COSI-230B: Natural Language Annotation for Machine Learning

Lecture 12: Adjudication & Gold Standards

Jin Zhao

Brandeis University

Spring 2026

Today's Agenda

1. Why resolve annotator disagreements?
2. Adjudication strategies: majority voting, expert review, discussion
3. Weighted and probabilistic adjudication
4. LLM-assisted adjudication
5. Creating gold standard datasets
6. Gold standard data splits and documentation
7. From annotations to trained models
8. Annotation quality and model performance

Theme: Disagreements are data. Gold standards are decisions about that data.

Why Resolve Disagreements?

Disagreements are inevitable

But ML models need single labels:

- Supervised learning requires ground truth
- Multiple labels per instance is problematic
- Evaluation needs clear correct answers

Creating gold standard:

- Authoritative, adjudicated labels
- Used for training and evaluation
- Represents "correct" annotation

Important

Disagreements carry information. Don't just discard them — they reveal task difficulty and data quality.

Adjudication Strategies

Common approaches:

1. **Majority voting:**
   - Most common label wins
   - Simple, scalable
   - Requires odd number of annotators
2. **Expert adjudication:**
   - Expert reviews disagreements
   - Higher quality, more expensive
3. **Discussion:**
   - Annotators discuss until consensus
   - Time-intensive but educational
4. **Probabilistic:**
   - Weight by annotator reliability
   - More sophisticated

Majority Voting

Simple and effective

Process:

1. Collect labels from all annotators
2. For each item, select most frequent label
3. Handle ties (random, expert review, or skip)

Advantages:

- Easy to implement
- Scalable to large datasets
- No additional annotation needed

Disadvantages:

- Ignores annotator quality differences
- Majority can be wrong
- Doesn't improve guidelines

Expert Adjudication

For higher quality gold standard

Process:

1. Identify items with disagreement
2. Expert reviews each case
3. Expert makes final decision
4. Optionally: update guidelines based on patterns

When to use:

- High-stakes tasks
- Creating evaluation benchmarks
- Complex annotation requiring domain expertise

Cost: More expensive and slower, but produces higher-quality labels

Weighted & Probabilistic Adjudication

Not all annotators are equal

Annotator weighting:

- Estimate each annotator's reliability from agreement data
- Weight votes by reliability score
- Experienced annotators count more than novices

Dawid–Skene model (1979):

- EM algorithm to estimate true labels and annotator error rates simultaneously
- Models each annotator's confusion matrix
- Iterates: estimate labels $\rightarrow$ update error rates $\rightarrow$ re-estimate labels

When to use:

- Large-scale crowdsourcing (many annotators, variable quality)
- When you have enough data to estimate reliability

Discussion: Choosing a Strategy

For each scenario, which adjudication strategy would you choose? Why?

Scenario 1

You have 10,000 tweets labeled for sentiment by 5 crowdworkers each. Budget is limited. You need training data, not a benchmark.

Scenario 2

You are building a clinical NER dataset for FDA submission. Three expert annotators labeled 500 medical records. Accuracy is critical.

Scenario 3

Two annotators labeled 2,000 items. They agree on 85% but disagree sharply on the rest. One annotator has twice the experience.

LLM-Assisted Adjudication

Emerging approach for 2025+

Options:

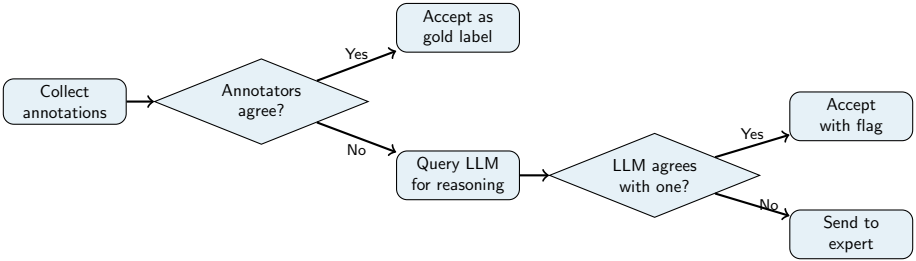
1. **LLM as tie-breaker:**
   - When annotators disagree, ask LLM
   - Use as third "vote"
2. **LLM reasoning:**
   - Ask LLM to explain which label is correct
   - Human reviews LLM reasoning
3. **LLM confidence:**
   - Accept human majority if LLM confidence low
   - Review if LLM disagrees with high confidence

Caution

LLM shouldn't be sole adjudicator for evaluation data — circular evaluation risk if the same model family is later tested on that data.

LLM Adjudication in Practice

A practical workflow:



Benefits:

- Reduces expert review load by 40–60%
- LLM reasoning helps experts decide faster
- Creates audit trail of disagreement rationales

Creating Gold Standard Dataset

Complete workflow:

1. Multiple annotators label data
2. Calculate IAA to verify quality
3. Identify disagreements
4. Apply adjudication strategy
5. Verify adjudicated labels
6. Document process

Best practices:

- Keep original annotations (for analysis)
- Document adjudication decisions
- Track problematic patterns
- Update guidelines for future annotation

Gold Standard Data Splits

Splitting your gold standard for ML:

Critical rules:

- **Test set:** Highest quality adjudication (expert review preferred)
- **Training set:** Majority voting acceptable (noise is tolerable)
- **Dev set:** Used for tuning — moderate quality needed
- **Never** let adjudication leakage cross split boundaries

Stratify: Ensure label distribution is balanced across splits

Documenting Adjudication Decisions

Every adjudication decision should be traceable:

Item	Labels	Strategy	Decision	Rationale
sent_042	A: Pos, B: Neg	Expert review	Negative	Rhetorical structure signals negative
sent_107	A: Neu, B: Pos, C: Pos	Majority vote	Positive	2/3 agree
sent_215	A: Neg, B: Neg, LLM: Pos	Human majority	Negative	LLM misread sarcasm

Why document?

- Reproducibility: others can audit your gold standard
- Error analysis: patterns in disagreements reveal task difficulty
- Guideline improvement: recurring issues inform future revisions

From Annotations to Models

The ML pipeline:

Annotation quality directly affects model quality

- Noisy labels → confused model → lower accuracy
- Clean labels → clear signal → better generalization
- Biased labels → biased model → unfair predictions

Training on Annotated Data

Classification tasks:

- Input: text features (BoW, embeddings)
- Output: predicted label
- Algorithms: Logistic Regression, SVM, Neural Networks

Sequence labeling tasks:

- Input: token sequences
- Output: BIO label sequences
- Algorithms: CRF, BiLSTM-CRF, BERT-NER

Key insight: Model learns patterns from annotations

- If annotators labeled sarcasm inconsistently, the model will too
- The model can only be as good as the data it trains on

Evaluation Metrics

Classification:

- Accuracy, Precision, Recall, F1
- Confusion matrix

Sequence labeling:

- Token-level accuracy
- Entity-level Precision/Recall/F1
- Exact match vs. partial match

Important:

- Evaluate on held-out test set
- Report multiple metrics
- Compare to baseline
- Consider human upper bound (IAA as ceiling)

Annotation Quality and Model Performance

The relationship:

- Higher IAA $\rightarrow$ cleaner training signal
- Cleaner signal $\rightarrow$ better model
- Better model $\rightarrow$ higher accuracy

Upper bound:

- Human agreement is ceiling for model
- If humans agree 80%, model likely $\leq$ 80%
- (Though models can sometimes beat individuals)

Rule of thumb: Expect model F1 $\approx$ Human IAA

Common Modeling Pitfalls

Data issues:

- Training on unadjudicated data
- Test set contamination
- Class imbalance

Evaluation issues:

- Overfitting to dev set
- Cherry-picking metrics
- Not comparing to baseline

Process issues:

- Not documenting preprocessing
- Inconsistent tokenization
- Not releasing data/code

Discussion: Quality vs. Scale

A common tension in annotation projects:

High Quality	High Scale
Expert adjudication on all items	Majority voting only
Small dataset (1,000 items)	Large dataset (50,000 items)
$\kappa = 0.85$	$\kappa = 0.65$
High confidence in labels	Some label noise

Questions to discuss:

- Which would you prefer for training a model? For evaluation?
- Can you have both? What strategies would help?
- How does the task complexity affect this tradeoff?

Key Takeaways

1. **Adjudication** resolves disagreements to create gold standard
2. **Majority voting** is simple but may ignore annotator quality
3. **Expert adjudication** produces higher quality but costs more
4. **Weighted methods** (Dawid–Skene) model annotator reliability
5. **LLMs can assist** adjudication but shouldn't be sole judge
6. **Document** your adjudication process for reproducibility
7. **Model quality** is bounded by annotation quality

Questions?

Questions?

Office Hours: Wednesdays 1–3pm, Volen 109

jinzhao@brandeis.edu