

RLHF BOOK: CONSOLIDATED MASTER GUIDE

PART 1: CHAPTER STRUCTURE & FLOW

FOUNDATION (Ch 1-3)

What you need to know first

Ch 1: Introduction | Ch 2: Key Works | Ch 3: Definitions

PROBLEM SETUP (Ch 4-6)

Why RLHF exists & how to get data

Ch 4: Training Overview → Ch 5: Preferences → Ch 6: Data Collection

CORE BACKBONE (Ch 9→7→8→11) The Main 3-Stage Pipeline

STAGE 1: Ch 9
Instruction Tuning

↓

STAGE 2: Ch 7
Reward Modeling
(uses Ch 6 data)

↓

Ch 8: Regularization
(KL constraints)

↓

STAGE 3: Ch 11
RL Optimization
(PPO/GRPO/REINFORCE)

ALTERNATIVE ROUTES

Ch 10: Rejection Sampling
(RM + filter, no RL)

OR

Ch 12: Direct Alignment (DPO)
(skip reward model)

ADVANCED EXTENSIONS (Ch 13-16)

Building on the core pipeline

| | |
|---------------------------|-----------------------------------|
| Ch 13: Constitutional AI | Uses Ch 16 → replaces Ch 6 |
| Ch 14: Reasoning Training | Uses Ch 11 algorithms differently |
| Ch 15: Tool Use | Extends Ch 9 formatting |
| Ch 16: Synthetic Data | Replaces/augments Ch 6 |

MEASUREMENT & DEPLOYMENT (Ch 17-20)

Understanding and evaluating the output

| | |
|---------------------|--------------------------|
| Ch 17: Evaluation | Ch 18: Over-optimization |
| Ch 19: Style & Info | Ch 20: Product/UX |

PART 2: THE TWO-LOOP SYSTEM

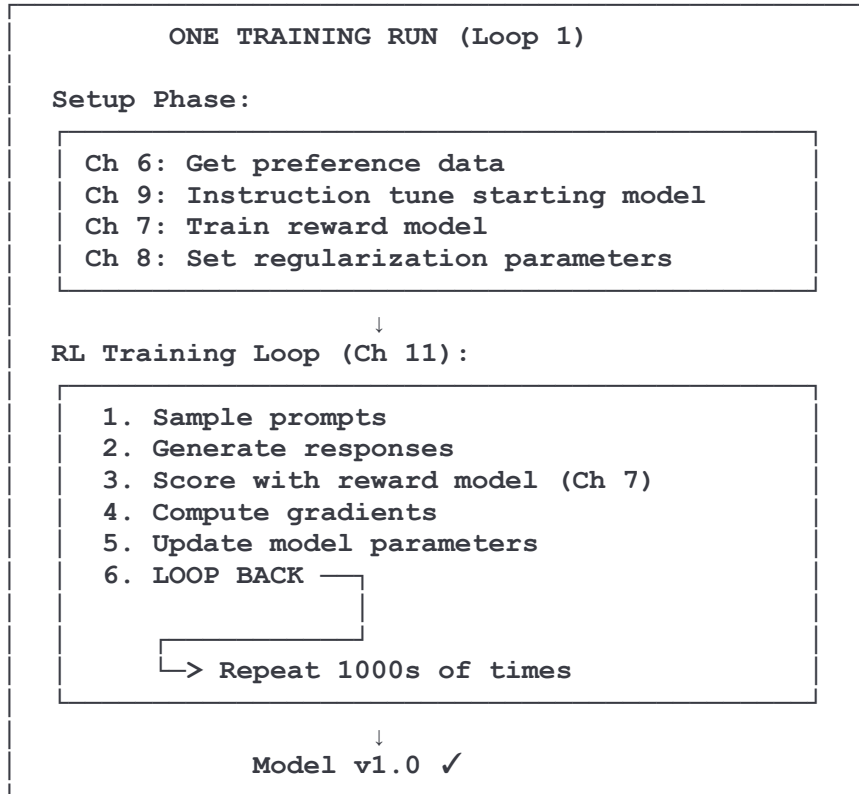
LOOP 1: INTERNAL FEEDBACK (*Automatic, Inside Training*)

Location: Inside Ch 11 (RL Optimization)

Duration: Hours to days

Automated: YES

Frequency: Thousands of iterations



This is what Ch 1-12 teaches you to do!

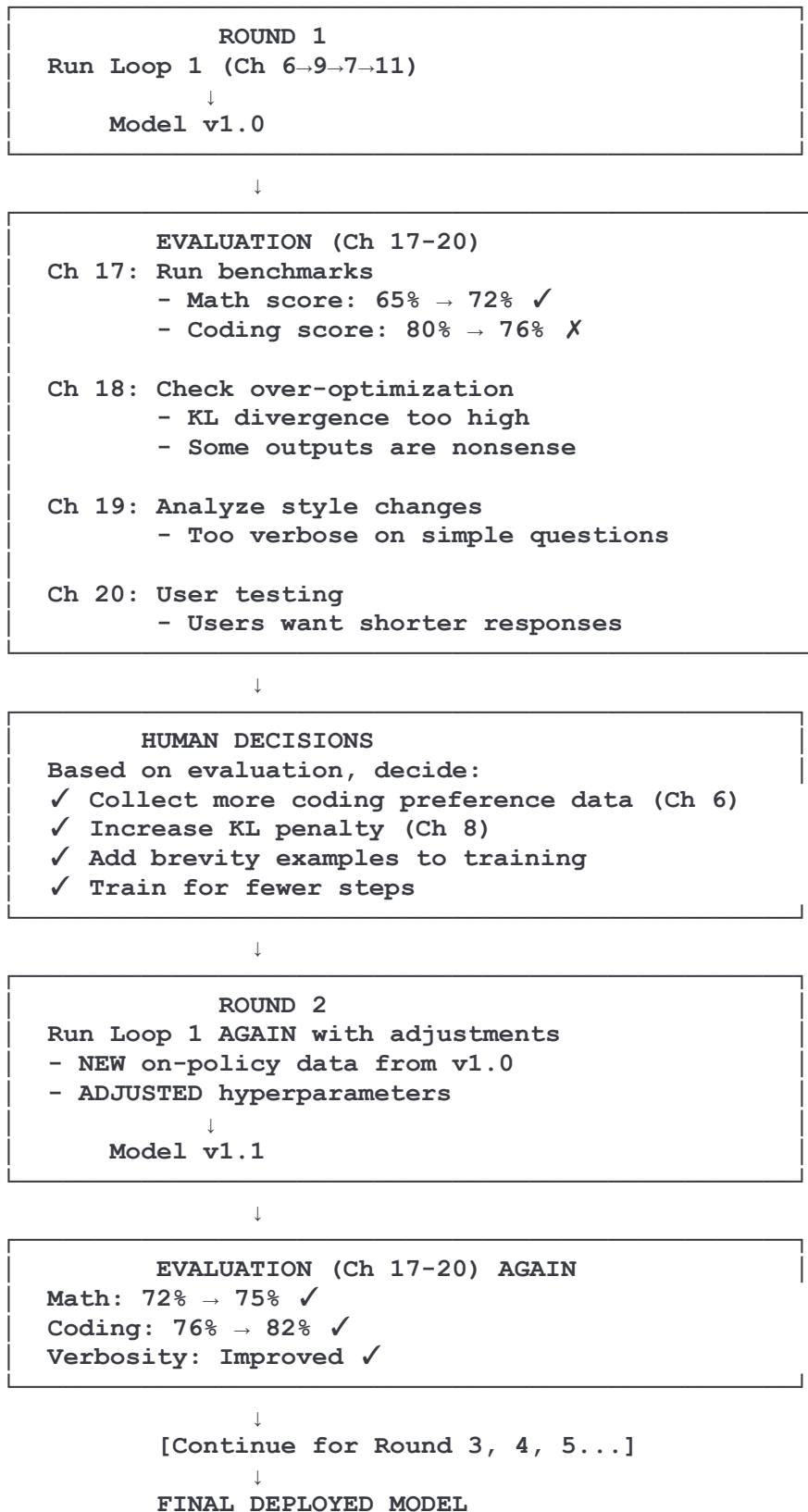
LOOP 2: EXTERNAL FEEDBACK (Human-guided, Across Rounds)

Location: Across multiple training runs

Duration: Weeks to months

Automated: NO - requires human decisions

Frequency: 3-14+ major iterations for modern models



This is how Ch 17-20 feeds back to improve Ch 1-12!

PART 3: INTEGRATED VIEW - THE COMPLETE PICTURE

START: Base Language Model

LEARN THE FOUNDATIONS (Ch 1-3)
Understand what RLHF is and why it exists

↓

LOOP 2 ITERATION 1

LOOP 1: Run One Training Cycle

Problem Setup (Ch 4-6)

↓

Choose Path:

- Core: Ch 9→7→8→11 (main route)
- Alt: Ch 10 (simpler)
- Alt: Ch 12 (different)

↓

Optional Extensions (Ch 13-16)

↓

Trained Model v1.0

↓

Measure & Analyze (Ch 17-20)

- What works?
- What broke?
- What changed?

↓

DECISION: Good enough? NO → adjust recipe

↓

LOOP 2 ITERATION 2

[Same structure, with improvements from iteration 1]

↓

LOOP 2 ITERATION 3

[Continue refining...]

↓

...

↓

FINAL DEPLOYED MODEL (after 3-14+ iterations)

PART 4: KEY CONNECTIONS SUMMARY

BACKBONE CHAPTERS (The Core Path):

Ch 6 (Data) → Ch 9 (SFT) → Ch 7 (Reward Model) → Ch 8 (Regularization)
→ Ch 11 (RL)

ALTERNATIVE ROUTES:

- Ch 10 (Rejection Sampling) : Uses Ch 7 + Ch 9, skips Ch 11 (simpler)
- Ch 12 (DPO) : Uses Ch 6 directly, skips Ch 7 (different approach)

EXTENSIONS:

- Ch 16 (Synthetic Data) → replaces Ch 6
- Ch 13 (Constitutional AI) → uses Ch 16, feeds into Ch 7 or Ch 12
- Ch 14 (Reasoning) → uses Ch 11 algorithms with different rewards
- Ch 15 (Tool Use) → extends Ch 9 with special formatting

FEEDBACK MECHANISMS:

- Ch 8 (Regularization) → prevents issues detected by Ch 18
- Ch 17 (Evaluation) → measures output of Ch 9/10/11/12
- Ch 18 (Over-optimization) → explains why Ch 8 exists
- Ch 19-20 → understand what the model actually learned

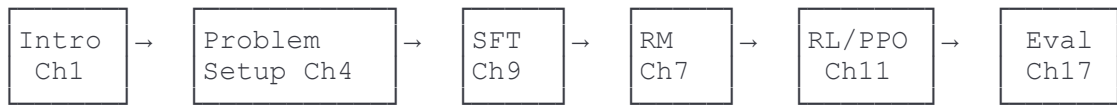
LOOP STRUCTURE:

- Loop 1 (Internal): Inside Ch 11, automatic RL updates
- Loop 2 (External): Ch 17-20 evaluation → adjust Ch 1-12 → repeat

PART 5: PATHS FOR DIFFERENT GOALS

PATH 1: Quick Understanding (Minimum viable knowledge)

"Understand the basic 3-stage pipeline and how to evaluate it"



PATH 2: Full Implementation (Learn to do it yourself)

"Master the core techniques and alternatives"

Foundation: Ch 1 (Intro) → Ch 2 (Key Works) → Ch 3 (Definitions)
Problem: Ch 4 (Problem Setup) → Ch 5 (Preferences) → Ch 6 (Data Collection)
Core Pipeline: Ch 9 (SFT) → Ch 7 (RM) → Ch 8 (Regularization/KL) → Ch 11 (RL/PPO)
Alternatives: Ch 10 (Rejection Sampling) → Ch 12 (DPO/Direct Alignment)
Measurement: Ch 17 (Evaluation) → Ch 18 (Over-optimization)

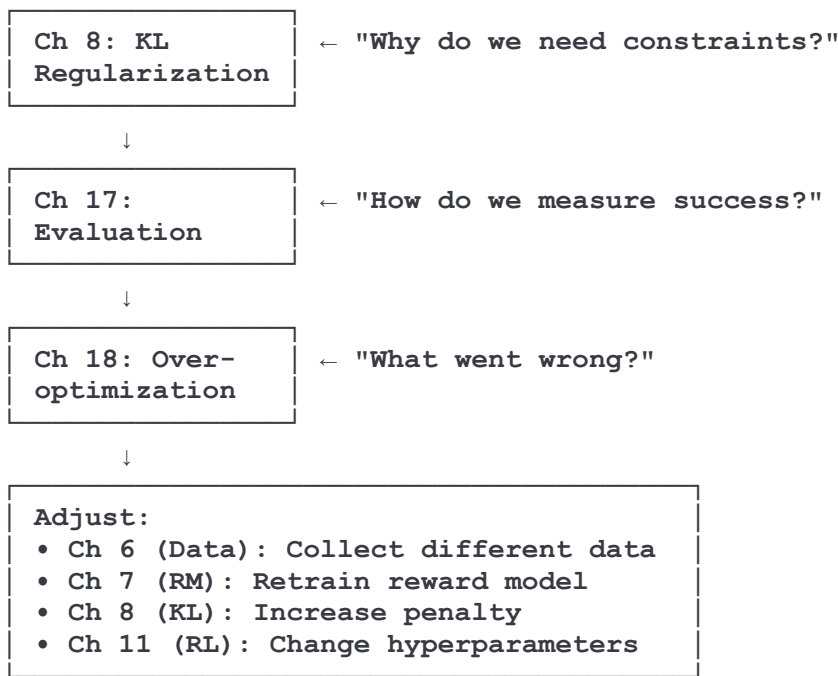
PATH 3: Modern Production Systems (State of the art)

"Understand how frontier labs train their models"

Complete Flow:
Ch 1-12, 17-18 (Full Implementation)
Extensions:
Ch 16 (Synthetic Data) → Ch 13 (Constitutional AI/RLAIF)
Ch 14 (Reasoning/ol) [uses same RL as Ch 11]
Ch 15 (Tool Use) [extends Ch 9]
Understanding Output:
Ch 19 (Style & Info) → Ch 20 (Product/UX)

PATH 4: Debugging & Troubleshooting (Fix what's broken)

"Focus on what goes wrong and how to fix it"



QUICK REFERENCE: ALL CHAPTERS WITH TERMS

FOUNDATIONS

- Ch 1: Introduction (Intro)
- Ch 2: Key Related Works (Key Works)
- Ch 3: Definitions & Background (Definitions)

PROBLEM SETUP

- Ch 4: Training Overview (Problem Setup)
- Ch 5: The Nature of Preferences (Preferences)
- Ch 6: Preference Data (Data Collection)

CORE OPTIMIZATION TOOLS

- Ch 7: Reward Modeling (Reward Model / RM)
- Ch 8: Regularization (Regularization / KL Penalty)
- Ch 9: Instruction Finetuning (Instruction Tuning / SFT)
- Ch 10: Rejection Sampling (Rejection Sampling)
- Ch 11: Reinforcement Learning (RL / PPO / Policy Gradients)
- Ch 12: Direct Alignment Algorithms (DPO / Direct Alignment)

ADVANCED TECHNIQUES

- Ch 13: Constitutional AI & AI Feedback (Constitutional AI / RLAIIF)
- Ch 14: Reasoning Training (Reasoning / o1 / RLVR)
- Ch 15: Tool Use & Function Calling (Tool Use)
- Ch 16: Synthetic Data & Distillation (Synthetic Data)

MEASUREMENT & DEPLOYMENT

- Ch 17: Evaluation (Evaluation / Eval)
- Ch 18: Over-optimization (Over-optimization)
- Ch 19: Style & Information (Style & Info)
- Ch 20: Product, UX, & Character (Product / UX)

ABBREVIATED NOTATION SYSTEM

When discussing RLHF informally, you can use these shorthand terms:

CORE PIPELINE:

SFT (Ch 9) → RM (Ch 7) → RL (Ch 11)

WITH PREREQUISITES:

Data (Ch 6) → SFT (Ch 9) → RM (Ch 7) → KL (Ch 8) → RL (Ch 11)

ALTERNATIVES:

- Rejection Sampling (Ch 10) = RM (Ch 7) + SFT (Ch 9) filtering
- DPO (Ch 12) = Skip RM (Ch 7), optimize directly from Data (Ch 6)

EXTENSIONS:

- RLAIIF (Ch 13) = AI-generated Data (Ch 6)
- o1-style (Ch 14) = RL (Ch 11) with binary rewards
- Tools (Ch 15) = Enhanced SFT (Ch 9)
- Synthetic (Ch 16) = Generated Data (Ch 6)

MEASUREMENT:

Eval (Ch 17) → Over-opt Check (Ch 18) → Iterate

PART 6: THE ULTIMATE ONE-PAGE SUMMARY

WHAT IS RLHF?

A technique to make language models follow human preferences using:

1. Instruction tuning (teach format)
2. Reward modeling (capture preferences)
3. RL optimization (optimize toward preferences)

HOW DOES IT WORK?

Two nested loops:

- Inner (Loop 1): RL algorithm updates model thousands of times
- Outer (Loop 2): Humans evaluate and adjust recipe 3-14+ times

WHAT DO THE CHAPTERS TEACH?

- Ch 1-6: Why RLHF exists and how to set it up
- Ch 7-12: How to execute one training run (Loop 1)
- Ch 13-16: Advanced variations and improvements
- Ch 17-20: How to measure success and iterate (Loop 2)

KEY INSIGHT:

Modern RLHF is NOT a single training run!

It's an iterative process: Train → Evaluate → Adjust → Repeat

The book teaches you:

- How to do ONE pass of Loop 1 (Ch 1-12)
- How to measure results for Loop 2 iterations (Ch 17-20)
- How advanced techniques fit into this framework (Ch 13-16)