

Brief Contents

Endorsement	iv
Preface	v
Acknowledgement	ix
Foreword	x
1 Introduction	1
2 An Overview of Natural Language Processing and Neural Networks	29
3 Word Embedding	69
4 Statistical Language Model	95
5 Neural Language Models	125
6 Transformers	151
7 Language Model Pretraining	197
8 Fine-Tuning and Alignment of LLMs	225
9 Prompting Strategies in LLMs	251
10 Efficient Methods for Fine-Tuning LLMs	291
11 Augmented Large Language Models	317
12 Multilingual and Multimodal LLMs	343
13 Responsible LLMs	373
14 ★Advanced Topics in Large Language Models	389
15 LLMs in Action	423
Index	453

Contents

Endorsement	iv
Preface	v
Acknowledgement	ix
Foreword	x
1 Introduction	1
1.1 What Is a Language Model?	2
1.2 Evolution of Language Modelling Technologies	4
1.3 Scaling Laws in Language Models	6
1.4 Evolution of LLMs	8
1.4.1 The Emergence and Development of LLMs	8
1.4.2 Implications of Encoder-Decoder in LLM Development	9
1.4.3 Optimising Scale and Resource Efficiency in LLMs	12
1.5 Organisation of the Book	14
Additional Resources	16
Bibliography	16
2 An Overview of Natural Language Processing and Neural Networks	29
Part I: Natural Language Processing	30
2.1 Computational Linguistics and Natural Language Processing	31
2.2 Overview of the Natural Language Processing Pipeline	32
2.3 Morphology	35
2.3.1 Morphemes	36
2.3.2 Stemming	37
2.3.3 Lemmatisation	38
2.3.4 Lexicon	38
2.4 Tokenisation	39
2.4.1 Advanced Techniques: Subword Tokenisation	40
2.5 Syntactics	43

2.6	Semantics	45
2.7	Introduction to Language Modelling	46
	Part II: Neural Networks	48
2.8	The Perceptron	48
	2.8.1 Definition	49
	2.8.2 Implementing AND, OR, and XOR Logic	49
2.9	Multilayer Perceptron	51
	2.9.1 Neural Networks	52
	2.9.2 Types of Activation Functions	53
2.10	Training Neural Networks	55
	2.10.1 Backpropagation	56
	2.10.2 Batching	58
	2.10.3 Hyperparameters	60
	2.10.4 Regularisation	61
2.11	Vanishing and Exploding Gradients	62
2.12	Evaluation Metrics	63
2.13	Summary	65
	Additional Resources	66
	Exercises	66
	Bibliography	68
3	Word Embedding	69
3.1	Distributional Hypothesis	70
3.2	Vector Semantics	70
	3.2.1 Defining and Measuring Semantic Similarity	71
3.3	Types of Word Embedding	72
	3.3.1 Frequency-Based Embeddings	72
	3.3.2 Word2Vec	77
	3.3.3 Global Vectors for Word Representation	83
	3.3.4 FastText	86
3.4	Bias in Word Embedding	88
3.5	Limitations of Word Embedding Methods	89
3.6	Applications of Word Embeddings	90
3.7	Summary	90
	Additional Resources	91
	Exercises	91
	Bibliography	93

4	Statistical Language Model	95
4.1	Statistical Language Model	97
4.1.1	The Conditional Probability	98
4.1.2	The Chain Rule of Probability	98
4.1.3	The Markov Assumption	99
4.1.4	Unigram Language Model	100
4.1.5	Bigram Language Model	101
4.2	Smoothing	102
4.2.1	The Unknown Tokens	103
4.2.2	Smoothing	104
4.2.3	Back-Off	105
4.2.4	Interpolation	106
4.2.5	★Good-Turing	106
4.3	Evaluation of Language Model	107
4.3.1	Extrinsic Evaluation	108
4.3.2	Intrinsic Evaluation	109
4.3.3	Human Evaluation	111
4.3.4	Evaluation Metrics	112
4.3.5	Benchmark Suits	114
4.4	Limitations of Statistical Language Models	118
4.5	Summary	119
	Additional Resources	120
	Exercises	121
	Bibliography	122
5	Neural Language Models	125
5.1	Convolutional Neural Networks	126
5.1.1	Components of CNNs: Kernel, Stride, Pooling, and Padding	127
5.1.2	Hierarchical and Dilated Convolutions	129
5.1.3	Applications of CNNs in NLP	130
5.2	Recurrent Neural Networks	130
5.2.1	Training RNNs	131
5.2.2	Applications of RNNs	133
5.2.3	Challenges in Sequence Modelling	133
5.2.4	RNN Variants: LSTM, GRU, and Bidirectional RNNs	134
5.3	Sequence-to-Sequence Models	139
5.3.1	Training Sequence-to-Sequence Models	139

5.3.2	Inference Decoding	140
5.3.3	Applications of Sequence-to-Sequence Models	142
5.4	Attention Mechanisms	142
5.4.1	Introduction to Attention	143
5.4.2	Advantages of Attention	145
5.4.3	Variants of Attention	146
5.5	Limitations of Neural Language Models	146
5.6	Summary	147
	Additional Resources	148
	Exercises	148
	Bibliography	149
6	Transformers	151
6.1	Self-Attention	152
6.1.1	Multi-Head Self-Attention	156
6.2	Transformer Encoder Block	160
6.2.1	Components of the Transformer Encoder Block	160
6.2.2	Feed-Forward Neural Network	161
6.2.3	Layer Normalisation	162
6.2.4	Residual Connections	164
6.3	Transformer Decoder Block	165
6.3.1	Masked Multi-Head Self-Attention	166
6.3.2	Cross-Attention (Encoder-Decoder Attention)	168
6.4	Positional Embeddings	168
6.4.1	Types of Positional Embeddings	169
6.4.2	★Rotary Position Embedding	172
6.5	★ Efficient Attention Mechanisms	176
6.5.1	KV Caching in Multi-Head Self-Attention	176
6.5.2	★Multi-Query Attention	180
6.5.3	★Grouped-Query Attention	182
6.5.4	★Sliding Window Attention	185
6.6	An Alternate Formulation of Transformers	187
6.6.1	Residual Stream Perspective of Transformers	187
6.6.2	Attention Heads: Reading and Writing	187
6.6.3	Feed-Forward Networks: Transformation of Residual Streams	189
6.6.4	Prediction Head: Generating the Next Token	190
6.6.5	Decomposing the Transformer: Attention and Feed-Forward Contributions	190

6.6.6	Residual Networks as Shallow Ensembles	190
6.6.7	★Interpreting the Mechanism of LLMs	192
6.7	Summary	192
	Additional Resources	193
	Exercises	193
	Bibliography	195
7	Language Model Pretraining	197
7.1	Embeddings from Language Model	199
7.1.1	Architecture and Training of ELMo	199
7.1.2	Applications of ELMo	201
7.1.3	Limitations of ELMo	202
7.2	Evaluation Datasets	202
7.3	Encoder-Based Pretraining	203
7.3.1	Fundamentals of Encoder-Based Models	204
7.3.2	Training Paradigm	204
7.3.3	BERT Pretraining	206
7.3.4	Applications and Limitations	207
7.4	Decoder-Based Pretraining	208
7.4.1	Decoder-Based Architecture	208
7.4.2	Training Paradigm	209
7.4.3	GPT Pretraining	210
7.4.4	Applications and Limitations	210
7.5	Encoder-Decoder Based Pretraining	211
7.5.1	Architecture	211
7.5.2	Joint Pretraining Strategy	212
7.5.3	T5 Pretraining	213
7.5.4	Applications and Limitations	215
7.6	Emergence of Large Language Models	216
7.7	Limitations of Pretraining	217
7.8	Summary	218
	Additional Resources	220
	Exercises	220
	Bibliography	221
8	Fine-Tuning and Alignment of LLMs	225
8.1	Moving from Pretraining to Fine-Tuning	226

8.2	Fine-Tuning on Various Task-Specific Applications	227
8.2.1	Sequence Classification	228
8.2.2	Pairwise Sequence Classification	230
8.2.3	Sequence Labelling	232
8.2.4	Learning Spans	233
8.2.5	Challenges in Classical Fine-Tuning Methods	234
8.3	Instruction Tuning	235
8.4	Alignment Methods	237
8.4.1	Reinforcement Learning from Human Feedback	238
8.4.2	★Direct Preference Optimisation	243
8.5	Summary	245
	Additional Resources	246
	Exercises	246
	Bibliography	248
9	Prompting Strategies in LLMs	251
9.1	Prompt Engineering	253
9.1.1	Prompt Shape	254
9.1.2	Manual Template Engineering	255
9.1.3	Automated Template Learning	257
9.1.4	Continuous Prompts	261
9.2	Prompt Application	263
9.2.1	In-Context Learning	263
9.2.2	Knowledge Probing	265
9.2.3	Classification-Based Tasks	268
9.2.4	Information Extraction	269
9.2.5	Reasoning in Natural Language Processing	272
9.2.6	Question Answering	274
9.2.7	Text Generation	275
9.2.8	Automatic Evaluation of Text Generation	276
9.3	Chain-of-Thoughts	276
9.4	★Tree-of-Thoughts	279
9.5	★Graph-of-Thoughts	281
9.6	Summary	284
	Additional Resources	285
	Exercises	285
	Bibliography	287

10	Efficient Methods for Fine-Tuning LLMs	291
10.1	Model Compression with Knowledge Distillation	293
10.1.1	White-Box Knowledge Distillation	294
10.1.2	Meta Knowledge Distillation	296
10.1.3	Black-Box Knowledge Distillation	297
10.2	Model Compression Techniques	297
10.2.1	Model Pruning	297
10.2.2	Model Quantisation	299
10.3	Parameter-Efficient Fine-Tuning	301
10.3.1	Adapters	301
10.3.2	Prefix Tuning	303
10.3.3	Prompt Tuning	303
10.3.4	Selective PEFT Techniques	303
10.3.5	Reparameterisation-Based PEFT Techniques	306
10.3.6	Hybrid Approaches for Efficient Fine-Tuning	307
10.4	★Efficient Strategies for Fine-Tuning LLMs	307
10.4.1	Mixed-Precision Tuning	308
10.4.2	Data Selection for Efficient Fine-Tuning	309
10.4.3	Prompt Compression	310
10.5	Summary	311
	Additional Resources	312
	Exercises	313
	Bibliography	314
11	Augmented Large Language Models	317
11.1	Retrieval-Augmented Generation	319
11.1.1	Indexing in RAGs	319
11.1.2	Context Searching in RAGs	320
11.1.3	Prompting in RAGs	322
11.1.4	Inferencing in RAGs	323
11.1.5	Comparison of RAGs with LLMs	325
11.2	Evaluation of RAGs	325
11.2.1	Assessing of Retrieval Quality	326
11.2.2	Generation Quality	326
11.2.3	Knowledge Integration and Factuality Evaluation	327
11.2.4	Response Time and Efficiency	328

11.2.5	User Satisfaction	328
11.2.6	RAGAs Framework for RAG Evaluation	329
11.3	Tool Calling with LLMs	330
11.3.1	Autonomously Determining Which Tools to Use and Where	330
11.3.2	Examples of Different Tools	331
11.3.3	Evaluation of Code Generation Capabilities of Agents	333
11.3.4	Error Handling and Optimisation	333
11.4	LLM Augmentation with Agents	334
11.4.1	Reasoning in LLM Agents	335
11.4.2	Planning in LLM Agents	335
11.4.3	Handling Memory in LLM Agents	336
11.5	Summary	337
	Additional Resources	338
	Exercises	338
	Bibliography	340
12	Multilingual and Multimodal LLMs	343
12.1	Multilingual Language Models	344
12.1.1	The Evolution of Multilingual NLP	344
12.1.2	The Need for Multilingual LLMs	345
12.1.3	Cross-Lingual Representation Learning	346
12.1.4	Applications	348
12.2	Multimodal Language Models	350
12.2.1	Integration of Diverse Modalities	350
12.2.2	Applications	352
12.3	Training Multilingual and Multimodal LLMs	354
12.3.1	Efficient Data Collection and Preprocessing	354
12.3.2	Model Training Strategies	356
12.4	Addressing Challenges in Multilingual and Multimodal LLMs	362
12.4.1	Challenges in Multilingual LLMs	362
12.4.2	Challenges in Multimodal LLMs	363
12.5	Future Directions and Emerging Trends	363
12.6	Limitations of Multilingual and Multimodal LLMs	365
12.7	Summary	365
	Additional Resources	366
	Exercises	367
	Bibliography	368

13	Responsible LLMs	373
13.1	Inaccurate, Inappropriate, and Unethical Behaviour of LLMs	374
13.2	Responsible AI	375
13.3	Bias	376
13.3.1	Visibility of Bias	376
13.3.2	Source of Bias	380
13.4	Bias Mitigation	382
13.5	Summary	384
	Additional Resources	384
	Exercises	385
	Bibliography	386
14	*Advanced Topics in Large Language Models	389
14.1	Reasoning with LLMs	391
14.1.1	Advancements in Reasoning Capabilities	391
14.1.2	Challenges in Reasoning with LLMs	392
14.1.3	Types of Reasoning Tasks	392
14.1.4	How Do LLMs Approach Reasoning?	393
14.1.5	Evaluating Reasoning Abilities in LLMs	394
14.2	Handling Long Context in LLMs	395
14.2.1	Challenges in Processing Long Context	396
14.2.2	Training and Fine-Tuning Approaches to Extend Context Length	400
14.2.3	Evaluation of Long-Context LLMs	401
14.3	Model Editing	402
14.3.1	Conditions for Successful Editing	403
14.3.2	Methods for Model Editing	404
14.3.3	Metrics for Evaluation of Model Editing	409
14.4	Hallucination in LLMs	410
14.4.1	Definition	410
14.4.2	Sources of Hallucination	411
14.4.3	Metrics Measuring Hallucination	412
14.4.4	Hallucination Mitigation	413
14.5	Self-Evolving LLMs	414
14.5.1	Conceptual Framework	414
14.5.2	Evolution Objectives and Techniques	417
14.5.3	Challenges	418
14.6	Summary	418

Additional Resources	419
Exercises	420
Bibliography	421
15 LLMs in Action	423
15.1 An Overview of the Landscape	424
15.1.1 Tracing the Evolution and Importance of LLMs in Contemporary AI	424
15.1.2 Open-Source vs Closed-Source Paradigms: Benefits and Trade-offs	425
15.2 A Panoramic View of LLMs	425
15.2.1 General-Purpose Large Language Models	426
15.2.2 Language-Specific LLMs	439
15.2.3 Domain-Specific LLMs	440
15.2.4 Task-Specific LLMs	441
15.3 Diverse Applications of LLMs	443
15.3.1 Healthcare: Enhancing Diagnostics and Patient Care	444
15.3.2 Finance: Transforming Data Analysis and Risk Management	444
15.3.3 Legal: Streamlining Research and Case Management	444
15.3.4 Education: Personalised Learning and Academic Support	445
15.4 Emerging Trends and Future Directions in LLMs	445
15.4.1 Beyond Text: The Advent of Multimodal LLMs	446
15.4.2 Autonomous Agents: The LLM Leap in AI Evolution (AutoGPT)	446
15.5 Summary	447
Additional Resources	447
Exercises	447
Bibliography	449
Index	453