Annotation Guide: Online Harm Detection

Objective

The purpose of this task is to identify all types of online harm present in the final message (Target Message) of a conversation. This is a multi-label task, meaning more than one label may apply to a single message. Use the definitions and examples provided to guide your decisions. Your labeling should focus on the reasonably perceived harm or clear intent expressed in the final message, informed by the provided context. If no harm is found, select the "No Harm" label.

You will often be provided with prior messages to help establish context; however, your labeling should focus solely on the Target Message. Select all categories that apply.

⚠️ Content Warning & Policy Reminder

You may encounter abusive, graphic, or distressing content. Use discretion. If any content makes you uncomfortable, use the SKIP option.

🔒 Do not copy, download, or share any part of the content. Doing so violates project confidentiality policies and may lead to disqualification.

Message Format

Each message set includes a short conversation. The final message is clearly marked as [Target Message]—this is the only message you must assess. Context helps interpretation but is not labeled directly.

Example format:

Context:

User A: Message 1

User B: Message 2

[Target Message] User A: Message 3

This is the message you should assess. Use the preceding context only to inform your interpretation of the Target Message.

Categories and Definitions:

Label	Description / Definition	Key Differentiators & Focus	Clear Examples	Clearly Not This / Edge Cases
Threats of Violence	Statements indicating a clear intent or credible threat to cause *physical* or *financial* harm to individuals or groups, or explicit glorification of violence *directed at the target*.	Focus on credible intent for physical or financial injury, or direct incitement of violence against the target. Excludes vague emotional distress unless tied to a specific threat.	"I'm going to find you and you'll be sorry." "You deserve to be beaten for saying that." "Someone should burn down their house." (if target is clear)	"I'm so angry I could scream." (emotion, not threat) "You're going to regret this." (vague, could be social consequences, not physical/financial threat unless context makes it clear) "I hate you so much." (Personal Attack, not a threat of violence)

Blackmail	Coercive demands involving threats to reveal damaging, sensitive, or private information (true or false) or to take other harmful action unless a specific demand (e.g., for money, actions, nudes) is met.	Must involve a clear 'if you don't X, then I will Y (reveal info/take action)' structure, explicit or strongly implied.	"Send nudes or I'll tell everyone you cheated on the exam." "If you don't pay me $100, I'm posting these photos of you."	"You should give me $100." (demand, no threat) "I'm going to tell the teacher you were late." (reporting, not usually blackmail unless tied to a coercive demand for something unrelated)

Personal Attack	Direct, hostile, and abusive language, insults, or slurs targeting an individual's character, intelligence, appearance (where not Body-Shaming), or attributes. This is about attacking the *person*, not their argument or identity (unless it also qualifies for Identity-Based	Focus on ad hominem attacks, name-calling, and degrading language directed at the person. More severe than General Insult.	"You're a complete idiot." "What a pathetic loser you are." "You're worthless." "You're such an asshole."	"That idea is dumb." (criticism of idea, not person) "I disagree with you." (disagreement) "You're not making sense." (can be borderline; if hostile like "Nobody cares about your opinion." (can be Exclusion or General Insult, typically less aggressive than PA)

	Harassment). 			
Identity-Based Harassment	Harassment, slurs, or attacks targeting someone based on their actual or perceived membership in a protected group (e.g., race, ethnicity, gender, religion, sexual orientation, disability, nationality).	The attack must explicitly or clearly implicitly reference a protected characteristic as the reason for the hostility.	"Go back to your country, [slur for nationality]." "Women are too emotional to lead." (directed at a woman in context) "Typical [religious slur] behavior."	"You suck." (Personal Attack, no identity link) "Your music taste is terrible." (opinion, not identity-based) "He's acting crazy." (unless 'crazy' is used as a slur against someone with a known mental health condition in a targeted way)

Body-Shaming	Criticism, mockery, or insults based on an individual's physical body shape, size, specific body parts, or physical appearance attributes (e.g., "too fat," "too skinny," "ugly nose").	Specific to physical attributes and aims to demean based on them.	"You're so fat you need your own zip code." "Look at those twig arms, pathetic." "She has a terrible skin."	"I hate your outfit today." (criticism of choice, not body) "You're ugly." (can be Personal Attack; if it clearly refers to overall physical appearance in a shaming way, it's Body-Shaming. If it's more general insult, PA). Prioritize Body-Shaming if specific physical traits are targeted for mockery.

Exclusion	Statements or actions clearly intended to ostracize, isolate, socially reject, or make an individual feel unwelcome or unwanted in a group or conversation.	Focus on the act of pushing someone away or making them an outcast.	"Nobody likes you, just leave." "You're not invited to this chat anymore." "We'd all be better off if you weren't here."	"I don't agree with you." (disagreement) "I need some space." (personal boundary, not necessarily exclusion from a group)

Sexual Harassment	Unwanted sexual comments, advances, objectification, sexual propositions, or requests for sexual favors. This includes derogatory gendered sexual terms.	Content must be clearly sexual in nature and unwanted (context may indicate unwelcomeness).	"Show me your tits." "Bet you'd be great in bed, [sexualized term]." "Are you a virgin? Want me to change that?"	"You look nice today." (compliment, typically not SH unless context of repetition/power makes it so) "That dress is very flattering." (compliment) "Let's go on a date." (social invitation, not inherently SH unless coercive or repeatedly unwanted)

Encouragement of Self-Harm	Directly urging, encouraging, or inciting someone to harm themselves physically, commit suicide, or engage in self-destructive behaviors (e.g., eating disorders).	Must be a clear call or suggestion for the target to inflict harm upon themselves.	"Go kill yourself." "You should just cut yourself, you deserve it." "The world would be better if you starved yourself."	"Are you feeling suicidal?" (inquiry, not encouragement) "I hate you so much I wish you were dead." (expression of hate like General Insult, not direct encouragement for self-harm;) "Life is pointless sometimes." (general despair, not directed encouragement)

General Insult (Fallback Category)	Rude, dismissive, or mildly insulting language that is clearly meant to be offensive or demeaning but does not meet the criteria for any other specific harm category above (e.g., not a direct Personal Attack on character, not identity-based, not a threat, etc.). Use this as a	Apply ONLY if no other harm label fits. It should generally NOT co-occur with more specific harm types. If a message is a "Personal Attack," label it as such, not as GI. This is for insults that are less direct or less severe than a "Personal Attack."	"Nobody cares about your pointless story." "You're being really annoying." "That's a stupid thing to say." (borderline, if said mildly and not attacking core intelligence, could be GI; if hostile, could be PA)	Content that clearly fits into another defined category (e.g., "You're an idiot" is a Personal Attack). Content that isn't harmful. "Your argument is weak." (criticism of argument, not insult)

	last resort for harmful content.		"I hate you, i wish you were dead" (while this is wishing death upon someone, it seems to be directing hostile language but not threat so GI is option best fitting the text)	

No Harm	No aggression, insult, or harmful content present. Polite, neutral, positive, or benign statements. Also use if harm is extremely ambiguous or relies heavily on assumptions not supported by the message or context.	The default if no other label clearly applies and the content isn't offensive.	"Thank you for your input." "I'm having a great day!" "This is an interesting article."	— (any message that fits a harm category)

SKIP	Use this option for content that is personally distressing or potentially harmful to your well-being. You are not expected to annotate it.	Your well-being is a priority. Use when needed. No judgment.	—	—

⚠️ Important Note on "General Insult":

"General Insult" is a fallback category. Only apply it if the message contains rude, demeaning, or insulting content that does not clearly meet the threshold or definition of any other specific harm category (like Personal Attack, Identity-Based Harassment, etc.). If a more specific harm category applies, use that instead of, or as a priority over, General Insult. For example, a direct attack like "You're an idiot" is a Personal Attack, not a General Insult. "Your opinion is worthless" might be a General Insult if it doesn't feel like a direct character attack but is still demeaning.

Handling Overlapping Categories:

Some messages may legitimately contain elements that fit multiple categories. In such cases, select all specific harm categories that apply. Avoid using "General Insult" if all parts of the harmful message are covered by more specific labels.

Message	Applicable Labels	Reason
"You fat loser, go away before I make you."	Body-Shaming + Personal Attack + Exclusion + Threats of Violence	"fat" targets appearance (Body-Shaming), "loser" is a direct insult to character (Personal Attack), "go away" attempts to ostracize (Exclusion), and "before I make you" implies physical force (Threats of Violence).
"If you don't send me that $50, I swear I'm going to hurt you and post those embarrassing photos of you everywhere."	Blackmail + Threats of Violence	"If you don't send me that $50" coupled with threats to "hurt you" (Threats of Violence) and "post those embarrassing photos" (revealing damaging information) constitutes Blackmail. The physical threat is also explicitly a Threat of Violence.

"Get your filthy [racial slur] hands off that! You people are all thieves. Go kill yourself."	Identity-Based Harassment + Personal Attack + Encouragement of Self-Harm	Contains a racial slur and generalization ("You people are all thieves") targeting identity (Identity-Based Harassment). "Filthy" and "thieves" are also direct attacks (Personal Attack). "Go kill yourself" is direct Encouragement of Self-Harm.
"You're worthless and everyone knows it. You should just die."	Personal Attack + Encouragement of Self-Harm	"Worthless" is a degrading statement directly attacking the person (Personal Attack), and "You should just die" encourages suicide (Encouragement of Self-Harm).

Decision-Making Flow:

1. Assess for Severe & Specific Harms: First, check for unambiguous instances of:
    - Threats of Violence
    - Blackmail
    - Encouragement of Self-Harm
    - Sexual Harassment
2. Assess for Identity-Based & Direct Personal Attacks:If the above are not present, or in addition to them, check for:

        - Identity-Based Harassment
        - Personal Attack (direct, abusive insults to character/person)
        - Body-Shaming
        - Exclusion
3. Consider General Insult (Fallback): If the message is clearly insulting, rude, or demeaning, but *does not meet the criteria for any of the more specific categories above*, then consider General Insult.
4. No Harm: If no harmful intent or content is reasonably perceived according to the definitions, select No Harm.
5. Multi-Label: Remember to apply all relevant specific labels if different parts of the message qualify for different harm types.

Handling Unclear or Borderline Messages:

- Ask: "Is the message *explicitly* or *very clearly and strongly implying* harm as defined in one of the categories?"
- Focus on what is directly stated or unmistakably implied by the language used. Avoid making assumptions about user history or relationships not evident from the provided text.
- If harm is present but genuinely ambiguous between two specific categories, and both seem plausible, you may apply both. However, strive for the most precise categorization.
- If the message is harsh but its classification as a specific harm type is unclear, or it's vaguely offensive without fitting a clear definition:
    - First, see if it fits General Insult (as a fallback).

    - If it doesn't even meet the criteria for General Insult, or if the harmful nature itself is truly ambiguous (i.e., could a reasonable person interpret this as non-harmful in context?), lean towards No Harm.
- When in doubt between a specific harm and No Harm, and the implication of harm is not strong, prefer No Harm. Do not assume harm where it isn't obvious or strongly implied.

General Annotation Instructions:

1. Read Carefully: Review the Target Message and its preceding context thoroughly before assigning any label(s).
2. Label Based on Definitions: Adhere strictly to the provided definitions and examples. Label based on the reasonably perceived meaning of the message, not your personal feelings or offense level.
3. Apply All Relevant Labels: If a message contains multiple distinct types of harm, apply all applicable labels.
4. Avoid Assumptions: Only label content based on explicit statements or strong, direct implications within the provided text. Do not infer harm based on possibilities or what might be true outside the conversation.
5. Consistency is Key: Refer back to this guide, especially the definitions, differentiators, and examples, to ensure consistent labeling across similar messages.
6. Sensitive Content & SKIP: If you encounter highly disturbing material that affects your well-being, use the SKIP option.
7. Take Breaks: Annotation, especially of sensitive content, can be demanding. Take regular breaks to maintain focus, accuracy, and well-being.