
Table of Contents

Preface.....	ix
1. Introduction to Building AI Applications with Foundation Models.....	1
The Rise of AI Engineering	2
From Language Models to Large Language Models	2
From Large Language Models to Foundation Models	8
From Foundation Models to AI Engineering	12
Foundation Model Use Cases	16
Coding	20
Image and Video Production	22
Writing	22
Education	24
Conversational Bots	26
Information Aggregation	26
Data Organization	27
Workflow Automation	28
Planning AI Applications	28
Use Case Evaluation	29
Setting Expectations	32
Milestone Planning	33
Maintenance	34
The AI Engineering Stack	35
Three Layers of the AI Stack	37
AI Engineering Versus ML Engineering	39
AI Engineering Versus Full-Stack Engineering	46
Summary	47

2. Understanding Foundation Models.....	49
Training Data	50
Multilingual Models	51
Domain-Specific Models	56
Modeling	58
Model Architecture	58
Model Size	67
Post-Training	78
Supervised Finetuning	80
Preference Finetuning	83
Sampling	88
Sampling Fundamentals	88
Sampling Strategies	90
Test Time Compute	96
Structured Outputs	99
The Probabilistic Nature of AI	105
Summary	111
3. Evaluation Methodology.....	113
Challenges of Evaluating Foundation Models	114
Understanding Language Modeling Metrics	118
Entropy	119
Cross Entropy	120
Bits-per-Character and Bits-per-Byte	121
Perplexity	121
Perplexity Interpretation and Use Cases	122
Exact Evaluation	125
Functional Correctness	126
Similarity Measurements Against Reference Data	127
Introduction to Embedding	134
AI as a Judge	136
Why AI as a Judge?	137
How to Use AI as a Judge	138
Limitations of AI as a Judge	141
What Models Can Act as Judges?	145
Ranking Models with Comparative Evaluation	148
Challenges of Comparative Evaluation	152
The Future of Comparative Evaluation	155
Summary	156

4. Evaluate AI Systems.....	159
Evaluation Criteria	160
Domain-Specific Capability	161
Generation Capability	163
Instruction-Following Capability	172
Cost and Latency	177
Model Selection	179
Model Selection Workflow	179
Model Build Versus Buy	181
Navigate Public Benchmarks	191
Design Your Evaluation Pipeline	200
Step 1. Evaluate All Components in a System	200
Step 2. Create an Evaluation Guideline	202
Step 3. Define Evaluation Methods and Data	204
Summary	208
5. Prompt Engineering.....	211
Introduction to Prompting	212
In-Context Learning: Zero-Shot and Few-Shot	213
System Prompt and User Prompt	215
Context Length and Context Efficiency	218
Prompt Engineering Best Practices	220
Write Clear and Explicit Instructions	220
Provide Sufficient Context	223
Break Complex Tasks into Simpler Subtasks	224
Give the Model Time to Think	227
Iterate on Your Prompts	229
Evaluate Prompt Engineering Tools	230
Organize and Version Prompts	233
Defensive Prompt Engineering	235
Proprietary Prompts and Reverse Prompt Engineering	236
Jailbreaking and Prompt Injection	238
Information Extraction	243
Defenses Against Prompt Attacks	248
Summary	251
6. RAG and Agents.....	253
RAG	253
RAG Architecture	256
Retrieval Algorithms	257
Retrieval Optimization	268

RAG Beyond Texts	273
Agents	275
Agent Overview	276
Tools	278
Planning	281
Agent Failure Modes and Evaluation	298
Memory	300
Summary	305
7. Finetuning.....	307
Finetuning Overview	308
When to Finetune	311
Reasons to Finetune	311
Reasons Not to Finetune	312
Finetuning and RAG	316
Memory Bottlenecks	319
Backpropagation and Trainable Parameters	320
Memory Math	322
Numerical Representations	325
Quantization	328
Finetuning Techniques	332
Parameter-Efficient Finetuning	333
Model Merging and Multi-Task Finetuning	347
Finetuning Tactics	357
Summary	361
8. Dataset Engineering.....	363
Data Curation	365
Data Quality	368
Data Coverage	370
Data Quantity	372
Data Acquisition and Annotation	377
Data Augmentation and Synthesis	380
Why Data Synthesis	381
Traditional Data Synthesis Techniques	383
AI-Powered Data Synthesis	386
Model Distillation	395
Data Processing	396
Inspect Data	397
Deduplicate Data	399
Clean and Filter Data	401

Format Data	401
Summary	403
9. Inference Optimization.....	405
Understanding Inference Optimization	406
Inference Overview	406
Inference Performance Metrics	412
AI Accelerators	419
Inference Optimization	426
Model Optimization	426
Inference Service Optimization	440
Summary	447
10. AI Engineering Architecture and User Feedback.....	449
AI Engineering Architecture	449
Step 1. Enhance Context	450
Step 2. Put in Guardrails	451
Step 3. Add Model Router and Gateway	456
Step 4. Reduce Latency with Caches	460
Step 5. Add Agent Patterns	463
Monitoring and Observability	465
AI Pipeline Orchestration	472
User Feedback	474
Extracting Conversational Feedback	475
Feedback Design	480
Feedback Limitations	490
Summary	492
Epilogue.....	495
Index.....	497