

à partir de zéro

Construire des grands modèles de langage à partir de zéro

Intelligence Artificielle

Concevoir un LLM avec PyTorch

12 juin
2026

Mis à jour le 28 juillet
2026

18 min de
lecture

1992
mots

252
vues



Télécharger le
PDF

Telechargez gratuitement l'ouvrage complet sur la construction d'un LLM from scratch avec PyTorch : tokenizer BPE, RoPE, GQA, MoE, RLHF et kernels CUDA. 342 pages, ISBN 978-2-9580154-1-1.

L'ouvrage **Construire des grands modèles de langage à partir de zéro** est le guide technique francophone de référence pour comprendre, implémenter et entraîner un LLM (*Large Language*

Model) complet — du tokenizer BPE à l'inférence en production. Rédigé par **Ayi NEDJIMI**, expert en cybersécurité et intelligence artificielle avec plus de 25 ans d'expérience terrain, ce livre vous guide pas à pas dans la construction d'un grand modèle de langage avec PyTorch, en couvrant les architectures modernes : mécanismes d'attention RoPE, Grouped Query Attention (GQA), fenêtre glissante, Mixture of Experts, SwiGLU, RMSNorm et kernels CUDA personnalisés. Avec 342 pages de code commenté, de schémas mathématiques et de retours d'expérience, cet ouvrage s'adresse aux développeurs, ingénieurs ML et chercheurs qui veulent comprendre les LLM de l'intérieur et non plus les utiliser comme boîtes noires.



Ce que vous apprendrez

Fondations des LLM et mecanismes d attention

Comprenez ce qu'est réellement un grand modèle de langage, ses fondements mathématiques, le mécanisme d'attention (self-attention,
