

Contents

<i>List of Contributors</i>	<i>page x</i>
<i>Preface</i>	<i>xiii</i>
<i>Introduction</i>	<i>1</i>
I.1 Why We Wrote This Book	1
I.2 How to Read This Book	2
I.3 What Sets This Book Apart	3
I.4 Common Pitfalls and How to Avoid Them	4
I.5 Looking Ahead	5
PART I READY, AIM, FIRE, AIM, FIRE, ...	7
1 Planning	9
1.1 Scoping the Project	10
1.2 Prioritization of Projects	11
1.3 Lifecycle of an ML Product	12
1.4 KPIs for ML Projects	21
1.5 Team Structure: Evolving with Team Topologies	25
1.6 Recap and Checklist	28
2 Data	29
2.1 Exploration	34
2.2 Sourcing	65
2.3 Enrichment	74
2.4 Preparation	117
2.5 Cataloging	133
2.6 Quality	154
2.7 Compliance	160
2.8 Data in Live Systems: Data-Centric AI	172
2.9 Recap and Checklist	180

3	Model Development	184
3.1	Ideation	185
3.2	Representation	203
3.3	Evaluation	223
3.4	Error Analysis	264
3.5	Training	268
3.6	Recap and Checklist	280
4	Model Deployment and Beyond	283
4.1	Types of Model Delivery	284
4.2	Inference	288
4.3	Large Language Model Operations	323
4.4	Recap and Checklist	333
5	Compute Optimizations	334
5.1	Planning for Compute	334
5.2	Compute Hardware	336
5.3	Hardware Tricks	338
5.4	Training Tricks	346
5.5	Inference Tricks	351
5.6	Advanced Topics	353
5.7	Example: QLoRA Training and Inference for LLaMa-3.1-8B	365
5.8	How Does This All Fit into the Flow?	368
5.9	Recap and Checklist	370
PART II	CASE STUDIES	373
6	Nauto: Data and Model Management	375
6.1	Problem Background	375
6.2	Data Management	376
6.3	Data Quality	379
6.4	Model Deployment and Monitoring	380
6.5	Takeaways	382
7	Kavak: ML Serverless Architecture for Car Sales	384
7.1	Problem Background	384
7.2	Designing ML Platform for Kavak	385
7.3	Serverless Architecture on AWS	385
7.4	Data Management and Continual Training	386
7.5	Feature Store	386
7.6	Model Deployment and Monitoring	387
7.7	Start with Batch Process	387

7.8	Continuous Model Improvements	387
7.9	Takeaways	388
8	Instacart: Journey in Building Griffin	390
8.1	Problems Background	390
8.2	Requirements for ML Platform	391
8.3	Griffin Systems Architecture	392
8.4	Takeaways	397
9	WhatsApp: Enhancing ML Operations for Fraud and Abuse Detection Model	399
9.1	Problem Background	399
9.2	Model Development, Feature Engineering, and Training	400
9.3	Model Deployment	400
9.4	Continual Model Improvement	401
9.5	Takeaways	402
10	ShortlyAI: Your AI Writing Partner	403
10.1	Problem Background	403
10.2	The Many Skills of Large Language Models	404
10.3	Prompt Engineering	406
10.4	Sampling Temperature	408
10.5	Concept Drift	409
10.6	Takeaways	410
	<i>References</i>	411
	<i>Index</i>	427