Index

A

- accuracy
  - confusion matrices, Using a Task-Specific Model
  - output verification, Output Verification
- adaptive pretraining, Using TSDAE for Domain Adaptation
- agents, Agents: Creating a System of LLMs-ReAct in LangChain
  - agentic RAG, Agentic RAG
  - ReAct in LangChain, ReAct in LangChain-ReAct in LangChain
  - step-by-step reasoning, The Driving Power Behind Agents: Step-by-step Reasoning-The Driving Power Behind Agents: Step-by-step Reasoning
- AI (artificial intelligence)
  - accelerated development of, An Introduction to Large Language Models
  - defined, What Is Language AI?
- ALBERT, Model Selection
- align_labels function, Preparing Data for Named-Entity Recognition
- all-MiniLM-L6-v2 model, Supervised
- all-mpnet-base-v2 model, Fine-Tuning for Few-Shot Classification
- Annoy, Nearest neighbor search versus vector databases
- Anthropic Claude, Proprietary, Private Models
- APIs (application programming interfaces), Proprietary, Private Models
  - Cohere, API Keys, Dense retrieval example
  - external, ChatGPT for Classification
  - generating embeddings, Supervised Classification
  - OpenAI, API Keys, ChatGPT for Classification
- artificial intelligence (see AI)
- ArXiv, ArXiv's Articles: Computation and Language

attention, Encoding and Decoding Context with Attention-Attention Is All You Need
overview of, Encoding and Decoding Context with Attention-Encoding and Decoding Context with Attention
Transformer architecture, Attention Is All You Need-Attention Is All You Need
attention calculation, How attention is calculated-How attention is calculated
attention layer, The Components of the Forward Pass, Inside the Transformer Block, The attention layer at a glance, Summary
Flash Attention, Flash Attention
grouped-query attention, Multi-query and grouped-query attention-Optimizing attention: From multi-head to multi-query to grouped query
local attention, Local/sparse attention
multi-query attention, Multi-query and grouped-query attention-Optimizing attention: From multi-head to multi-query to grouped query
optimizing attention, Optimizing attention: From multi-head to multi-query to grouped query
self-attention and relevance scoring, Self-attention: Relevance scoring-Self-attention: Combining information
sparse attention, Local/sparse attention
attention heads, How attention is calculated-How attention is calculated, Summary
audience, in text-generation prompts, The Potential Complexity of a Prompt
Augmented SBERT, Augmented SBERT-Augmented SBERT
autoregressive architecture, Encoding and Decoding Context with Attention, Attention Is All You Need, Generative Models: Decoder-Only Models, The Inputs and Outputs of a Trained Transformer LLM, Local/sparse attention

B

bag-of-words model, Representing Language as a Bag-of-Words-Representing Language as a Bag-of-Words
embeddings, Types of Embeddings

topic modeling, From Text Clustering to Topic Modeling-BERTopic: A Modular Topic Modeling Framework, Adding a Special Lego Block

benchmarks, in generative model evaluation, Benchmarks

BERT (Bidirectional Encoder Representations from Transformers), Representation Models: Encoder-Only Models-Generative Models: Decoder-Only Models

adoption by search engines, Semantic Search and Retrieval-Augmented Generation

BERT-like models, Model Selection

comparing to other trained tokenizers, BERT base model (uncased) (2018)

fine-tuning pretrained BERT models, Fine-Tuning a Pretrained BERT Model-Fine-Tuning a Pretrained BERT Model

masked language modeling, Supervised

Transformer blocks versus, Local/sparse attention

BERTopic, BERTopic: A Modular Topic Modeling Framework-BERTopic: A Modular Topic Modeling Framework

algorithmic variants, BERTopic: A Modular Topic Modeling Framework

modularity of, BERTopic: A Modular Topic Modeling Framework

representation blocks, Adding a Special Lego Block-The Text Generation Lego Block

KeyBERTInspired, KeyBERTInspired

maximal marginal relevance, Maximal marginal relevance

text generation, The Text Generation Lego Block-The Text Generation Lego Block

BERTScore, Word-Level Metrics

bias and fairness, Responsible LLM Development and Usage

Bidirectional Encoder Representations from Transformers (see BERT)

bitsandbytes package, Model Quantization

BLEU, Word-Level Metrics

BLIP-2 (Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation 2)

chat-based prompting, Use Case 2: Multimodal Chat-Based Prompting-Use Case 2: Multimodal Chat-Based Prompting

image captioning, Use Case 1: Image Captioning

preprocessing text, Preprocessing text
Q-Former, BLIP-2: Bridging the Modality Gap-BLIP-2: Bridging the Modality Gap
BM25 algorithm, Search the index
BPE (byte pair encoding), How Does the Tokenizer Break Down Text?, Tokenization methods
byte tokens, Word Versus Subword Versus Character Versus Byte Tokens

C

c-TF-IDF, BERTopic: A Modular Topic Modeling Framework, KeyBERTInspired, The Text Generation Lego Block
capitalization, Tokenizer parameters
captioning, Use Case 1: Image Captioning-Use Case 1: Image Captioning
centroid-based algorithms, Cluster the Reduced Embeddings
chains, Chain Prompting: Breaking up the Problem-Chain Prompting: Breaking up the Problem, Chains: Extending the Capabilities of LLMs-A Chain with Multiple Prompts
chain-of-thought, Chain-of-Thought: Think Before Answering-Chain-of-Thought: Think Before Answering
chaining single prompt, A Single Link in the Chain: Prompt Template-A Single Link in the Chain: Prompt Template
sequential chaining of multiple prompts, A Chain with Multiple Prompts-A Chain with Multiple Prompts
character tokens, Word Versus Subword Versus Character Versus Byte Tokens
chat tokens, Phi-3 (and Llama 2)
chat-based prompting, Use Case 2: Multimodal Chat-Based Prompting-Use Case 2: Multimodal Chat-Based Prompting
Chatbot Arena, Human Evaluation
ChatGPT, Model I/O: Loading Quantized Models with LangChain, Making Text Generation Models Multimodal, Reward model training step, Templating Alignment Data
release of, An Introduction to Large Language Models
text classification, ChatGPT for Classification-ChatGPT for Classification
chatgpt_generation function, ChatGPT for Classification

chat_history input variable, Conversation Buffer
classification reports, Using a Task-Specific Model-Using a Task-Specific Model
classification step, embedding model, Supervised Classification
CLIP, CLIP: Connecting Text and Images-How Can CLIP Generate Multimodal Embeddings?
  connecting text and images, CLIP: Connecting Text and Images
  generating multimodal embeddings, How Can CLIP Generate Multimodal Embeddings?-How Can CLIP Generate Multimodal Embeddings?
  OpenCLIP, OpenCLIP-OpenCLIP
closed-source LLMs, Proprietary, Private Models
[CLS] token, Representation Models: Encoder-Only Models, BERT base model (uncased) (2018), Creating Contextualized Word Embeddings with Language Models, OpenCLIP, SBERT, Transformer-Based Sequential Denoising Auto-Encoder, Preparing Data for Named-Entity Recognition
cluster model, Cluster the Reduced Embeddings-Cluster the Reduced Embeddings
clustering (see text clustering)
CNNs (convolutional neural networks), Transformers for Vision
Cohere, Example: Grounded Generation with an LLM API
  Command R+, Open Models, Agentic RAG
  creating accounts, API Keys, Dense retrieval example
  generating embeddings, Supervised Classification
  query rewriting, Query rewriting
  Rerank endpoint, Reranking example
completion models, Generative Models: Decoder-Only Models
compute_metrics function, Preparing Data for Named-Entity Recognition
confusion matrices, Using a Task-Specific Model
CoNLL-2003 dataset, Preparing Data for Named-Entity Recognition
constrained sampling, Grammar: Constrained Sampling-Grammar: Constrained Sampling
context
  attention and, The attention layer at a glance
  prompt engineering, The Potential Complexity of a Prompt

training datasets, The Word2vec Algorithm and Contrastive Training

context length

completion models, Generative Models: Decoder-Only Models

token processing limits, Parallel Token Processing and Context Size

context window, in completion models, Generative Models: Decoder-Only Models

contrastive learning

text embedding models, What Is Contrastive Learning?-What Is Contrastive Learning?, Generating Contrastive Examples

word2vec algorithm and, The Word2vec Algorithm and Contrastive Training-The Word2vec Algorithm and Contrastive Training

Contrastive Tension (CT), Unsupervised Learning

conversation buffer memory, Conversation Summary

overview of, Conversation Buffer-Conversation Buffer

windowed, Windowed Conversation Buffer-Windowed Conversation Buffer

conversation summary memory, Conversation Summary-Conversation Summary

convert_ids_to_tokens function, OpenCLIP

convolutional neural networks (CNNs), Transformers for Vision

cosine similarity, What If We Do Not Have Labeled Data?, Cosine similarity-Cosine similarity

CountVectorizer, BERTopic: A Modular Topic Modeling Framework

cross-encoders, How reranking models work

cross-entropy loss, Multiple negatives ranking loss

CT (Contrastive Tension), Unsupervised Learning

D

data outliers, Cluster the Reduced Embeddings

DataCollator class, Fine-Tuning a Pretrained BERT Model, Continued Pretraining with Masked Language Modeling

datamapplot package, The Text Generation Lego Block

DBSCAN (Density-Based Spatial Clustering), Cluster the Reduced Embeddings

DeBERTa, Creating Contextualized Word Embeddings with Language Models, Model Selection decoder-only models (see generative models)

decoding strategy, Choosing a Single Token from the Probability Distribution (Sampling/Decoding)-Choosing a Single Token from the Probability Distribution (Sampling/Decoding), Summary

dense retrieval, Overview of Semantic Search and RAG, Dense Retrieval-Fine-tuning embedding models for dense retrieval

- caveats of, Caveats of dense retrieval
- example of, Dense retrieval example-Search the index
- fine-tuning embedding models for, Fine-tuning embedding models for dense retrieval
- nearest neighbor search versus vector databases, Nearest neighbor search versus vector databases
- text chunking, Chunking long texts-Multiple vectors per document

density-based algorithms, Cluster the Reduced Embeddings

Density-Based Spatial Clustering (DBSCAN), Cluster the Reduced Embeddings

dimensionality reduction model, Reducing the Dimensionality of Embeddings-Reducing the Dimensionality of Embeddings

DistilBERT, Model Selection, Using a Task-Specific Model

domain adaptation, Using TSDAE for Domain Adaptation

do_sample parameter, Generating Your First Text

DPO (Direct Preference Optimization), Training No Reward Model-Training

- fine-tuning, Training
- model quantization, Model Quantization
- templating alignment data, Templating Alignment Data
- training configuration, Training Configuration

DSPy, Advanced Text Generation Techniques and Tools

E

easy negatives, Multiple negatives ranking loss

Elo rating system, Human Evaluation

embeddings, Better Representations with Dense Vector Embeddings-Types of Embeddings, Tokens and Embeddings, Token Embeddings-Training a Song Embedding Model

- dense retrieval, Overview of Semantic Search and RAG, Dense Retrieval-Fine-tuning embedding models for dense retrieval

- embedding models, defined, Text Classification with Representation Models, Embedding Models
- multimodality, Multimodal Embedding Models-OpenCLIP
  - CLIP, CLIP: Connecting Text and Images-How Can CLIP Generate Multimodal Embeddings?
  - OpenCLIP, OpenCLIP-OpenCLIP
- overview of, Better Representations with Dense Vector Embeddings-Better Representations with Dense Vector Embeddings, Embedding Models-Embedding Models
- positional embeddings, Positional Embeddings (RoPE)-Positional Embeddings (RoPE)
- recommendation systems, Embeddings for Recommendation Systems-Training a Song Embedding Model
- text classification tasks that leverage, Classification Tasks That Leverage Embeddings-What If We Do Not Have Labeled Data?
  - supervised classification, Supervised Classification-Supervised Classification
  - zero-shot classification, What If We Do Not Have Labeled Data?-What If We Do Not Have Labeled Data?
- text clustering pipeline, A Common Pipeline for Text Clustering-Inspecting the Clusters
  - cluster model, Cluster the Reduced Embeddings-Cluster the Reduced Embeddings
  - dimensionality reduction model, Reducing the Dimensionality of Embeddings-Reducing the Dimensionality of Embeddings
  - embedding model, Embedding Documents
  - inspecting clusters, Inspecting the Clusters-Inspecting the Clusters
- text embedding models, Text Embeddings (for Sentences and Whole Documents)-Text Embeddings (for Sentences and Whole Documents), Creating Text Embedding Models-Summary
  - contrastive learning, What Is Contrastive Learning?-What Is Contrastive Learning?
  - creating, Creating an Embedding Model-Multiple negatives ranking loss
  - fine-tuning, Fine-Tuning an Embedding Model-Augmented SBERT

SBERT, SBERT-SBERT

unsupervised learning, Unsupervised Learning

token embeddings, Token Embeddings-Creating Contextualized Word Embeddings with Language Models

creating contextualized word embeddings, Creating Contextualized Word Embeddings with Language Models-Creating Contextualized Word Embeddings with Language Models

tokenizer's vocabulary and, A Language Model Holds Embeddings for the Vocabulary of Its Tokenizer

types of, Types of Embeddings

word embeddings, Word Embeddings Beyond LLMs-The Word2vec Algorithm and Contrastive Training

pretrained, Using pretrained Word Embeddings

word2vec algorithm and contrastive training, The Word2vec Algorithm and Contrastive Training-The Word2vec Algorithm and Contrastive Training

encoder-decoder models, Using the Text-to-Text Transfer Transformer

encoder-only models (see representation models)

ethics, validating output, Output Verification

exponential backoff, ChatGPT for Classification

F

F1 score, confusion matrices, Using a Task-Specific Model

FAISS, Nearest neighbor search versus vector databases

Falcon, Choosing a Text Generation Model

feature extraction step, embedding model, Supervised Classification

feedforward layer, Inside the Transformer Block

feedforward neural networks, Attention Is All You Need, The feedforward neural network at a glance, Summary, Fine-Tuning a Pretrained BERT Model

few-shot classification, Few-Shot Classification-Fine-Tuning for Few-Shot Classification

fine-tuning for classification, Fine-Tuning for Few-Shot Classification-Fine-Tuning for Few-Shot Classification

SetFit, SetFit: Efficient Fine-Tuning with Few Training Examples-SetFit: Efficient Fine-Tuning with Few Training Examples

- few-shot prompting, In-Context Learning: Providing Examples-In-Context Learning: Providing Examples, Providing Examples-Providing Examples
- find_topics() function, BERTopic, BERTopic: A Modular Topic Modeling Framework
- fine-tuning
  - embedding models for dense retrieval, Fine-tuning embedding models for dense retrieval
  - generative models, Output Verification, Fine-Tuning Generation Models-Summary
    - evaluating, Evaluating Generative Models-Human Evaluation
    - preference tuning, The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning, Preference-Tuning / Alignment / RLHF-Training
    - supervised fine-tuning, The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning-Merge Weights
    - training steps, The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning-The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning
  - overview of, The Training Paradigm of Large Language Models
  - representation models, Fine-Tuning Representation Models for Classification-Summary
    - few-shot classification, Few-Shot Classification-Fine-Tuning for Few-Shot Classification
    - masked language modeling, Continued Pretraining with Masked Language Modeling-Continued Pretraining with Masked Language Modeling
    - named-entity recognition, Named-Entity Recognition-Fine-Tuning for Named-Entity Recognition
    - supervised classification, Supervised Classification-Freezing Layers
  - T5 model, Using the Text-to-Text Transfer Transformer
  - text embedding models, Fine-Tuning an Embedding Model-Augmented SBERT
    - Augmented SBERT, Augmented SBERT-Augmented SBERT

supervised, Supervised-Supervised

Flan-T5 model, Flan-T5 (2022), Using the Text-to-Text Transfer Transformer-Using the Text-to-Text Transfer Transformer

Flash Attention, Flash Attention, Summary

forward pass, Summary

components of, The Components of the Forward Pass-The Components of the Forward Pass

defined, The Inputs and Outputs of a Trained Transformer LLM

foundation models, The Year of Generative AI

fp16 parameter, Train Model

freezing layers, Train Model, Freezing Layers-Freezing Layers

frozen (nontrainable) models, Text Classification with Representation Models, Supervised Classification, Supervised Classification

G

Ġ symbol, Preprocessing text

Galactica, Galactica

General Language Understanding Evaluation (GLUE) benchmark, Generating Contrastive Examples, In-Depth Evaluation, Adapters, Benchmarks

generated_text variable, Use Case 1: Image Captioning, Use Case 1: Image Captioning

generation_output variable, Downloading and Running an LLM

generative models, Generative Models: Decoder-Only Models-Generative Models: Decoder-Only Models

evaluating, Evaluating Generative Models-Human Evaluation

automated evaluation, Automated Evaluation

benchmarks, Benchmarks

human evaluation, Human Evaluation

leaderboards, Leaderboards

word-level metrics, Word-Level Metrics

fine-tuning, Fine-Tuning Generation Models-Summary

evaluation, Evaluating Generative Models-Human Evaluation

preference tuning, The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning, Preference-Tuning / Alignment / RLHF-Training

    supervised fine-tuning, The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning-Merge Weights
    training steps, The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning-The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning
  reasoning, Reasoning with Generative Models-Tree-of-Thought: Exploring Intermediate Steps
    self-consistency, Self-Consistency: Sampling Outputs
    tree-of-thought, Tree-of-Thought: Exploring Intermediate Steps-Tree-of-Thought: Exploring Intermediate Steps
  representation models versus, Representation Models: Encoder-Only Models
  text classification, Text Classification with Generative Models-ChatGPT for Classification
    ChatGPT, ChatGPT for Classification-ChatGPT for Classification
    T5, Using the Text-to-Text Transfer Transformer-Using the Text-to-Text Transfer Transformer
generative pre-trained transformers (see GPTs)
Generative Pseudo-Labeling (GPL), Unsupervised Learning
Gensim library, Using pretrained Word Embeddings
get_topic function, BERTopic: A Modular Topic Modeling Framework
get_topic_info() method, BERTopic: A Modular Topic Modeling Framework
GGUF model, Grammar: Constrained Sampling, Model I/O: Loading Quantized Models with LangChain
GitHub, Using Code Examples
GloVe, SBERT
GLUE (General Language Understanding Evaluation) benchmark, Generating Contrastive Examples, In-Depth Evaluation, Adapters, Benchmarks
gold datasets, Augmented SBERT
Goodhart's Law, Human Evaluation
Google Colab, Prerequisites, Hardware and Software Requirements, Limited Resources Are All You Need, Generating Your First Text, Speeding Up Generation by Caching Keys and Values, Training

Google Gemini, Retrieval-Augmented Generation (RAG)

Google Search, Semantic Search and Retrieval-Augmented Generation

GPL (Generative Pseudo-Labeling), Unsupervised Learning

GPT2Tokenizer, Preprocessing text

GPTs (generative pre-trained transformers), Generative Models: Decoder-Only Models, Prompt Engineering

- (see also text generation)
- GPT-1, Generative Models: Decoder-Only Models
- GPT-2, An Introduction to Large Language Models, Generative Models: Decoder-Only Models, Word Versus Subword Versus Character Versus Byte Tokens, GPT-2 (2019)
- GPT-3, Generative Models: Decoder-Only Models, The feedforward neural network at a glance, Local/sparse attention, Low-Rank Adaptation (LoRA)
- GPT-3.5, The Year of Generative AI, ChatGPT for Classification, A Single Link in the Chain: Prompt Template, ReAct in LangChain
- GPT-4, The Year of Generative AI, Proprietary, Private Models, How Tokenizers Prepare the Inputs to the Language Model, GPT-4 (2023), The feedforward neural network at a glance

GPUs

- Flash Attention, Flash Attention
- requirements, Hardware and Software Requirements, Limited Resources Are All You Need
- SRAM and HBM, Flash Attention

grammar, Output Verification, Grammar: Constrained Sampling-Grammar: Constrained Sampling

greedy decoding, Choosing a Single Token from the Probability Distribution (Sampling/Decoding)

grounded generation, From Search to RAG-Example: Grounded Generation with an LLM API

grouped-query attention, Multi-query and grouped-query attention-Optimizing attention: From multi-head to multi-query to grouped query, Summary

GSM8k, Benchmarks

Guardrails, Grammar: Constrained Sampling

Guidance, Grammar: Constrained Sampling

H

hallucination
avoiding in in instruction-based prompting, Instruction-Based Prompting
text generation models, Semantic Search and Retrieval-Augmented Generation
hard negatives, Multiple negatives ranking loss
harmful content, generating, Responsible LLM Development and Usage
Haystack, Advanced Text Generation Techniques and Tools
HBM (high bandwidth memory), Flash Attention
HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), Cluster the Reduced Embeddings, BERTopic: A Modular Topic Modeling Framework
HellaSwag, Benchmarks
high bandwidth memory (HBM), Flash Attention
Hugging Face, Generating Your First Text, The Sentiment of Movie Reviews
creating accounts, API Keys
evaluate package, Preparing Data for Named-Entity Recognition
tokenizers, The domain of the data
human evaluation, Inspecting the Clusters, Human Evaluation
HumanEval, Benchmarks
hybrid search, Caveats of dense retrieval, Reranking example

I

Idefics 2, BLIP-2: Bridging the Modality Gap
images (see multimodality)
in-context learning, In-Context Learning: Providing Examples-In-Context Learning: Providing Examples
indexes, Building the search index
InfoNCE, Multiple negatives ranking loss
input_ids variable, Downloading and Running an LLM
instruction-based prompting, Instruction-Based Prompting-Instruction-Based Prompting

intellectual property, Responsible LLM Development and Usage
intuition-first philosophy, An Intuition-First Philosophy
invoke function, A Single Link in the Chain: Prompt Template

J

Jupyter, Speeding Up Generation by Caching Keys and Values

K

k-means algorithm, Cluster the Reduced Embeddings, BERTopic: A Modular Topic Modeling Framework, BERTopic: A Modular Topic Modeling Framework
KeyBERTInspired, KeyBERTInspired, Maximal marginal relevance
keyword search
  reranking, Reranking example-Reranking example
  verifying semantic search with, Search the index
kv (keys and values) cache, Speeding Up Generation by Caching Keys and Values-Speeding Up Generation by Caching Keys and Values

L

LangChain, Advanced Text Generation Techniques and Tools
  (see also chains)
  loading quantized models with, Model I/O: Loading Quantized Models with LangChain-Model I/O: Loading Quantized Models with LangChain
  ReAct in, ReAct in LangChain-ReAct in LangChain
Language AI (Language Artificial Intelligence), An Introduction to Large Language Models-The Year of Generative AI
  defining, What Is Language AI?
  recent history of, A Recent History of Language AI-The Year of Generative AI
    attention, Encoding and Decoding Context with Attention-Attention Is All You Need
    bag-of-words model, Representing Language as a Bag-of-Words-Representing Language as a Bag-of-Words
    embeddings, Better Representations with Dense Vector Embeddings-Types of Embeddings

generative models, Generative Models: Decoder-Only Models-Generative Models: Decoder-Only Models

representation models, Representation Models: Encoder-Only Models-Representation Models: Encoder-Only Models

Year of Generative AI, The Year of Generative AI-The Year of Generative AI

language modeling, The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning

language modeling head (LM head), The Components of the Forward Pass-The Components of the Forward Pass

large language models (see LLMs)

latent Dirichlet allocation, From Text Clustering to Topic Modeling

LayerNorm, The Transformer Block

leaderboards, in generative model evaluation, Leaderboards

learning_rate parameter, Training Configuration

Llama, Choosing a Text Generation Model

Llama 2, API Keys, Limited Resources Are All You Need, Phi-3 (and Llama 2), Multi-query and grouped-query attention, Choosing a Text Generation Model, Making Text Generation Models Multimodal

llama-cpp-python library, Grammar: Constrained Sampling

LLaVA, BLIP-2: Bridging the Modality Gap

LLM-as-a-judge, RAG Evaluation

LLMs (large language models), An Introduction to Large Language Models-Summary

code examples and exercises, Using Code Examples

embeddings, Tokens and Embeddings, Token Embeddings-Training a Song Embedding Model

recommendation systems, Embeddings for Recommendation Systems-Training a Song Embedding Model

text embeddings, Text Embeddings (for Sentences and Whole Documents)-Text Embeddings (for Sentences and Whole Documents)

token embeddings, Token Embeddings-Creating Contextualized Word Embeddings with Language Models

word embeddings, Word Embeddings Beyond LLMs-The Word2vec Algorithm and Contrastive Training fine-tuning generative models, Fine-Tuning Generation Models-Summary

evaluation, Evaluating Generative Models-Human Evaluation

preference tuning, The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning, Preference-Tuning / Alignment / RLHF-Training

supervised fine-tuning, The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning-Merge Weights

training steps, The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning-The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning

fine-tuning representation models, Fine-Tuning Representation Models for Classification-Summary

few-shot classification, Few-Shot Classification-Fine-Tuning for Few-Shot Classification

masked language modeling, Continued Pretraining with Masked Language Modeling-Continued Pretraining with Masked Language Modeling

named-entity recognition, Named-Entity Recognition-Fine-Tuning for Named-Entity Recognition

supervised classification, Supervised Classification-Freezing Layers

generating text, Generating Your First Text-Generating Your First Text

generative models, Generative Models: Decoder-Only Models-Generative Models: Decoder-Only Models

hardware and software requirements, Hardware and Software Requirements, Limited Resources Are All You Need

high-level view, How Tokenizers Prepare the Inputs to the Language Model

history of Language AI, A Recent History of Language AI-The Year of Generative AI

interfacing with, Interfacing with Large Language Models-Open Source Frameworks

closed-source models, Proprietary, Private Models open models, Open Models-Open Source Frameworks

intuition-first philosophy, An Intuition-First Philosophy

moving definition of, The Moving Definition of a “Large Language Model”

multimodality, Multimodal Large Language Models-Summary

  embedding models, Multimodal Embedding Models-OpenCLIP

  text generation models, Making Text Generation Models Multimodal-Use Case 2: Multimodal Chat-Based Prompting

  Vision Transformer, Transformers for Vision-Transformers for Vision

prompt engineering, Prompt Engineering-Summary

  chain prompting, Chain Prompting: Breaking up the Problem-Chain Prompting: Breaking up the Problem

  in-context learning, In-Context Learning: Providing Examples-In-Context Learning: Providing Examples

  instruction-based prompting, Instruction-Based Prompting-Instruction-Based Prompting

  output verification, Output Verification-Grammar: Constrained Sampling

  potential complexity of prompts, The Potential Complexity of a Prompt-The Potential Complexity of a Prompt

  prompt components, The Basic Ingredients of a Prompt-The Basic Ingredients of a Prompt

  reasoning with generative models, Reasoning with Generative Models-Tree-of-Thought: Exploring Intermediate Steps

  text generation models, Using Text Generation Models-top_p

representation models, Representation Models: Encoder-Only Models-Representation Models: Encoder-Only Models

responsible development and usage of, Responsible LLM Development and Usage

retrieval-augmented generation, Overview of Semantic Search and RAG, Retrieval-Augmented Generation (RAG)-RAG Evaluation

  agentic RAG, Agentic RAG

  converting search system to, From Search to RAG

  evaluating results, RAG Evaluation

  grounded generation, Example: Grounded Generation with an LLM API

multi-hop RAG, Multi-hop RAG
multi-query RAG, Multi-query RAG
query rewriting, Query rewriting
query routing, Query routing
with local models, Example: RAG with Local Models, The RAG prompt
semantic search, Semantic Search and Retrieval-Augmented Generation-Scoring across multiple queries with mean average precision
dense retrieval, Overview of Semantic Search and RAG, Dense Retrieval-Fine-tuning embedding models for dense retrieval
reranking, Overview of Semantic Search and RAG, Reranking-How reranking models work
retrieval evaluation metrics, Retrieval Evaluation Metrics-Scoring across multiple queries with mean average precision
text classification, Text Classification-Summary
with generative models, Text Classification with Generative Models-ChatGPT for Classification
movie reviews, The Sentiment of Movie Reviews, The Sentiment of Movie Reviews
with representation models, Text Classification with Representation Models-What If We Do Not Have Labeled Data?
text clustering, Text Clustering and Topic Modeling-Inspecting the Clusters
text embedding models, Creating Text Embedding Models-Summary
contrastive learning, What Is Contrastive Learning?-What Is Contrastive Learning?
creating, Creating an Embedding Model-Multiple negatives ranking loss
fine-tuning, Fine-Tuning an Embedding Model-Augmented SBERT
SBERT, SBERT-SBERT
unsupervised learning, Unsupervised Learning
text generation, Advanced Text Generation Techniques and Tools-Summary
agents, Agents: Creating a System of LLMs-ReAct in LangChain

- chains, Chains: Extending the Capabilities of LLMs-A Chain with Multiple Prompts
- memory of conversations, Memory: Helping LLMs to Remember Conversations-Conversation Summary
- model I/O, Model I/O: Loading Quantized Models with LangChain-Model I/O: Loading Quantized Models with LangChain

tokens and tokenizers, Tokens and Embeddings-Creating Contextualized Word Embeddings with Language Models

- comparing trained tokenizers, Comparing Trained LLM Tokenizers-Phi-3 (and Llama 2)
- downloading and running LLMs, Downloading and Running an LLM-Downloading and Running an LLM
- input preparation, How Tokenizers Prepare the Inputs to the Language Model
- text breakdown, How Does the Tokenizer Break Down Text?
- token embeddings, Token Embeddings-Creating Contextualized Word Embeddings with Language Models
- tokenization schemes, Word Versus Subword Versus Character Versus Byte Tokens-Word Versus Subword Versus Character Versus Byte Tokens
- tokenizer properties, Tokenizer Properties-The domain of the data

topic modeling, Text Clustering and Topic Modeling, From Text Clustering to Topic Modeling-The Text Generation Lego Block

training paradigm of, The Training Paradigm of Large Language Models-The Training Paradigm of Large Language Models

Transformer architecture, Looking Inside Large Language Models-Summary

- decoding strategy, Choosing a Single Token from the Probability Distribution (Sampling/Decoding)-Choosing a Single Token from the Probability Distribution (Sampling/Decoding)
- forward pass components, The Components of the Forward Pass-The Components of the Forward Pass
- inputs and outputs of, The Inputs and Outputs of a Trained Transformer LLM-The Inputs and Outputs of a Trained Transformer LLM

keys and values cache, Speeding Up Generation by Caching Keys and Values-Speeding Up Generation by Caching Keys and Values

parallel token processing and context size, Parallel Token Processing and Context Size-Parallel Token Processing and Context Size

recent improvements to, Recent Improvements to the Transformer Architecture-Other Architectural Experiments and Improvements

Transformer blocks, Inside the Transformer Block-Self-attention: Combining information

utility of, Large Language Model Applications: What Makes Them So Useful?

LM head (language modeling head), The Components of the Forward Pass-The Components of the Forward Pass

LMQL, Grammar: Constrained Sampling

local attention, Local/sparse attention

LoRA (low-rank adaptation), Low-Rank Adaptation (LoRA)-Low-Rank Adaptation (LoRA), Model Quantization

(see also QLoRA)

lora_alpha parameter, LoRA Configuration

loss functions, Loss Functions-Multiple negatives ranking loss

cosine similarity loss, Cosine similarity-Cosine similarity

multiple negatives ranking loss, Multiple negatives ranking loss-Multiple negatives ranking loss

lr_scheduler_type parameter, Training Configuration

M

Mamba, The Year of Generative AI

MAP (mean average precision), Retrieval Evaluation Metrics-Scoring across multiple queries with mean average precision

MarginMSE loss, Loss Functions

masked language modeling (MLM), Continued Pretraining with Masked Language Modeling-Continued Pretraining with Masked Language Modeling

mask_token [MASK], BERT base model (uncased) (2018)

Massive Text Embedding Benchmark (MTEB), Model Selection, Embedding Documents, Loading the embedding model, In-Depth

Evaluation

matplotlib library, Inspecting the Clusters

maximal marginal relevance (MMR), Maximal marginal relevance

max_new_tokens parameter, Generating Your First Text

McCarthy, John, What Is Language AI?

mean average precision (MAP)

memory of conversations, Memory: Helping LLMs to Remember Conversations-Conversation Summary

- conversation buffer, Conversation Buffer-Conversation Buffer
- conversation summary, Conversation Summary-Conversation Summary
- windowed conversation buffer, Windowed Conversation Buffer-Windowed Conversation Buffer

Meta Llama model, Open Models

Microsoft Bing, Semantic Search and Retrieval-Augmented Generation, Reranking

Microsoft Bing AI, Retrieval-Augmented Generation (RAG)

Microsoft Phi model, Open Models

microsoft/mpnet-base model, Train Model

min_cluster_size parameter, Cluster the Reduced Embeddings

min_dist parameter, Reducing the Dimensionality of Embeddings

MIRACL, Reranking example

Mistral, Open Models, Choosing a Text Generation Model, BLIP-2: Bridging the Modality Gap

MLM (masked language modeling), Continued Pretraining with Masked Language Modeling-Continued Pretraining with Masked Language Modeling

MMLU, Benchmarks, Leaderboards

MMR (maximal marginal relevance), Maximal marginal relevance

MNLI (Multi-Genre Natural Language Inference) corpus, Generating Contrastive Examples, Multiple negatives ranking loss

MNR (multiple negatives ranking) loss, Multiple negatives ranking loss-Multiple negatives ranking loss

model I/O, Model I/O: Loading Quantized Models with LangChain-Model I/O: Loading Quantized Models with LangChain

monoBERT, How reranking models work

MTEB (Massive Text Embedding Benchmark), Model Selection, Embedding Documents, Loading the embedding model, In-Depth Evaluation

Multi-Genre Natural Language Inference (MNLI) corpus, Generating Contrastive Examples, Multiple negatives ranking loss

multi-hop RAG, Multi-hop RAG

multi-query attention, Multi-query and grouped-query attention-Optimizing attention: From multi-head to multi-query to grouped query

multi-query RAG, Multi-query RAG

multilevel perceptrons, The Components of the Forward Pass

(see also feedforward neural networks)

multimodality, Multimodal Large Language Models-Summary

defined, Multimodal Large Language Models

embedding models, Multimodal Embedding Models-OpenCLIP

CLIP, CLIP: Connecting Text and Images-How Can CLIP Generate Multimodal Embeddings?

OpenCLIP, OpenCLIP-OpenCLIP

text generation models, Making Text Generation Models Multimodal-Use Case 2: Multimodal Chat-Based Prompting

BLIP-2, BLIP-2: Bridging the Modality Gap-BLIP-2: Bridging the Modality Gap

chat-based prompting, Use Case 2: Multimodal Chat-Based Prompting-Use Case 2: Multimodal Chat-Based Prompting

image captioning, Use Case 1: Image Captioning-Use Case 1: Image Captioning

preprocessing images, Preprocessing images

preprocessing text, Preprocessing text

Vision Transformer, Transformers for Vision-Transformers for Vision

multiple negatives ranking (MNR) loss, Multiple negatives ranking loss-Multiple negatives ranking loss

N

named-entity recognition (see NER)

natural language inference (NLI), Generating Contrastive Examples natural language processing (NLP), What Is Language AI?, What Is Contrastive Learning?

nDCG (normalized discounted cumulative gain), Reranking example, Scoring across multiple queries with mean average precision

nearest neighbor search

- pretrained word embeddings, Using pretrained Word Embeddings
- recommendation system embeddings, Recommending Songs by Embeddings
- vector databases versus, Nearest neighbor search versus vector databases

negative sampling, The Word2vec Algorithm and Contrastive Training

NER (named-entity recognition), Named-Entity Recognition-Fine-Tuning for Named-Entity Recognition, Adapters

- fine-tuning for, Fine-Tuning for Named-Entity Recognition
- preparing data for, Preparing Data for Named-Entity Recognition-Preparing Data for Named-Entity Recognition

neural networks, Better Representations with Dense Vector Embeddings

NLI (natural language inference), Generating Contrastive Examples

NLP (natural language processing), What Is Language AI?, What Is Contrastive Learning?

noise-contrastive estimation, The Word2vec Algorithm and Contrastive Training

nonplayable characters (NPCs), What Is Language AI?

nontrainable (frozen) models, Text Classification with Representation Models, Supervised Classification, Supervised Classification

normalization, Transformer block, The Transformer Block

normalized discounted cumulative gain (nDCG), Reranking example, Scoring across multiple queries with mean average precision

NPCs (nonplayable characters), What Is Language AI?

NTXentLoss, Multiple negatives ranking loss

nucleus sampling, top_p

NumPy, Nearest neighbor search versus vector databases

num_train_epochs parameter, Train Model, Training Configuration

NVIDIA GPUs, Hardware and Software Requirements, Generating Your First Text n_components parameter, Reducing the Dimensionality of Embeddings

O

Odds Ratio Preference Optimization (ORPO), Training

one-shot prompting, In-Context Learning: Providing Examples

chain-of-thought versus, Chain-of-Thought: Think Before Answering

in-context learning, In-Context Learning: Providing Examples

Open LLM Leaderboard, Model I/O: Loading Quantized Models with LangChain, Leaderboards

open-source LLMs, Open Models-Open Source Frameworks

OpenAI, ChatGPT for Classification

(see also ChatGPT; GPTs)

creating accounts, API Keys, ChatGPT for Classification

generating embeddings, Supervised Classification

OpenCLIP, OpenCLIP-OpenCLIP

optim parameter, Training Configuration

ORPO (Odds Ratio Preference Optimization), Training

output verification, Output Verification-Grammar: Constrained Sampling

constrained sampling, Grammar: Constrained Sampling-Grammar: Constrained Sampling

providing examples, Providing Examples

P

pad_token [PAD], BERT base model (uncased) (2018)

parallel processing, Attention is all you need, Summary

parallel prompts, Chain Prompting: Breaking up the Problem

PCA (Principal Component Analysis), Reducing the Dimensionality of Embeddings

PEFT (parameter-efficient fine-tuning), Parameter-Efficient Fine-Tuning (PEFT)-Compressing the model for (more) efficient training

adapters, Adapters-Adapters

compression, Compressing the model for (more) efficient training-Compressing the model for (more) efficient training

LoRA, Low-Rank Adaptation (LoRA)-Low-Rank Adaptation (LoRA)
peft library, LoRA Configuration
peft_config parameter, LoRA Configuration
Perplexity, Retrieval-Augmented Generation (RAG)
persona, in text-generation prompts, The Potential Complexity of a Prompt
per_device_eval_batch_size argument, Train Model
per_device_train_batch_size argument, Train Model
Phi-3
comparing to other trained tokenizers, Phi-3 (and Llama 2)
forward pass, The Components of the Forward Pass
loading quantized models, Model I/O: Loading Quantized Models with LangChain
prompt template, A Single Link in the Chain: Prompt Template
quantization, Model I/O: Loading Quantized Models with LangChain
Phi-3-mini, Generating Your First Text, Choosing a Text Generation Model
Pinecone, Nearest neighbor search versus vector databases
positional embeddings, Positional Embeddings (RoPE)-Positional Embeddings (RoPE)
PPO (Proximal Policy Optimization), Reward model training step
precision predictions, confusion matrices, Using a Task-Specific Model
predictions, task-specific model, Using a Task-Specific Model
preference tuning, ChatGPT for Classification, The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning, Preference-Tuning / Alignment / RLHF-Training
Direct Preference Optimization, Training No Reward Model-Training
fine-tuning, Training
model quantization, Model Quantization
templating alignment data, Templating Alignment Data
training configuration, Training Configuration
reward models, Automating Preference Evaluation Using Reward Models-Reward model training step
inputs and outputs of, The Inputs and Outputs of a Reward Model

    - training, Training a Reward Model-Reward model training step
- pretraining, defined, The Training Paradigm of Large Language Models
- primacy effect, Instruction-Based Prompting
- Principal Component Analysis (PCA), Reducing the Dimensionality of Embeddings
- projection matrices, How attention is calculated
- prompt engineering, Text Classification with Generative Models, Prompt Engineering-Summary
  - chain prompting, Chain Prompting: Breaking up the Problem-Chain Prompting: Breaking up the Problem
  - in-context learning, In-Context Learning: Providing Examples-In-Context Learning: Providing Examples
  - instruction-based prompting, Instruction-Based Prompting-Instruction-Based Prompting
  - output verification, Output Verification-Grammar: Constrained Sampling
    - constrained sampling, Grammar: Constrained Sampling-Grammar: Constrained Sampling
    - providing examples, Providing Examples
  - potential complexity of prompts, The Potential Complexity of a Prompt-The Potential Complexity of a Prompt
  - prompt components, The Basic Ingredients of a Prompt-The Basic Ingredients of a Prompt
  - reasoning with generative models, Reasoning with Generative Models-Tree-of-Thought: Exploring Intermediate Steps
  - text generation models, Using Text Generation Models-top_p
    - choosing, Choosing a Text Generation Model
    - controlling output, Controlling Model Output-top_p
    - loading, Loading a Text Generation Model-Loading a Text Generation Model
- Proximal Policy Optimization (PPO), Reward model training step
- Python, learning about, Prerequisites

Q

- Q-Former (Querying Transformer), BLIP-2: Bridging the Modality Gap-BLIP-2: Bridging the Modality Gap

Q8 model, Model I/O: Loading Quantized Models with LangChain

QLoRA (quantized low-rank adaptation), Instruction Tuning with QLoRA-Merge Weights

- fine-tuning, Training
- LoRA configuration, LoRA Configuration
- merging weights, Merge Weights
- model quantization, Model Quantization
- templating instruction data, Templating Instruction Data
- training configuration, Training Configuration

quantization, Model I/O: Loading Quantized Models with LangChain, Compressing the model for (more) efficient training-Compressing the model for (more) efficient training

quantization_config parameter, LoRA Configuration

quantized low-rank adaptation (see QLoRA)

Querying Transformer (Q-Former), BLIP-2: Bridging the Modality Gap-BLIP-2: Bridging the Modality Gap

R

r parameter, LoRA Configuration

RAG (retrieval-augmented generation), Token Embeddings, Text Embeddings (for Sentences and Whole Documents), Overview of Semantic Search and RAG, Retrieval-Augmented Generation (RAG)-RAG Evaluation

- agentic RAG, Agentic RAG
- basic pipeline, Retrieval-Augmented Generation (RAG)
- converting search system to, From Search to RAG
- evaluating results, RAG Evaluation
- grounded generation, Example: Grounded Generation with an LLM API
- with local models, Example: RAG with Local Models-The RAG prompt
- multi-hop RAG, Multi-hop RAG
- multi-query RAG, Multi-query RAG
- query rewriting, Query rewriting
- query routing, Query routing

Ragas, RAG Evaluation

random_state parameter, Reducing the Dimensionality of Embeddings rate limit errors, ChatGPT for Classification

ReAct

in LangChain, ReAct in LangChain-ReAct in LangChain

step-by-step reasoning, The Driving Power Behind Agents: Step-by-step Reasoning

reasoning

with generative models, Reasoning with Generative Models-Tree-of-Thought: Exploring Intermediate Steps

chain-of-thought, Chain-of-Thought: Think Before Answering-Chain-of-Thought: Think Before Answering

self-consistency, Self-Consistency: Sampling Outputs

tree-of-thought, Tree-of-Thought: Exploring Intermediate Steps-Tree-of-Thought: Exploring Intermediate Steps

step-by-step, The Driving Power Behind Agents: Step-by-step Reasoning, The Driving Power Behind Agents: Step-by-step Reasoning

recall predictions, confusion matrices, Using a Task-Specific Model

recency effect, Instruction-Based Prompting

recommendation systems, Embeddings for Recommendation Systems-Training a Song Embedding Model

recurrent neural networks (RNNs), Encoding and Decoding Context with Attention

reduce_outliers() function, BERTopic: A Modular Topic Modeling Framework

regulation, Responsible LLM Development and Usage

relevance scoring, Attention is all you need, How attention is calculated-Self-attention: Relevance scoring, Positional Embeddings (RoPE)

repository, Hardware and Software Requirements

representation models, Representation Models: Encoder-Only Models-Representation Models: Encoder-Only Models

defined, Representing Language as a Bag-of-Words

fine-tuning for classification, Fine-Tuning Representation Models for Classification-Summary

few-shot classification, Few-Shot Classification-Fine-Tuning for Few-Shot Classification

masked language modeling, Continued Pretraining with Masked Language Modeling-Continued Pretraining with Masked

Language Modeling
named-entity recognition, Named-Entity Recognition-Fine-Tuning for Named-Entity Recognition
supervised classification, Supervised Classification-Freezing Layers
generative models versus, Representation Models: Encoder-Only Models
text classification, Text Classification with Representation Models-What If We Do Not Have Labeled Data?
classification tasks that leverage embeddings, Classification Tasks That Leverage Embeddings-What If We Do Not Have Labeled Data?
model selection, Model Selection-Model Selection
task-specific models, Using a Task-Specific Model
representation_model parameter, The Text Generation Lego Block
reranking, Overview of Semantic Search and RAG, Reranking-How reranking models work
BERTopic, Adding a Special Lego Block
example of, Reranking example-Reranking example
function of reranking models, How reranking models work
sentence transformers, Open source retrieval and reranking with sentence transformers
response validation, in chain prompting, Chain Prompting: Breaking up the Problem
retrieval evaluation metrics, Retrieval Evaluation Metrics-Scoring across multiple queries with mean average precision
scoring multiple queries with mean average precision, Scoring across multiple queries with mean average precision
scoring single queries with average precision, Scoring a single query with average precision
retrieval-augmented generation (see RAG)
return_full_text parameter, Generating Your First Text
reward models, Automating Preference Evaluation Using Reward Models-Reward model training step
inputs and outputs of, The Inputs and Outputs of a Reward Model
training, Training a Reward Model-Reward model training step
RMSNorm, The Transformer Block

RNNs (recurrent neural networks), Encoding and Decoding Context with Attention

RoBERTa, Word Versus Subword Versus Character Versus Byte Tokens, Model Selection

RoPE (rotary positional embeddings), Positional Embeddings (RoPE)-Positional Embeddings (RoPE)

Rorschach test, Use Case 1: Image Captioning

Rotten Tomatoes dataset, The Sentiment of Movie Reviews, Fine-Tuning a Pretrained BERT Model

ROUGE, Word-Level Metrics

RWKV, The Year of Generative AI

S

SBERT, SBERT-SBERT, Augmented SBERT-Augmented SBERT

self-attention, Attention Is All You Need, Self-attention: Relevance scoring-Self-attention: Combining information

self-consistency, Self-Consistency: Sampling Outputs

semantic search, Semantic Search and Retrieval-Augmented Generation-Scoring across multiple queries with mean average precision

- defined, Semantic Search and Retrieval-Augmented Generation
- dense retrieval, Overview of Semantic Search and RAG, Dense Retrieval-Fine-tuning embedding models for dense retrieval
  - caveats of, Caveats of dense retrieval
  - example of, Dense retrieval example-Search the index
  - fine-tuning embedding models for, Fine-tuning embedding models for dense retrieval
  - nearest neighbor search versus vector databases, Nearest neighbor search versus vector databases
  - text chunking, Chunking long texts-Multiple vectors per document
- reranking, Overview of Semantic Search and RAG, Reranking-How reranking models work
  - example of, Reranking example-Reranking example
  - function of reranking models, How reranking models work
  - sentence transformers, Open source retrieval and reranking with sentence transformers

  - retrieval evaluation metrics, Retrieval Evaluation Metrics-Scoring across multiple queries with mean average precision
    - scoring multiple queries with mean average precision, Scoring across multiple queries with mean average precision
    - scoring single queries with average precision, Scoring a single query with average precision
- Semantic Textual Similarity Benchmark (STSB), Train Model
- semi-hard negatives, Multiple negatives ranking loss
- sentence-transformers, Text Embeddings (for Sentences and Whole Documents), Supervised Classification, BERTopic: A Modular Topic Modeling Framework, Open source retrieval and reranking with sentence transformers, OpenCLIP, SBERT-SBERT, Supervised, SetFit: Efficient Fine-Tuning with Few Training Examples
- SentencePiece, Flan-T5 (2022)
- [SEP] token, BERT base model (uncased) (2018), Creating Contextualized Word Embeddings with Language Models, Preparing Data for Named-Entity Recognition
- sequence-to-sequence models, Text Classification with Generative Models, Using the Text-to-Text Transfer Transformer
- SetFit, Fine-Tuning Representation Models for Classification, SetFit: Efficient Fine-Tuning with Few Training Examples-SetFit: Efficient Fine-Tuning with Few Training Examples
- SFT (supervised fine-tuning), The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning-Merge Weights
  - full fine-tuning, Full Fine-Tuning
  - parameter-efficient fine-tuning, Parameter-Efficient Fine-Tuning (PEFT)-Compressing the model for (more) efficient training
    - adapters, Adapters-Adapters
    - compression, Compressing the model for (more) efficient training-Compressing the model for (more) efficient training
    - LoRA, Low-Rank Adaptation (LoRA)-Low-Rank Adaptation (LoRA)
  - QLoRA, Instruction Tuning with QLoRA-Merge Weights
    - fine-tuning, Training
    - LoRA configuration, LoRA Configuration
    - merging weights, Merge Weights

    - model quantization, Model Quantization
    - templating instruction data, Templating Instruction Data
    - training configuration, Training Configuration
- shared memory (SRAM), Flash Attention
- shortlisting, Reranking example
- silver datasets, Augmented SBERT
- SimCSE (Simple Contrastive Learning of Sentence Embeddings), Unsupervised Learning
- skip-gram, The Word2vec Algorithm and Contrastive Training
- softmax loss function, Loss Functions
- song recommendation systems, Embeddings for Recommendation Systems-Training a Song Embedding Model
- sparse attention, Local/sparse attention
- special tokens, Comparing Trained LLM Tokenizers, Tokenizer parameters
- specificity, in instruction-based prompting, Instruction-Based Prompting
- SRAM (shared memory), Flash Attention
- StableLM, Choosing a Text Generation Model
- StarCoder2, StarCoder2 (2024)
- step-by-step reasoning, The Driving Power Behind Agents: Step-by-step Reasoning-The Driving Power Behind Agents: Step-by-step Reasoning
- structured output, validating, Output Verification
- STSB (Semantic Textual Similarity Benchmark), Train Model
- subword tokens, Word Versus Subword Versus Character Versus Byte Tokens
- supervised classification, Supervised Classification-Supervised Classification
  - fine-tuning representation models for, Supervised Classification-Freezing Layers
    - freezing layers, Freezing Layers-Freezing Layers
    - pretrained BERT models, Fine-Tuning a Pretrained BERT Model-Fine-Tuning a Pretrained BERT Model
- supervised fine-tuning (see SFT)
- system 1 and 2 thinking processes, Reasoning with Generative Models

T

T5 (Text-to-Text Transfer Transformer), Using the Text-to-Text Transfer Transformer-Using the Text-to-Text Transfer Transformer

target_modules parameter, LoRA Configuration

task-specific models, Text Classification with Representation Models

temperature parameter, Temperature-top_p, Self-Consistency: Sampling Outputs

Tesla T4, Training

test splits, The Sentiment of Movie Reviews, Using a Task-Specific Model

text chunking, Getting the text archive and chunking it, Chunking long texts-Multiple vectors per document

- approaches for, Multiple vectors per document
- multiple vectors per document, Multiple vectors per document
- one vector per document, One vector per document

text classification, Text Classification-Summary

- with generative models, Text Classification with Generative Models-ChatGPT for Classification
  - ChatGPT, ChatGPT for Classification-ChatGPT for Classification
  - T5, Using the Text-to-Text Transfer Transformer-Using the Text-to-Text Transfer Transformer
- movie reviews, The Sentiment of Movie Reviews-The Sentiment of Movie Reviews
- with representation models, Text Classification with Representation Models-What If We Do Not Have Labeled Data?
  - classification tasks that leverage embeddings, Classification Tasks That Leverage Embeddings-What If We Do Not Have Labeled Data?
  - model selection, Model Selection-Model Selection
  - task-specific models, Using a Task-Specific Model-Using a Task-Specific Model

text clustering, Text Clustering and Topic Modeling-Inspecting the Clusters

- CLIP embedding model and, CLIP: Connecting Text and Images
- common pipeline for, A Common Pipeline for Text Clustering-Inspecting the Clusters

cluster model, Cluster the Reduced Embeddings-Cluster the Reduced Embeddings

dimensionality reduction model, Reducing the Dimensionality of Embeddings-Reducing the Dimensionality of Embeddings

embedding model, Embedding Documents

inspecting clusters, Inspecting the Clusters-Inspecting the Clusters

text embedding models, Text Embeddings (for Sentences and Whole Documents)-Text Embeddings (for Sentences and Whole Documents), Creating Text Embedding Models-Summary

contrastive learning, What Is Contrastive Learning?-What Is Contrastive Learning?

creating, Creating an Embedding Model-Multiple negatives ranking loss

evaluating, In-Depth Evaluation

generating contrastive examples, Generating Contrastive Examples

loss functions, Loss Functions-Multiple negatives ranking loss

training, Train Model-Train Model

fine-tuning, Fine-Tuning an Embedding Model

Augmented SBERT, Augmented SBERT-Augmented SBERT

supervised, Supervised-Supervised

SBERT, SBERT-SBERT

unsupervised learning, Unsupervised Learning

text generation, Generating Your First Text-Generating Your First Text, Advanced Text Generation Techniques and Tools-Summary

agents, Agents: Creating a System of LLMs-ReAct in LangChain

ReAct in LangChain, ReAct in LangChain-ReAct in LangChain

step-by-step reasoning, The Driving Power Behind Agents: Step-by-step Reasoning-The Driving Power Behind Agents: Step-by-step Reasoning

chains, Chains: Extending the Capabilities of LLMs-A Chain with Multiple Prompts

chaining single prompt, A Single Link in the Chain: Prompt Template-A Single Link in the Chain: Prompt Template

sequential chaining of multiple prompts, A Chain with Multiple Prompts-A Chain with Multiple Prompts

- memory of conversations, Memory: Helping LLMs to Remember Conversations-Conversation Summary
  - conversation buffer, Conversation Buffer-Conversation Buffer
  - conversation summary, Conversation Summary-Conversation Summary
  - windowed conversation buffer, Windowed Conversation Buffer-Windowed Conversation Buffer
- model I/O, Model I/O: Loading Quantized Models with LangChain-Model I/O: Loading Quantized Models with LangChain
- multimodality, Making Text Generation Models Multimodal-Use Case 2: Multimodal Chat-Based Prompting
  - BLIP-2, BLIP-2: Bridging the Modality Gap-BLIP-2: Bridging the Modality Gap
  - chat-based prompting, Use Case 2: Multimodal Chat-Based Prompting-Use Case 2: Multimodal Chat-Based Prompting
  - image captioning, Use Case 1: Image Captioning-Use Case 1: Image Captioning
  - preprocessing images, Preprocessing images
  - preprocessing text, Preprocessing text
- prompt engineering, Using Text Generation Models-top_p
  - choosing models, Choosing a Text Generation Model
  - controlling output, Controlling Model Output-top_p
  - loading models, Loading a Text Generation Model-Loading a Text Generation Model
- topic modeling, The Text Generation Lego Block-The Text Generation Lego Block

text-in-text-out model, The Inputs and Outputs of a Trained Transformer LLM

Text-to-Text Transfer Transformer (T5), Using the Text-to-Text Transfer Transformer-Using the Text-to-Text Transfer Transformer

thenlper/gte-small model, Embedding Documents

%%timeit magic command, Speeding Up Generation by Caching Keys and Values

TinyLlama, Instruction Tuning with QLoRA, Preference Tuning with DPO

tokenization-free encoding, Word Versus Subword Versus Character Versus Byte Tokens

tokens and tokenizers, Generating Your First Text, Tokens and Embeddings-Creating Contextualized Word Embeddings with Language Models

- bag-of-words model, Representing Language as a Bag-of-Words
- comparing trained tokenizers, Comparing Trained LLM Tokenizers-Phi-3 (and Llama 2)
  - BERT base model (cased), BERT base model (cased) (2018)
  - BERT base model (uncased), BERT base model (uncased) (2018)
  - Flan-T5, Flan-T5 (2022)
  - Galactica, Galactica
  - GPT-2, GPT-2 (2019)
  - GPT-4, GPT-4 (2023)
  - Phi-3 and Llama 2, Phi-3 (and Llama 2)
  - StarCoder2, StarCoder2 (2024)
- decoding strategy, Choosing a Single Token from the Probability Distribution (Sampling/Decoding)-Choosing a Single Token from the Probability Distribution (Sampling/Decoding)
- downloading and running LLMs, Downloading and Running an LLM-Downloading and Running an LLM
- forward pass, The Components of the Forward Pass-The Components of the Forward Pass
- input preparation, How Tokenizers Prepare the Inputs to the Language Model
- masked language modeling, Using the Text-to-Text Transfer Transformer
- parallel token processing and context size, Parallel Token Processing and Context Size-Parallel Token Processing and Context Size
- special tokens, Comparing Trained LLM Tokenizers
- task-specific representation model, Using a Task-Specific Model, Text Classification with Generative Models
- text breakdown, How Does the Tokenizer Break Down Text?
- text-focused versus code-focused models, The domain of the data
- token embeddings, Token Embeddings-Creating Contextualized Word Embeddings with Language Models, The Components of the Forward Pass, Summary

creating contextualized word embeddings, Creating Contextualized Word Embeddings with Language Models-Creating Contextualized Word Embeddings with Language Models

tokenizer’s vocabulary and, A Language Model Holds Embeddings for the Vocabulary of Its Tokenizer

token spans, Using the Text-to-Text Transfer Transformer

tokenization schemes, Word Versus Subword Versus Character Versus Byte Tokens-Word Versus Subword Versus Character Versus Byte Tokens

byte tokens, Word Versus Subword Versus Character Versus Byte Tokens

character tokens, Word Versus Subword Versus Character Versus Byte Tokens

subword tokens, Word Versus Subword Versus Character Versus Byte Tokens

word tokens, Word Versus Subword Versus Character Versus Byte Tokens

tokenizer properties, Tokenizer Properties-The domain of the data

datasets, The domain of the data

methods, Tokenization methods

parameters, Tokenizer parameters

white space characters, GPT-2 (2019)

tone of voice, in text-generation prompts, The Potential Complexity of a Prompt

topic modeling, Text Clustering and Topic Modeling, From Text Clustering to Topic Modeling-The Text Generation Lego Block

BERTopic, BERTopic: A Modular Topic Modeling Framework-BERTopic: A Modular Topic Modeling Framework

representation blocks, Adding a Special Lego Block-The Text Generation Lego Block

top_k parameter, top_p

top_p parameter, top_p, Self-Consistency: Sampling Outputs

train splits, The Sentiment of Movie Reviews

TrainingArguments class, Fine-Tuning a Pretrained BERT Model

transfer learning, Representation Models: Encoder-Only Models

- Transformer architecture, Attention Is All You Need-Attention Is All You Need, Looking Inside Large Language Models-Summary
  - attention layer, The Components of the Forward Pass, Inside the Transformer Block
  - decoding strategy, Choosing a Single Token from the Probability Distribution (Sampling/Decoding)-Choosing a Single Token from the Probability Distribution (Sampling/Decoding)
  - feedforward layer, Inside the Transformer Block
  - forward pass components, The Components of the Forward Pass-The Components of the Forward Pass
  - inputs and outputs of, The Inputs and Outputs of a Trained Transformer LLM-The Inputs and Outputs of a Trained Transformer LLM
  - keys and values cache, Speeding Up Generation by Caching Keys and Values-Speeding Up Generation by Caching Keys and Values
  - optimizing attention, Optimizing attention: From multi-head to multi-query to grouped query
  - parallel token processing and context size, Parallel Token Processing and Context Size-Parallel Token Processing and Context Size
  - recent improvements to, Recent Improvements to the Transformer Architecture-Other Architectural Experiments and Improvements
    - more efficient attention, More Efficient Attention-Flash Attention
    - positional embeddings, Positional Embeddings (RoPE)-Positional Embeddings (RoPE)
    - Transformer blocks, The Transformer Block
  - Transformer blocks, Inside the Transformer Block-Self-attention: Combining information
    - attention calculation, How attention is calculated-How attention is calculated
    - attention layer, The attention layer at a glance
    - attention mechanism, Attention is all you need-Attention is all you need
    - feedforward neural networks, The feedforward neural network at a glance

        - self-attention and relevance scoring, Self-attention: Relevance scoring-Self-attention: Combining information
    - Vision Transformer, Transformers for Vision-Transformers for Vision
- transparency and accountability, Responsible LLM Development and Usage
- tree-of-thought, Tree-of-Thought: Exploring Intermediate Steps-Tree-of-Thought: Exploring Intermediate Steps
- TruthfulQA, Benchmarks, Leaderboards
- TSDAE (Transformer-Based Sequential Denoising Auto-Encoder)
    - for domain adaptation, Using TSDAE for Domain Adaptation
    - overview of, Transformer-Based Sequential Denoising Auto-Encoder-Transformer-Based Sequential Denoising Auto-Encoder

U

- UltraChat dataset, Templating Instruction Data
- UMAP (Uniform Manifold Approximation and Projection), Reducing the Dimensionality of Embeddings
- unigram language model, Flan-T5 (2022)
- unk_token [UNK], BERT base model (uncased) (2018)
- use_cache parameter, Speeding Up Generation by Caching Keys and Values

V

- valid output, verifying, Output Verification
- validation splits, The Sentiment of Movie Reviews
- vector databases
    - dense retrieval, Dense Retrieval
    - nearest neighbor search versus, Nearest neighbor search versus vector databases
    - retrieval-augmented generation, Loading the embedding model
- video random-access memory (VRAM), Hardware and Software Requirements, Limited Resources Are All You Need
- visualization
    - BERTopic, BERTopic: A Modular Topic Modeling Framework
    - cluster analysis, Inspecting the Clusters

dimensionality reduction and, Inspecting the Clusters
ViT (Vision Transformer), Transformers for Vision-Transformers for Vision, BLIP-2: Bridging the Modality Gap
vocabulary, of tokenizers, Tokenizer parameters, A Language Model Holds Embeddings for the Vocabulary of Its Tokenizer, The Components of the Forward Pass, Summary
VRAM (video random-access memory), Hardware and Software Requirements, Limited Resources Are All You Need

W

warmup_ratio parameter, Training Configuration
warmup_steps argument, Train Model
Weaviate, Nearest neighbor search versus vector databases
whitespace characters, GPT-2 (2019)
windowed conversation buffer memory, Windowed Conversation Buffer-Windowed Conversation Buffer, Conversation Summary
word embeddings, Word Embeddings Beyond LLMs-The Word2vec Algorithm and Contrastive Training
pretrained, Using pretrained Word Embeddings
word2vec algorithm and contrastive training, The Word2vec Algorithm and Contrastive Training-The Word2vec Algorithm and Contrastive Training
word tokens, Word Versus Subword Versus Character Versus Byte Tokens
word-level metrics, in generative model evaluation, Word-Level Metrics
word2vec algorithm, Better Representations with Dense Vector Embeddings, Types of Embeddings-Encoding and Decoding Context with Attention
contrastive training and, The Word2vec Algorithm and Contrastive Training-The Word2vec Algorithm and Contrastive Training, What Is Contrastive Learning?
embedding songs, Recommending Songs by Embeddings
WordPiece, How Does the Tokenizer Break Down Text?
cased BERT base model, BERT base model (cased) (2018)
uncased BERT base model, BERT base model (uncased) (2018)
<work> token, Galactica

Y

Year of Generative AI, The Year of Generative AI-The Year of Generative AI

Z

zero-shot classification, What If We Do Not Have Labeled Data?-What If We Do Not Have Labeled Data?

- CLIP, CLIP: Connecting Text and Images
- SetFit, Fine-Tuning for Few-Shot Classification

zero-shot prompting

- chain-of-thought, Chain-of-Thought: Think Before Answering
- in-context learning, In-Context Learning: Providing Examples