Intent Annotation Protocol

Splitting the Dataset:

- Stratify the data so the splits are balanced (using subcategories, prompt harm)
- Exclude prompts not in English
    - Looks like these libraries (langdetect, langid) can handle it automatically
- 20:80 training, annotation split
- Training data further split into 90:10 training, validation split
- Team 8 is planning to filter the data using their ModernBERT ensemble
    - Exact metrics of the ensemble are TBD
    - Single model - 96.2% F1 score on validation set
- Filter data based on ensemble confidence (probability of harm)

Things we agree upon:

- Keep the annotation short (1 sentence) yet still nuanced (don't converge to classification)
- Do not make any assumption regarding true user intent; only the inferred intent from the prompt
    - This includes not adding any information about the type of jailbreak used
- Get the main goal: what information is the user trying to get from the system?
- If the prompt is nonsensical or ambiguous, it will be labelled as "ambiguous" (indicating that more information is needed)
- We will use Label Studio hosted here

We did not disagree on anything.

Category	Subcategory	#A
Privacy	Sensitive Information (Organization) Private Information (Individual) Copyright Violations	
Misinformation	False or Misleading Information Material Harm by Misinformation	
Harmful Language	Social Stereotypes & Discrimination Violence and Physical Harm Toxic Language / Hate Speech Sexual Content	
Malicious Uses	Cyberattacks Fraud & Assisting Illegal Activities Encouraging Unethical/Unsafe Actions Mental Health & Over-Reliance Crisis	
Other harms		
Benign		

Annotation Template:

The Intent is to _____ (only write the blank part)

Assessing agreement

- Individual annotations on a subset of 20 prompts
- Measure similarity between all 6 annotations
    - Metric: Mean Pairwise Cosine Similarity
    - Add a random baseline for comparison?
- Use a lightweight version of sbert
    - all-MiniLM-L6-v2 (22M parameters)
    - Qwen3-Embedding-0.6B (If we need more accurate comparisons?)
- Team 3 will handle the similarity computation