# CS336 Lecture 15: Alignment - SFT/RLHF

From Pre-training to Post-training

Stanford CS336 - Spring 2025

**Abstract**

This lecture covers the transition from pre-training to post-training, focusing on Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF). We explore how to transform a pre-trained model like GPT-3 into an instruction-following system like ChatGPT, examining data collection strategies, training methodologies, and the challenges of alignment at scale.

## Contents

# 1 Introduction: From Pre-training to Post-training

The fundamental challenge in modern language modeling is the transition from powerful but unaligned pre-trained models to useful, safe, instruction-following systems. This lecture addresses the critical transformation that enabled the shift from GPT-3 to ChatGPT.

Key Concept

**The GPT-3 to ChatGPT Transformation**
While GPT-3 demonstrated remarkable capabilities, it lacked instruction-following abilities and safety guardrails. Post-training techniques enable models to:

- Follow complex, nested instructions
- Maintain safety and content moderation standards
- Provide helpful, truthful, and harmless responses
- Exhibit consistent behavioral patterns

## 1.1 Post-training Pipeline Overview

The modern post-training pipeline follows the InstructGPT framework:

1. **Supervised Fine-Tuning (SFT)**: Train on expert demonstrations
2. **Reward Modeling**: Learn from pairwise feedback
3. **Reinforcement Learning**: Optimize for reward maximization

# 2 Part I: Supervised Fine-Tuning (SFT)

## 2.1 Training Data Paradigms

Three primary approaches exist for constructing instruction-tuning datasets:

### 2.1.1 FLAN: Aggregated NLP Tasks

FLAN aggregates existing NLP datasets into a unified instruction-tuning corpus:

- Natural Instructions V2 (question answering)
- T0-SF (task formatting)
- Adversarial QA
- Topic classification tasks

**Characteristics:**

- Benchmark-centric task structure
- Often short, phrase-based responses
- Visible ”surgery” to convert datasets into instruction format
- High data volume but potentially unnatural interactions

### 2.1.2 OpenAssistant: Human-Written Data

Community-driven effort producing high-quality human annotations:

- Complex, diverse queries
- Detailed, citation-rich responses
- High quality but labor-intensive
- Strong educational content with references

### 2.1.3 Alpaca: AI-Generated Data

Language model-generated instruction-tuning data:

- Seed set of human instructions
- LM-generated instruction expansion
- InstructGPT-generated responses
- ChatGPT-style interactions but limited diversity

## 2.2 Data Quality Considerations

Challenge

**The Annotation Challenge**
Real-world annotation faces several challenges:

- Time constraints (often 1-5 minutes per example)
- Quality vs. quantity trade-offs
- Length bias in human and AI evaluations
- Difficulty of factual verification
- Risk of AI-generated responses in crowdsourcing

### 2.2.1 Length and Evaluation Biases

Research shows strong preferences for longer outputs:

- 60-70% preference for longer responses in human evaluation
- AI judges also exhibit length bias
- Potential optimization for style over substance
- Benchmark performance often independent of response length

### 2.2.2 The Citation Hallucination Problem

High-quality training data with citations can paradoxically encourage hallucination:

**Two competing learning mechanisms:**

1. **Knowledge Association**: Model learns correct fact-citation pairs
2. **Format Mimicking**: Model learns to generate plausible citations regardless of knowledge

When the model lacks knowledge of the cited facts, it may learn the format pattern rather than the content, leading to systematic hallucination.

## 2.3 Safety Tuning

Safety tuning addresses the trade-off between refusing harmful requests and maintaining helpfulness:

- Small amounts of safety data (500 examples) can provide significant improvements
- Challenge: distinguishing truly harmful requests from benign ones (e.g., ”How to kill a Python process?”)
- Requires carefully curated datasets balancing safety and utility

## 2.4 Modern SFT: Mid-Training Integration

Method

**Mid-Training Pipeline**

Modern instruction tuning integrates with pre-training:

1. **Pure Pre-training**: Standard pre-training on web data
2. **Mid-training**: Mix instruction data into pre-training during learning rate decay
3. **Final SFT**: Short instruction tuning phase on remaining data

**Benefits:**

- Reduces catastrophic forgetting
- Enables scaling without hyperparameter sensitivity
- Integrates instruction following into core model capabilities

# 3 Part II: Reinforcement Learning from Human Feedback (RLHF)

## 3.1 Paradigm Shift: From Generative Modeling to Policy Optimization

RLHF represents a fundamental shift in perspective:

**Generative Modeling (SFT):**

$$\text{Goal: } p_\theta(y|x) \approx p^*(y|x) \tag{1}$$

**Policy Optimization (RLHF):**

$$\text{Goal: } \max_{p_\theta} \mathbb{E}_{x,y\sim p_\theta}[R(x,y)] \tag{2}$$

## 3.2 Motivation for RLHF

Two primary drivers for RLHF adoption:

1. **Cost Efficiency**: Pairwise comparisons cheaper than expert demonstrations
2. **Generator-Verifier Gap**: People often prefer AI outputs to their own writing

## 3.3 Pairwise Feedback Collection

### 3.3.1 Annotation Guidelines

Successful RLHF requires comprehensive annotation guidelines covering:

**InstructGPT Criteria:**

- **Helpful**: Clear language, addresses intended question, sensitive to context
- **Truthful**: Accurate information, minimal hallucination
- **Harmless**: Non-toxic, appropriate content

**Practical Challenges:**

- Extremely tight time constraints (1 minute per comparison)
- Difficulty of factual verification
- Need for domain expertise
- Scale requirements (millions of comparisons)

### 3.3.2 Annotation Quality Issues

Real-world annotation faces several systematic problems:

Challenge

**Annotation Challenges**

- **Time Pressure**: Insufficient time for thorough evaluation
- **Fact-Checking**: Nearly impossible to verify complex claims quickly
- **AI Contamination**: Annotators may use AI tools, defeating the purpose
- **Cultural Bias**: Annotator demographics influence model alignment
- **Attention Patterns**: Crowdworkers focus on formatting over factuality

## 3.4 AI Feedback as Alternative

Given annotation challenges, AI feedback has become increasingly popular:

**Advantages:**

- High agreement with human feedback (comparable to human-human agreement)

- Significantly lower cost
- Scalable to millions of examples
- Consistent evaluation criteria

**Examples:**

- Constitutional AI (Anthropic)
- UltraFeedback datasets
- Tulu3 project methodology

### 3.5 RLHF Training Process

The standard RLHF pipeline:

1. **Rollout Generation**: Sample responses from current policy
2. **Preference Collection**: Gather pairwise comparisons
3. **Reward Model Training**: Learn scalar reward function
4. **Policy Optimization**: Maximize expected rewards (PPO/DPO)

## 4 Key Insights and Takeaways

### 4.1 Instruction Tuning Effectiveness

- Surprisingly powerful with modest data requirements
- Effective even with standard datasets and reasonable hyperparameters
- Small amounts of targeted data can create significant behavioral changes

### 4.2 Data Quality Complexity

- ”High quality” is not simply more detailed or longer
- Must balance model capabilities with training objectives
- Risk of teaching hallucination patterns through well-intentioned data

### 4.3 Scale and Integration

- Modern post-training increasingly resembles pre-training
- Mid-training blurs traditional boundaries
- ”Base models” may already incorporate significant instruction tuning

### 4.4 Evaluation Considerations

- Chat-style evaluations susceptible to length and format biases
- Benchmark performance provides important grounding
- Need diverse evaluation strategies to avoid optimization pitfalls

## 5 Mathematical Formulations

### 5.1 SFT Objective

$$\mathcal{L}_{SFT} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{SFT}}[\log p_\theta(y|x)] \tag{3}$$

### 5.2 Reward Model Training

$$\mathcal{L}_{RM} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{comp}}[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))] \tag{4}$$

### 5.3 RLHF Objective (PPO)

$$\mathcal{L}_{RLHF} = \mathbb{E}_{x\sim\mathcal{D},y\sim p_\theta(\cdot|x)}[r_\phi(x, y)] - \beta\mathbb{D}_{KL}[p_\theta(\cdot|x)\|p_{ref}(\cdot|x)] \tag{5}$$

## 6 Future Directions

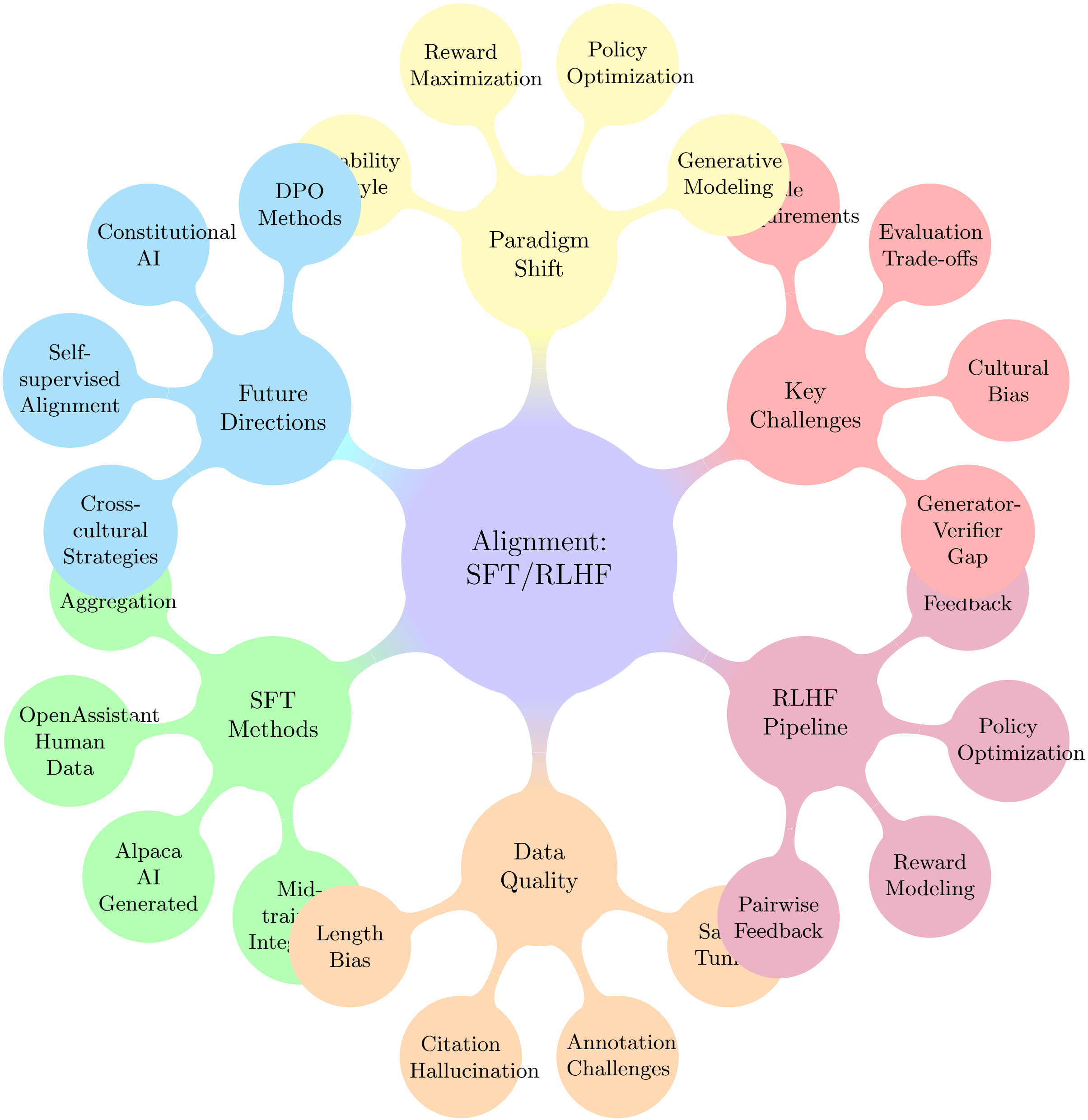
### 6.1 Emerging Techniques

- Direct Preference Optimization (DPO)
- Constitutional AI approaches
- Self-supervised alignment methods
- Tool-augmented training

### 6.2 Open Research Questions

- Optimal balance between human and AI feedback
- Scaling laws for post-training data
- Cross-cultural alignment strategies
- Factual accuracy vs. instruction following trade-offs

## 7 Lecture 15 Mind Map



**Mind Map Description:**

- **SFT Methods (Green):** Covers the different approaches to supervised fine-tuning, from FLAN's task aggregation to modern mid-training integration
- **Data Quality (Orange):** Highlights critical data challenges including length bias, citation hallucination, and annotation difficulties
- **RLHF Pipeline (Purple):** Shows the reinforcement learning workflow from pairwise feedback collection to policy optimization

- **Key Challenges (Red):** Identifies major obstacles in alignment including the generator-verifier gap and evaluation trade-offs
- **Paradigm Shift (Yellow):** Illustrates the fundamental transition from generative modeling to policy optimization perspectives
- **Future Directions (Cyan):** Points to emerging techniques and research directions in alignment methodology