LLM Course

LLM Course documentation

Summary

LLM Course

■ View all resources

Agents Course

Audio Course

Community Computer Vision Course

Deep RL Course

Diffusion Course

LLM Course

MCP Course

ML for 3D Course

ML for Games Course

Open-Source AI Cookbook

Robotics Course

a smol course

Search documentation

AR

BN

DE

EN

ES

FA

FR

GJ

HE

HI

ID

IT

JA

KO

MY

NE

PL

PT

RO

RU

RUM

TE

TH

TR

VI

ZH-CN

ZH-TW

Join the Hugging Face community

and get access to the augmented documentation experience

Collaborate on models, datasets and Spaces

Faster examples with accelerated inference

Switch between documentation themes

Sign Up

to get started

Copy page

Summary

In this chapter, you’ve been introduced to the fundamentals of Transformer models, Large Language Models (LLMs), and how they’re revolutionizing AI and beyond.

Key concepts covered

Natural Language Processing and LLMs

We explored what NLP is and how Large Language Models have transformed the field. You learned that:

NLP encompasses a wide range of tasks from classification to generation

LLMs are powerful models trained on massive amounts of text data

These models can perform multiple tasks within a single architecture

Despite their capabilities, LLMs have limitations including hallucinations and bias

Transformer capabilities

You saw how the

pipeline()

function from ■ Transformers makes it easy to use pre-trained models for various tasks:

Text classification, token classification, and question answering

Text generation and summarization

Translation and other sequence-to-sequence tasks

Speech recognition and image classification

Transformer architecture

We discussed how Transformer models work at a high level, including:

The importance of the attention mechanism

How transfer learning enables models to adapt to specific tasks

The three main architectural variants: encoder-only, decoder-only, and encoder-decoder

Model architectures and their applications

A key aspect of this chapter was understanding which architecture to use for different tasks:

Model

Examples

Tasks

Encoder-only

BERT, DistilBERT, ModernBERT

Sentence classification, named entity recognition, extractive question answering

Decoder-only

GPT, LLaMA, Gemma, SmolLM

Text generation, conversational AI, creative writing

Encoder-decoder

BART, T5, Marian, mBART

Summarization, translation, generative question answering

Modern LLM developments

You also learned about recent developments in the field:

How LLMs have grown in size and capability over time

The concept of scaling laws and how they guide model development

Specialized attention mechanisms that help models process longer sequences

The two-phase training approach of pretraining and instruction tuning

Practical applications

Throughout the chapter, you’ve seen how these models can be applied to real-world problems:

Using the Hugging Face Hub to find and use pre-trained models

Leveraging the Inference API to test models directly in your browser

Understanding which models are best suited for specific tasks

Looking ahead

Now that you have a solid understanding of what Transformer models are and how they work at a high level, you’re ready to dive deeper into how to use them effectively. In the next chapters, you’ll learn how to:

Use the Transformers library to load and fine-tune models

Process different types of data for model input

Adapt pre-trained models to your specific tasks

Deploy models for practical applications

The foundation you’ve built in this chapter will serve you well as you explore more advanced topics and techniques in the coming sections.

Update

on GitHub

←

Bias and limitations

Certification exam

→