COSI-230B: Natural Language Annotation for Machine Learning

Lecture 10: Guideline Iteration & Annotator Feedback

Jin Zhao

Brandeis University

Spring 2026

Today's Agenda

1. Confusion vs. disagreement: why the distinction matters
2. Wording vs. theory revision: two tools for two problems
3. Guideline overfitting: when adding rules makes things worse
4. The iteration cycle: a principled workflow
5. Discussion: diagnose real cases
6. LLMs as simulated annotators for guideline testing
7. Revision rationales: documenting why guidelines changed

Theme: If your guidelines didn't change after testing, you didn't actually test them.

Why Guidelines Must Change

The Myth

"Good guidelines are written once, correctly, and then followed."

The reality:

- Guidelines are **hypotheses** about how to carve up a phenomenon
- Pilots are **experiments** that test those hypotheses
- Low agreement is **data**, not failure

Analogy

Guidelines are like software. Version 1.0 ships with bugs. If v1.0 and v2.0 are identical, nobody ran the tests.

Confusion vs. Genuine Disagreement

When annotators disagree, exactly one of two things is happening:

Confusion

Annotators *misunderstand* the guidelines

- Ambiguous wording
- Missing procedure step
- Undefined term
- Example gaps

Fix: Revise the guidelines

Genuine Disagreement

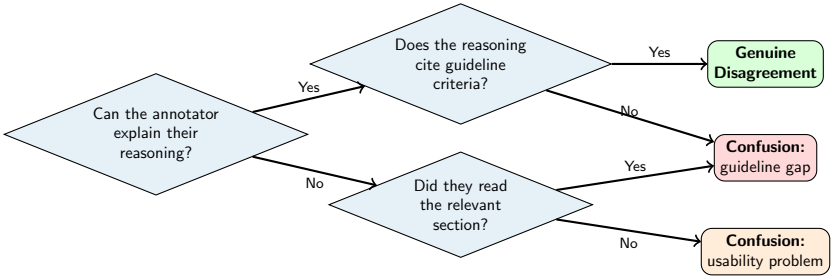
Annotators *understand* but the phenomenon is ambiguous

- Vague language in the data
- Subjective constructs
- Context-dependent meaning
- Cultural variation

Fix: Revise the *theory* or accept variance

You cannot choose the right fix until you know which one you're dealing with.

A Diagnostic Protocol



Key: If the annotator cites your criteria and still disagrees → genuine disagreement. If they can't explain why → confusion, and the fix is in your document.

Example: Confusion

Task: Label whether a sentence expresses *uncertainty*
Sentence: "The treatment might work for some patients."

Annotator A: Uncertain
"It says 'might' — that's hedging."

Annotator B: Not Uncertain
"It's stating an objective possibility."

Diagnosis: The guidelines didn't distinguish:

- **Epistemic uncertainty:** speaker doesn't know ("I think it might work")
- **Objective possibility:** the world is variable ("It might rain in April")

This is **confusion**. Add the distinction → both annotators would agree.

Example: Genuine Disagreement

Task: Label movie review as *positive* or *negative*
Review: "The acting was phenomenal, the cinematography breathtaking, and yet I left the theater feeling absolutely nothing."

Annotator A: Positive

"Two explicit praise statements outweigh one vague complaint."

Annotator B: Negative

"The 'and yet' signals praise is setup for a negative conclusion."

Both understand the guidelines. They disagree about which signal dominates: **surface praise** vs. **rhetorical structure**.

Options: Add priority rule ("structure overrides surface"), add a **Mixed** label, or accept the variance.

Two Kinds of Revision

	Revising Wording	**Revising Theory**
What changes	How you explain the categories	What the categories *are*
Trigger	Annotators are confused	Schema doesn't fit the data
Example	Clarify "might" = epistemic only	Split "Negative" into Anger, Sadness, Fear
Scope	Local: one definition, one example	Global: label set, procedure structure

Most teams default to wording revisions because they're cheaper.
But some problems require theory revisions — no amount of clearer wording fixes a broken schema.

When Theory Revision Is Needed

Four signals that wording fixes won't help:

1. **One label catches too many phenomena**
   "Negative" collapses anger, sadness, boredom → split
2. **Two labels overlap systematically**
   "Suggestion" and "Request" used interchangeably → merge or sharpen
3. **Data has patterns you didn't anticipate**
   Binary schema meets a ternary phenomenon → add a label
4. **Agreement stays low after wording fixes**
   Three rounds of clarification, still 30%+ disagreement → the categories don't match

Case: Emotion Annotation

v1: {Positive, Negative, Neutral} → $\kappa = .52$ on Negative. Three rounds of rewriting the definition. No improvement.

Fix: Split Negative → {Anger, Sadness, Fear}. Agreement rose to $\kappa = .71$ immediately.

Guideline Overfitting

Definition

Adding so many specific rules that guidelines only work on the exact examples you tested on.

Rule-Based (overfitting risk)

- "`might` + medical → Uncertain"
- "`might` + weather → Not Uncertain"
- "`might` + legal → Uncertain"
- ...

New context = new rule. Grows without bound.

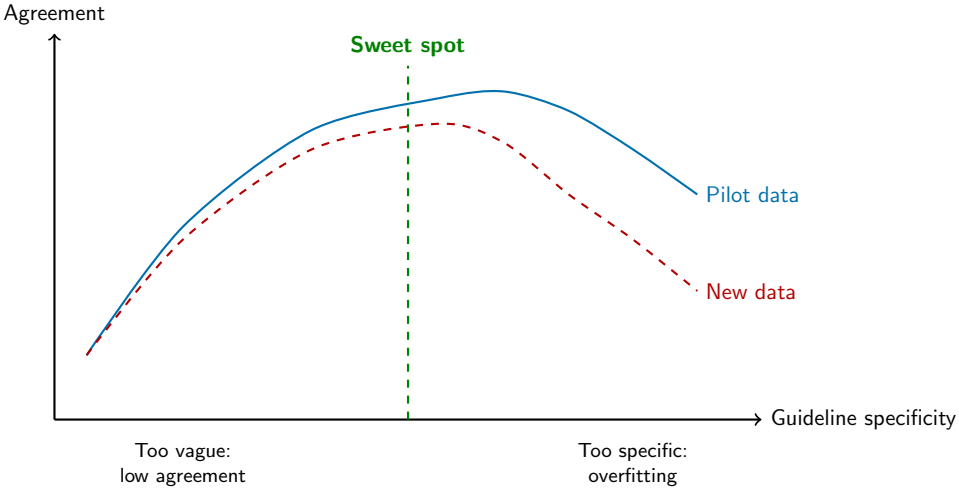
Principle-Based (generalizable)

- "Uncertain = speaker doesn't know (epistemic)"
- "Not Uncertain = objective variability"
- Test: "Could the speaker resolve it with more info?"

One principle covers all domains.

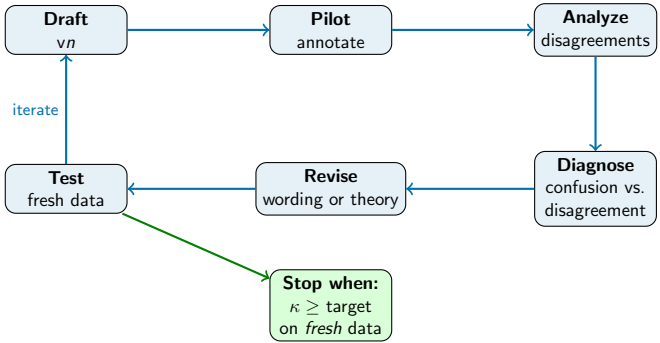
Warning sign: If you're writing increasingly long exceptions for edge cases, the problem is in the theory, not the wording.

The Overfitting Curve



Detect overfitting: (1) Fresh-data test — does agreement drop >10 κ points on new data?

The Iteration Cycle



Round	Typical Focus
v1 $\rightarrow$ v2	Wording clarity: ambiguous terms, missing steps, under-specified scope
v2 $\rightarrow$ v3	Boundary cases: priority conflicts, edge cases wording can't resolve
v3 $\rightarrow$ v4	Theory: labels that need splitting or merging
v4+	Consolidation: reducing exceptions to principles

Discussion: Confusion or Disagreement?

For each case: is this confusion or genuine disagreement? What's your fix?

Case 1: NER

"The White House issued a statement." Annotator A: **LOC** (building). Annotator B: **ORG** (administration). Guideline says: "Mark real-world entities with proper names."

Case 2: Sentiment

"The food was decent." Annotator A: **Positive**. Annotator B: **Negative**. Guideline says: "Positive = reviewer liked the product."

Case 3: Toxicity

"Men are trash." Annotator A: **Toxic** (targets a group). Annotator B: **Not toxic** (cultural expression, not literal). Guideline lists "targeting a group" as toxic.

Discussion: What Would You Fix?

Guideline excerpt: "Label the emotion expressed in the tweet. Labels: Happy, Sad, Angry, Other."

Annotator questions from the pilot:

1. "The tweet says 'I can't believe this happened again.' Is that Angry or Sad?"
2. "What if the tweet is about someone *else's* emotion, not the author's?"
3. "A tweet says 'lmao this is terrible.' Happy because laughing, or Sad because 'terrible'?"
4. "40% of my labels are Other. Is that normal?"

Questions to discuss:

- Which questions indicate confusion? Which indicate a theory problem?
- What's wrong with the "Other" rate, and what does it tell you?
- Propose one wording fix and one theory fix for this guideline.

LLMs as Simulated Annotators

LLMs and humans face the same guideline problems — but express them differently.

Three uses:

1. **Question generation:** Prompt the LLM to list what's unclear *before* labeling
2. **Assumption surfacing:** Prompt it to state assumptions not in the guidelines
3. **Failure mode detection:** Find cases where it confidently gives *wrong* labels

Key Insight

LLMs catch 40–60% of the same guideline gaps that human pilots find. But they miss usability problems (information that exists but is hard to find). Use both.

Two Prompts for Testing Guidelines

Prompt 1: "What Questions Would You Ask?"

"You are a new annotator. Read these guidelines: [paste]. Before labeling anything, list every question you would want to ask the project lead. For each, explain why the guidelines don't answer it."

→ Surfaces undefined terms, scope gaps, priority conflicts, missing cases.

Prompt 2: "State Your Assumptions"

"Read these guidelines: [paste]. Label this example: [paste]. Before giving your label, explicitly state every assumption you are making that is NOT in the guidelines."

→ Example: "I assume 'the text' means the entire post, not just the first sentence."

Every stated assumption is a guideline gap.

LLM Variance as a Diagnostic

Protocol: Run each ambiguous item 5× through the same LLM (temperature 0.7)

Example	R1	R2	R3	R4	R5	Diagnosis
"You're so smart" (sincere)	+	+	+	+	+	Clear
"You're so smart" (sarcastic)	−	−	+	−	−	Mild ambiguity
"Interesting take"	+	−	+	+	−	High ambiguity
"Whatever you say"	+	−	−	+	−	Guideline gap

Interpretation:

- 5/5 → clear 4/5 → needs boundary example 3/5 or worse → guideline gap

This is not measuring LLM quality. It's using LLM variance as a **proxy for guideline ambiguity**.

The Revision Rationale

When you revise, document why each change was made:

What Changed	Evidence	Type	Expected Impact
Added epistemic vs. objective distinction	3/8 pilot items disagreed on "might"	Wording: clarified definition	Resolves "might" disagreements
Split "Negative" $\rightarrow$ Anger, Sadness	$\kappa = .38$ on negative items	Theory: split label	Agreement should rise to $\sim .70$

Why? Prevents regressions, helps new collaborators understand design logic, creates an auditable trail.

Key Takeaways

1. **Confusion $\neq$ disagreement.**
   Confusion = fixable by revising guidelines. Disagreement = revise theory or accept variance.
2. **Wording fixes confusion. Theory fixes structure.**
   Three rounds of wording fixes that don't help $\rightarrow$ it's a theory problem.
3. **Guidelines can overfit.**
   Replace clusters of exceptions with principles. Test on fresh data.
4. **Every annotator question is a bug report.**
   Triage systematically: already answered, should have been, needs a decision, unanswerable.
5. **LLMs surface gaps but hide confusion.**
   Use variance and assumption-prompting. Don't trust confidence.
6. **Document your revisions.**
   A revision rationale prevents regressions and builds institutional memory.

Questions?

Office Hours: Wednesdays 1–3pm, Volen 109

jinzhao@brandeis.edu