Preference Data & RLHF Annotation

Aligning Language Models with Human Values

Jin Zhao

Brandeis University

April 20, 2025

Today's Agenda

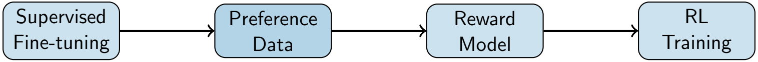
1. Introduction to RLHF
2. The preference annotation task
3. Reward modeling data collection
4. Constitutional AI
5. Direct Preference Optimization (DPO)
6. Practical considerations

Assignment: HW 4 due this week

What is RLHF?

Reinforcement Learning from Human Feedback

Goal: Align LLM behavior with human preferences



Key insight: Preferences are easier to annotate than demonstrations

Why Preferences?

Comparing is easier than generating

Demonstration annotation:

- “Write the ideal response to this prompt”
- Requires expertise to write good responses
- Time-consuming
- Single point of reference

Preference annotation:

- “Which response is better?”
- Easier cognitive task
- Faster annotation
- Captures relative quality

Preference Annotation Task

Given: A prompt and two (or more) responses
Task: Indicate which response is better

Example

Prompt: "Explain photosynthesis simply."
Response A: "Plants convert sunlight into food..." (accurate, clear)
Response B: "It's a complex biochemical process involving..." (accurate, verbose)

Annotation: A $>$ B (A is preferred)

Preference Data Formats

Common formats:

1. **Binary comparison:**
   - A wins, B wins, or tie
   - Simple, clear
2. **Rating scale:**
   - Rate each response 1-5
   - More information but harder to calibrate
3. **Ranking:**
   - Order multiple responses
   - Rich signal, more complex annotation
4. **Best-of-N:**
   - Select best from N options
   - Efficient for data collection

What Makes a Response "Better"?

Common evaluation criteria (InstructGPT):

1. **Helpful:** Provides useful information
2. **Harmless:** Avoids dangerous or unethical content
3. **Honest:** Accurate and doesn't hallucinate

Additional criteria:

- Follows instructions
- Appropriate length
- Clear and well-organized
- Appropriate tone

Challenge: Criteria can conflict (helpful vs. safe)

Reward Model Training

Converting preferences to reward function

Process:

1. Collect preference pairs: (x, y_w, y_l)
2. Train model: $r(x, y_w) > r(x, y_l)$
3. Use Bradley-Terry model or similar

Loss function:

$$L = -\log \sigma(r(x, y_w) - r(x, y_l))$$

Result: Model that scores response quality

Constitutional AI

Anthropic's approach to AI alignment

Key idea: Define principles, not just preferences

1. Define constitution (principles like "be helpful", "be harmless")
2. AI critiques its own outputs against principles
3. AI revises based on self-critique
4. Human feedback on principles, not instances

Annotation role:

- Validate that principles are correct
- Check self-critique quality
- Identify edge cases

Direct Preference Optimization (DPO)

Skip the reward model

Traditional RLHF:
Preferences $\rightarrow$ Reward Model $\rightarrow$ RL Training

DPO:
Preferences $\rightarrow$ Direct LLM Training

Benefits:

- Simpler pipeline
- No reward model to train
- More stable training
- Same preference data requirements

Key insight: Preference data is the bottleneck, not the algorithm

Data Requirements for DPO

What you need:

- Prompts (from target distribution)
- Chosen responses (preferred)
- Rejected responses (dis-preferred)

Data format (Hugging Face):

- `prompt`: The input query
- `chosen`: Preferred response
- `rejected`: Dis-preferred response

Scale:

- 10K-100K pairs for good results
- Quality matters more than quantity
- Diverse prompts important

Annotation Challenges

Common issues in preference annotation:

Subjectivity:

- Different annotators have different preferences
- Cultural variation
- Personal style preferences

Biases:

- Position bias (prefer first/second)
- Length bias (prefer longer/shorter)
- Verbosity bias (more words = better?)

Fatigue:

- Annotation quality degrades over time
- Need breaks and quality checks

Best Practices

For preference annotation:

1. **Clear criteria:** Define what "better" means
2. **Randomize order:** Avoid position bias
3. **Multiple annotators:** 3+ per pair recommended
4. **Calibration:** Regular alignment sessions
5. **Quality control:** Monitor agreement over time
6. **Fair pay:** Preference annotation is cognitively demanding

Tools for Preference Annotation

Specialized platforms:

- **Argilla:** Built for RLHF data collection
- **Label Studio:** Customizable for comparison
- **Scale AI:** Commercial platform
- **Surge AI:** Managed workforce

Key features needed:

- Side-by-side response display
- Randomization of order
- Multiple response comparison
- Tie/skip options

Next Class: Safety & Red Teaming

Lecture 24 (Apr 22): Safety & Red Teaming Annotation

Topics:

- Red teaming and adversarial annotation
- Harmfulness and toxicity annotation
- Content moderation datasets
- Safety evaluation benchmarks
- Ethical considerations

Key Takeaways

1. **RLHF** aligns LLMs using human preference data
2. **Preferences are easier** to annotate than demonstrations
3. **Reward models** learn to score response quality
4. **Constitutional AI** uses principles instead of instance-level feedback
5. **DPO** simplifies training by skipping reward model
6. **Quality preference data** is the key bottleneck

Questions?

Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

jinzhao@brandeis.edu