datacamp

COURSE GLOSSARY

Introduction to LLMs in Python

Auto classes: Convenience constructors (e.g., AutoModel, AutoTokenizer) in the transformers library that automatically load the appropriate model and tokenizer implementations from a model identifier

Bias: Systematic and unfair patterns a model learns or amplifies from training data that lead to prejudiced, stereotyped, or unequal outputs across individuals or groups

BLEU and ROUGE: Automatic text-similarity metrics where BLEU measures n-gram precision often used for translation, and ROUGE measures n-gram recall and longest-common-subsequence overlap commonly used for summarization

Decoder-only architecture: A model design optimized for autoregressive generation that predicts the next token from prior context, making it well suited for text generation tasks, exemplified by GPT

Encoder-decoder architecture: Also called sequence-to-sequence, this architecture encodes input sequences and then decodes target sequences, which is ideal for tasks like translation and summarization (e.g., T5, BART)

Encoder-only architecture: A model design that focuses on encoding input text into contextual representations for understanding tasks like classification and extractive QA, exemplified by BERT

Evaluate library: A toolkit that provides standardized implementations of common NLP evaluation metrics and utilities for computing task-specific measures like accuracy, BLEU, ROUGE, and perplexity

Hallucination: The tendency of a model to generate false, fabricated, or nonsensical information presented as factual output, often due to gaps or biases in training data

Hugging Face Hub: An online repository and platform for sharing, discovering, and downloading pre-trained models, datasets, and model cards used in NLP and other ML tasks

Large Language Model (LLM): A deep learning model trained on massive text corpora to understand and generate human-like language, typically containing millions or billions of parameters and capable of tasks like summarization, translation, and question answering

N-shot learning: A paradigm describing how many examples a model sees for a new task—zero-shot uses none, one-shot uses a single example, and few-shot uses a small number of examples to guide generalization

Padding (pad_token_id): Padding adds special tokens to shorter sequences so batches share the same length for efficient processing, and pad_token_id is the numeric id used to represent those padding tokens

Perplexity: A metric that quantifies how well a language model predicts a set of tokens by computing the inverse geometric mean of predicted token probabilities, where lower values indicate better predictive confidence

Pipeline: A high-level API in the transformers library that wires together a model and tokenizer to perform common tasks (e.g., text-generation, summarization, translation) with minimal code

Pre-trained model: A model whose weights have been previously trained on large general-purpose datasets to learn language patterns and that can be reused or adapted for downstream tasks

Prompt engineering: The practice of designing and refining input prompts to steer LLM behavior and obtain more accurate, relevant, or controlled outputs from a model

Subword tokenization: A tokenization approach that splits words into smaller subword units (e.g., prefixes, suffixes) to handle rare words and reduce vocabulary size while preserving meaningful pieces

Tokenizer and token ids: A tokenizer converts raw text into discrete tokens and maps them to numeric token ids that models consume, handling vocabulary lookup, special tokens, and text normalization

Trainer: A high-level training loop in the transformers library that orchestrates optimization, evaluation, checkpointing, and the interaction between model, tokenizer, datasets, and TrainingArguments

TrainingArguments: A configuration object in transformers that specifies training hyperparameters and behaviors such as output directory, learning rate, batch sizes, number of epochs, and evaluation strategy

Transfer learning: Reusing knowledge learned by a model on one task or domain to improve performance on a related task or domain, often achieved via fine-tuning

Transformer: A neural network architecture that uses self-attention to process entire sequences in parallel, enabling efficient modeling of long-range dependencies in text

Truncation: The process of shortening input sequences that exceed a specified maximum length by removing tokens (commonly from the end) so they fit model input constraints