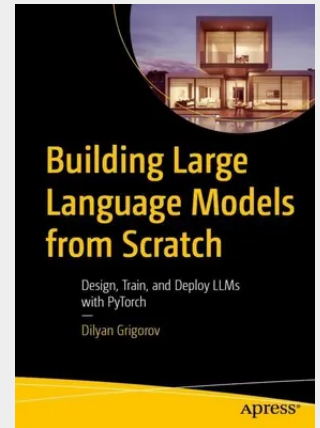


Building Large Language Models from Scratch

Design, Train, and Deploy LLMs with PyTorch

This book is a complete, hands-on guide to designing, training, and deploying your own Large Language Models (LLMs)—from the foundations of tokenization to the advanced stages of fine-tuning and reinforcement learning. Written for developers, data scientists, and AI practitioners, it bridges core principles and state-of-the-art techniques, offering a rare, transparent look at how modern transformers truly work beneath the surface. Starting from the essentials, you'll learn how to set up your environment with Python and PyTorch, manage datasets, and implement critical fundamentals such as tensors, embeddings, and gradient descent. You'll then progress through the architectural heart of modern models, covering RMS normalization, rotary positional embeddings (RoPE), scaled dot-product attention, Grouped Query Attention (GQA), Mixture of Experts (MoE), and SwiGLU activations, each explored in depth and built step by step in code. As you advance, the book introduces custom CUDA kernel integration, teaching you how to optimize key components for speed and memory efficiency at the GPU level—an essential skill for scaling real-world LLMs. You'll also gain mastery over the phases of training that define today's leading models: Pretraining - Building general linguistic and semantic understanding. Midtraining - Expanding domain-specific capabilities and adaptability. Supervised Fine-Tuning (SFT) - Aligning behavior with curated, task-driven data. Reinforcement Learning from Human Feedback (RLHF) - Refining responses through reward-based optimization for human alignment. The final chapters guide you through dataset preparation, filtering, deduplication, and training optimization, culminating in model evaluation and real-world prompting with a custom TokenGenerator for text generation and inference. By the end of this book, you'll have the knowledge and confidence to architect, train, and deploy your own transformer-based models, equipped with both the theoretical depth and practical expertise to innovate in the rapidly evolving world of AI. What You'll Learn How to configure and optimize your development environment using PyTorch The mechanics of tokenization, embeddings, normalization, and attention mechanisms. How to implement transformer components like RMSNorm, RoPE, GQA, MoE, and SwiGLU from scratch. How to integrate custom CUDA kernels to accelerate transformer computations. The full LLM training pipeline: pretraining, midtraining, supervised fine-tuning, and RLHF. Techniques for dataset preparation, deduplication, model debugging, and GPU memory management. How to train, evaluate, and deploy a complete GPT-like architecture for real-world tasks. Who this book is for: Software developers, data scientists, machine learning engineers and AI enthusiasts looking to build their models from scratch.

This book is a complete, hands-on guide to designing, training, and deploying your own Large Language Models (LLMs)—from the foundations of tokenization to the advanced stages of fine-tuning and reinforcement learning. Written for developers, data scientists, and AI practitioners, it bridges core principles and state-of-the-art techniques, offering a rare, transparent look at how modern transformers truly work beneath the surface. Starting from the essentials, you'll learn how to set up your environment with Python and PyTorch, manage datasets, and implement critical fundamentals such as tensors, embeddings, and gradient descent. You'll then progress through the architectural heart of modern models, covering RMS normalization, rotary positional embeddings (RoPE), scaled dot-product attention, Grouped Query Attention (GQA), Mixture of Experts (MoE), and SwiGLU activations, each explored in depth and built step by step in code. As you advance, the book introduces custom CUDA kernel integration, teaching you how to optimize key components for speed and memory efficiency at the GPU level—an essential skill for scaling real-world LLMs. You'll also gain mastery over the phases of training that define today's leading models:- Pretraining - Building general linguistic and semantic understanding.- Midtraining - Expanding domain-specific capabilities and adaptability.- Supervised Fine-Tuning (SFT) - Aligning behavior with curated, task-driven data. Reinforcement Learning from Human Feedback (RLHF)



64,19 €

59,99 € (zzgl. MwSt.)

Nicht mehr lieferbar

Artikelnummer: 9798868822964

Medium: Buch

ISBN: 979-8-8688-2296-4

Verlag: APRESS L.P.

Erscheinungstermin: 28.04.2026

Sprache(n): Englisch

Auflage: 1. Auflage 2026

Produktform: Kartoniert

Gewicht: 1032 g

Seiten: 530

Format (B x H): 178 x 254 mm

