Contents

1 Introduction 5
1.1 What Does RLHF Do? 6
1.2 An Intuition for Post-Training 7
1.3 How We Got Here 8
1.4 Scope of This Book 10
1.4.1 Chapter Summaries 10
1.4.2 Target Audience 11
1.4.3 How to Use This Book 11
1.4.4 About the Author 11
1.5 Future of RLHF 11

2 Key Related Works 13
2.1 Origins to 2018: RL on Preferences 13
2.2 2019 to 2022: RL from Human Preferences on Language Models 13
2.3 2023 to Present: ChatGPT Era 14

3 Definitions & Background 15
3.1 Language Modeling Overview 15
3.2 ML Definitions 15
3.3 NLP Definitions 16
3.4 RL Definitions 16
3.5 RLHF Only Definitions 17
3.6 Extended Glossary 17

4 Training Overview 19
4.1 Problem Formulation 19
4.2 Manipulating the Standard RL Setup 19
4.3 Optimization Tools 20
4.4 RLHF Recipe Example 21
4.5 Finetuning and Regularization 21

5 The Nature of Preferences 23
5.1 The path to optimizing preferences 24
5.1.1 Quantifying preferences 24
5.1.2 On the possibility of preferences 25

6 Preference Data 26
6.1 Why We Need Preference Data 26
6.2 Collecting Preference Data 26
6.2.1 Interface 26
6.2.2 Rankings vs. Ratings 31
6.2.3 Structured Preference Data 31
6.2.4 Sourcing and Contracts 33
6.3 Are the Preferences Expressed in the Models? 34

7 Reward Modeling 36
7.1 Training Reward Models 36

7.2 Architecture . . . 37
7.3 Implementation Example . . . 37
7.4 Variants . . . 38
7.4.1 Preference Margin Loss . . . 38
7.4.2 Balancing Multiple Comparisons Per Prompt . . . 38
7.4.3 K-wise Loss Function . . . 38
7.5 Outcome Reward Models . . . 39
7.6 Process Reward Models . . . 39
7.7 Reward Models vs. Outcome RMs vs. Process RMs vs. Value Functions . . . 40
7.8 Generative Reward Modeling . . . 41
7.9 Further Reading . . . 42

8 Regularization 43
8.1 KL Distances in RL Optimization . . . 43
8.1.1 Reference Model to Generations . . . 43
8.1.2 Implementation Example . . . 44
8.2 Pretraining Gradients . . . 44
8.3 Other Regularization . . . 45

9 Instruction Finetuning 46
9.1 Chat templates and the structure of instructions . . . 46
9.2 Best practices of instruction tuning . . . 48

10 Rejection Sampling 49
10.1 Training Process . . . 49
10.1.1 Generating Completions . . . 49
10.1.2 Selecting Top-N Completions . . . 50
10.1.3 Fine-tuning . . . 52
10.1.4 Details . . . 52
10.2 Related: Best-of-N Sampling . . . 53

11 Policy Gradient Algorithms 54
11.1 Policy Gradient Algorithms . . . 54
11.1.1 Vanilla Policy Gradient . . . 55
11.1.2 REINFORCE . . . 56
11.1.3 Proximal Policy Optimization . . . 58
11.1.4 Group Relative Policy Optimization . . . 60
11.2 Implementation . . . 62
11.2.1 Policy Gradient Basics . . . 63
11.2.2 Loss Aggregation . . . 63
11.2.3 Proximal Policy Optimization . . . 66
11.2.4 Group Relative Policy Optimization . . . 68
11.3 Auxiliary Topics . . . 70
11.3.1 Generalized Advantage Estimation (GAE) . . . 70
11.3.2 Double Regularization . . . 71
11.3.3 Further Reading . . . 71

12 Direct Alignment Algorithms 73
12.1 Direct Preference Optimization (DPO) . . . 73

12.1.1 How DPO Works 73
12.1.2 DPO Derivation 74
12.2 Numerical Concerns, Weaknesses, and Alternatives 78
12.3 Implementation Considerations 80
12.4 DAAs vs. RL: Online vs. Offline Data 81

13 Constitutional AI & AI Feedback 82
13.1 Constitutional AI 82
13.2 Specific LLMs for Judgement 83
13.3 Further Reading 83

14 Reasoning Training & Inference-Time Scaling 84
14.1 Why Does RL Work Now? 86
14.2 RL Training vs. Inference Time Scaling 87
14.3 The Future (Beyond Reasoning) of Reinforcement Finetuning 88

15 Synthetic Data & Distillation 89

16 Evaluation 91
16.1 Prompting Formatting: From Few-shot to Zero-shot to CoT 91
16.2 Using Evaluations vs. Observing Evaluations 95
16.3 Contamination 97
16.4 Tooling 97

17 Over Optimization 98
17.1 Qualitative Over-optimization 98
17.1.1 Managing Proxy Objectives 98
17.1.2 Over-refusal and “Too Much RLHF” 100
17.2 Quantitative over-optimization 101
17.3 Misalignment and the Role of RLHF 102

18 Style and Information 104
18.1 The Chattiness Paradox 104
18.1.1 How Chattiness Emerges 105

19 Product, UX, and Model Character 107
19.1 Character Training 107
19.2 Model Specifications 108
19.3 Product Cycles, UX, and RLHF 108

Bibliography 110