Writing Annotation Guidelines

From Human Guidelines to LLM Prompts

Jin Zhao

Brandeis University

February 25, 2025

Today's Agenda

1. Why guidelines matter
2. Anatomy of good annotation guidelines
3. Positive and negative examples
4. Handling edge cases and ambiguity
5. Iterative guideline refinement
6. Guidelines for humans vs. prompts for LLMs
7. Prompt engineering principles for annotation

Why Guidelines Matter

Guidelines are the bridge between your conception and annotators' execution

Without good guidelines:

- Annotators interpret task differently
- Low inter-annotator agreement
- Inconsistent training data
- Poor model performance

With good guidelines:

- Consistent annotation across team
- Higher IAA scores
- Reproducible datasets
- Better ML models

Investment: Time spent on guidelines saves time in annotation and revision

Anatomy of Annotation Guidelines

Essential components:

1. **Task overview:** What are we annotating and why?
2. **Scope:** What text/units are we annotating?
3. **Label definitions:** Clear description of each category
4. **Positive examples:** Clear cases for each label
5. **Negative examples:** What does NOT qualify
6. **Edge cases:** Difficult cases with resolutions
7. **Decision procedures:** Steps for ambiguous cases
8. **Annotation workflow:** How to use the tool

Task Overview Section

Sets context for annotators

Example

Task: Named Entity Recognition for biomedical literature
Goal: Identify mentions of genes, proteins, diseases, and drugs in scientific abstracts. This data will be used to train an information extraction system for drug-disease relationship mining.
Annotation unit: Individual PubMed abstracts
Expected time: 5-10 minutes per abstract

Why it helps: Annotators make better decisions when they understand the purpose

Label Definitions

Be precise, not vague

Bad definition:
PERSON: Names of people

Problems:

- What about nicknames?
- What about titles?
- What about fictional characters?

Good definition:
PERSON: Proper names referring to specific individuals, including:

- Full names (John Smith)
- Partial names (John)
- Nicknames (Johnny)
- Fictional characters (Sherlock Holmes)

Exclude titles unless part of the proper name.

Positive Examples

Show clear cases for each label

PERSON examples

Annotate as PERSON:

- "Barack Obama was elected president." – Full name
- "Dr. Smith performed the surgery." – Title + name
- "Einstein's theory changed physics." – Single name
- "They call him The Rock." – Stage name

Best practices:

- Include diverse examples
- Show the annotation in context
- Explain why each qualifies

Negative Examples

Equally important as positive examples

NOT PERSON examples

Do NOT annotate as PERSON:

- "The president gave a speech." – Generic title, no name
- "A doctor examined the patient." – Generic profession
- "The author writes well." – Generic reference
- "People gathered in the square." – Generic noun

Why negative examples help:

- Define boundaries clearly
- Prevent over-annotation
- Address common mistakes

Handling Edge Cases

Edge cases are where disagreements happen

Example edge cases for PERSON

- **"President Biden":** Annotate as PERSON (title + name together)
- **"the Biden administration":** Annotate "Biden" only
- **"McDonald's":** NOT PERSON – it's a company name now
- **"@realDonaldTrump":** Annotate as PERSON (username = name)
- **"God":** Annotate as PERSON (specific entity)
- **"the defendant":** NOT PERSON (legal role, not name)

Process: Edge cases emerge from pilot annotation – add them to guidelines

Decision Procedures

Give annotators a process for difficult cases

Decision tree for PERSON

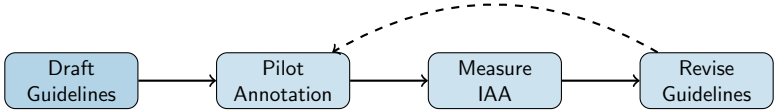
1. Is it a proper noun referring to a human? → If no, skip
2. Is it a specific individual (not generic)? → If no, skip
3. Does the mention include a name (not just title/role)?
   - If name + title: annotate full span
   - If title only: do not annotate
   - If name only: annotate

Benefits:

- Reduces cognitive load
- Increases consistency
- Makes disagreements traceable

Iterative Guideline Refinement

Guidelines evolve through the MAMA cycle



Each iteration:

1. Identify disagreements
2. Discuss with annotators
3. Add clarifications and examples
4. Re-annotate problematic cases

Rule of thumb: Plan for at least 3 guideline versions

Common Guideline Problems

Issues that lead to low IAA:

1. **Vague definitions:** "Annotate important entities"
2. **Missing edge cases:** No guidance for ambiguous instances
3. **Inconsistent examples:** Examples contradict rules
4. **Too complex:** Too many rules to remember
5. **Too simple:** Not enough guidance for real data
6. **Implicit assumptions:** Assumed knowledge not stated

Fix: Have someone unfamiliar with the task read guidelines and ask questions

Guidelines for Humans vs. Prompts for LLMs

Same goal, different formats

Human guidelines:

- Document format (PDF, Word)
- Detailed explanations
- Training sessions possible
- Can ask clarifying questions
- Learn from feedback

LLM prompts:

- Text in context window
- Concise instructions
- No interactive training
- Cannot ask questions
- Same response each time

Key insight: Good guidelines translate to good prompts
If you can explain it clearly to a human, you can prompt an LLM

Converting Guidelines to Prompts

Example transformation:

Human guideline (excerpt):
"Annotate named entities of type PERSON. This includes full names, partial names, and nicknames. Do not include generic references like 'the president' unless accompanied by a name."

LLM prompt:

Return as JSON: {"entities": [{"text": "...", "label": "PERSON"}]}" *"Extract all PERSON entities from the text. PERSON entities are proper names of specific individuals. Include full names (Barack Obama), partial names (Obama), and nicknames (The Rock). Do NOT include generic titles without names (the president, the doctor).*
Return as JSON: {"entities": [{"text": "...", "label": "PERSON"}]}"

Prompt Engineering for Annotation

Key principles:

1. **Be explicit:** State exactly what you want
2. **Show examples:** Few-shot learning is powerful
3. **Specify format:** Request structured output (JSON)
4. **Include edge cases:** Handle ambiguity in prompt
5. **Request reasoning:** Chain-of-thought can help

Template structure:

1. Task description
2. Label definitions
3. Examples (2-5 per label)
4. Output format specification
5. The text to annotate

Few-Shot Prompting Example

Include annotated examples in the prompt:

```
Task: Extract PERSON entities from the text.

Examples:
Text: "Barack Obama visited Paris."
Output: {"entities": [{"text": "Barack Obama",
         "label": "PERSON"}]}

Text: "The president spoke at the UN."
Output: {"entities": []}

Text: "Dr. Fauci recommended masks."
Output: {"entities": [{"text": "Dr. Fauci",
         "label": "PERSON"}]}
```

Validating LLM Annotation Quality

Always verify LLM annotations

Validation approaches:

1. **Sample review:** Human check random sample
2. **Agreement measurement:** Compare to human annotations
3. **Format validation:** Ensure output is well-formed
4. **Consistency checks:** Same input should give same output

Red flags:

- Annotations outside the text
- Inconsistent formatting
- Hallucinated entities
- Systematic errors on certain categories

Guideline Document Structure

Recommended sections for your semester project:

1. **Introduction:** Task goal and motivation
2. **Data description:** Source, format, scope
3. **Annotation schema:** All tags/labels with definitions
4. **Examples:** Positive, negative, and edge cases for each
5. **Decision procedures:** How to handle ambiguity
6. **Annotation workflow:** Tool instructions, shortcuts
7. **Quality control:** How to handle uncertainty
8. **Changelog:** Record of guideline revisions

Semester Project: Guidelines Due Soon

Upcoming deadlines:

- **Week 9 (Mar 11):** Draft 1 guidelines + pilot annotations
- **Week 10 (Mar 18):** Draft 2 guidelines

Draft 1 should include:

- Task overview and motivation
- Complete label definitions
- At least 3 examples per label (positive and negative)
- Initial edge case documentation
- Tool configuration

Pilot annotations: 50-100 instances annotated by all group members

Next Class: Group Presentations Continue

Lecture 12 (Mar 2): Group Task Presentations

Topics:

- Continue group task presentations
- Peer feedback on annotation designs
- Discussion of common challenges

Prepare:

- Finalize your task presentation if not done
- Bring questions for other groups
- Think about how feedback applies to your project

Key Takeaways

1. **Good guidelines** are the foundation of quality annotation
2. **Essential components:** Definitions, positive/negative examples, edge cases
3. **Iterate:** Guidelines improve through pilot annotation cycles
4. **Decision procedures** help annotators handle ambiguity consistently
5. **Prompts** are guidelines for LLMs – same principles apply
6. **Always validate** LLM annotations against human judgments

Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

jinzhao@brandeis.edu