Index

A

accuracy
- confusion matrices, 120
- output verification, 192

adaptive pretraining, 320
agents, 218-223
- agentic RAG, 256
- ReAct in LangChain, 221-223
- step-by-step reasoning, 219-221

AI (artificial intelligence)
- accelerated development of, 3
- defined, 4

ALBERT, 115
align_labels function, 350
all-MiniLM-L6-v2 model, 309
all-mpnet-base-v2 model, 337
Annoy, 239
Anthropic Claude, 29
APIs (application programming interfaces), 30
- Cohere, xv, 230
- external, 134
- generating embeddings, 123
- OpenAI, xv, 133

artificial intelligence (see AI)
ArXiv, 138
attention, 11-18
- overview of, 11-14
- Transformer architecture, 15-18
  - attention calculation, 91-93
  - attention layer, 79, 86, 88, 106
  - Flash Attention, 100
  - grouped-query attention, 98-100
  - local attention, 96
  - multi-query attention, 98-100
  - optimizing attention, 98
  - self-attention and relevance scoring, 93-94
  - sparse attention, 96

attention heads, 91-93, 107
audience, in text-generation prompts, 178
Augmented SBERT, 311-315
autoregressive architecture, 12, 15, 22, 76, 97

B

bag-of-words model, 6-7
- embeddings, 10
- topic modeling, 148-150, 156

benchmarks, in generative model evaluation, 374
BERT (Bidirectional Encoder Representations from Transformers), 18-20
- adoption by search engines, 225
- BERT-like models, 115
- comparing to other trained tokenizers, 47
- fine-tuning pretrained BERT models, 325-328
- masked language modeling, 311
- Transformer blocks versus, 97

BERTopic, 148-155
- algorithmic variants, 152
- modularity of, 151
- representation blocks, 156-163
  - KeyBERTInspired, 158
  - maximal marginal relevance, 159
  - text generation, 160-163

BERTScore, 374
bias and fairness, 28

Bidirectional Encoder Representations from Transformers (see BERT)
bitsandbytes package, 369
BLEU, 374
BLIP-2 (Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation 2)
  chat-based prompting, 283-286
  image captioning, 280
  preprocessing text, 280
  Q-Former, 273-277
BM25 algorithm, 233
BPE (byte pair encoding), 43, 55
byte tokens, 45

C

c-TF-IDF, 149, 158, 161
capitalization, 56
captioning, 280-283
centroid-based algorithms, 143
chains, 182-184, 202-209
  chain-of-thought, 185-188
  chaining single prompt, 203-205
  sequential chaining of multiple prompts, 206-209
character tokens, 45
chat tokens, 54
chat-based prompting, 283-286
Chatbot Arena, 376
ChatGPT, 202, 273, 383, 386
  release of, 3
  text classification, 132-135
chatgpt_generation function, 133
chat_history input variable, 211
classification reports, 119-120
classification step, embedding model, 121
CLIP, 265-268
  connecting text and images, 265
  generating multimodal embeddings, 265-268
  OpenCLIP, 268-272
closed-source LLMs, 29
[CLS] token, 18, 48, 60, 270, 294, 318, 351
cluster model, 142-144
clustering (see text clustering)
CNNs (convolutional neural networks), 260
Cohere, 252
  Command R+, 30, 256
  creating accounts, xv, 230
  generating embeddings, 123
  query rewriting, 255
  Rerank endpoint, 241
completion models, 22
compute_metrics function, 351
confusion matrices, 119
CoNLL-2003 dataset, 347
constrained sampling, 194-197
context
  attention and, 88
  prompt engineering, 178
  training datasets, 64
context length
  completion models, 22
  token processing limits, 81
context window, in completion models, 22
contrastive learning
  text embedding models, 291-293, 296
  word2vec algorithm and, 64-67
Contrastive Tension (CT), 316
conversation buffer memory, 217
  overview of, 210-212
  windowed, 212-214
conversation summary memory, 214-217
convert_ids_to_tokens function, 269
convolutional neural networks (CNNs), 260
cosine similarity, 125, 302-304
CountVectorizer, 150
cross-encoders, 244
cross-entropy loss, 305
CT (Contrastive Tension), 316

D

data outliers, 143
DataCollator class, 326, 342
datamapplot package, 163
DBSCAN (Density-Based Spatial Clustering), 143
DeBERTa, 59, 115
decoder-only models (see generative models)
decoding strategy, 79-81, 106
dense retrieval, 226, 228-240
  caveats of, 234
  example of, 230-234
  fine-tuning embedding models for, 239
  nearest neighbor search versus vector databases, 238
  text chunking, 235-237
density-based algorithms, 143

- Density-Based Spatial Clustering (DBSCAN), 143
- dimensionality reduction model, 140-142
- DistilBERT, 115, 120
- domain adaptation, 320
- do_sample parameter, 34
- DPO (Direct Preference Optimization), 384-389
  - fine-tuning, 388
  - model quantization, 386
  - templating alignment data, 386
  - training configuration, 387
- DSPy, 200

E

- easy negatives, 308
- Elo rating system, 377
- embeddings, 8-10, 37, 57-70
  - dense retrieval, 226, 228-240
  - embedding models, defined, 114, 290
  - multimodality, 263-272
    - CLIP, 265-268
    - OpenCLIP, 268-272
  - overview of, 8-10, 289-291
  - positional embeddings, 102-104
  - recommendation systems, 67-70
  - text classification tasks that leverage, 120-126
    - supervised classification, 121-123
    - zero-shot classification, 123-126
  - text clustering pipeline, 139-146
    - cluster model, 142-144
    - dimensionality reduction model, 140-142
    - embedding model, 139
    - inspecting clusters, 144-146
  - text embedding models, 61-63, 289-321
    - contrastive learning, 291-293
    - creating, 296-308
    - fine-tuning, 309-315
    - SBERT, 293-296
    - unsupervised learning, 316
  - token embeddings, 57-61
    - creating contextualized word embeddings, 58-61
    - tokenizer's vocabulary and, 57
  - types of, 10
  - word embeddings, 63-67
    - pretrained, 63
    - word2vec algorithm and contrastive training, 64-67
- encoder-decoder models, 128
- encoder-only models (see representation models)
- ethics, validating output, 192
- exponential backoff, 134

F

- F1 score, confusion matrices, 120
- FAISS, 239
- Falcon, 168
- feature extraction step, embedding model, 121
- feedforward layer, 86
- feedforward neural networks, 15, 87, 106, 326
- few-shot classification, 333-339
  - fine-tuning for classification, 337-339
  - SetFit, 333-336
- few-shot prompting, 181-182, 192-194
- find_topics() function, BERTopic, 154
- fine-tuning
  - embedding models for dense retrieval, 239
  - generative models, 192, 355-389
    - evaluating, 373-377
    - preference tuning, 356, 378-389
    - supervised fine-tuning, 356-373
    - training steps, 355-357
  - overview of, 26
  - representation models, 323-354
    - few-shot classification, 333-339
    - masked language modeling, 340-345
    - named-entity recognition, 345-353
    - supervised classification, 323-332
  - T5 model, 129
  - text embedding models, 309-315
    - Augmented SBERT, 311-315
    - supervised, 309-311
- Flan-T5 model, 50, 130-131
- Flash Attention, 100, 107
- forward pass, 106
  - components of, 76-79
  - defined, 74
- foundation models, 23
- fp16 parameter, 299
- freezing layers, 298, 328-332
- frozen (nontrainable) models, 114, 121, 324

G

- Ġ symbol, 280

Galactica, 52
General Language Understanding Evaluation (GLUE) benchmark, 297, 300, 359, 374
generated_text variable, 281, 283
generation_output variable, 42
generative models, 20-22
- evaluating, 373-377
  - automated evaluation, 376
  - benchmarks, 374
  - human evaluation, 376
  - leaderboards, 376
  - word-level metrics, 374
- fine-tuning, 355-389
  - evaluation, 373-377
  - preference tuning, 356, 378-389
  - supervised fine-tuning, 356-373
  - training steps, 355-357
- reasoning, 184-191
  - self-consistency, 188
  - tree-of-thought, 189-191
- representation models versus, 20
- text classification, 127-135
  - ChatGPT, 132-135
  - T5, 128-131

generative pre-trained transformers (see GPTs)
Generative Pseudo-Labeling (GPL), 316
Gensim library, 63
get_topic function, 153
get_topic_info() method, 152
GGUF model, 195, 200
GitHub, xvi
GloVe, 294
GLUE (General Language Understanding Evaluation) benchmark, 297, 300, 359, 374
gold datasets, 312
Goodhart's Law, 377
Google Colab, xii, xiv, 29, 33, 85, 372
Google Gemini, 250
Google Search, 225
GPL (Generative Pseudo-Labeling), 316
GPT2Tokenizer, 280
GPTs (generative pre-trained transformers), 20, 167
- (see also text generation)
- GPT-1, 20
- GPT-2, 3, 20, 46, 49
- GPT-3, 20, 87, 96, 364
- GPT-3.5, 23, 132, 205, 221
- GPT-4, 23, 29, 38, 50, 87

GPUs
- Flash Attention, 100
- requirements, xiv, 28
- SRAM and HBM, 100

grammar, 192, 194-197
greedy decoding, 80
grounded generation, 250-252
grouped-query attention, 98-100, 107
GSM8k, 374
Guardrails, 194
Guidance, 194

H

hallucination
- avoiding in in instruction-based prompting, 177
- text generation models, 225

hard negatives, 308
harmful content, generating, 28
Haystack, 200
HBM (high bandwidth memory), 100
HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), 143, 153
HellaSwag, 374
high bandwidth memory (HBM), 100
Hugging Face, 32, 112
- creating accounts, xv
- evaluate package, 351
- tokenizers, 56

human evaluation, 145, 376
HumanEval, 375
hybrid search, 235, 242

I

Idefics 2, 277
images (see multimodality)
in-context learning, 180-182
indexes, 232
InfoNCE, 304
input_ids variable, 40
instruction-based prompting, 175-177
intellectual property, 28
intuition-first philosophy, xi
invoke function, 204

J

Jupyter, 85

K

k-means algorithm, 143, 151, 153
KeyBERTInspired, 158, 160
keyword search
  reranking, 242-243
  verifying semantic search with, 233
kv (keys and values) cache, 83-85

L

LangChain, 200
  (see also chains)
  loading quantized models with, 200-202
  ReAct in, 221-223
Language AI (Language Artificial Intelligence), 3-24
  defining, 4
  recent history of, 5-24
    attention, 11-18
    bag-of-words model, 6-7
    embeddings, 8-10
    generative models, 20-22
    representation models, 18-20
    Year of Generative AI, 23-24
language modeling, 355
language modeling head (LM head), 76-79
large language models (see LLMs)
latent Dirichlet allocation, 147
LayerNorm, 101
leaderboards, in generative model evaluation, 376
learning_rate parameter, 371
Llama, 168
Llama 2, xv, 29, 53, 98, 168, 273
llama-cpp-python library, 195
LLaVA, 277
LLM-as-a-judge, 257
LLMs (large language models), 3-35
  code examples and exercises, xvi
  embeddings, 37, 57-70
    recommendation systems, 67-70
    text embeddings, 61-63
    token embeddings, 57-61
    word embeddings, 63-67
  fine-tuning generative models, 355-389
    evaluation, 373-377
    preference tuning, 356, 378-389
    supervised fine-tuning, 356-373
    training steps, 355-357
  fine-tuning representation models, 323-354
    few-shot classification, 333-339
    masked language modeling, 340-345
    named-entity recognition, 345-353
    supervised classification, 323-332
  generating text, 32-34
  generative models, 20-22
  hardware and software requirements, xiv, 28
  high-level view, 38
  history of Language AI, 5-24
  interfacing with, 29-32
    closed-source models, 29
    open models, 30-32
  intuition-first philosophy, xi
  moving definition of, 25
  multimodality, 259-286
    embedding models, 263-272
    text generation models, 273-286
    Vision Transformer, 260-262
  prompt engineering, 167-198
    chain prompting, 182-184
    in-context learning, 180-182
    instruction-based prompting, 175-177
    output verification, 191-197
    potential complexity of prompts, 177-179
    prompt components, 173-175
    reasoning with generative models, 184-191
    text generation models, 167-172
  representation models, 18-20
  responsible development and usage of, 28
  retrieval-augmented generation, 227, 249-257
    agentic RAG, 256
    converting search system to, 250
    evaluating results, 257
    grounded generation, 252
    multi-hop RAG, 256
    multi-query RAG, 255
    query rewriting, 255
    query routing, 256
    with local models, 252, 254
  semantic search, 225-249
    dense retrieval, 226, 228-240
    reranking, 226, 240-244
    retrieval evaluation metrics, 244-249
  text classification, 111-135
    with generative models, 127-135
    movie reviews, 112, 113

with representation models, 113-126
text clustering, 137-146
text embedding models, 289-321
contrastive learning, 291-293
creating, 296-308
fine-tuning, 309-315
SBERT, 293-296
unsupervised learning, 316
text generation, 199-224
agents, 218-223
chains, 202-209
memory of conversations, 209-217
model I/O, 200-202
tokens and tokenizers, 37-61
comparing trained tokenizers, 46-54
downloading and running LLMs, 39-42
input preparation, 38
text breakdown, 43
token embeddings, 57-61
tokenization schemes, 44-46
tokenizer properties, 55-56
topic modeling, 138, 146-163
training paradigm of, 25-26
Transformer architecture, 73-107
decoding strategy, 79-81
forward pass components, 76-79
inputs and outputs of, 74-76
keys and values cache, 83-85
parallel token processing and context size, 81-83
recent improvements to, 95-105
Transformer blocks, 85-94
utility of, 27
LM head (language modeling head), 76-79
LMQL, 194
local attention, 96
LoRA (low-rank adaptation), 361-364, 387
(see also QLoRA)
lora_alpha parameter, 370
loss functions, 301-308
cosine similarity loss, 302-304
multiple negatives ranking loss, 304-308
lr_scheduler_type parameter, 371

M

Mamba, 24
MAP (mean average precision), 244-249
MarginMSE loss, 302
masked language modeling (MLM), 340-345
mask_token [MASK], 48
Massive Text Embedding Benchmark (MTEB), 116, 140, 253, 300
matplotlib library, 145
maximal marginal relevance (MMR), 159
max_new_tokens parameter, 34
McCarthy, John, 4
mean average precision (MAP)
memory of conversations, 209-217
conversation buffer, 210-212
conversation summary, 214-217
windowed conversation buffer, 212-214
Meta Llama model, 30
Microsoft Bing, 225, 241
Microsoft Bing AI, 250
Microsoft Phi model, 30
microsoft/mpnet-base model, 297
min_cluster_size parameter, 144
min_dist parameter, 142
MIRACL, 243
Mistral, 30, 168, 277
MLM (masked language modeling), 340-345
MMLU, 374, 376
MMR (maximal marginal relevance), 159
MNLI (Multi-Genre Natural Language Inference) corpus, 297, 305
MNR (multiple negatives ranking) loss, 304-308
model I/O, 200-202
monoBERT, 244
MTEB (Massive Text Embedding Benchmark), 116, 140, 253, 300
Multi-Genre Natural Language Inference (MNLI) corpus, 297, 305
multi-hop RAG, 256
multi-query attention, 98-100
multi-query RAG, 255
multilevel perceptrons, 79
(see also feedforward neural networks)
multimodality, 259-286
defined, 259
embedding models, 263-272
CLIP, 265-268
OpenCLIP, 268-272
text generation models, 273-286
BLIP-2, 273-277
chat-based prompting, 283-286
image captioning, 280-283
preprocessing images, 278

preprocessing text, 279
Vision Transformer, 260-262
multiple negatives ranking (MNR) loss, 304-308

N

named-entity recognition (see NER)
natural language inference (NLI), 296
natural language processing (NLP), 4, 293
nDCG (normalized discounted cumulative gain), 243, 249
nearest neighbor search
pretrained word embeddings, 63
recommendation system embeddings, 68
vector databases versus, 238
negative sampling, 65
NER (named-entity recognition), 345-353, 360
fine-tuning for, 352
preparing data for, 347-351
neural networks, 8
NLI (natural language inference), 296
NLP (natural language processing), 4, 293
noise-contrastive estimation, 65
nonplayable characters (NPCs), 4
nontrainable (frozen) models, 114, 121, 324
normalization, Transformer block, 101
normalized discounted cumulative gain (nDCG), 243, 249
NPCs (nonplayable characters), 4
NTXentLoss, 304
nucleus sampling, 171
NumPy, 238
num_train_epochs parameter, 299, 371
NVIDIA GPUs, xiv, 33
n_components parameter, 142

O

Odds Ratio Preference Optimization (ORPO), 389
one-shot prompting, 182
chain-of-thought versus, 186
in-context learning, 181
Open LLM Leaderboard, 202, 376
open-source LLMs, 30-32
OpenAI, 132
(see also ChatGPT; GPTs)
creating accounts, xv, 133
generating embeddings, 123
OpenCLIP, 268-272
optim parameter, 372
ORPO (Odds Ratio Preference Optimization), 389
output verification, 191-197
constrained sampling, 194-197
providing examples, 192

P

pad_token [PAD], 47
parallel processing, 91, 106
parallel prompts, 184
PCA (Principal Component Analysis), 141
PEFT (parameter-efficient fine-tuning), 359-367
adapters, 359-361
compression, 364-367
LoRA, 361-364
peft library, 370
peft_config parameter, 371
Perplexity, 250
persona, in text-generation prompts, 178
per_device_eval_batch_size argument, 299
per_device_train_batch_size argument, 299
Phi-3
comparing to other trained tokenizers, 53
forward pass, 79
loading quantized models, 202
prompt template, 204
quantization, 201
Phi-3-mini, 32, 168
Pinecone, 239
positional embeddings, 102-104
PPO (Proximal Policy Optimization), 383
precision predictions, confusion matrices, 119
predictions, task-specific model, 118
preference tuning, 132, 356, 378-389
Direct Preference Optimization, 384-389
fine-tuning, 388
model quantization, 386
templating alignment data, 386
training configuration, 387
reward models, 379-383
inputs and outputs of, 380
training, 380-383
pretraining, defined, 26
primacy effect, 177
Principal Component Analysis (PCA), 141
projection matrices, 92
prompt engineering, 127, 167-198

chain prompting, 182-184
in-context learning, 180-182
instruction-based prompting, 175-177
output verification, 191-197
constrained sampling, 194-197
providing examples, 192
potential complexity of prompts, 177-179
prompt components, 173-175
reasoning with generative models, 184-191
text generation models, 167-172
choosing, 167
controlling output, 170-172
loading, 168-169
Proximal Policy Optimization (PPO), 383
Python, learning about, xii

Q

Q-Former (Querying Transformer), 274-277
Q8 model, 201
QLoRA (quantized low-rank adaptation), 367-373
fine-tuning, 372
LoRA configuration, 370
merging weights, 373
model quantization, 369
templating instruction data, 367
training configuration, 371
quantization, 201, 364-367
quantization_config parameter, 371
quantized low-rank adaptation (see QLoRA)
Querying Transformer (Q-Former), 274-277

R

r parameter, 370
RAG (retrieval-augmented generation), 57, 63, 227, 249-257
agentic RAG, 256
basic pipeline, 249
converting search system to, 250
evaluating results, 257
grounded generation, 252
with local models, 252-254
multi-hop RAG, 256
multi-query RAG, 255
query rewriting, 255
query routing, 256
Ragas, 257
random_state parameter, 142
rate limit errors, 134
ReAct
in LangChain, 221-223
step-by-step reasoning, 219
reasoning
with generative models, 184-191
chain-of-thought, 185-188
self-consistency, 188
tree-of-thought, 189-191
step-by-step, 219, 221
recall predictions, confusion matrices, 119
recency effect, 177
recommendation systems, 67-70
recurrent neural networks (RNNs), 11
reduce_outliers() function, 153
regulation, 28
relevance scoring, 90, 92-94, 104
repository, xiv
representation models, 18-20
defined, 7
fine-tuning for classification, 323-354
few-shot classification, 333-339
masked language modeling, 340-345
named-entity recognition, 345-353
supervised classification, 323-332
generative models versus, 20
text classification, 113-126
classification tasks that leverage embeddings, 120-126
model selection, 115-116
task-specific models, 116
representation_model parameter, 161
reranking, 226, 240-244
BERTopic, 156
example of, 241-243
function of reranking models, 244
sentence transformers, 243
response validation, in chain prompting, 184
retrieval evaluation metrics, 244-249
scoring multiple queries with mean average precision, 248
scoring single queries with average precision, 247
retrieval-augmented generation (see RAG)
return_full_text parameter, 33
reward models, 379-383
inputs and outputs of, 380
training, 380-383
RMSNorm, 101
RNNs (recurrent neural networks), 11

RoBERTa, 46, 115
RoPE (rotary positional embeddings), 102-104
Rorschach test, 282
Rotten Tomatoes dataset, 112, 325
ROUGE, 374
RWKV, 24

S

SBERT, 293-296, 311-315
self-attention, 15, 93-94
self-consistency, 188
semantic search, 225-249
  defined, 225
  dense retrieval, 226, 228-240
    caveats of, 234
    example of, 230-234
    fine-tuning embedding models for, 239
    nearest neighbor search versus vector databases, 238
    text chunking, 235-237
  reranking, 226, 240-244
    example of, 241-243
    function of reranking models, 244
    sentence transformers, 243
  retrieval evaluation metrics, 244-249
    scoring multiple queries with mean average precision, 248
    scoring single queries with average precision, 247
Semantic Textual Similarity Benchmark (STSB), 298
semi-hard negatives, 308
sentence-transformers, 62, 122, 151, 243, 272, 293-295, 309, 333
SentencePiece, 50
[SEP] token, 47, 60, 351
sequence-to-sequence models, 127, 128
SetFit, 323, 333-336
SFT (supervised fine-tuning), 356-373
  full fine-tuning, 357
  parameter-efficient fine-tuning, 359-367
    adapters, 359-361
    compression, 364-367
    LoRA, 361-364
  QLoRA, 367-373
    fine-tuning, 372
    LoRA configuration, 370
    merging weights, 373
    model quantization, 369
    templating instruction data, 367
    training configuration, 371
shared memory (SRAM), 100
shortlisting, 242
silver datasets, 312
SimCSE (Simple Contrastive Learning of Sentence Embeddings), 316
skip-gram, 65
softmax loss function, 301
song recommendation systems, 67-70
sparse attention, 96
special tokens, 47, 55
specificity, in instruction-based prompting, 176
SRAM (shared memory), 100
StableLM, 168
StarCoder2, 51
step-by-step reasoning, 219-221
structured output, validating, 191
STSB (Semantic Textual Similarity Benchmark), 298
subword tokens, 44
supervised classification, 121-123
  fine-tuning representation models for, 323-332
    freezing layers, 328-332
    pretrained BERT models, 325-328
supervised fine-tuning (see SFT)
system 1 and 2 thinking processes, 185

T

T5 (Text-to-Text Transfer Transformer), 128-131
target_modules parameter, 370
task-specific models, 113
temperature parameter, 171-172, 188
Tesla T4, 372
test splits, 113, 118
text chunking, 230, 235-237
  approaches for, 237
  multiple vectors per document, 236
  one vector per document, 236
text classification, 111-135
  with generative models, 127-135
    ChatGPT, 132-135
    T5, 128-131
  movie reviews, 112-113
  with representation models, 113-126
    classification tasks that leverage embeddings, 120-126

model selection, 115-116
task-specific models, 116-120
text clustering, 137-146
CLIP embedding model and, 265
common pipeline for, 139-146
cluster model, 142-144
dimensionality reduction model, 140-142
embedding model, 139
inspecting clusters, 144-146
text embedding models, 61-63, 289-321
contrastive learning, 291-293
creating, 296-308
evaluating, 300
generating contrastive examples, 296
loss functions, 301-308
training, 297-300
fine-tuning, 309
Augmented SBERT, 311-315
supervised, 309-311
SBERT, 293-296
unsupervised learning, 316
text generation, 32-34, 199-224
agents, 218-223
ReAct in LangChain, 221-223
step-by-step reasoning, 219-221
chains, 202-209
chaining single prompt, 203-205
sequential chaining of multiple prompts, 206-209
memory of conversations, 209-217
conversation buffer, 210-212
conversation summary, 214-217
windowed conversation buffer, 212-214
model I/O, 200-202
multimodality, 273-286
BLIP-2, 273-277
chat-based prompting, 283-286
image captioning, 280-283
preprocessing images, 278
preprocessing text, 279
prompt engineering, 167-172
choosing models, 167
controlling output, 170-172
loading models, 168-169
topic modeling, 160-163
text-in-text-out model, 74
Text-to-Text Transfer Transformer (T5), 128-131
thenlper/gte-small model, 140
%%timeit magic command, 85
TinyLlama, 367, 385
tokenization-free encoding, 45
tokens and tokenizers, 33, 37-61
bag-of-words model, 6
comparing trained tokenizers, 46-54
BERT base model (cased), 48
BERT base model (uncased), 47
Flan-T5, 50
Galactica, 52
GPT-2, 49
GPT-4, 50
Phi-3 and Llama 2, 53
StarCoder2, 51
decoding strategy, 79-81
downloading and running LLMs, 39-42
forward pass, 76-77
input preparation, 38
masked language modeling, 129
parallel token processing and context size, 81-83
special tokens, 47
task-specific representation model, 117, 127
text breakdown, 43
text-focused versus code-focused models, 56
token embeddings, 57-61, 77, 106
creating contextualized word embeddings, 58-61
tokenizer's vocabulary and, 57
token spans, 129
tokenization schemes, 44-46
byte tokens, 45
character tokens, 45
subword tokens, 44
word tokens, 44
tokenizer properties, 55-56
datasets, 56
methods, 55
parameters, 55
white space characters, 50
tone of voice, in text-generation prompts, 178
topic modeling, 138, 146-163
BERTopic, 148-155
representation blocks, 156-163
top_k parameter, 172
top_p parameter, 171, 188
train splits, 113
TrainingArguments class, 327

transfer learning, 19
Transformer architecture, 15-18, 73-107
  attention layer, 79, 86
  decoding strategy, 79-81
  feedforward layer, 86
  forward pass components, 76-79
  inputs and outputs of, 74-76
  keys and values cache, 83-85
  optimizing attention, 98
  parallel token processing and context size, 81-83
  recent improvements to, 95-105
    more efficient attention, 96-100
    positional embeddings, 102-104
    Transformer blocks, 101
  Transformer blocks, 85-94
    attention calculation, 91-93
    attention layer, 88
    attention mechanism, 89-91
    feedforward neural networks, 87
    self-attention and relevance scoring, 93-94
  Vision Transformer, 260-262
transparency and accountability, 28
tree-of-thought, 189-191
TruthfulQA, 374, 376
TSDAE (Transformer-Based Sequential Denoising Auto-Encoder)
  for domain adaptation, 320
  overview of, 316-320

U

UltraChat dataset, 368
UMAP (Uniform Manifold Approximation and Projection), 141
unigram language model, 50
unk_token [UNK], 47
use_cache parameter, 84

V

valid output, verifying, 192
validation splits, 113
vector databases
  dense retrieval, 229
  nearest neighbor search versus, 238
  retrieval-augmented generation, 253
video random-access memory (VRAM), xiv, 29
visualization
  BERTopic, 155
  cluster analysis, 144
  dimensionality reduction and, 145
ViT (Vision Transformer), 260-262, 274
vocabulary, of tokenizers, 55, 57, 77, 106
VRAM (video random-access memory), xiv, 29

W

warmup_ratio parameter, 387
warmup_steps argument, 299
Weaviate, 239
whitespace characters, 50
windowed conversation buffer memory, 212-214, 217
word embeddings, 63-67
  pretrained, 63
  word2vec algorithm and contrastive training, 64-67
word tokens, 44
word-level metrics, in generative model evaluation, 374
word2vec algorithm, 8, 10-12
  contrastive training and, 64-67, 293
  embedding songs, 67
WordPiece, 43
  cased BERT base model, 48
  uncased BERT base model, 47
<work> token, 53

Y

Year of Generative AI, 23-24

Z

zero-shot classification, 123-126
  CLIP, 265
  SetFit, 339
zero-shot prompting
  chain-of-thought, 187
  in-context learning, 181

About the Authors

Jay Alammar is Director and Engineering Fellow at Cohere (pioneering provider of large language models as an API). In this role, he advises and educates enterprises and the developer community on using language models for practical use cases. Through his popular AI/ML blog, Jay has helped millions of researchers and engineers visually understand machine learning tools and concepts from the basic (ending up in the documentation of packages like NumPy and pandas) to the cutting-edge (Transformers, BERT, GPT-3, Stable Diffusion). Jay is also a co-creator of popular machine learning and natural language processing courses on Deeplearning.ai and Udacity.

Maarten Grootendorst is a Senior Clinical Data Scientist at IKNL (Netherlands Comprehensive Cancer Organization). He holds master's degrees in organizational psychology, clinical psychology, and data science, which he leverages to communicate complex machine learning concepts to a wide audience. With his popular blogs, he has reached millions of readers by explaining the fundamentals of artificial intelligence—often from a psychological point of view. He is the author and maintainer of several open source packages that rely on the strength of large language models, such as BERTopic, PolyFuzz, and KeyBERT. His packages are downloaded millions of times and used by data professionals and organizations worldwide.

Colophon

The animal on the cover of *Hands-On Large Language Models* is a red kangaroo (*Osphranter rufus*). They are the largest of all kangaroos, with a body length that can get up to a little over 5 feet and a tail as long as 3 feet. They are very fast and can hop to speeds over 35 miles per hour. They can jump 6 feet high and leap a distance of 25 feet in a single bound. The position of their eyes allows them see up to 300 degrees.

Red kangaroos are named after the color of their fur. While the name makes sense for the males—they have short, red-brown fur—females are typically more of a blue-grey color with a tinge of brown throughout. The red color in their fur comes from a red oil excreted from the glands in their skin. Because of their color, Australians refer to male red kangaroos as "big reds." However, because females are faster than males, they are often called "blue fliers."

Preferring open, dry areas with some trees for shade, red kangaroos can be found across Australia's mainland except in the upper north, lower southwest, and east coast regions of the country. Surrounding environmental conditions can affect reproduction. Because of this, females can pause or postpone pregnancy or birth until conditions are better. They often use this ability to delay birth of a new baby (joey) until the previous one has left their pouch.

The cover illustration is by Karen Montgomery, based on an antique line engraving from *Cassell's Popular Natural History*. The series design is by Edie Freedman, Ellie Volckhausen, and Karen Montgomery. The cover fonts are Gilroy Semibold and Guardian Sans. The text font is Adobe Minion Pro; the heading font is Adobe Myriad Condensed; and the code font is Dalton Maag's Ubuntu Mono.

O'REILLY®

Learn from experts.
Become one yourself.

Books | Live online courses
Instant answers | Virtual events
Videos | Interactive learning

Get started at oreilly.com.

©2023 O'Reilly Media, Inc. O'Reilly is a registered trademark of O'Reilly Media, Inc. 175 7x9.1975