

Annotation Guide: Intent of Online Conversations


Annotation Guide: Intent of Online Conversations-----	1
Objective:-----	1
Message Format:-----	2
Categories and Definitions:-----	3
Sarcasm-----	3
Gaslighting-----	4
Blaming-----	5
Abet / Instigate-----	6
Attack-----	7
Defend-----	9
No Intent-----	10
Unclear-----	11
SKIP-----	12
Handling Overlapping Labels-----	12
Handling Ambiguity & Unclear Messages-----	13
General Annotation Instructions:-----	14

Objective:

The purpose of this task is to identify the primary **intentions** or **communicative functions** of the final message (Target Message) in a short online conversation. You will assess whether the message serves one or more of the defined functions below. You will often be provided with prior messages to help establish context; however, your labeling should focus solely on the **Target Message**. Select all applicable categories for that message.

Content Warning & Policy Reminder

Messages may include manipulation, emotional abuse, or inflammatory language. Proceed with caution and use **SKIP** if the content causes discomfort.

 Do not copy or share any part of the content. Doing so violates policy and may result in disqualification and forfeiture of compensation.

Message Format:

Each item includes a conversation with a clearly marked final message ([Target Message]).

Use the context to understand tone or target, but assess **only** the Target Message.

Example format:

None

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	Examples	Non-Examples
Sarcasm	Messages that appear to state one thing (often polite or positive) but, due to context or tone, are intended to mean the opposite, typically to mock, ridicule, or criticize indirectly.	Focus on the contrast between literal meaning and intended meaning. Often relies on context, exaggeration, or understatement.	<p><i>"Trump is the most innocent man wrongly accused since O.J. Simpson"</i></p> <p><i>"You're a real genius for that idea." (said after a very bad idea)</i></p>	<p><i>"This is genuinely a great idea!" (sincere praise)</i></p> <p><i>"I'm very busy today." (factual statement)</i></p>

Gaslighting	<p>Messages that manipulate someone by distorting facts, denying past events, or questioning their sanity/memory to make them doubt their own perception, judgment, or reality.</p>	<p>Focus on the manipulative act of undermining someone's grasp of reality or memory related to specific events or statements.</p>	<p><i>"You're crazy, I never said that.</i> <i>You're imagining things again."</i></p> <p><i>"Everyone knows you're too sensitive; that's not how it happened at all."</i></p>	<p><i>"Actually, I think you might be mistaken; my memory is that we agreed on Tuesday." (polite correction without intent to make someone doubt their sanity)</i></p> <p><i>"I forgot I said that, sorry." (owning a mistake, not denying reality)</i></p>
--------------------	---	--	--	---

Blaming	<p>Messages that explicitly or implicitly assign responsibility to someone (or a group) for a negative outcome, problem, harm, or for provoking abusive behavior.</p>	<p>Focus on the act of attributing fault. Can be direct or indirect.</p>	<p><i>"This is all your fault! If you hadn't been late, we wouldn't have missed it."</i></p> <p><i>"She wouldn't get teased if she didn't dress that way."</i></p>	<p><i>"We should all try to be more careful next time."</i></p> <p><i>(constructive, shared responsibility)</i></p> <p><i>"I'm upset that this happened."</i></p> <p><i>(expressing emotion, not assigning fault to another)</i></p>
----------------	---	--	--	--

Abet / Instigate	<p>Messages that actively encourage, urge, provoke, or incite others to engage in aggression, hostility, harmful actions, or conflict towards an individual or group.</p>	<p>Focus on the call to action for negative behavior. Can be direct commands or suggestive endorsements of hostility.</p>	<p><i>"Yeah, you should go tell them off! They deserve it."</i></p> <p><i>"Let's all report their account until it's banned." (if malicious and coordinated harassment intent)</i></p> <p><i>"Someone needs to teach him a lesson."</i></p>	<p><i>"Let's report the post if it violates the rules." (following procedure)</i></p> <p><i>"I agree that what they did was wrong." (expressing opinion, not inciting action)</i></p> <p><i>"You should stand up for yourself." (general advice, not necessarily instigating aggression depending on context)</i></p>
-----------------------------	---	---	---	---

Attack	<p>Messages that directly express aggression towards an individual or group through insults, name-calling, direct threats of harm (not covered by Abet/Instigate), hostile mockery, or overt attempts to discredit or demean.</p>	<p>Focus on direct, hostile communication .</p> <p>Distinction from Sarcasm:</p> <p>While sarcasm can be used to mock, if the mockery is indirect and relies on contradicting literal meaning, label Sarcasm. If Sarcasm is also a clear, hostile insult, label both.</p> <p>Attack is more for overt aggression.</p> <p>Distinction from Blaming:</p> <p>Blaming</p>	<p><i>"You're an absolute idiot."</i></p> <p><i>"I'm going to make you regret saying that." (direct threat from speaker)</i></p> <p><i>"Only a moron would believe that."</i></p>	<p><i>"I strongly disagree with your statement." (disagreement)</i></p> <p><i>"That was not a smart move." (criticism of action, milder than direct insult to person)</i></p> <p><i>"Your argument is weak." (critiquing argument, not person)</i></p>
---------------	---	---	---	--

		<p>assigns fault.</p> <p>Attack aims to harm/insult directly. They can co-occur if blaming is done via insult.</p>		
--	--	--	--	--

Defend	<p>Messages that aim to protect, support, or shield someone (or oneself, or a group) from a perceived attack, criticism, or blame within the conversation.</p> <p>Can be non-aggressive or retaliatory.</p>	<p>Focus on the protective function. May co-occur with Attack if defense is aggressive.</p>	<p><i>"Stop picking on her, she didn't do anything wrong."</i></p> <p><i>"Actually, I was there, and that's not what happened. He's telling the truth."</i></p> <p><i>"Leave him alone! You're the one who started it, you bully!"</i></p> <p><i>(Defend + Attack)</i></p>	<p><i>"Let's try to hear both sides before judging."</i></p> <p><i>(mediating, neutral stance)</i></p> <p><i>"I understand why you feel that way." (empathy, not necessarily defense of a specific party)</i></p>
---------------	---	---	--	---

No Intent	<p>The message is clear and understandable , but it does not primarily serve any of the specific communicative functions listed above (Sarcasm, Gaslighting, etc.). This includes simple factual statements, questions, greetings, expressions of neutral emotion, or other common conversational exchanges.</p>	<p>Use when the message is clear, but none of the other defined labels accurately describe its primary communicative role.</p>	<p>"The meeting is scheduled for 3 PM tomorrow."</p> <p>"What time does the movie start?"</p> <p>"Thanks for sharing this."</p> <p>"I'm feeling tired today."</p>	<p><i>A message that clearly fits one of the other defined categories.</i></p> <p><i>A message that is genuinely ambiguous in its meaning or intent (see Unclear).</i></p>
------------------	--	--	---	--

<p>Unclear</p>	<p>The message's meaning, target, or communicative function is genuinely ambiguous, making it impossible to confidently assign any other label, including "No Intent". This could be due to lack of context (despite provided context), extreme vagueness, or indecipherable language.</p>	<p>Use sparingly.</p> <p>Must be accompanied by a comment explaining the ambiguity.</p> <p>Different from "Neutral" where the message is clear but out of scope for other labels.</p>	<p>Context:</p> <p>User A: "Did you see it?"</p> <p>User B: "Green."</p> <p>[Target Message]</p> <p>User A: "Seven."</p> <p>(Without more context, "Seven" is Unclear).</p> <p>"Idk maybe lol but not really."</p> <p>(Very vague, hard to pin down function)</p>	<p><i>A message that is clear but simply doesn't fit other categories (use "No Intent").</i></p> <p><i>A message where an intent is subtly implied but still reasonably discernible.</i></p>
-----------------------	--	---	---	--

SKIP	Use when content is personally distressing or uncomfortable to engage with. Your well-being takes priority.	Use as needed. No explanation required.	—	—
-------------	---	---	---	---

Handling Overlapping Labels

Some messages may serve more than one communicative function simultaneously. In such cases, apply all relevant labels.


Message	Applicable Labels	Reason for Multiple Labels
"Oh, because it's always MY fault, right? Of course, I'm the designated screw-up."	Sarcasm + Blaming	The message uses clear sarcasm to emphasize the point, while the underlying function is to make the speaker (or someone else) feel they are being unfairly blamed, or to assign blame sarcastically.

"You're just imagining things, I never said that. You always have to make me the bad guy."	Gaslighting + Blaming	It denies the recipient's reality/memory (Gaslighting) and simultaneously accuses them of wrongfully casting the speaker as the villain (Blaming).
"Hey, don't listen to them! They're just jealous. If they say one more word, let 'em have it!"	Defend + Abet / Instigate	The speaker defends someone ("Don't listen to them! They're just jealous.") and then encourages an aggressive or retaliatory response ("let 'em have it!").
"Wow, you so huge you caused a solar eclipse" (said to someone who just embarrassed another person)	Sarcasm + Attack	The message uses sarcasm ("real hero," "so brave") to indirectly deliver a hostile criticism or mockery of the person's actions, functioning as a covert attack.

Handling Ambiguity & Unclear Messages

When the communicative function of the target message is not immediately obvious:

- First, determine if the message is clear in its literal meaning but simply doesn't fit the defined functional categories (Sarcasm, Gaslighting, etc.). If so, use **No Intent**.
- If the message itself is vague, its target is unknown, its tone is indecipherable even with context, or it's open to multiple conflicting interpretations of its basic meaning (not just multiple functions), apply the **Unclear** label.

-  **Important:** Every use of the **Unclear** label must be accompanied by a brief comment explaining why the message's function could not be confidently determined.

General Annotation Instructions:

1. **Focus on Target Message Function:** Carefully read the Target Message and the provided context. Determine the primary communicative function(s) the Target Message serves in the conversation based on the definitions.
2. **Interpret Tone and Context Carefully:** Tone (e.g., sarcastic, aggressive) and context are crucial for distinguishing between literal statements and specific communicative functions. For example, sarcasm can mask other functions like Blaming or Attack.
3. **Select All Applicable Labels:** If a message clearly serves multiple defined functions, select all relevant labels. Refer to the "Handling Overlapping Labels" section for guidance.
4. **Avoid Assumptions Beyond Provided Text:** Label based on what is explicitly stated or very strongly and directly implied by the language and context. Do not infer functions based on possibilities not well-supported by the text.
5. **Maintain Consistency:** Regularly refer to the definitions, key differentiators, and examples to ensure consistent labeling across all messages.
6. **Use SKIP for Distressing Content:** If you encounter material that is highly disturbing or makes you uncomfortable, use the **SKIP** option. Your well-being is paramount.
7. **Take Regular Breaks:** Annotation tasks, especially those involving nuanced interpretation or potentially sensitive content, can be mentally demanding. Take short, regular breaks to maintain focus and accuracy.