Human-AI Collaborative Annotation

Combining Human Expertise with LLM Efficiency

Jin Zhao

Brandeis University

March 18, 2025

Today's Agenda

1. The human-in-the-loop paradigm
2. LLM pre-annotation + human correction
3. Active learning with LLMs
4. Efficiency gains from hybrid approaches
5. When to trust LLM annotations
6. Measuring productivity gains
7. Best practices

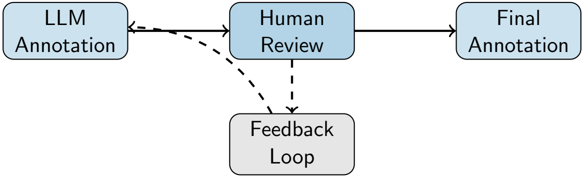
The Annotation Landscape in 2025

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

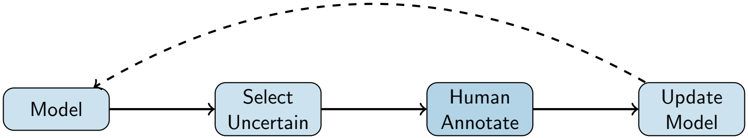
Cognitive load reduction:

- **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

Active Learning with LLMs

Using LLM confidence for selection:

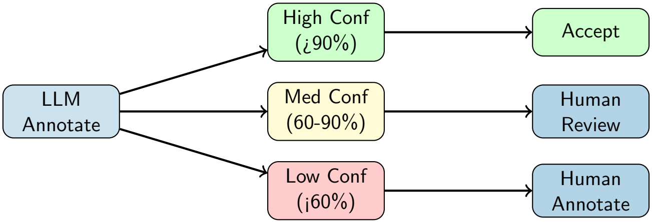
1. LLM annotates with confidence scores
2. **High confidence:** Accept automatically
3. **Low confidence:** Route to human
4. Humans annotate uncertain cases
5. Use human labels to improve prompt

Challenge: LLM confidence may not reflect true uncertainty
Mitigation:

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

Confidence-Based Routing

Example workflow:



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

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
3. **Track changes:** Monitor what humans correct
4. **Iterate prompts:** Improve based on common errors
5. **Maintain standards:** Don't let AI lower quality bar
6. **Document:** Record the hybrid process

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

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

Next Class: Inter-Annotator Agreement I

Lecture 18 (Mar 23): Inter-Annotator Agreement I

Topics:

- Why measure agreement?
- Observed agreement and its limitations
- Cohen's Kappa (2 annotators)
- Interpreting Kappa values

Project: IAA evaluation due soon
Assignment: HW 3 assigned

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

Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

jinzhao@brandeis.edu