COSI-230B: Natural Language Annotation for Machine Learning

Lecture 17: Human-AI Collaborative Annotation

Jin Zhao

Brandeis University

Spring 2026

Today's Agenda

1. The human-in-the-loop paradigm
2. LLM pre-annotation and human correction
3. Active learning with LLMs
4. Confidence-based routing
5. Efficiency gains from hybrid approaches
6. When to trust LLM annotations
7. Quality assurance and best practices

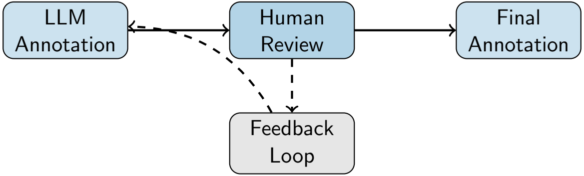
The Annotation Landscape in 2026

Three approaches:

1. **Pure human annotation:**
   - High quality, high cost, slow
2. **Pure LLM annotation:**
   - Low cost, fast, variable quality
3. **Human-AI collaborative:**
   - Best of both worlds?

Key insight: The optimal approach depends on task, budget, and quality requirements

Human-in-the-Loop Paradigm



Workflow:

1. LLM generates initial annotations
2. Humans review and correct
3. Corrections can improve LLM prompts
4. Iterate until quality is satisfactory

LLM Pre-Annotation Workflow

Step-by-step process:

1. **Design prompt:** Create annotation prompt with examples
2. **Run LLM:** Generate annotations for all items
3. **Import to tool:** Load pre-annotations into Label Studio/etc.
4. **Human review:** Annotators verify and correct
5. **Quality check:** Measure human changes
6. **Iterate:** Improve prompt based on common corrections

Key benefit: Correcting is faster than creating from scratch

Why Pre-Annotation Helps

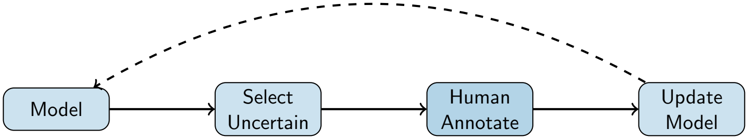
Cognitive load reduction:

- **Without pre-annotation:**
  - Read text
  - Identify all relevant items
  - Decide labels
  - Create annotations
- **With pre-annotation:**
  - Read text
  - Verify existing annotations
  - Fix errors

Typical speedup: 2–5x faster annotation

Active Learning

Smart selection of what to annotate



Idea: Annotate examples where model is most uncertain
Result: More learning per annotation dollar spent

Active Learning with LLMs

Using LLM confidence for selection:

1. LLM annotates with confidence scores
2. **High confidence:** Accept automatically
3. **Low confidence:** Route to human
4. Humans annotate uncertain cases
5. Use human labels to improve prompt

Challenge: LLM confidence may not reflect true uncertainty
Mitigation:

- Run multiple times, check consistency
- Ask LLM to explain reasoning
- Calibrate on validation set

Uncertainty Metrics

For classification, uncertainty can be measured by:

- **Entropy:** High entropy = predictions spread across classes
- **Margin:** Difference between top two probabilities; small margin = uncertain
- **Least confidence:** Low top probability = uncertain

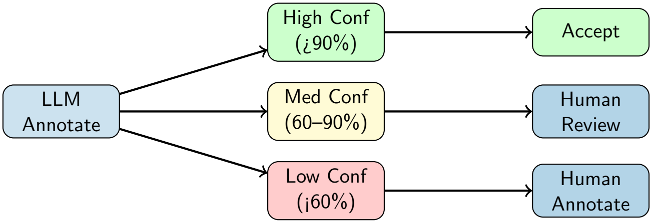
Examples:

- $[0.98, 0.01, 0.01]$ — very confident
- $[0.35, 0.33, 0.32]$ — very uncertain
- $[0.60, 0.38, 0.02]$ — uncertain (small margin)

For LLMs: Use token probabilities, self-consistency, or explicit confidence prompting

Confidence-Based Routing

Example workflow:



Setting Thresholds

Threshold selection is critical:

- **Too high "accept" threshold:** Humans review everything — no efficiency gain
- **Too low "accept" threshold:** Errors slip through — quality suffers

Finding the right threshold:

1. Create a validation set with gold-standard human labels
2. Run LLM with confidence scores on validation set
3. Measure error rate at each threshold
4. Choose threshold that meets your quality requirements

Typical thresholds: Accept if confidence > 0.95, human verify if 0.7–0.95, full annotation if < 0.7

Efficiency Gains

Measuring the benefit:

Metrics:

- Annotations per hour (with vs. without pre-annotation)
- Cost per annotation
- Time to complete dataset
- Quality maintained (IAA, accuracy)

Typical results:

- 2–5x speedup with pre-annotation
- 50–80% cost reduction
- Quality comparable or higher (fewer oversights)

When to Trust LLM Annotations

Factors that increase trustworthiness:

- ✓ Clear, objective task
- ✓ Well-defined categories
- ✓ Common knowledge domain
- ✓ High-resource language
- ✓ Consistent outputs across runs

Factors that decrease trustworthiness:

- × Subjective judgments
- × Domain expertise required
- × Novel or specialized concepts
- × Low-resource language
- × Inconsistent outputs

Quality Assurance Strategies

Ensuring hybrid annotation quality:

1. **Validation set:** Compare LLM to human on gold set
2. **Sample review:** Human checks random subset
3. **Agreement tracking:** Monitor LLM-human agreement
4. **Error analysis:** Categorize LLM mistakes
5. **Iterative improvement:** Fix systematic errors in prompt

Red flags:

- Human corrections $>$ 30% of cases
- Systematic bias in LLM outputs
- Decreasing agreement over time

Case Study: NER Pre-Annotation

Scenario: Named entity annotation for 10,000 sentences

Pure human:

- 2 min/sentence
- 333 hours total
- $5,000 cost

LLM + human:

- LLM pre-annotate: $50
- Human correct: 0.5 min/sent
- 83 hours human time
- $1,300 cost

Result: 74% cost reduction, similar quality

Challenges in Hybrid Systems

Key challenges to be aware of:

- **Feedback loops:** Only annotating uncertain items may leave gaps in model coverage
- **Calibration drift:** Model confidence becomes uncalibrated over time
- **Human trust issues:**
  - If AI is often wrong $\rightarrow$ humans lose trust, second-guess everything
  - If AI is often right $\rightarrow$ humans get complacent, miss errors
- **Evaluation difficulty:** Hard to evaluate system performance when humans and AI contribute differently

Automation Bias

Annotators may over-rely on LLM suggestions, accepting errors they would catch if annotating from scratch.

Ethical Considerations

Transparency:

- Document use of LLM assistance
- Report what was AI-generated vs. human-verified

Labor implications:

- Changing role of human annotators
- From creators to reviewers
- Skill requirements may change

Bias propagation:

- LLM biases can enter dataset
- Human review should catch these
- Document and measure

Best Practices

For successful human-AI collaboration:

1. **Start with validation:** Test LLM quality before full run
2. **Clear instructions:** Tell humans how to review pre-annotations
3. **Track changes:** Monitor what humans correct
4. **Iterate prompts:** Improve based on common errors
5. **Maintain standards:** Don't let AI lower the quality bar
6. **Document everything:** Record the hybrid process for reproducibility

Remember: The goal is quality data, not just cheap data

Key Takeaways

1. **Human-in-the-loop** combines LLM efficiency with human quality
2. **Pre-annotation** can speed up annotation 2–5x
3. **Active learning** focuses human effort on uncertain cases
4. **Confidence routing** automates easy cases, escalates hard ones
5. **Quality assurance** is essential — always validate
6. **Document** your hybrid process for reproducibility

Discussion

What annotation tasks in your experience would benefit most from human-AI collaboration? What tasks would be hardest to automate?

Questions?

Office Hours: Wednesdays 1–3pm, Volen 109

✉ jinzhao@brandeis.edu