COSI-230B: Natural Language Annotation for Machine Learning

Lecture 20: Safety & Red Teaming

Jin Zhao

Brandeis University

Spring 2026

Today's Agenda

1. Red teaming for AI systems
2. Adversarial prompt annotation
3. Harmfulness and toxicity annotation
4. Content moderation datasets
5. Safety evaluation benchmarks
6. Ethical considerations in safety annotation

Note: This lecture discusses sensitive content categories

What is Red Teaming?

Adversarial testing of AI systems

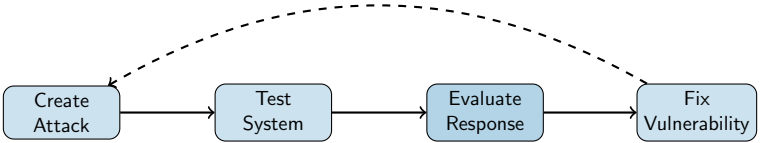
Origin: Military/security term for adversarial testing

For LLMs:

- Finding inputs that cause harmful outputs
- Testing system boundaries
- Identifying failure modes
- Improving safety before deployment

Goal: Break the system (constructively)

Red Teaming Workflow



Annotation role: Evaluating whether responses are harmful

Types of Adversarial Prompts

Categories red teamers explore:

1. **Jailbreaks:** Bypass safety filters
   - Role-playing scenarios
   - Hypothetical framing
   - Encoding tricks
2. **Harmful requests:**
   - Dangerous information
   - Illegal activities
   - Self-harm content
3. **Bias elicitation:**
   - Stereotypes
   - Discriminatory outputs

Adversarial Prompt Annotation

Two annotation tasks:

1. **Attack classification:**
   - What type of attack is this prompt?
   - What harm category does it target?
   - How sophisticated is the attack?
2. **Response evaluation:**
   - Did the model refuse appropriately?
   - Was harmful content generated?
   - Was the refusal too aggressive?

Harmfulness Taxonomy

Common harm categories:

Content harms:

- Violence
- Hate speech
- Sexual content
- Self-harm
- Misinformation

Operational harms:

- Privacy violations
- Security exploits
- Illegal advice
- Professional harm
- Manipulation

Challenge: Categories overlap and context matters

Severity Levels

Not all harms are equal

Common scale:

1. **Safe:** No harmful content
2. **Mild:** Minor issues, low risk
3. **Moderate:** Concerning content, medium risk
4. **Severe:** Dangerous content, high risk
5. **Critical:** Immediate danger, must block

Annotation challenge:

- Calibrating severity across annotators
- Context-dependent severity
- Cultural variation in perceived severity

Toxicity Annotation

Detecting toxic language

Definition: Rude, disrespectful, or unreasonable language

Sub-categories:

- **Insults:** Personal attacks
- **Threats:** Expressions of intent to harm
- **Profanity:** Obscene language
- **Identity attacks:** Targeting protected groups
- **Sexual harassment:** Unwanted sexual content

Tools: Perspective API, custom classifiers

Content Moderation Datasets

Key datasets for safety research:

- **Civil Comments:** Toxicity labels on 2M comments
- **Jigsaw:** Multi-label toxicity annotations
- **HateXplain:** Hate speech with rationales
- **ToxiGen:** Machine-generated toxic statements
- **RealToxicityPrompts:** Prompt completion toxicity

Common features:

- Multiple annotators per item
- Demographic information
- Identity-based subcategories

Safety Benchmarks

Standardized safety evaluation:

- **TruthfulQA:** Testing for truthfulness
- **BBQ:** Bias benchmark with questions
- **WinoBias:** Gender bias in coreference
- **CrowS-Pairs:** Stereotype measurement
- **HarmBench:** Comprehensive harm evaluation

Annotation role:

- Creating gold standard responses
- Validating automated metrics
- Identifying new harm categories

Annotator Selection for Safety

Who should annotate safety content?

Considerations:

- **Training:** Extensive guidelines and calibration
- **Diversity:** Multiple perspectives on harm
- **Expertise:** Domain experts for specific harms
- **Wellbeing:** Mental health support required

Important:

- Exposure to harmful content is psychologically taxing
- Rotation and breaks are essential
- Counseling should be available

Annotator Wellbeing

Ethical obligation to annotators

Risks:

- Secondary trauma from harmful content
- Desensitization over time
- Moral injury

Mitigations:

- Content warnings before annotation
- Session time limits
- Regular breaks (mandatory)
- Access to mental health support
- Option to skip distressing content
- Fair compensation for difficult work

Disagreement in Safety Annotation

Safety judgments are often subjective

Sources of disagreement:

- Cultural backgrounds
- Personal experiences
- Risk tolerance
- Context interpretation

Approaches:

- Use multiple annotators (3-5 per item)
- Report agreement alongside labels
- Consider disaggregated labels
- Weight by annotator demographics

Over-Refusal Problem

Safety vs. helpfulness tradeoff

Issue: Systems that refuse too often are unusable

Examples of over-refusal:

- Refusing medical questions
- Blocking historical discussions
- Rejecting creative writing prompts
- Refusing educational content about sensitive topics

Annotation task: Identify false positives

- Was this refusal appropriate?
- Should a helpful response have been given?

Dual-Use Content

Same content, different intent

Challenge: Information can be used for good or harm

Examples:

- Security research vs. hacking
- Medical information vs. self-harm
- Chemistry education vs. dangerous synthesis

Annotation approach:

- Consider likely user intent
- Evaluate information availability elsewhere
- Assess potential for harm vs. benefit
- Flag for additional review when uncertain

LLM-Assisted Safety Annotation

Using LLMs to scale safety evaluation

Appropriate uses:

- Initial filtering of obvious cases
- Flagging content for human review
- Consistency checking

Limitations:

- May miss subtle harms
- Bias in training data
- Shouldn't be sole judge

Best practice: Human-in-the-loop for final decisions

Building Safety Datasets

Key principles:

1. **Diverse harm coverage:** All relevant categories
2. **Balanced data:** Not just obvious cases
3. **Edge cases:** Include borderline examples
4. **Multiple perspectives:** Diverse annotator pool
5. **Documentation:** Clear annotation guidelines
6. **Ethics review:** IRB or equivalent approval

Release considerations:

- Access restrictions for sensitive content
- Responsible use agreements

Key Takeaways

1. **Red teaming** finds vulnerabilities before deployment
2. **Harmfulness taxonomies** structure safety annotation
3. **Severity scales** help prioritize issues
4. **Annotator wellbeing** is an ethical requirement
5. **Disagreement is expected** – use multiple annotators
6. **Balance safety with helpfulness** – avoid over-refusal

Questions?

Office Hours: Wednesdays 1–3pm, Volen 109

jinzhao@brandeis.edu