COSI-230B: Natural Language Annotation for Machine Learning

Lecture 11: Disagreement as Data

Jin Zhao

Brandeis University

Spring 2026

Today's Agenda

Part I: Rethinking Disagreement

1. Why agreement is **not** the goal
2. A taxonomy of disagreement types
3. When ambiguity is signal, not noise

Part II: Working with Disagreement

4. Soft labels and label distributions
5. Preserving disagreement in datasets

Part III: The LLM Dimension

6. LLM ensemble disagreement as ambiguity proxy
7. Why majority vote + LLMs erase minorities
8. When **not** to collapse disagreement

The Standard Pipeline — and Its Assumption

Collect Annotations → Adjudicate Disagreements → Produce "Gold" Labels → Train Model

The Hidden Assumption

Disagreement = error. The standard pipeline treats annotator disagreement as noise to be resolved, not as meaningful information about the task.

- Adjudication strategies: majority vote, expert tiebreaker, discussion until consensus
- All assume there is a single correct answer waiting to be found
- **But what if some items genuinely have multiple valid interpretations?**

Why Agreement Is Not the Goal

Agreement-Centric View

- High κ = good schema
- Low κ = fix guidelines or fire annotators
- Single gold label per instance
- Disagreement is always a problem

Disagreement-Aware View

- High κ on some, low on others — **both informative**
- Low agreement may reflect **genuine ambiguity**
- Label distributions capture richer information
- Disagreement patterns reveal task structure

Example: "This movie is not bad."

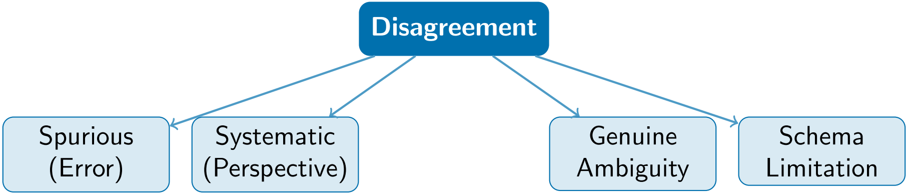
Annotator	Sentiment
A1	Positive
A2	Neutral
A3	Positive
A4	Neutral
A5	Negative

Majority vote $\rightarrow$ Positive

But is **Positive** really the "truth"?

The **distribution** (0.4, 0.4, 0.2) tells us more than a single label.

A Taxonomy of Disagreement — Sources



Spurious (Error-Based)

Inattention, misunderstanding guidelines, fatigue. **This** is noise — what training and QC address.

Systematic (Perspective)

Different backgrounds lead to different but **defensible** judgments (e.g., cultural norms).

Genuine Ambiguity

The text supports multiple readings. "I saw her duck" — syntactic ambiguity is real.

Schema Limitation

Categories don't cover the phenomenon well. Annotators improvise when forced into ill-fitting labels.

Disagreement Types — Concrete NLP Examples

Type	Example Text	What Happens	Correct Response
Spurious	"The food was great!"	One annotator labels Negative (wrong button)	QC filtering
Systematic	"That's so ghetto"	Offensive vs. Not offensive (in-group usage)	Preserve perspectives
Ambiguity	"I could care less"	Sarcasm? Idiom misuse? Indifference?	Soft labels
Schema	"I'm happy for you *but*..."	Mixed sentiment—forced binary fails	Revise schema

Key Insight

Only the first type (spurious) should be treated as noise. The other three carry **information** that collapsing to a single label destroys.

When Ambiguity Is Signal, Not Noise

Ambiguity as a Linguistic Phenomenon

- **Lexical**: "bank" (river vs. financial)
- **Syntactic**: "I saw the man with the telescope"
- **Pragmatic**: "Can you pass the salt?"

In Annotation Tasks

- Sentiment of hedged statements
- Toxicity of sarcasm and irony
- Hate speech: in-group vs. out-group

Case Study: Toxicity Annotation

Text: "Women belong in the kitchen — and in the boardroom, the lab, and everywhere."

Ann.	**Label**
A1	Toxic
A2	Not Toxic
A3	Not Toxic
A4	Toxic
A5	Not Toxic

First clause triggers toxicity frame; full text subverts it. **Both readings are valid.**

The Cost of Collapsing Disagreement

Step 1: 5 annotators label 1,000 items

Step 2: Majority vote → single gold label

⚠ **Lost:** Minority perspectives on 300+ items

Step 3: Train model on gold labels

⚠ **Lost:** Model learns "there is always one answer"

Step 4: Evaluate against gold labels

⚠ **Lost:** Penalizing valid minority interpretations

Soft Labels and Label Distributions

Hard Label

$$y_{\text{hard}} = \arg\max_c \text{count}(c)$$

One-hot encoding: $[0, 1, 0]$

Information: which class "won"

Soft Label (Distribution)

$$y_{\text{soft}} = \left[\frac{\text{count}(c_i)}{\sum_j \text{count}(c_j)}\right]$$

Distribution: $[0.2, 0.6, 0.2]$

Information: how annotators distributed across classes

Same Majority, Different Stories

Item A: 5/5 annotators say Positive
Hard: $[0, 0, 1]$ Soft: $[0.0, 0.0, 1.0]$
→ Clear, unambiguous positive

Item B: 3/5 Positive, 2/5 Negative
Hard: $[0, 0, 1]$ Soft: $[0.4, 0.0, 0.6]$
→ Contested! Near the boundary

Item C: 3/5 Positive, 1 Neutral, 1 Neg
Hard: $[0, 0, 1]$ Soft: $[0.2, 0.2, 0.6]$
→ Genuinely ambiguous

All three get the same hard label but carry very different information.

Training with Soft Labels — Methods

Why Use Soft Labels for Training?

Models trained on soft labels learn that some items are inherently uncertain, producing better-calibrated and fairer predictions.

Loss Functions

Hard labels: Cross-entropy with one-hot targets

$$\mathcal{L} = -\log p(y_{\text{gold}} \mid x)$$

Soft labels: KL divergence from distribution

$$\mathcal{L} = \mathrm{KL}(y_{\text{soft}} \| p(\cdot \mid x))$$

Or cross-entropy with soft targets as a label smoothing variant.

Practical Approaches

1. **Distribution matching**: Predict the full annotator distribution
2. **Multi-annotator**: Keep all labels; train with repeated items
3. **Mixture of experts**: Route ambiguous items to specialized heads
4. **Calibration**: Use disagreement as confidence signal

Preserving Disagreement in Datasets

What to Store

- **All individual annotations**
- **Annotator IDs** (pseudonymized)
- **Annotator metadata** (when appropriate)
- **Disagreement flags** per item

Exemplar Datasets

Dataset	**Practice**
ChaosNLI	100 annotators; full distributions
Social Bias	Annotators + demographic info
DICES	350+ raters with demographics
HS-Brexit	Perspectives preserved

Principle

Always release individual annotations alongside aggregated labels — collapsed data **cannot** be uncollapsed.

Measuring Disagreement Beyond Kappa

Item-Level Disagreement Metrics

- **Entropy of label distribution:**

$$H(y_i) = -\sum_c p(c) \log p(c)$$

  High entropy = high disagreement

- **Annotator agreement ratio:**

$$\mathrm{AR}(i) = \frac{\max_c \mathrm{count}(c)}{n}$$

- **Variance of annotations** (for continuous scales)

Disagreement Profiles

Partition items by agreement level:

Zone	AR	Interpretation
Consensus	> 0.8	Clear cases
Majority	0.6–0.8	Leans one way
Contested	0.4–0.6	Near boundary
Chaotic	< 0.4	No dominant view

Analyzing **what falls in each zone** reveals task structure.

Systematic Disagreement and Annotator Identity

When Identity Shapes Annotation

- **AAE**: Speakers judge AAE text as non-toxic more often than non-speakers (Sap et al., 2019)
- **Gender**: Women annotators identify subtle harassment men may miss (Al Kuwatly et al., 2020)
- **Political stance**: Leaning predicts disagreement on political speech

The Representational Problem

If your annotator pool is predominantly one demographic, "majority vote" encodes that demographic's perspective, not objective truth.

Case Study: Hate Speech (Kennedy et al., 2020)

- 39,565 comments; avg 5.8 annotators with demographic self-reports
- **African American annotators** rated anti-Black content as more hateful
- **LGBTQ+ annotators** more sensitive to anti-LGBTQ+ content
- Majority vote **systematically underestimated** severity for targeted communities

Mitigation: Stratified recruitment, perspectival protocols, results per demographic group.

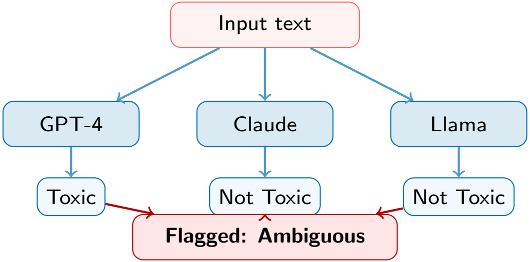
LLM Ensemble Disagreement as Ambiguity Proxy

The Idea

1. Query **multiple LLMs** (or same LLM, different prompts/temperatures)
2. Record label distribution across runs
3. High LLM disagreement ≈ high ambiguity

Three Strategies

- **Multi-Model**: Query 3–5 LLMs; captures training data diversity
- **Temperature Sampling**: Same model, varying T; measures uncertainty
- **Prompt Variation**: Rephrase prompt 5–10 ways; tests framing robustness



Caveat: LLM disagreement is a **proxy**, not a replacement. Always validate against human patterns.

Why Majority Vote + LLMs Can Erase Minority Interpretations

The Compounding Problem

1. Human majority vote already loses minority views
2. LLM training data reflects majority perspectives
3. LLM annotation adds another majority-default layer
4. Majority vote over LLM outputs: **triple majority filtering**

Concrete Harm

- In-group reclaimed slurs labeled as toxic
- AAVE labeled as “ungrammatical”
- Sarcasm in marginalized communities misclassified

Diverse human perspectives

↓ loss

Human majority vote

↓ loss

LLM trained on majority labels

↓ loss

LLM majority vote output

Minority views fully erased

When NOT to Collapse Disagreement for Training

Collapse Is Acceptable When:

- Task has objectively verifiable answers (e.g., POS tagging)
- Disagreement is clearly spurious (QC issues)
- Downstream task requires crisp decisions

Do NOT Collapse When:

- **Subjective tasks**: sentiment, toxicity, humor
- **Culturally sensitive**: hate speech, harassment
- **Safety-critical**: content moderation at scale
- **Systematic disagreement**: correlates with demographics

A Decision Framework

1. Measure item-level disagreement (entropy, agreement ratio)
2. Partition into zones (consensus / contested / chaotic)
3. Analyze *why* contested items disagree — only collapse when uninformative

Modeling Approaches That Preserve Disagreement

Approach	Description	Advantages	Limitations
Multi-task per annotator	One head per annotator; shared representation	Models individual views	Doesn't generalize
Distribution prediction	Predict label distribution	Captures ambiguity	Needs many annotators
Jury Learning	Learn which "jury" for which items	Flexible	Complex setup
Curricula by agreement	Train on high-agreement items first	Better convergence	No disagreement at inference

Emerging Trend: Perspectival AI

Rather than one answer, models output **multiple perspectives** with likelihoods: "From perspective A, toxic (0.7); from B, not toxic (0.8)."

LLMs and Disagreement — What the Research Shows

Where LLMs **Agree** with Majority

- Clear-cut factual tasks: NER, POS tagging
- Unambiguous sentiment (strongly positive/negative)
- Items in the "consensus zone" (AR > 0.8)

Where LLMs **Diverge** from Humans

- Sarcasm detection (especially cultural)
- Implicit toxicity and microaggressions
- Items requiring lived experience

Key Empirical Findings

1. LLMs correlate with majority vote at $r \approx 0.75$–0.85 on subjective tasks (Gilardi et al., 2023)
2. But they **fail to capture** the variance structure of human annotations
3. Prompt sensitivity creates "artificial disagreement" unrelated to true ambiguity

Bottom line: LLM disagreement signals are useful but must be validated against human patterns.

A Framework for Disagreement-Aware Annotation

A Six-Step Process

1. **Design:** Build schema expecting disagreement — include "ambiguous" category
2. **Recruit:** Diverse annotator pool — stratify by relevant demographics
3. **Annotate:** Collect 5+ annotations per item; record individual labels + metadata
4. **Analyze:** Compute item-level disagreement; classify type (spurious / systematic / ambiguity / schema)
5. **Decide:** Consensus → hard label; contested → soft label; chaotic → investigate
6. **Release:** Publish individual annotations + metadata for downstream flexibility

Key Principle

This framework requires treating disagreement as **data**, not a problem. The step most projects skip is Step 4 (Analyze) — arguably the most important.

Common Pitfalls in Handling Disagreement

Pitfall 1: Forced Agreement

Requiring discussion until consensus. This **suppresses** legitimate differences; the loudest voice wins.

Pitfall 2: Expert Override

"The expert says toxic, so it's toxic." On subjective tasks, expertise $\neq$ objectivity.

Pitfall 3: Filtering "Bad" Annotators

Removing low-agreement annotators. If disagreement is systematic, you remove a valid perspective.

Pitfall 4: Ignoring Patterns

Reporting only aggregate κ without analyzing *which* items and annotators drive disagreement.

Pitfall 5: Assuming LLMs "Solve" It

Using LLMs to break ties. LLMs default to majority perspectives in their training data.

Instead...

- Let disagreement stand where informative
- Investigate patterns before resolving
- Design for multiplicity, not unanimity

Key Takeaways

Conceptual Shifts

1. **Disagreement $\neq$ noise.** Much of it is signal about the task, text, or annotators.
2. **Agreement is not the goal.** Understanding the *structure* of disagreement is.
3. **Gold labels are lossy compression** of richer annotation data.
4. **Majority vote encodes majority perspectives**, which may disadvantage minority viewpoints.

Practical Actions

1. **Preserve** individual annotations alongside aggregated labels
2. **Analyze** disagreement before resolving it
3. Use **soft labels** for subjective tasks
4. **Validate** LLM signals against human patterns
5. Build **disagreement-aware** pipelines

The One-Liner

"If all your annotators agree on everything, either the task is trivial or you've suppressed the interesting signal."

Questions?

Questions?

Office Hours: Wednesdays 1–3pm, Volen 109

jinzhao@brandeis.edu