

CONTENTS

<i>Preface</i>	<i>xxvii</i>
<i>About the Contributor</i>	<i>xxxiii</i>
Chapter 1: The Generative AI Landscape	1
What is Generative AI?	2
Key Features of Generative AI	2
Popular Techniques in Generative AI	2
What Makes Generative AI Unique	3
Conversational AI Versus Generative AI	4
Primary Objective	4
Applications	4
Technologies Used	4
Training and Interaction	5
Evaluation	5
Data Requirements	5
What are Generative AI Models?	5
Is DALL-E Part of Generative AI?	9
Are ChatGPT-3 and GPT-4 Part of Generative AI?	10
Generative AI Versus ML, DL, and NLP	11
Which Fields Benefit the Most from Generative AI?	12
Generative AI for Enterprise	14
The Effect of Generative AI on Jobs	16

What is Artificial General Intelligence (AGI)?	18
When Will AGI Arrive?	20
What is the Path to AGI?	21
How Can We Prepare for AGI?	22
Will AGI Control the World?	25
Should Humans Fear AGI?	26
What is Beyond AGI?	28
Artificial General Intelligence Versus Generative AI	30
What are LLMs?	31
What is the Purpose of LLMs?	32
Recent, Powerful LLMs	34
Do LLMs Understand Language?	36
Caveats Regarding LLMs	37
Model Size Versus Training Set Size	38
Memory Requirements for LLMs	38
Memory Types in LLMs	40
LLMs Versus Deep Learning Models	42
Cost Comparison among LLMs	44
LLMs and Deception	46
Deceptive Completions in LLMs	47
LLMs and Intentional Deception	48
Selecting an LLM: Factors to Consider	50
Pitfalls of Working with LLMs	52
A Brief History of Modern LLMs	54
Aspects of LLM Development	56
LLM Size Versus Performance	58
Emergent Abilities of LLMs	59
Skepticism Regarding Emergent Abilities	60
What are Hallucinations?	62
Why do LLMs Hallucinate?	64
Hallucination Types in LLMs	65
Can LLMs Detect Errors in Prompts?	66
Are Intentional Hallucinations Possible?	67
Reducing Hallucinations	69
Causes of Hallucinations in LLMs	70

Intrinsic Versus Extrinsic Hallucinations	72
Hallucination Detection	74
Model Calibration	76
Kaplan and Under-Trained Models	78
Success Stories in Generative AI	79
Real-World Use Cases for Generative AI	81
Summary	84
Chapter 2: Prompt Engineering (1)	85
LLMs and Context Length	85
Batch Size and Context Length	88
Python Code for Batch Size and Context Length	89
Common Context Length Values	91
Lost-in-the-Middle Challenge	93
Self-Exploring Language Models (SELMs)	94
Overview of Prompt Engineering	96
What is a Prompt?	98
The Components of a Prompt	98
The Purpose of Prompt Engineering	99
Designing Prompts	100
Prompt Categories	100
Hard Prompts	101
Prompts and Completions	103
Guidelines for Effective Prompts	103
Effective Prompts for ChatGPT	104
Concrete Versus Subjective Words in Prompts	105
Prompts and Politeness	106
Negative Prompting	106
Self-Criticism Prompting	108
Using Flattery or a Sense of Urgency	110
Unethical or Dishonest Prompts	112
Prompts with Confessions of a Crime	114
Prompt Hijacking	116
What is Prompt Caching?	119
Python Code for Client-Side Prompt Caching	120

Common Types of Prompts	122
“Shot” Prompts	123
Instruction Prompts	123
Reverse Prompts	124
Sequential Prompt Chaining	124
System Prompts Versus Agent Prompts	124
Prompt Templates	125
Prompts for Different LLMs	126
Prompt Optimization	127
Poorly Worded Prompts	130
Prompts with Slang and Idiomatic English	131
Distribution of Users’ Prompts	133
Overly Complicated Prompts	135
Prompt Injections	136
Accidental Prompt Injections	139
How to Refine Prompts	141
Chain of Thought (CoT) Prompts	143
Self-Consistency and CoT	143
Self-Consistency, CoT, and Unsupervised Datasets (LMSI)	144
Zero Shot CoT	144
Python Code for Zero-shot CoT	145
Auto Chain of Thought (AutoCoT)	148
CoT for Financial Forecasts	150
Tree of Thought (ToT) Prompts	150
Python Code Sample for ToT	152
Buffer of Thoughts (BoT) Prompting	155
Python Code Sample for BoT Prompting	156
Re-Reading Prompts: Better Completions?	159
Is Re-Reading Prompts Always Recommended?	160
Which Techniques Work Best for Re-Reading Prompts?	161
Assigning a Role in a Prompt	163
Assorted Prompts with Roles	166
Roles in CoT Prompts	168
Prompts with Roles: Better Completions?	170

Assorted Prompt Engineering Techniques	172
End-Goal Prompting	173
Chain-of-Verification (CoV) Prompting	173
Emotionally Expressed Prompting	174
Mega-Personas Prompting	174
Flipped Interaction Prompting	175
Trust Layers for Prompting	175
Step-Around Prompting Technique	176
Summary of Recommendations	176
What is Prompt Compression?	177
Use Cases for Prompt Compression	177
Prompt Compression Techniques	179
Python Code Sample	180
Anthropic Prompt Generator	182
Summary	184
Chapter 3: Prompt Engineering (2)	185
A Note About Google Collaboratory	185
Ranking Prompt Techniques	186
Recommended Prompt Techniques	188
Adversarial Prompting	189
Python Code Sample for Adversarial Prompting	190
Meta Prompting	192
Python Code for Meta Prompting	193
Advanced Meta Prompt Engineering	196
Python Code for Recursive Meta Prompting	196
Useful Links	199
Prompt Techniques to Avoid	200
GPT-4 and Prompt Samples	202
GPT-4 and Arithmetic Operations	203
Algebra and Number Theory	203
The Power of Prompts	203
Language Translation with GPT-4	205
Can GPT-4 Write Poetry?	206
GPT-4 and Humor	207

Question Answering with GPT-4	208
Stock-Related Prompts for GPT-4	209
Philosophical Prompts for GPT-4	210
Mathematical Prompts for GPT-4	210
DSPy and Prompt Engineering	211
DSPy Code Sample	214
Advanced Prompt Techniques	214
Omni Prompting	218
Python Code for Omni Prompting	219
Multimodal Prompting	222
Python Code for Multimodal Prompting	223
Omni Prompting Versus Multimodal Prompting	226
Multi-Model Prompting	228
Python Code for Multi-Model Prompting	229
Prompt Decomposition	232
Needle in a Haystack	234
What are Inference Parameters?	240
Temperature Inference Parameter	241
Temperature and the softmax() Function	242
The top-p Inference Parameter	243
Python Code Sample for the top-p Inference Parameter	244
The top-k Inference Parameter	246
Python Code Sample for the top-k Inference Parameter	247
Using top-k and top-p in LLMs	249
GPT-4o Overview of Inference Parameters	252
GPT-4o and the Temperature Inference Parameter	255
Python Code Sample for the Temperature Parameter	256
Overview of top-k Algorithms	260
GPT-4o Ranking of top-k Inference Parameters	264
Python Code Samples for top-k Algorithms	265
TF-IDF (Term Frequency-Inverse Document Frequency)	266
BM25 (Best Matching 25)	267
GPT-4o Ranking of top-k Algorithms	269
GPT-4o Ranking of Inference Parameters	271

GPT Mini	273
SearchGPT	275
CriticGPT	276
Important Yet Under-Utilized Prompt Techniques	277
Prompt Testing	279
Python Code Sample for Prompt Testing	279
Summary	282

Chapter 4: Well-Known LLMs and APIs **283**

The pytorch_model.bin File	284
The BERT Family	285
Are BERT Models Also LLMs?	286
ALBERT	288
The GPT-x Series of Models	288
Are GPT-x Models Also LLMs?	290
Language Models Versus Embedding Models	291
Python Code for Language Model and Text Generation	292
Python Code for Embedding Model and Text Similarity	293
OpenAI Models	295
What is GPT-3?	297
OpenAI Extensions of GPT-3	299
What is ChatGPT?	302
ChatGPT : GPT-3 “on steroids”?	303
ChatGPT: Google “Code Red”	303
ChatGPT Versus Google Search	304
ChatGPT Custom Instructions	304
ChatGPT on Mobile Devices and Browsers	305
ChatGPT and Prompts	305
GPTBot	306
ChatGPT Playground	306
Let’s Chat with ChatGPT	307
A Simple Chat Code Sample	307
Specify Multiple Roles	309
Specify max_tokens and stop Values	311
Specify Multiple Stop Values	314

Specify Temperature Values	314
Working with top-p Values	317
Other Inference Parameters	319
Plugins, Advanced Data Analytics, and Code Whisperer	321
Plugins	321
Advanced Data Analytics	323
Code Whisperer	323
Concerns about ChatGPT	324
Code Generation and Dangerous Topics	324
ChatGPT Strengths and Weaknesses	325
Sample Queries and Completions from ChatGPT	326
Detecting Generated Text	328
What is GPT-4?	329
GPT-4 and Test-Taking Scores	330
GPT-4 Parameters	330
Main Features of GPT-4	331
Main Features of GPT-4o	333
When is GPT-5 Available?	334
What is InstructGPT?	335
Some Well-Known LLMs	336
Google Gemini	336
Copilot (OpenAI/Microsoft)	337
Codex (OpenAI)	338
Apple GPT	338
PaLM-2	338
Claude 3 Sonnet, Opus, and Haiku	339
Grok 2	340
Llama 3.1 Models	342
Main Features of Llama 3.1	342
Main Features of Llama 3.1 405B	343
Limitations of Llama 3.1 405B	345
Llama 3.1 Versus Llama 3.1 405B	345
What About Llama 4?	346

Accessing OpenAI APIs	346
Accessing Hugging Face APIs	352
What are Small Language Models (SLMs)?	357
Top Computations of LLMs	357
GPUs and LLMs	358
Machine Learning Tasks and LLMs	360
What are LPUs?	362
LPUs Versus GPUs	363
What is an NPU?	364
NLP Tasks and LLMs	366
Metrics for NLP Tasks and LLMs	367
LLM Benchmarks	370
Benchmarks for Evaluating LLMs	371
What is Pruning in LLMs?	373
Python Code Sample	374
LLMs: Underrated, Overrated, Trends, and Performance	376
Smallest Useful Decoder-Only LLMs	377
Underrated LLMs	377
Overrated LLMs	379
Trends in LLMs: Larger or Smaller?	380
Best Performance LLMs	382
The Hugging Face Leaderboard	384
LLM Compressors	385
Ranking of LLM Compressors	386
Summary	387
Chapter 5: Fine-Tuning LLMs (1)	389
What is Pre-Training?	389
Time and Cost for Pre-Training	391
Pre-training Strategies	393
Additional Pre-training Topics	395
Outliers and Pre-Training LLMs	396
Three Techniques for Detecting Outliers	397
Python Code for Detecting Outliers	398
What to Do with Outliers	401

What is Model Collapse in Generative AI?	403
Training LLMs on LLM-Generated Data	405
What is Fine-Tuning?	407
Python Code Sample for Fine-Tuning GPT-2	408
Is Fine-Tuning Always Required for Pre-trained LLMs?	419
Well-Known Fine-Tuning Techniques	421
When is Fine-Tuning Recommended?	426
Fine-Tuning BERT for Sentiment Analysis	428
Fine-Tuning GPT-4 Models	434
Python Code Sample	435
Odds Ratio Preference Optimization (ORPO)	438
Python Code Sample	439
Instruction Fine-Tuning (IFT)	442
An Example of Instruction Fine-Tuning	444
Continual Instruction Tuning	447
Python Code for Continual Instruction Tuning	449
Fine-Tuning Embeddings	454
Generating Fine-Tuning Datasets	457
Representation Fine-Tuning (REFT) Versus PEFT	459
Fine-Tuning LLMs for Specific NLP Tasks	461
Preparing a Labeled Dataset for Sentiment Analysis	463
Preparing a Labeled Dataset for Text Classification	466
Loss Functions for LLMs	469
What is Few-Shot Learning?	473
Few-Shot Learning and Prompts	475
Fine-Tuning Versus Few-Shot Learning	475
In-Context Learning (ICL)	478
ICL Versus Other Prompt Techniques	479
Many-Shot In-Context Learning	481
How Do We Train LLMs with New Data?	483
Python Code with Regular Expressions	484
Disabling Greedy Matching	486
Local Directories for Downloaded LLMs	487
Hugging Face Local Cache for Downloaded LLMs	487

Ollama Local Cache for Downloaded LLMs	488
List of Downloaded LLMs via Ollama	489
Summary	490
Chapter 6: LLMs and Fine-Tuning (2)	491
Steps for Fine-Tuning LLMs	492
Alternatives to Fine-Tuning LLMs	497
Fine-Tuning Versus Prompt Engineering	500
Massive Prompts Versus LLM Fine-Tuning	503
Synthetic Data and Fine-Tuning	503
What is Prompt Tuning?	505
Parameter Efficient Fine-Tuning (PEFT)	508
Sparse Fine-Tuning Versus Supervised Fine-Tuning	511
Sparse Fine-Tuning (SFT) and PEFT	514
Representation Fine-Tuning	515
Python Code Sample	517
Step-by-Step Fine-Tuning	520
Fine-Tuning Tips	522
What is LoRA?	525
Python Code Sample with LoRA	525
When is LoRA Recommended for Fine-Tuning?	529
LoRA Versus Full Fine-Tuning	532
LoRA-based Algorithms for Fine-Tuning	534
LoRA-FA (2023)	535
AdaLoRA (2023)	537
Delta-LoRA (2023)	537
LoRA+ (2024)	538
LoRA-drop (2024)	539
What is QLoRA?	539
LoRA Versus QLoRA	541
Best GPU for LoRA, QLoRA, and Inference	544
What is DoRA?	545
The Impact of Fine-Tuning on LLMs	548
Fine-Tuned LLMs and General Capability	550

Unstructured Fine-Tuning	552
Fine-Tuning and Dataset Size	553
Model Quality and Dataset Size	555
GPT Model Specification for Fine-Tuning Behavior	557
What is Ollama?	559
Starting the Ollama Server and Command Line Options	560
Downloading and Launching LLMs	561
Working with Phi Models	562
Phi-Based Requests with the Ollama Server	563
Phi-Based Prompts in Raw Mode	564
Fine-Tuning Phi-3	565
Fine-Tuning Llama 2	569
Python Code Sample	570
Working with Nvidia Models	573
Working with Qwen2 Models	575
Working with Gemma Models	577
Working with Llama 3.1 (4.7B)	579
Working with Mistral Models	582
Mistral NeMo 12B	583
Mistral Large 2	583
Downloading Mistral Large 2	584
Ollama Server Details for mistral-large	586
Ollama with Other LLMs	591
Working with Hugging Face Models	592
Downloading Hugging Face Models	592
Managing LLMs with Command Line Tools	593
anythingLLM	594
Gemma.cpp	595
Jan.ai	596
llama.cpp	597
llm	598
LMStudio	598
Ollama	599

Working with gpt4all	600
Download and Install gpt4all	600
Download Llama-3-8B-Instruct	601
Summary	603

Chapter 7: What is Tokenization? 605

What is the Transformer Architecture?	606
Python Code Sample	607
Key Components of the Transformer Architecture	610
What is Pre-tokenization?	613
What is a Word?	613
Pre-tokenization Versus Tokenization	614
A Python Code Sample for Pre-tokenization	616
What is Tokenization?	619
Nuances of Tokenizers	619
A Generic Token-to-Index Mapping Algorithm	619
A Python Code Sample for Tokenization	622
Tokenization Tasks and Their Challenges	624
An Alternative to Tokenization: ByT5 Model	626
Word, Character, and Subword Tokenizers	627
Word-based Tokenizers	627
Limitations of Word Tokenizers	628
Tokenization for Languages with Multiple Alphabets	629
Trade-Offs with Character-based Tokenizers	630
Limitations of Character-based Tokenizers	630
Subword Tokenization	631
A Python Example of a Subword Tokenizer	631
Key Points Regarding BERT Tokenization	633
Subword Tokenization Algorithms	633
What is BPE?	634
What is WordPiece?	635
What is SentencePiece?	636
Hugging Face Tokenizers and Models	637
Loading and Saving Tokenizers	639

AutoTokenizer, BERTTokenizer, and GPT2Tokenizer	640
What are AutoClasses?	640
Hugging Face Tokenizers	641
Slow and Fast Tokenizers	641
Token Classification Pipelines	641
Python Code to Tokenize DistilBERT	643
Sentiment Analysis with DistilBERT	647
Sentence Completion with opt-125m	649
Three Types of Parameters	650
Tokenization Methods and Model Performance	651
Assorted Python Code Samples for Tokenization	654
Token Truncation in LLMs	659
Embedding Sizes of LLMs	660
Types of Embeddings for LLMs	661
Text, Audio, and Video Embeddings	661
Python Code Sample	662
Token, Positional, and Segment Embeddings	663
Word Embeddings for the Transformer Architecture	667
Python Code Sample for BERT Embeddings	667
Text Encoding Using the text-embedding-3-small LLM	670
Positional Encodings for the Transformer Architecture	672
Python Code Sample for Positional Encodings	672
Transformer Architecture Versus Mamba Architecture	675
Summary	678
Chapter 8: Attention Mechanism	679
What is Attention?	680
The Origin of Attention	680
Self-attention	681
GAtt (Ghost Attention)	681
Types of Attention and Algorithms	682
Attention in GPT-2 Versus BERT	683
What is FlashAttention-3?	683
Masked Attention	685
Python Code Sample	685

What is Tree Attention?	688
Calculating Attention with Q, K, and V	690
Python Code for Self-Attention	691
Python Code for BERT and Attention Values	694
Multi-Head Attention (MHA)	696
CNN Filters and Multi-Head Attention	697
Sliding Window Attention	698
Python Code Sample	699
Grouped-Query Attention	702
Python Code Sample	702
Paged Attention	705
Python Code Sample	706
Self-Attention and Quadratic Complexity	709
List of Attention Techniques for LLMs	711
Popular Types of Attention for LLMs	714
Self-Attention Code Sample	714
Scaled Dot-Product Attention Code Sample	716
Cross Attention Code Sample	718
Multi-Head Attention Code Sample	723
Masked Attention Code Sample	727
What is FlexAttention?	729
Python Code Sample	730
LLMs and Matrix Multiplication	736
Feed Forward Propagation in Neural Networks	738
LLMs are Often Decoder-only Architectures	738
Summary	741
Chapter 9: LLMs and Quantization (1)	743
What is Quantization?	744
Types of Quantization	744
LLM Server Frameworks	747
vllm	748
CTranslate2	748
DeepSpeed-MII	749

OpenLLM	750
Ray Serve	750
mlc-llm	751
Frameworks with Quantization Support	751
Quantization Types	752
1.58 Quantization	755
Python Code Sample	756
List of Quantization Formats for LLMs	758
Non-Uniform Quantization Schemes	759
Python Code Sample	760
GGUF and GGML Formats for Quantization	763
What is GGUF?	763
What is GGML?	765
GGUF Versus GGML Comparison	766
Converting TensorFlow Models to GGUF Format	767
Other File Formats for Quantizing LLMs	768
LLM Size Versus GGUF File Size	770
Recommended File Formats	771
Launching GGUF Files from the Command Line	772
Manual Calculation of Quantized Values	777
Weight-Based Quantization Techniques	779
Time Estimates for Quantization	781
Quantization Time Estimates in Minutes/Hours/Days	783
Fastest and Slowest Quantization Techniques	785
CPU/GPU-Intensive Quantization Techniques	787
Decrease in Accuracy in Quantization Techniques	789
Simple Quantization Code Sample	792
Min-Max Scaling (Normalization)	793
Linear Quantization	796
Python Code Sample	797
Uniform Quantization	799
Python Code Sample	799
Min-Max, Linear, and Uniform Quantization: A Comparison	801
Logarithmic Quantization	803
Python Code Sample	803

Exponential Quantization	806
Python Code Sample	807
K-Means Quantization	809
Python Code Sample	810
Lloyd-Max Quantization	813
Python Code Sample	814
Vector Quantization	816
Python Code Sample	816
Huffman Encoding	820
Python Code Sample	821
Entropy-Coded Quantization	824
Python Code Sample	825
Sigma-Delta Quantization	828
Python Code Sample	829
Companding Quantization	831
Python Code Sample	832
Finite State Vector Quantization	834
Python Code Sample	835
Adaptive Weight Quantization (AWQ)	839
Python Code Sample	840
Double Quantization	842
Python Code Sample	843
When is Quantization Recommended?	845
Significant Loss of Accuracy	846
Minimal Loss of Accuracy	848
Quantized Model Versus Full Model	849
Hardware Requirements	849
Naming Conventions for Quantization	850
Python Code Sample	851
Acronyms for Quantization Techniques	852
Characteristics of Good Quantization Algorithms	855
Quantization Versus Mixed Precision Training	857
Optimizing Model Inferences	859
Python Code with Mixed Precision Inference	861

Calibration Techniques in Quantization	865
Types of Calibration Techniques	866
Calculating Quantization Errors	868
Python Code Sample	869
Extensive List of Quantization Techniques	871
Quantization Techniques for Neural Network Optimization	873
What are the “Must Know” Quantization Techniques?	875
Summary	875
Chapter 10: LLMs and Quantization (2)	877
Georgi Gerganov Machine Learning Quantization (GGML)	878
Python Code Sample	878
Generalized Gradient-Based Uncertainty-Aware Filter Quantization (GGUF)	881
Python Code Sample	883
Intel’s AutoRound Quantization	887
Python Code Sample	887
AQLM Quantization	889
AQLM 2-bit Quantization	890
Python Code Sample	891
Generalized Precision Tuning Quantization (GPTQ)	893
When GPTQ Quantization is Recommended	898
Post Training Quantization (PTQ)	899
Quantization-Aware Training (QAT)	901
HQQ Quantization	903
Dynamic Quantization with a Neural Network	904
Quantized LLMs and Testing	905
Fine-Tuning Quantized LLMs for Sentiment Analysis	907
Practical Examples of Quantization	910
Quantization with TensorFlow (PTQ)	910
Quantization with TensorFlow (QAT)	912
Dynamic Quantization with PyTorch	914
Five LLMs and Five Quantization Techniques	916
Which Criteria are Significant?	918

RAM Requirements for Quantized LLMs	920
Largest Quantized LLM for 128GB RAM	920
Time Estimates for Quantization	921
Time Estimate for MacBook with M3 Pro Chip	922
Suitable Tasks for Quantized 7B and 13B LLMs	924
Selecting Models for Available RAM	925
Selecting LLMs for Quantization on 16 GB of RAM	925
Selecting LLMs for Quantization on 48 GB of RAM	927
Selecting LLMs for Quantization on 128 GB of RAM	928
Setting Up llama.cpp on Your MacBook	931
Quick Overview	931
Software Requirements	932
Installing Conda and lfs	933
Installing llama.cpp	933
Working with the llama.cpp Server	934
How to Start the Server	934
How to Stop the Server	935
How to Access the Server via a URL	935
How to Access the Server via Python Code	935
Further Exploration	938
Download and Quantize Mistral 7B LLM	938
Downloading the Mistral 7B LLM	938
Downloading the Mistral Instruct LLM	939
Quantizing the Mistral Instruct LLM	940
Test the Performance of Quantizations and Models	940
Llama Models from Meta	940
Llama 3 Models on Hugging Face	941
Download and Run the Llama 3.1 405B Model	941
A Quantized LLM: Now What Do I Do?	943
Testing Token Generation of a Quantized LLM	945
Quantized LLMs and Testing	945
Evaluating a Quantized LLM	947
Testing the Performance of a Quantized LLM	947
Measuring the Inference Speed and Memory Usage	950

Python Code to Measure Inference Speed	951
Probabilistic Quantization	952
Python Code Sample for PQ	953
Formulas for PQ	956
Popular Formulas for PQ	957
Probability Distributions and PQ	959
Kullback-Leibler Divergence and PQ	960
Probabilistic Quantization Versus Discretization	963
Python Code for Discretization	964
Is Discretization Used for Data in Histograms?	965
Distillation Versus Quantization	966
A Comparison of 2-bit versus 4-bit Quantization	969
Disk Space for 2-bit Quantization of GPT-3	971
2-bit Quantization of GPT-3: Limited Space Reduction	973
Recommendations for 1-bit Quantization	975
Time Estimates for 2-bit Versus 4-bit Quantization	978
Performing Both 2-bit and 4-bit Quantization for GPT-3	980
What is Generative Compression (GC)?	983
Generative Compression Versus Quantization	984
Quantization Versus Distillation	985
Clustering Algorithms and Quantization	986
Python Code Samples	987
Ranking of Clustering-Based Quantization Algorithms	990
Usage Frequency of Clustering Algorithms for Quantization	991
Classification Algorithms and Quantization	992
Python Code Samples	993
Reinforcement Learning and Quantization	996
Summary	998
Index	999