute). The goal is something more valuable: deep, first-principles understanding of how large language models work, what makes them tick, and what it actually takes to train one.

There is a practical reason for this approach as well. Every year, new architectures, training techniques, and optimization strategies emerge. Models that were cutting-edge in 2023 are baseline in 2025. If you only know how to call APIs, you are at the mercy of whatever the latest framework provides. If you understand the underlying mechanics, you can evaluate new ideas critically, adapt them to your own needs, and contribute to the field rather than merely consuming it.

What You Will Build

By the end of this book, you will have built a complete LLM training pipeline that includes:

- A byte-pair encoding (BPE) tokenizer that converts raw text into integer sequences
- An embedding layer that maps tokens to dense vector representations
- A multi-head attention mechanism with causal masking for autoregressive generation
- Rotary positional embeddings (RoPE) that encode sequence order information
- A complete decoder-only transformer architecture with RMSNorm normalization and residual connections
- A data preparation pipeline for curating, cleaning, and formatting training text
- A training loop with cross-entropy loss, AdamW optimizer, learning rate scheduling, and gradient clipping
- Memory-efficient training patterns including mixed precision (BF16), gradient accumulation, and activation checkpointing
- Distributed training support using PyTorch's Distributed Data Parallel
- Checkpointing, experiment tracking, and reproducibility infrastructure
- Parameter-efficient fine-tuning via LoRA, including 4-bit quantized QLoRA
- Alignment techniques including Direct Preference Optimization (DPO)
- An evaluation harness covering perplexity, benchmark suites, and safety testing
- A production-ready inference API with KV caching, quantization, and FastAPI serving

The capstone project ties everything together: you will train a small but fully functional language model on a curated dataset, evaluate it against baselines, and deploy it behind a live API endpoint. The model will be modest in size (roughly 30 to 125 million parameters, depending on your hardware), but it will demonstrate every architectural and training principle used by models with billions of parameters.

Prerequisites and How to Use This Book

This book assumes you are comfortable writing Python code, including working with classes, functions, list comprehensions, and basic data structures. You should have used PyTorch or a similar framework at least once, enough to understand what a tensor is, how `nn.Module` works, and what happens when you call `.backward()` on a loss. A basic familiarity with linear algebra (matrix multiplication, vector operations) and machine learning concepts (loss functions, gradient descent, overfitting) will help, but I will review each concept as it becomes relevant.

You do not need prior experience with transformer architectures, attention mechanisms, or language model training. That is precisely what this book teaches. Each chapter builds on the previous one, so the recommended reading order matches the table of contents. If you want to jump ahead to a specific topic, the chapters on architecture (Chapters 3 through 5) are relatively self-contained, and the chapters on fine-tuning and deployment (Chapters 11 through 14) can be read independently once you understand the basic training loop.

Each chapter follows a consistent structure. Conceptual explanations come first, establishing the "why" before the "how." Code implementations follow, with complete, runnable examples that you can copy, modify, and experiment with. Mathematical notation appears only where it clarifies the code, and every formula is accompanied by an intuitive explanation in plain English. Each chapter ends with exercises or mini-projects that reinforce the key ideas through hands-on practice.

A Note on Hardware Requirements

Training language models requires computational resources, but the barrier to entry has dropped significantly. Here is what you need for different stages of this book:

CPU-only exploration: You can follow every conceptual explanation and run the smallest code examples (character-level models, tiny vocabularies) on any modern laptop. Training will be slow, but the code will work. This is sufficient for understanding Chapters 1 through 5 in depth.

Single GPU training: A consumer GPU with at least 8 GB of VRAM (NVIDIA RTX 3060, 4070, or better) lets you train small models (10 to 50 million parameters) on modest datasets. An RTX 4090 (24 GB) or A100 (40-80 GB) opens up medium-sized models (100M+ parameters). Google Colab's free tier provides a T4 GPU (16 GB), which is sufficient for many exercises.

Cloud GPU access: Services like Lambda Labs, RunPod, and Vast.ai offer A100 and H100 GPUs at roughly $1 to $3 per hour. The capstone project can be completed on a single A100 40GB for approximately $20 to $50 in total compute cost.

Multi-GPU scaling: Chapters 9 and beyond discuss distributed training across multiple GPUs. This is optional but recommended if you want to train larger models or understand production-scale training infrastructure.

Throughout the book, I will note where hardware requirements increase and provide scaled-down alternatives when possible. The code is designed to work on the smallest feasible setup, with clear paths to scale up.

Chapter 1: Tokens, Vocabularies, and Tokenization

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

From Text to Numbers: The Tokenization Pipeline

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

Character-Level vs Word-Level vs Subword Tokenization

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

Building a Byte-Pair Encoding (BPE) Tokenizer

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

Vocabulary Size, Special Tokens, and Edge Cases

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

Exercise: Tokenize the Shakespeare Corpus

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

Chapter 2: Embeddings: Turning Tokens into Vectors

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

The Embedding Layer as a Lookup Table

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

Learning vs Pre-trained Embeddings

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

Vector Space Geometry and Similarity

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

Implementing Embedding Layers in PyTorch

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

Exercise: Visualize an Embedding Space

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

Chapter 3: The Attention Mechanism

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

What Is Attention and Why It Matters

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

Scaled Dot-Product Attention from First Principles

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

Multi-Head Attention: Parallelizing Understanding

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

Causal Masking for Decoder-Only Models

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

Exercise: Trace Attention Through a Sentence

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

Chapter 4: Positional Encoding and Sequence Structure

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

The Permutation Invariance Problem

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

Sinusoidal Absolute Positional Encodings

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

Learned Position Embeddings

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

Rotary Positional Embeddings (RoPE)

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

Exercise: Implement Three Position Encoding Schemes

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

Chapter 5: Building the Decoder-Only Transformer Architecture

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

The Decoder-Only Design Decision

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

Feed-Forward Networks and MLP Blocks

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

Residual Connections and Layer Normalization

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

Assembling the Full Transformer Block

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

The Complete Decoder-Only Model

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

Exercise: Build a 3-Layer Decoder from Scratch

Chapter 6: Data Preparation for Language Model Training

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

The Data Landscape: Where Does Training Data Come From?

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

Cleaning and Filtering Pipeline Design

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

Deduplication Strategies

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

Dataset Mixing and Domain Balancing

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

Synthetic Data Generation Strategies

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

Exercise: Build a Mini C4 Dataset

Chapter 7: The Training Loop: Loss, Optimizers, and Gradient Flow

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

Cross-Entropy Loss and Next-Token Prediction

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

The AdamW Optimizer and Why It Works

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

Learning Rate Schedules: Warmup, Cosine Decay, and Beyond

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

Gradient Clipping and Training Stability

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

The Complete Training Loop

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

Exercise: Train on a Tiny Dataset and Monitor Loss Curves

Chapter 8: Memory-Efficient Training Patterns

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

The Memory Wall: Why Models Don't Fit in GPU RAM

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

Mixed-Precision Training with BF16/FP16

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

Gradient Accumulation for Effective Batch Sizes

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

Activation Checkpointing and Recomputation

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

Exercise: Train a 10x Larger Model on the Same Hardware

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

Chapter 9: Distributed Training and Parallelism Strategies

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

Data Parallelism and Distributed Data Parallel (DDP)

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

Tensor Parallelism for Massive Models

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

Pipeline Parallelism: Splitting the Forward Pass

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

Fully Sharded Data Parallel (FSDP)

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

Exercise: Multi-GPU Training Setup

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

Chapter 10: Checkpointing, Experiment Tracking, and Reproducibility

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

Checkpointing Strategies and Recovery

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

Experiment Tracking: Metrics, Configs, and Artifacts

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

Reproducibility: Seeds, Determinism, and Hardware Variability

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

Logging Design for Long Training Runs

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

Exercise: Set Up a Production-Style Training Dashboard

Chapter 11: Fine-Tuning: LoRA, QLoRA, and Instruction Tuning

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

The Fine-Tuning Landscape: Full vs Parameter-Efficient

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

Low-Rank Adaptation (LoRA) from First Principles

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

Quantized LoRA (QLoRA) for Memory-Constrained Fine-Tuning

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

Instruction Tuning Dataset Design

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

Exercise: Fine-Tune a Model on a Custom Task

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

Chapter 12: Alignment: RLHF and Beyond

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

Why Raw Models Need Alignment

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

The RLHF Pipeline: Reward Models and PPO

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

Direct Preference Optimization (DPO) as a Simpler Alternative

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

Constitutional AI and Rule-Based Alignment

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

Exercise: Build a Simple Preference Dataset

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

Chapter 13: Evaluation: Metrics, Benchmarks, and Red Teaming

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

Perplexity as a Training Metric vs Real-World Performance

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

Standard Benchmark Suites (MMLU, GSM8K, HumanEval)

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

Qualitative Evaluation and LLM-as-Judge

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

Safety Testing and Red Teaming Methodologies

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

Exercise: Build an Evaluation Harness

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

Chapter 14: Deployment: Inference, Quantization, and Serving

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

Inference Optimization: KV Cache and Speculative Decoding

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

Quantization Strategies: FP16 -> INT8 -> INT4

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

Building a Serving API with FastAPI

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

Containerization and Production Deployment Patterns

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

Exercise: Deploy Your Model Behind a Live API

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

Capstone Project: From Raw Data to Live Inference

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

Project Setup and Architecture Overview

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

Data Preparation: Curating a 100M-Token Dataset

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

Training Run: Configuration, Execution, and Monitoring

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

Evaluation: Benchmarking Against Baselines

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

Deployment: Containerized API with Health Checks

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

Post-Mortem: What Went Well and What Would Be Different

Conclusion: The Road Ahead

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

What You Have Accomplished

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

Where to Go From Here: Scaling Up

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

The Future of Language Models

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

A Final Word on Building vs Using

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

References

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