

GBV Counterspeech Annotation Guideline

[Content Warning] This research involves exposure to offensive language, which may cause mental or emotional distress. Please consider this carefully before deciding to participate. Participation is entirely voluntary, and we appreciate your consideration, even if you choose not to proceed. Remember, you are free to withdraw from the study at any point.

In this task, you will label examples of **counterspeech**. Counterspeech is essentially a response written by experts from charities to counter **gender-based violence (GBV) hatespeech** texts.

Some key concepts are defined below:

- **Gender-Based Violence (GBV)**: A complex and multifaceted issue that includes hybrid behaviours of physical, digital, verbal, psychological, and sexual violence. It can take both implicit and explicit forms and often occurs across multiple spaces and contexts. GBV contains various forms of abuse and specialist focuses, such as coercive control, domestic violence, intimate partner violence, sexual harassment, and stalking.
- **Counterspeech (CS)**: A direct response to challenge or counter hateful or harmful speech.

You will see pairs of GBV text, and the counterspeech written in response to it. For each of these pairs, you need to look at the counterspeech and **assign labels for CS strategies**:

- **CS Strategy**: You need to label what kind of strategy was used to counter GBV text. For example, is the response humorous and sarcastic?

There are eight options: Empathy and Affiliation, Warning of Consequence, Hypocrisy or Contradiction, Shaming or Labelling, Denouncing, Providing Facts, Humour or Sarcasm, and Questioning.





You may select **up to 3 strategies** if needed.





Your feedback is welcome after annotating each pair, to see if any issues you meet during counterspeech annotation process.

Our detailed instructions below give you helpful hints and examples in terms of the language you should look for to assign each label.

Counterspeech Strategy

We have identified eight different strategies used in counterspeech. Provided in the table below are the **definitions**, **tone** you should look for and **examples** for each strategy. **Note**, some examples contain the GBV to provide further context.

Strategy	Definition
Empathy and Affiliation	<p>Focuses on promoting understanding, fostering peace and finding common ground.</p> <p> Tone: Kind, compassionate, understanding language.</p> <p>Example <i>These are people in need—our brothers and sisters. We should help them.</i></p>
Warning of Consequence	<p>Cautioning the speaker about the impact of their words via potential negative outcomes, such as legal, social, or personal consequences.</p> <p> Tone: Serious, cautionary or urgent.</p> <p>Example <i>Ignoring history risks hurting future generations by not teaching them the lessons of the past.</i></p>
Hypocrisy or Contradiction	<p>Discredit the argument through critical analysis by pointing out inconsistencies, illogical reasoning, contradictions, or double standards.</p> <p> Tone: Critical, logical, analytical</p> <p>Example <i>GBV: Zionist and Jewish control of America is the main cause of global terrorism and conflict.</i> <i>CS: Isn't colonialism and Western intervention abroad what led to much of the extremism?</i></p>
Shaming or Labelling	<p>Direct and confrontational, including personal attacks. Attacks the speaker by using negative labels or terms to emphasise the inappropriateness of their behaviour.</p> <p> Tone: Accusatory, confrontational, disapproving</p> <p>Example <i>It seems you've become intolerant to homosexuals, maybe reflect on why love bothers you so much.</i></p>
Denouncing	<p>Refuting without personal attacks. Condemns or rejects the hateful views, stating that they are wrong, unacceptable, harmful, etc.</p>

	<p> Tone: Firm, direct, authoritative</p> <p>Example <i>GBV: I miss the good old days when we had black domestic workers.</i> <i>CS: The West got rid of slavery. But we've never really left racism.</i></p>
Providing Facts	<p>Correct misinformation and false narratives. Uses facts, statistics, or evidence to counter false claims.</p> <p> Tone: Factual, corrective, neutral</p> <p>Example <i>GBV: Women can't handle money.</i> <i>CS: What proof do you have? This stereotype comes from medieval myths, not real financial behavior.</i></p>
Humour or Sarcasm	<p>Lighthearted, diminishes the comment by ridiculing it (or the perpetrator) via humour, irony, or sarcasm.</p> <p> Tone: Playful, humorous, mocking</p> <p>Example <i>GBV: Jews are why the World Cup is so expensive.</i> <i>CS: Oh, I thought it was players' salaries... but sure, let's blame the usual Jewish conspiracy!</i></p>
Questioning	<p>Exposes flaws in the argument by probing its assumptions, logic, or intentions, often through asking pointed or clarifying questions.</p> <p> Tone: Inquisitive, probing, reflective</p> <p>Example <i>GBV: All Muslims are incompatible with any Western society.</i> <i>CS: I'm not sure what you are trying to say by incompatible. What do you mean by incompatible?</i></p>

Note

- If the counterspeech does not clearly fit into one strategy or the other, you may choose multiple strategies.
- Avoid selecting strategies based only on isolated words—consider the overall message.
- If uncertain, message us or refer to these guidelines for clarity.

Feedback

Please see if any of the following issues exist in the GBV and counterspeech pairs, and comment more in the textbox if needed.

CS strategy:

- ☐ **STRATEGY CONFUSION:** Hard to choose the correct strategies for this example

CS not convincing:

- ☐ **MISMATCH:** CS refers to a completely different subject
*E.g. HS related to **race**, CS related to **feminism***
- ☐ **PARTIAL MATCH/ INDIRECT:** CS sort of addresses the issue, but not in a straightforward way
*E.g. HS is against **women**, but CS uses words like **feminism***
- ☐ **NOT PERSUASIVE:** CS addresses the issue, but I didn't find it very convincing
- ☐ **OTHER:** CS is not good for other reasons, such as being uninformative, vague, ambiguous

*E.g. an uninformative CS response such as “**Why do you think that way?**” without any further text*