
Table of Contents

Preface.....	xi
--------------	----

Part I. Understanding Language Models

1. An Introduction to Large Language Models.....	3
What Is Language AI?	4
A Recent History of Language AI	5
Representing Language as a Bag-of-Words	6
Better Representations with Dense Vector Embeddings	8
Types of Embeddings	10
Encoding and Decoding Context with Attention	11
Attention Is All You Need	15
Representation Models: Encoder-Only Models	18
Generative Models: Decoder-Only Models	20
The Year of Generative AI	23
The Moving Definition of a “Large Language Model”	25
The Training Paradigm of Large Language Models	25
Large Language Model Applications: What Makes Them So Useful?	27
Responsible LLM Development and Usage	28
Limited Resources Are All You Need	28
Interfacing with Large Language Models	29
Proprietary, Private Models	29
Open Models	30
Open Source Frameworks	31
Generating Your First Text	32
Summary	34

2. Tokens and Embeddings.	37
LLM Tokenization	38
How Tokenizers Prepare the Inputs to the Language Model	38
Downloading and Running an LLM	39
How Does the Tokenizer Break Down Text?	43
Word Versus Subword Versus Character Versus Byte Tokens	44
Comparing Trained LLM Tokenizers	46
Tokenizer Properties	55
Token Embeddings	57
A Language Model Holds Embeddings for the Vocabulary of Its Tokenizer	57
Creating Contextualized Word Embeddings with Language Models	58
Text Embeddings (for Sentences and Whole Documents)	61
Word Embeddings Beyond LLMs	63
Using pretrained Word Embeddings	63
The Word2vec Algorithm and Contrastive Training	64
Embeddings for Recommendation Systems	67
Recommending Songs by Embeddings	67
Training a Song Embedding Model	69
Summary	71
3. Looking Inside Large Language Models.	73
An Overview of Transformer Models	74
The Inputs and Outputs of a Trained Transformer LLM	74
The Components of the Forward Pass	76
Choosing a Single Token from the Probability Distribution (Sampling/ Decoding)	79
Parallel Token Processing and Context Size	81
Speeding Up Generation by Caching Keys and Values	83
Inside the Transformer Block	85
Recent Improvements to the Transformer Architecture	95
More Efficient Attention	96
The Transformer Block	101
Positional Embeddings (RoPE)	102
Other Architectural Experiments and Improvements	105
Summary	106

Part II. Using Pretrained Language Models

4. Text Classification.	111
The Sentiment of Movie Reviews	112
Text Classification with Representation Models	113

Model Selection	115
Using a Task-Specific Model	116
Classification Tasks That Leverage Embeddings	120
Supervised Classification	121
What If We Do Not Have Labeled Data?	123
Text Classification with Generative Models	127
Using the Text-to-Text Transfer Transformer	128
ChatGPT for Classification	132
Summary	135
5. Text Clustering and Topic Modeling.....	137
ArXiv's Articles: Computation and Language	138
A Common Pipeline for Text Clustering	139
Embedding Documents	139
Reducing the Dimensionality of Embeddings	140
Cluster the Reduced Embeddings	142
Inspecting the Clusters	144
From Text Clustering to Topic Modeling	146
BERTopic: A Modular Topic Modeling Framework	148
Adding a Special Lego Block	156
The Text Generation Lego Block	160
Summary	164
6. Prompt Engineering.....	167
Using Text Generation Models	167
Choosing a Text Generation Model	167
Loading a Text Generation Model	168
Controlling Model Output	170
Intro to Prompt Engineering	173
The Basic Ingredients of a Prompt	173
Instruction-Based Prompting	175
Advanced Prompt Engineering	177
The Potential Complexity of a Prompt	177
In-Context Learning: Providing Examples	180
Chain Prompting: Breaking up the Problem	182
Reasoning with Generative Models	184
Chain-of-Thought: Think Before Answering	185
Self-Consistency: Sampling Outputs	188
Tree-of-Thought: Exploring Intermediate Steps	189
Output Verification	191
Providing Examples	192
Grammar: Constrained Sampling	194

Summary	198
7. Advanced Text Generation Techniques and Tools	199
Model I/O: Loading Quantized Models with LangChain	200
Chains: Extending the Capabilities of LLMs	202
A Single Link in the Chain: Prompt Template	203
A Chain with Multiple Prompts	206
Memory: Helping LLMs to Remember Conversations	209
Conversation Buffer	210
Windowed Conversation Buffer	212
Conversation Summary	214
Agents: Creating a System of LLMs	218
The Driving Power Behind Agents: Step-by-step Reasoning	219
ReAct in LangChain	221
Summary	224
8. Semantic Search and Retrieval-Augmented Generation	225
Overview of Semantic Search and RAG	226
Semantic Search with Language Models	228
Dense Retrieval	228
Reranking	240
Retrieval Evaluation Metrics	244
Retrieval-Augmented Generation (RAG)	249
From Search to RAG	250
Example: Grounded Generation with an LLM API	252
Example: RAG with Local Models	252
Advanced RAG Techniques	255
RAG Evaluation	257
Summary	258
9. Multimodal Large Language Models	259
Transformers for Vision	260
Multimodal Embedding Models	263
CLIP: Connecting Text and Images	265
How Can CLIP Generate Multimodal Embeddings?	265
OpenCLIP	268
Making Text Generation Models Multimodal	273
BLIP-2: Bridging the Modality Gap	273
Preprocessing Multimodal Inputs	278
Use Case 1: Image Captioning	280
Use Case 2: Multimodal Chat-Based Prompting	283
Summary	286

Part III. Training and Fine-Tuning Language Models

10. Creating Text Embedding Models.....	289
Embedding Models	289
What Is Contrastive Learning?	291
SBERT	293
Creating an Embedding Model	296
Generating Contrastive Examples	296
Train Model	297
In-Depth Evaluation	300
Loss Functions	301
Fine-Tuning an Embedding Model	309
Supervised	309
Augmented SBERT	311
Unsupervised Learning	316
Transformer-Based Sequential Denoising Auto-Encoder	316
Using TSDAE for Domain Adaptation	320
Summary	321
11. Fine-Tuning Representation Models for Classification.....	323
Supervised Classification	323
Fine-Tuning a Pretrained BERT Model	325
Freezing Layers	328
Few-Shot Classification	333
SetFit: Efficient Fine-Tuning with Few Training Examples	333
Fine-Tuning for Few-Shot Classification	337
Continued Pretraining with Masked Language Modeling	340
Named-Entity Recognition	345
Preparing Data for Named-Entity Recognition	347
Fine-Tuning for Named-Entity Recognition	352
Summary	353
12. Fine-Tuning Generation Models.....	355
The Three LLM Training Steps: Pretraining, Supervised Fine-Tuning, and Preference Tuning	355
Supervised Fine-Tuning (SFT)	357
Full Fine-Tuning	357
Parameter-Efficient Fine-Tuning (PEFT)	359
Instruction Tuning with QLoRA	367
Templating Instruction Data	367
Model Quantization	369
LoRA Configuration	370

Training Configuration	371
Training	372
Merge Weights	373
Evaluating Generative Models	373
Word-Level Metrics	374
Benchmarks	374
Leaderboards	376
Automated Evaluation	376
Human Evaluation	376
Preference-Tuning / Alignment / RLHF	378
Automating Preference Evaluation Using Reward Models	379
The Inputs and Outputs of a Reward Model	380
Training a Reward Model	380
Training No Reward Model	384
Preference Tuning with DPO	385
Templating Alignment Data	386
Model Quantization	386
Training Configuration	387
Training	388
Summary	389
Afterword.....	391
Index.....	393