# Overview of Annotation Tasks IV

## LLM-Specific Annotation Tasks

Jin Zhao

Brandeis University

February 23, 2025

# Today's Agenda

1. Welcome back from February Break
2. Introduction to LLM-specific annotation
3. Preference annotation for RLHF
4. Safety and toxicity annotation
5. Instruction-following evaluation
6. Multi-turn conversation annotation
7. Challenges and best practices

**Semester Project:** Groups present their chosen tasks

# The LLM Era: New Annotation Needs

**Traditional annotation:**

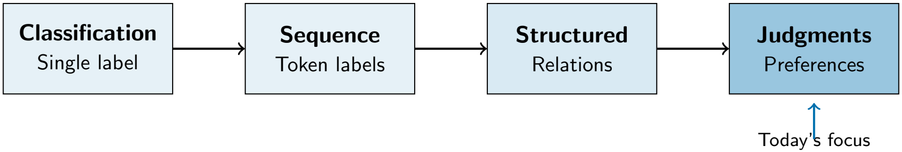
- Train models to perform specific tasks
- Classification, NER, parsing, etc.
- Ground truth labels for supervised learning

**LLM-era annotation:**

- Align models with human values and preferences
- Evaluate open-ended generation quality
- Ensure safety and helpfulness
- Human feedback as training signal

**Key shift:** From “is this correct?” to “is this better?”

# The Task Spectrum

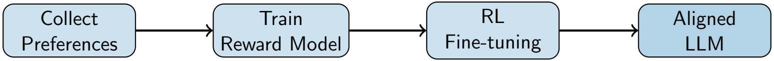



**Human Judgments:** Subjective assessments of quality, preference, safety

- Most subjective annotation type
- Highest annotator disagreement
- Critical for LLM alignment

# What is RLHF?

**Reinforcement Learning from Human Feedback**




**How ChatGPT, Claude, etc. are trained:**

1. Collect human preferences on model outputs
2. Train a reward model to predict preferences
3. Use reward model to fine-tune the LLM

**Annotation is the foundation of RLHF**

# Preference Annotation

**Core task:** Given two responses, which is better?

## Example

**Prompt:** "Explain quantum computing to a 10-year-old."
**Response A:** "Quantum computing uses quantum bits..."
**Response B:** "Imagine you have a magical coin..."

**Annotation:** $A > B$ or $B > A$ or $A \approx B$

**Annotation formats:**

- Binary choice: A or B
- Likert scale per response: 1-5 rating
- Ranking: Order multiple responses
- Continuous: How much better? (e.g., 60% prefer A)

# What Makes a Response “Better”?

**Typical evaluation criteria:**

**Helpfulness:**

- Answers the question
- Provides useful information
- Appropriate level of detail
- Well-organized

**Quality:**

- Factually accurate
- Coherent and clear
- Grammatically correct
- Appropriate tone

**Challenge:** These criteria can conflict!
A response can be helpful but contain minor errors, or accurate but unhelpful.

# Preference Annotation Challenges

**Subjectivity issues:**

- Different annotators have different preferences
- Cultural and individual variation
- Task interpretation differences

**Cognitive challenges:**

- Position bias (prefer first/second response)
- Length bias (prefer longer/shorter responses)
- Fatigue effects over time
- Inconsistency within annotators

**Mitigation strategies:**

- Clear rubrics with examples
- Randomize response order
- Calibration sessions
- Multiple annotators per pair

# Safety and Toxicity Annotation

**Goal:** Identify harmful or inappropriate content

**Categories of harm:**

- **Toxicity:** Hate speech, harassment, threats
- **Misinformation:** False claims, conspiracy theories
- **Dangerous content:** Instructions for harm
- **Adult content:** Sexual or violent material
- **Privacy:** Personal information exposure

**Annotation task:**

- Is this content harmful? (Binary)
- How harmful is it? (Severity scale)
- What type of harm? (Multi-label)
- To whom is it harmful? (Target identification)

# Safety Annotation Challenges

**Definitional challenges:**

- What counts as "harmful"?
- Context dependence (medical info vs. harm instructions)
- Cultural variation in acceptability
- Legal vs. ethical considerations

**Annotator wellbeing:**

- Exposure to disturbing content
- Psychological impact over time
- Need for support and breaks
- Ethical responsibility to annotators

**Best practices:**

- Clear content warnings
- Voluntary participation
- Mental health support
- Regular rotation off sensitive tasks

# Instruction-Following Evaluation

**Goal:** Did the model follow the user's instructions?

**Evaluation dimensions:**

- **Completeness:** Did it address all parts of the request?
- **Format compliance:** Did it follow format instructions?
- **Constraint satisfaction:** Did it respect constraints?
- **Task success:** Did it accomplish the goal?

**Example:**
**Prompt:** "Write a haiku about spring. Use exactly 3 lines."
**Response:** "Cherry blossoms fall / Gentle rain on fresh green leaves / Spring awakens earth"

**Annotation:** Format: ✓ Constraint: ✓ Quality: 4/5

# Multi-Turn Conversation Annotation

**Challenge:** Evaluating dialogue over multiple exchanges

**What to evaluate:**

- Individual turn quality
- Coherence across turns
- Context maintenance
- Goal progression
- Conversation-level success

**Annotation approaches:**

1. Rate each turn independently
2. Rate conversation as a whole
3. Compare two complete conversations
4. Identify specific failure points

**Complexity:** Earlier turns affect later ones

# Constitutional AI: Principle-Based Feedback

**Alternative to pure human preference:**

1. Define principles ("be helpful", "be harmless", "be honest")
2. Have AI critique its own outputs against principles
3. Revise based on self-critique
4. Use human feedback on principles, not instances

**Annotation role:**

- Validate that principles are correct
- Check that self-critique is reasonable
- Identify edge cases where principles conflict

**Benefit:** More scalable than instance-by-instance annotation

# Direct Preference Optimization (DPO)

**Recent alternative to full RLHF pipeline:**

**Traditional RLHF:**
Preferences → Reward Model → RL Training

**DPO:**
Preferences → Direct LLM Training

**Same annotation requirements:**

- Pairwise preferences (chosen vs. rejected)
- Quality of preferences still critical
- Annotation guidelines remain essential

**Key insight:** The preference data is the bottleneck, not the training method

# Tools for LLM Evaluation Annotation

**Specialized platforms:**

- **Argilla:** Built for RLHF data collection
  - Native preference annotation
  - Hugging Face integration
  - Feedback collection workflows
- **Label Studio:** General purpose, adaptable
  - Custom templates for comparison
  - LLM backend integration
- **Scale AI / Surge AI:** Commercial platforms
  - Managed annotator workforce
  - Quality control built-in

# IAA for Preference Annotation

**Measuring agreement on subjective tasks:**

**Metrics:**

- **Raw agreement:** % of pairs where annotators agree
- **Cohen's Kappa:** Agreement adjusted for chance
- **Krippendorff's Alpha:** For rankings/ordinal data

**Typical values:**

- Clear quality differences: $\kappa > 0.6$
- Similar quality responses: $\kappa \approx 0.3 - 0.5$
- Highly subjective: $\kappa < 0.3$

**Important:** Low agreement isn't always a problem – it may reflect genuine subjectivity that should be captured in the data

# Group Presentations Today

**Each group presents their chosen annotation task**

**Please cover:**

1. What task are you annotating?
2. What data will you use?
3. What is your initial schema?
4. Why is this task interesting/important?
5. What challenges do you anticipate?

**Format:** 5-10 minutes per group + questions

**Feedback:** Peer feedback will help refine your approach

# Next Class: Writing Annotation Guidelines

**Lecture 11 (Feb 25):** Writing Annotation Guidelines

**Topics:**

- Anatomy of good annotation guidelines
- Positive and negative examples
- Handling edge cases and ambiguity
- Guidelines for humans vs. prompts for LLMs
- Prompt engineering principles

**Reading:** Pustejovsky & Stubbs, Chapter 7

# Key Takeaways

1. **RLHF** uses human preferences to align LLMs
2. **Preference annotation** asks "which is better?" not "what is correct?"
3. **Safety annotation** requires clear definitions and annotator support
4. **Instruction-following** evaluates whether models follow user intent
5. **Multi-turn** annotation adds complexity of context across exchanges
6. **IAA** for subjective tasks may be naturally lower – that's okay

# Questions?

Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

jinzhao@brandeis.edu