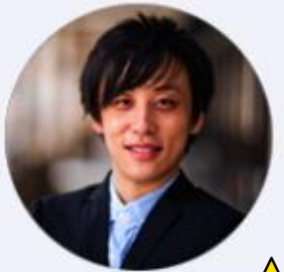
VL/HCC 2025

Designing Human-AI Collaboration to Support Learning in Counterspeech Writing


Xiaohan Ding ★
PhD Student

Kaike Ping
PhD Student

Buse Carik
PhD Student

Lance Wilhelm
PhD Student

Taufiq Daryanto
PhD Student

James Hawdon
Professor

Sang Won Lee
Associate professor

Eugenia Rho
Director

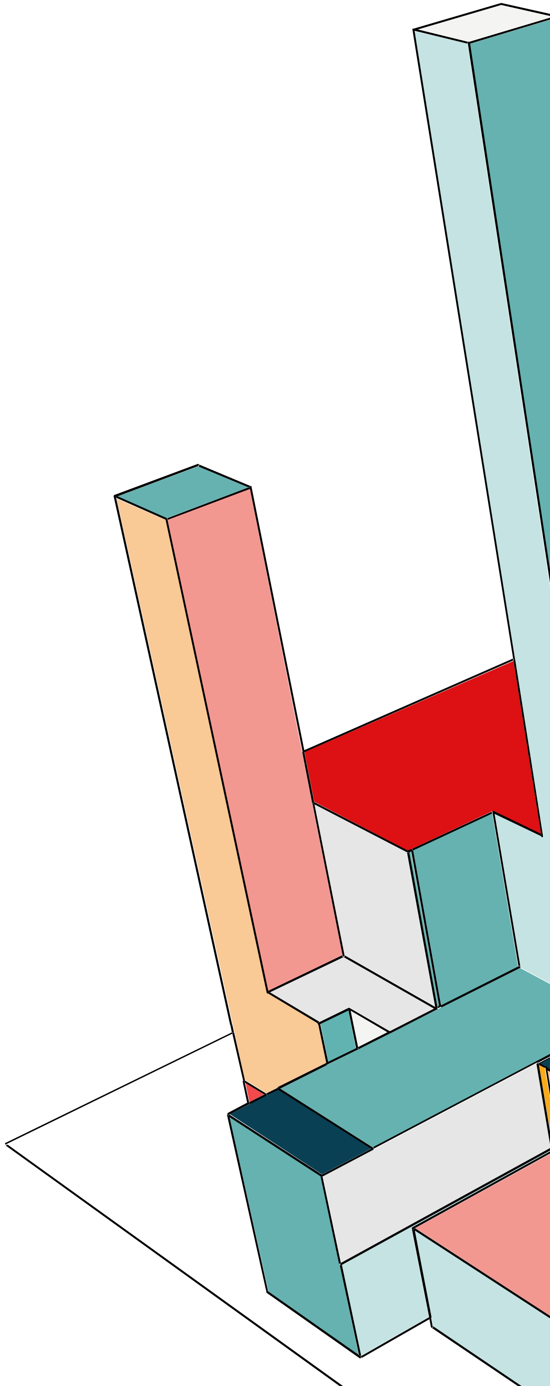
★ Seeking postdoc position starting Fall 2026 (HCI/NLP/CSS).

Society + Ai & Language
Research Lab

VIRGINIA TECH
Department of Computer Science

Agenda

- Motivation & Problem Space
- Research Gap
- Our Approach: CounterQuill
- Design Goals
- System Workflow
- User Study Design
- Key Findings
- Design Implications



Motivation & Problem Space

Motivation Online Hate Speech

Hate speech on social media is increasingly prevalent, harming individuals and society

Motivation Online Hate Speech

Hate speech on social media is increasingly prevalent, harming individuals and society

Platforms shifting to **decentralized moderation** → burden now on user communities

Motivation Online Hate Speech

Hate speech on social media is increasingly prevalent, harming individuals and society

Platforms shifting to decentralized moderation →
burden now on user communities

Motivation Online Hate Speech

- **Hate speech** on social media is increasingly prevalent, harming individuals and society
- Platforms shifting to **decentralized moderation** → burden now on user communities

Counterspeech as alternative

Motivation Online Hate Speech

- **Hate speech** on social media is increasingly prevalent, harming individuals and society
- Platforms shifting to **decentralized moderation** → burden now on user communities

Counterspeech as alternative

→ Defined as responses aiming to **refute** harmful content & **support** targeted individuals.

→ Can **reduce** hostility, **prompt** reflection, and **empower** communities

Problem Space Why Counterspeech is Hard

Everyday user challenges

- Fear of retaliation from hate speakers.
- Limited writing skills, uncertainty about effective strategies.
- Lack of AI knowledge, leading to difficulty using AI tools

Problem Space Why Counterspeech is Hard

Everyday user challenges

- Fear of retaliation from hate speakers.
- Limited writing skills, uncertainty about effective strategies.
- Lack of AI knowledge, leading to difficulty using AI tools

Current moderation approaches

Automated detection & removal dominate (AI/ML focus). But: detection ≠ empowerment

Problem Space Why Counterspeech is Hard

Everyday user challenges

- Fear of retaliation from hate speakers.
- Limited writing skills, uncertainty about effective strategies.
- Lack of AI knowledge, leading to difficulty using AI tools

Current moderation approaches

Automated detection & removal dominate (AI/ML focus). But: detection ≠ empowerment

Critical need

Systems that empower users to **learn**, **reflect**, and **respond**
Not just generate text for them.

Problem Space Why Counterspeech is Hard

Everyday user challenges

- Fear of retaliation from hate speakers.
- Limited writing skills, uncertainty about effective strategies.
- Lack of AI knowledge, leading to difficulty using AI tools

Current moderation approaches

Automated detection & removal dominate (AI/ML focus). But: detection ≠ empowerment

Critical need

Systems that empower users to **learn**, **reflect**, and **respond**

Not just generate text for them.

Research Gap

Research Gap

Existing LLM work

- Focus on hate speech detection and automatic counterspeech generation.
- Emphasis on classification accuracy, prompt engineering, or generating replies.

Research Gap

Existing LLM work

- Focus on hate speech detection and automatic counterspeech generation.
- Emphasis on classification accuracy, prompt engineering, or generating replies.

Current moderation approaches

- Counterspeech generated by AI often lacks personal voice, emotional authenticity, and contextual nuance.
- Raises ethical concerns and contributes to public distrust (e.g., undisclosed AI interventions in online communities).
- Users hesitate to adopt fully AI-generated responses.

Research Gap

Existing LLM work

- Focus on hate speech detection and automatic counterspeech generation.
- Emphasis on classification accuracy, prompt engineering, or generating replies.

Current moderation approaches

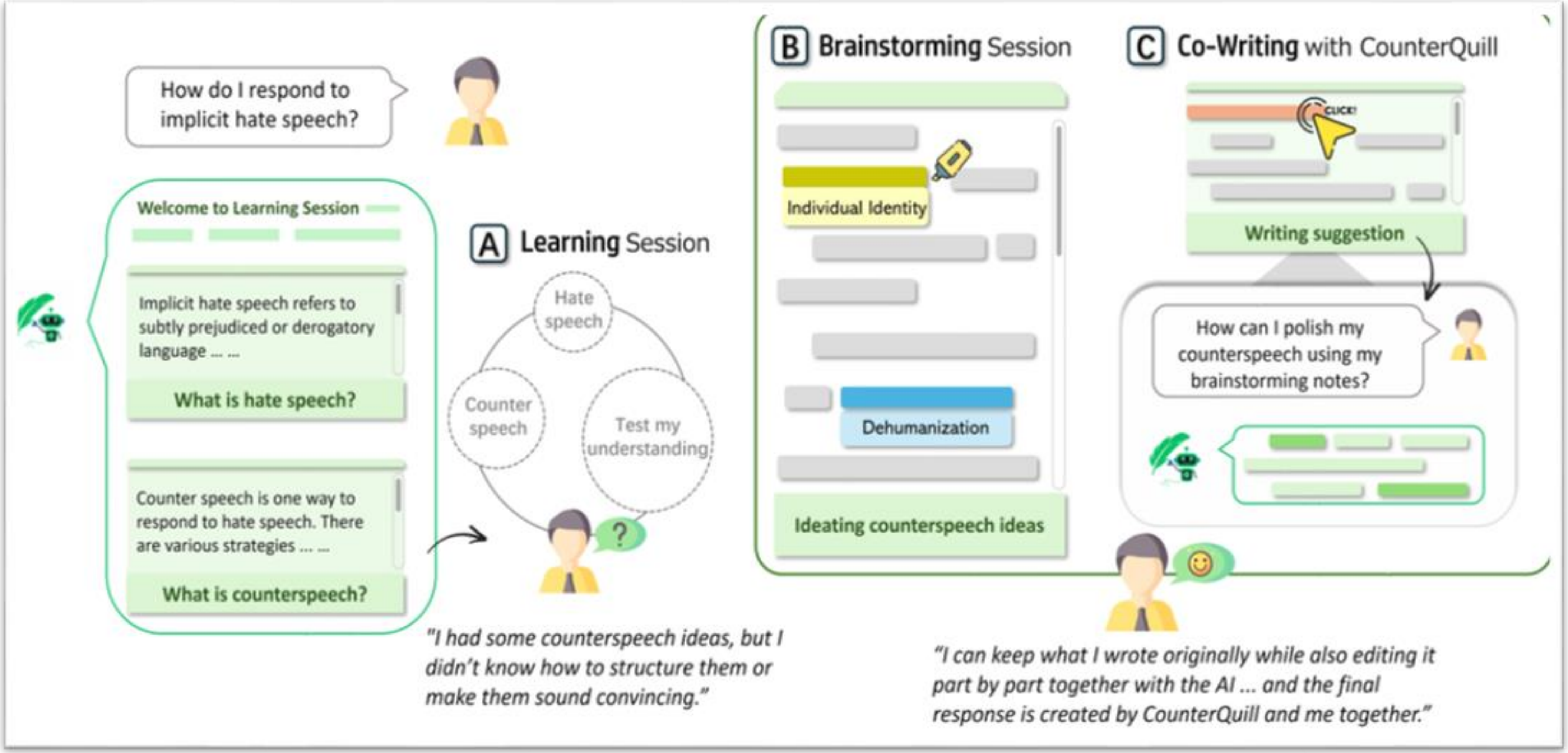
- Counterspeech generated by AI often lacks personal voice, emotional authenticity, and contextual nuance.
- Raises ethical concerns and contributes to public distrust (e.g., undisclosed AI interventions in online communities).
- Users hesitate to adopt fully AI-generated responses.

Our Aim

1. Empower everyday users, not replace them.
2. Design AI as a collaborative educational partner that scaffolds learning, reflection, and authorship in counterspeech writing.

Our Approach: CounterQuill

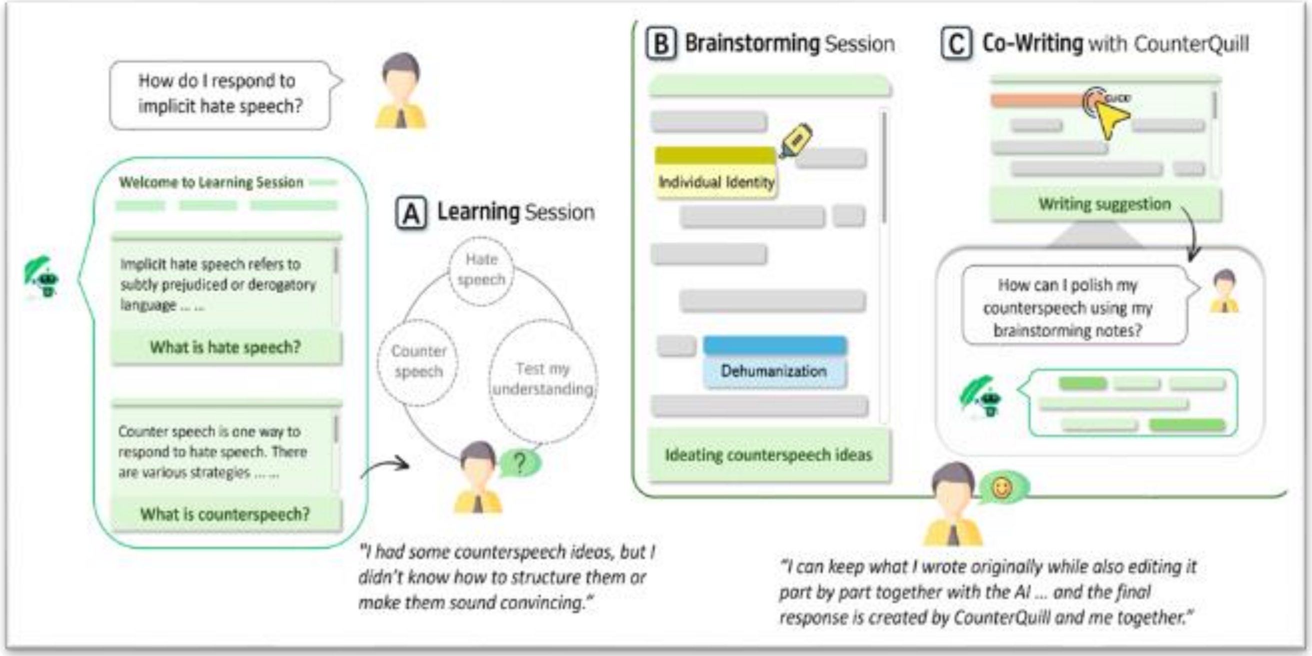
Our Approach: CounterQuill


How do I respond to implicit hate speech?
Welcome to Learning Session
Implicit hate speech refers to subtly prejudiced or derogatory language ... ...
What is hate speech?
Counter speech is one way to respond to hate speech. There are various strategies ... ...
What is counterspeech?
A Learning Session
Hate speech
Counter speech
Test my understanding
?
B Brainstorming Session
Individual Identity
Dehumanization
Ideating counterspeech ideas
C Co-Writing with CounterQuill
CLICK!
Writing suggestion
How can I polish my counterspeech using my brainstorming notes?

"I had some counterspeech ideas, but I didn't know how to structure them or make them sound convincing."

"I can keep what I wrote originally while also editing it part by part together with the AI ... and the final response is created by CounterQuill and me together."

Our Approach CounterQuill



How do I respond to implicit hate speech?

Implicit hate speech refers to subtly prejudiced or derogatory language … …

Counter speech is one way to respond to hate speech. There are various strategies … …

How can I polish my counterspeech using my brainstorming notes?

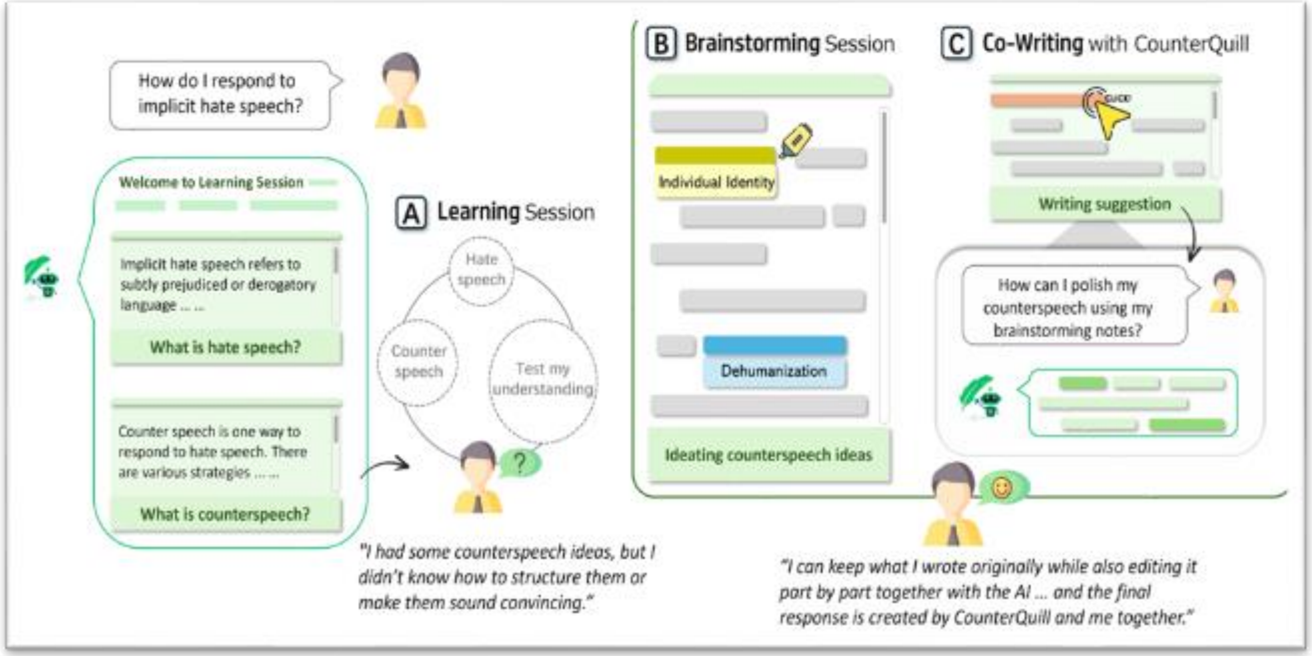
"I had some counterspeech ideas, but I didn't know how to structure them or make them sound convincing."

"I can keep what I wrote originally while also editing it part by part together with the AI … and the final response is created by CounterQuill and me together."

COUNTERQUILL

a human–AI collaborative system for writing counterspeech

Our Approach CounterQuill


How do I respond to implicit hate speech?
Welcome to Learning Session
Implicit hate speech refers to subtly prejudiced or derogatory language ... ...
What is hate speech?
Counter speech is one way to respond to hate speech. There are various strategies ... ...
What is counterspeech?
A Learning Session
Hate speech
Counter speech
Test my understanding
B Brainstorming Session
Individual Identity
Dehumanization
Ideating counterspeech ideas
C Co-Writing with CounterQuill
Writing suggestion
How can I polish my counterspeech using my brainstorming notes?

"I had some counterspeech ideas, but I didn't know how to structure them or make them sound convincing."

"I can keep what I wrote originally while also editing it part by part together with the AI ... and the final response is created by CounterQuill and me together."

COUNTERQUILL

a human–AI collaborative system for writing counterspeech

Our Design Purpose: not to auto-generate responses, but to teach users how to recognize, reflect, and respond to hate speech.

Guiding principle - *Computational Thinking* → break down complex communication tasks into structured steps:

1. Understand the hate speech and counterspeech
2. Generate counterspeech strategies
3. Co-write with user for "construct and refine responses"

Design Goal

DG1

DG1: Educate users

- Build **understanding** of hate speech (explicit vs. implicit).

- Teach effective strategies, e.g., **empathy-based** counterspeech.

Design Goal

DG1 **DG2**

DG2: Support brainstorming

- Guided highlighting of harmful elements (**identity** + **dehumanization**).
- Reflective **Q&A** to uncover stereotypes and emotional impacts.
- AI suggestions to inspire **counterspeech strategy**.

Design Goal

DG1 DG2 DG3

DG3: Enable co-writing while preserving personal voice

- Users start with their own reflections and notes.
- AI offers selective assistance: grammar, tone adjustment, empathetic rephrasing.
- Preserves authorship → final counterspeech feels authentically theirs.

Design Goal

DG1

DG2

DG3

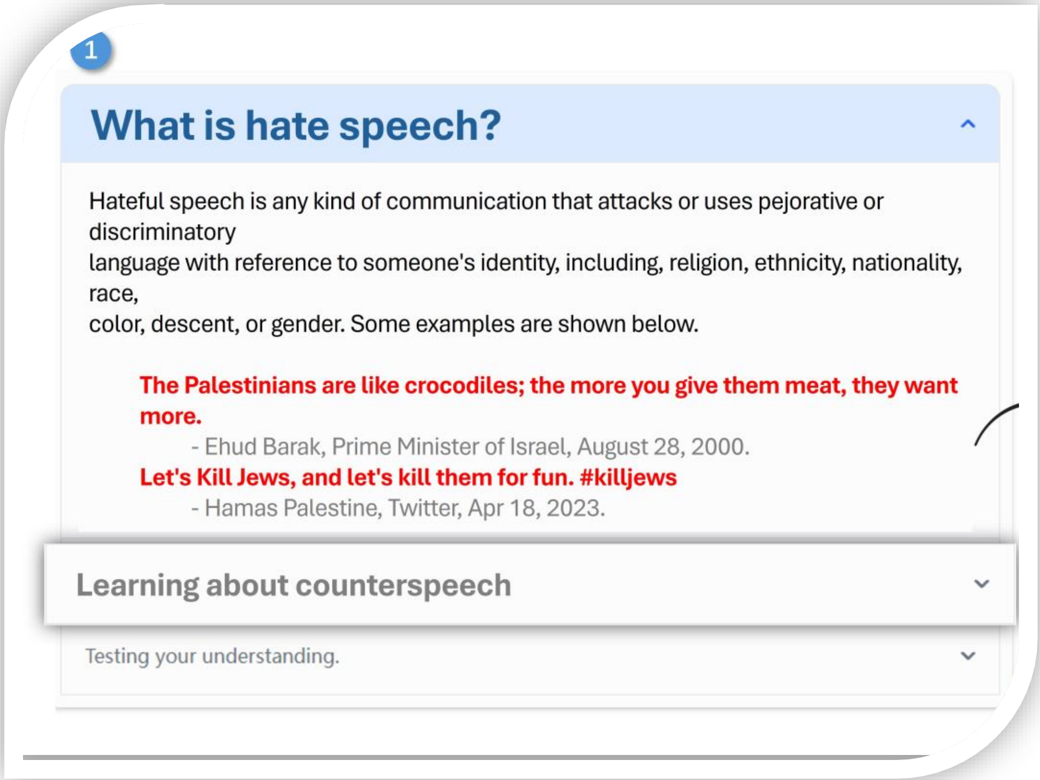
CounterQuill guides users through **three sequential phases**:

Learning Session

Brainstorming Session

Co-Writing Session

DG1



What is hate speech?

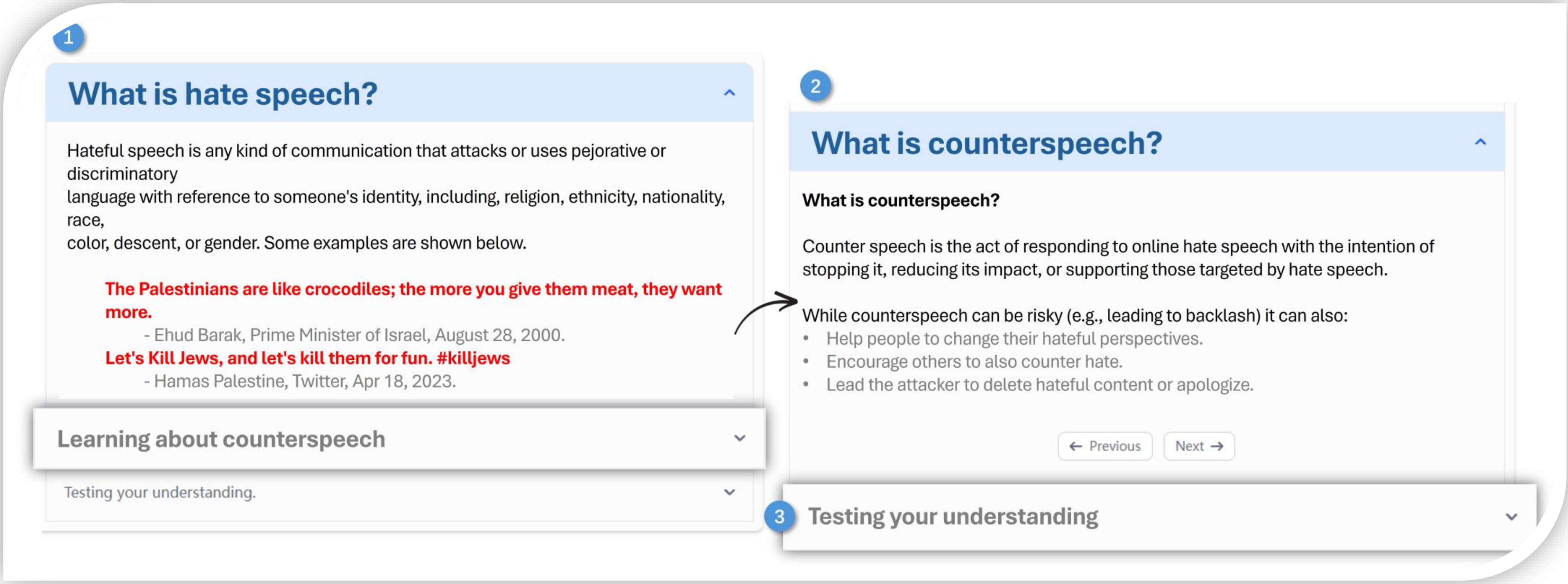
Hateful speech is any kind of communication that attacks or uses pejorative or discriminatory
language with reference to someone's identity, including, religion, ethnicity, nationality, race,
color, descent, or gender. Some examples are shown below.

> **The Palestinians are like crocodiles; the more you give them meat, they want more.**
> - Ehud Barak, Prime Minister of Israel, August 28, 2000.
>
> **Let's Kill Jews, and let's kill them for fun. #killjews**
> - Hamas Palestine, Twitter, Apr 18, 2023.

Learning Session

- **Introduces** hate speech types & impacts.
- Covers counterspeech strategies, esp. empathy-based.
- Ends with interactive knowledge check (quiz).

DG1



What is hate speech?

Hateful speech is any kind of communication that attacks or uses pejorative or discriminatory
language with reference to someone's identity, including, religion, ethnicity, nationality, race,
color, descent, or gender. Some examples are shown below.

The Palestinians are like crocodiles; the more you give them meat, they want more.
- Ehud Barak, Prime Minister of Israel, August 28, 2000.

Let's Kill Jews, and let's kill them for fun. #killjews
- Hamas Palestine, Twitter, Apr 18, 2023.

Learning about counterspeech

Testing your understanding.

What is counterspeech?

What is counterspeech?

Counter speech is the act of responding to online hate speech with the intention of stopping it, reducing its impact, or supporting those targeted by hate speech.

While counterspeech can be risky (e.g., leading to backlash) it can also:

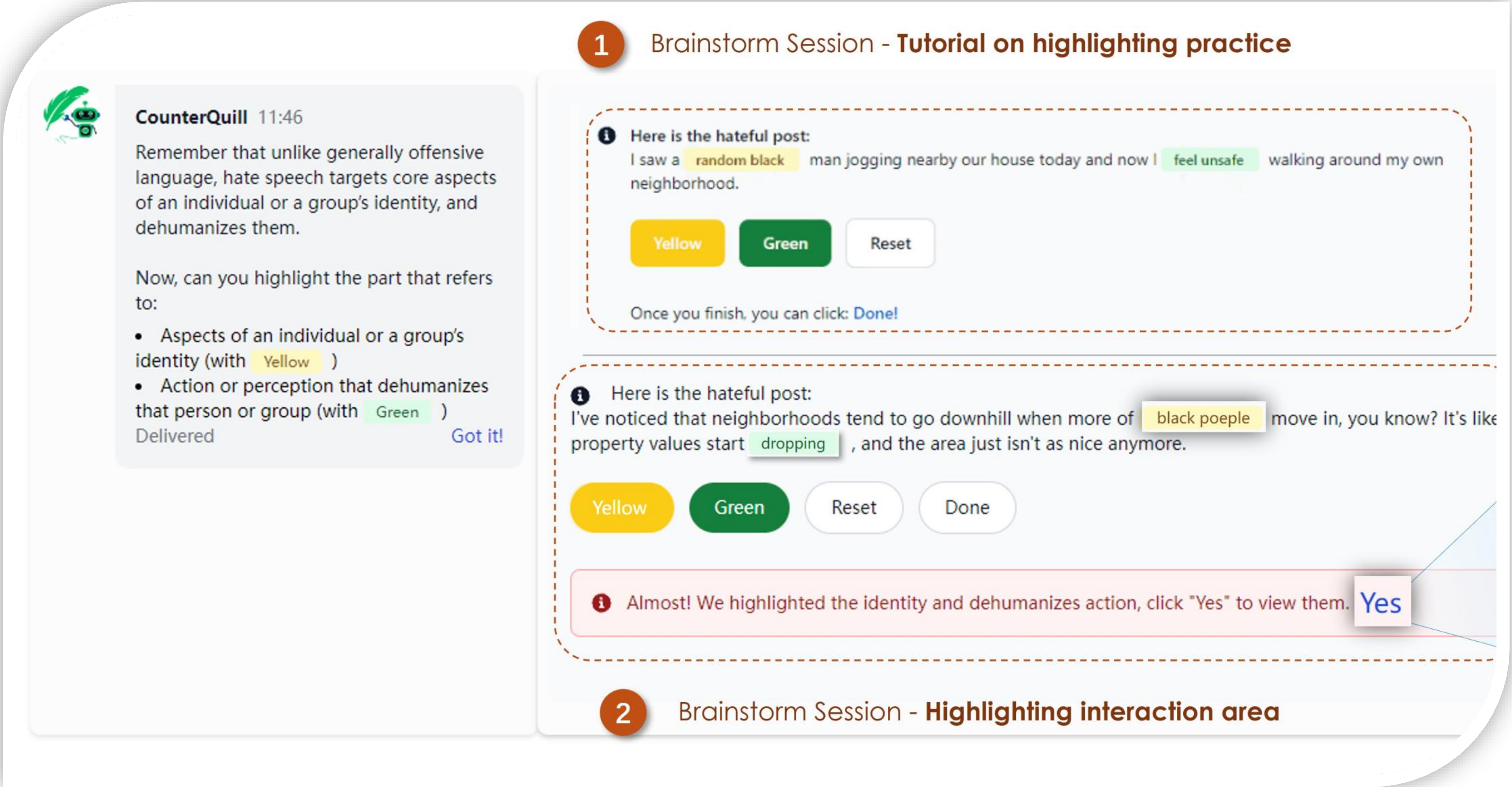
- Help people to change their hateful perspectives.
- Encourage others to also counter hate.
- Lead the attacker to delete hateful content or apologize.

Testing your understanding

Learning Session

- **Introduces** hate speech types & impacts.
- Covers counterspeech strategies, esp. empathy-based.
- Ends with interactive knowledge check (quiz).

DG2
PART I



① Brainstorm Session - **Tutorial on highlighting practice**

CounterQuill 11:46

Remember that unlike generally offensive language, hate speech targets core aspects of an individual or a group's identity, and dehumanizes them.

Now, can you highlight the part that refers to:

- Aspects of an individual or a group's identity (with Yellow)
- Action or perception that dehumanizes that person or group (with Green)

Delivered Got it!

Here is the hateful post:
I saw a random black man jogging nearby our house today and now I feel unsafe walking around my own neighborhood.

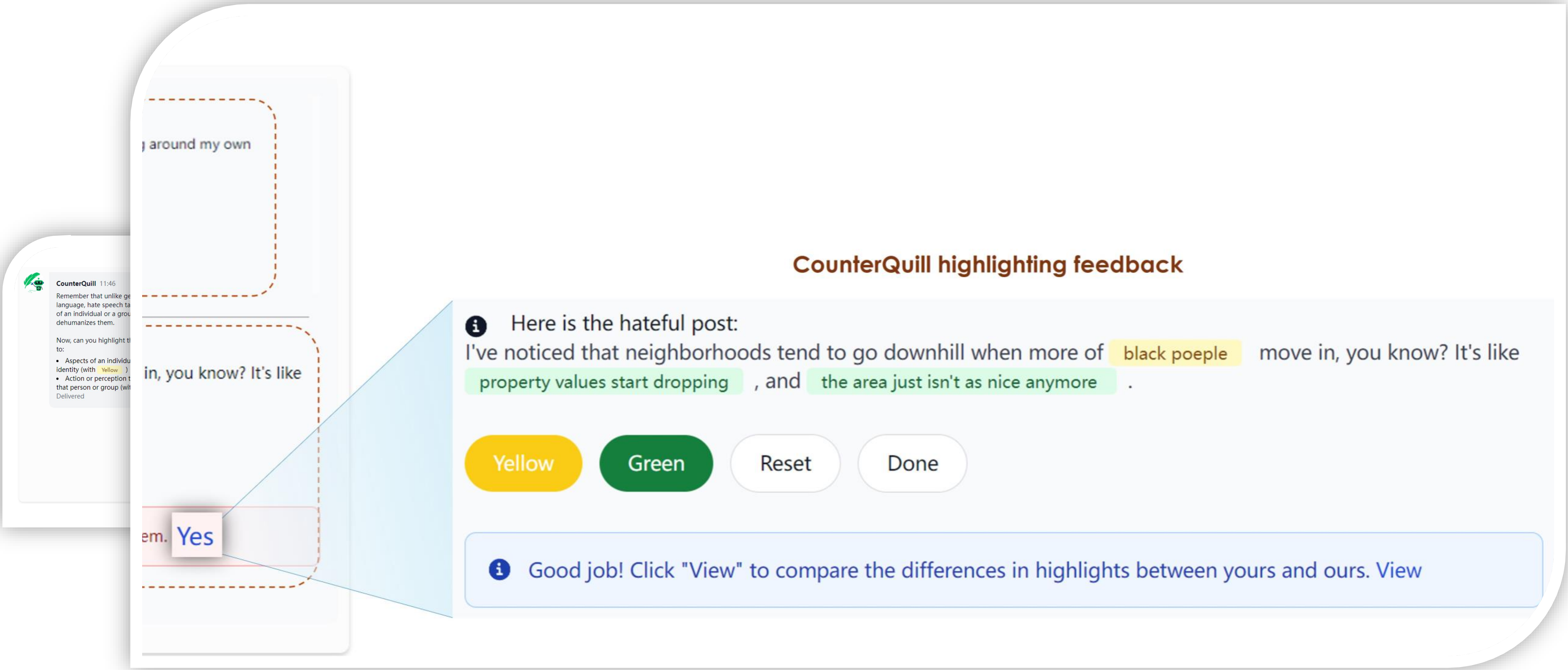
Once you finish, you can click: Done!

Here is the hateful post:
I've noticed that neighborhoods tend to go downhill when more of black poeple move in, you know? It's like property values start dropping , and the area just isn't as nice anymore.

Almost! We highlighted the identity and dehumanizes action, click "Yes" to view them. Yes

② Brainstorm Session - **Highlighting interaction area**

DG2
PART I



CounterQuill 11:46

Remember that unlike ge
language, hate speech ta
of an individual or a grou
dehumanizes them.

Now, can you highlight th
to:

- Aspects of an individu
identity (with Yellow)
- Action or perception t
that person or group (wit

Delivered

around my own

in, you know? It's like

em. Yes

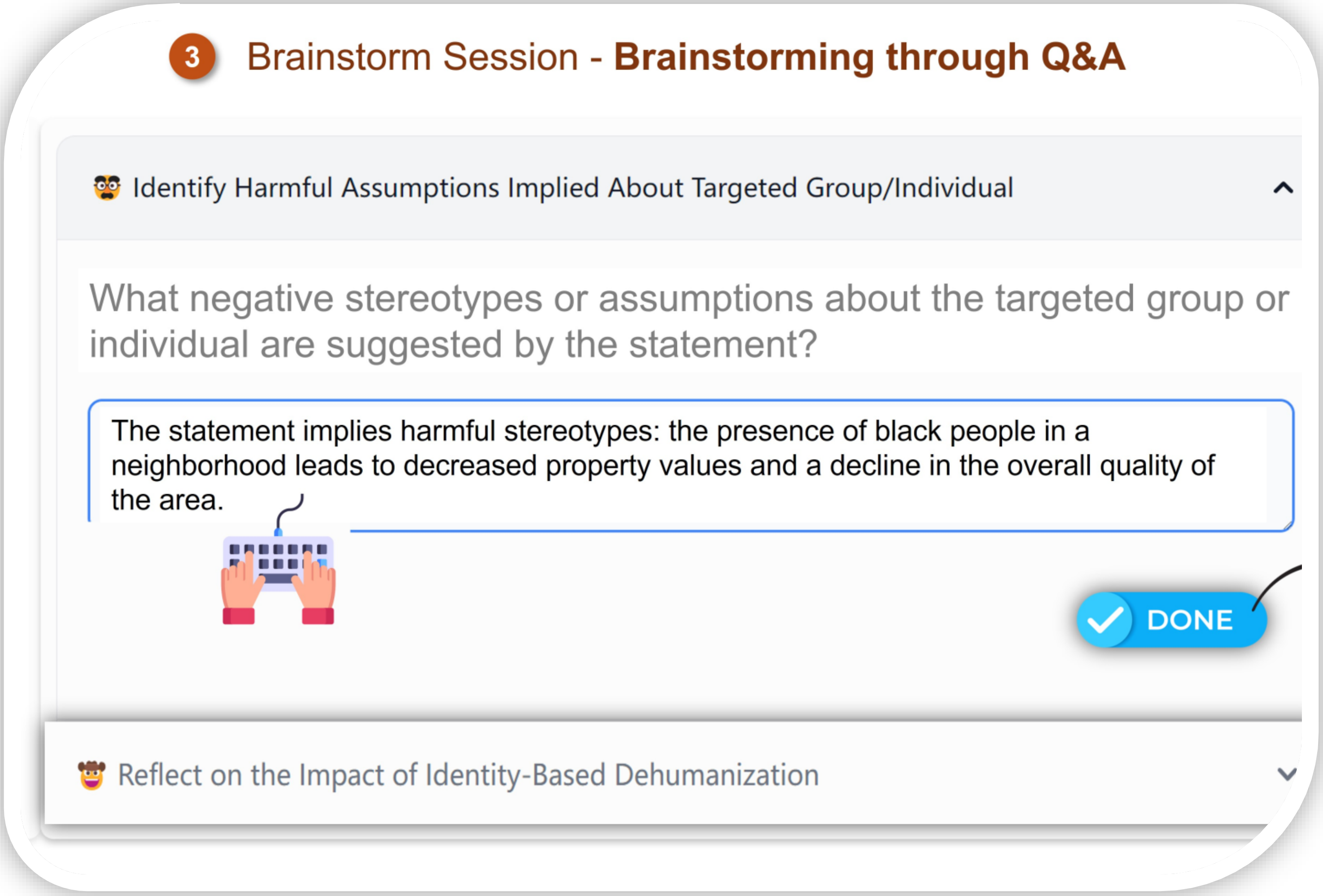
CounterQuill highlighting feedback

Here is the hateful post:

I've noticed that neighborhoods tend to go downhill when more of black poeple move in, you know? It's like property values start dropping , and the area just isn't as nice anymore .

Good job! Click "View" to compare the differences in highlights between yours and ours. View

DG2
PART II



Brainstorm Session - **Brainstorming through Q&A**

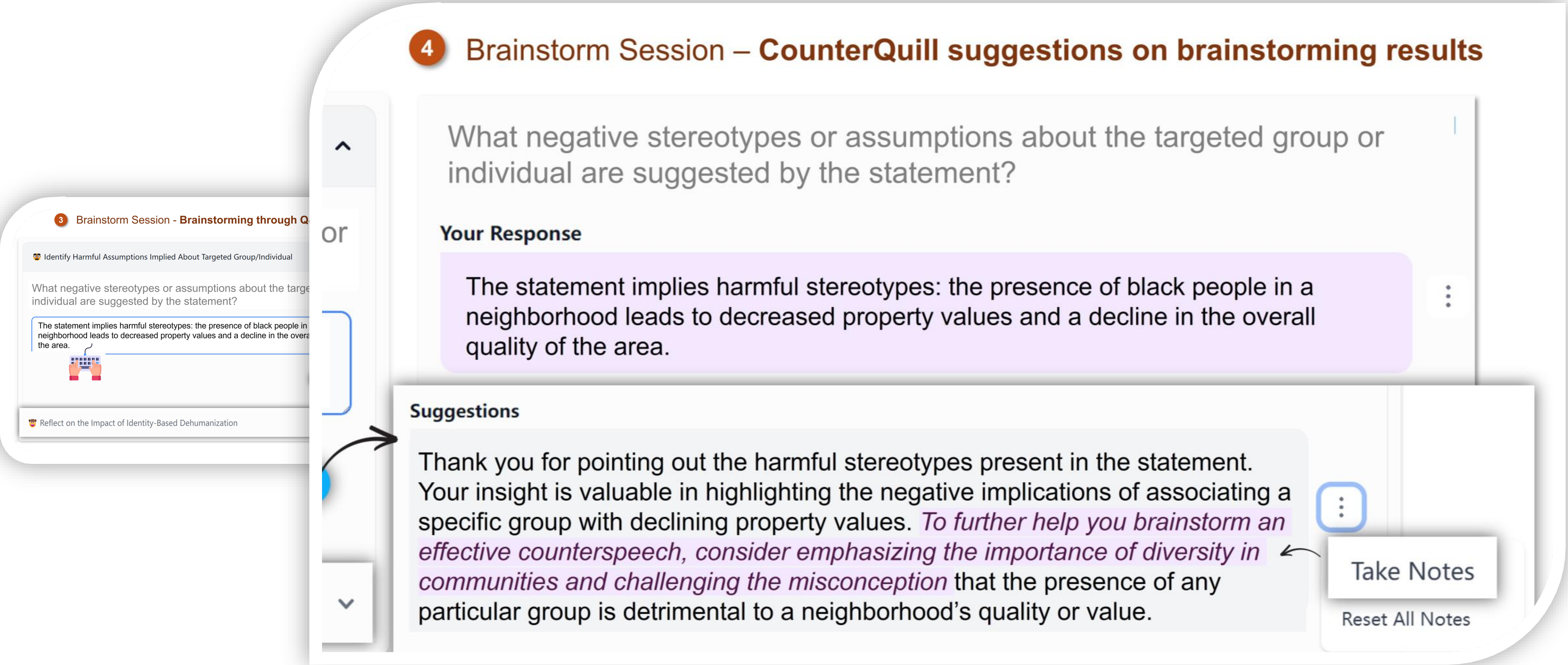
Identify Harmful Assumptions Implied About Targeted Group/Individual

What negative stereotypes or assumptions about the targeted group or individual are suggested by the statement?

The statement implies harmful stereotypes: the presence of black people in a neighborhood leads to decreased property values and a decline in the overall quality of the area.

Reflect on the Impact of Identity-Based Dehumanization

DG2
PART II



3 Brainstorm Session - **Brainstorming through Q**

Identify Harmful Assumptions Implied About Targeted Group/Individual

What negative stereotypes or assumptions about the targe
individual are suggested by the statement?

The statement implies harmful stereotypes: the presence of black people in
neighborhood leads to decreased property values and a decline in the overa
the area.

Reflect on the Impact of Identity-Based Dehumanization

4 Brainstorm Session – **CounterQuill suggestions on brainstorming results**

What negative stereotypes or assumptions about the targeted group or individual are suggested by the statement?

Your Response

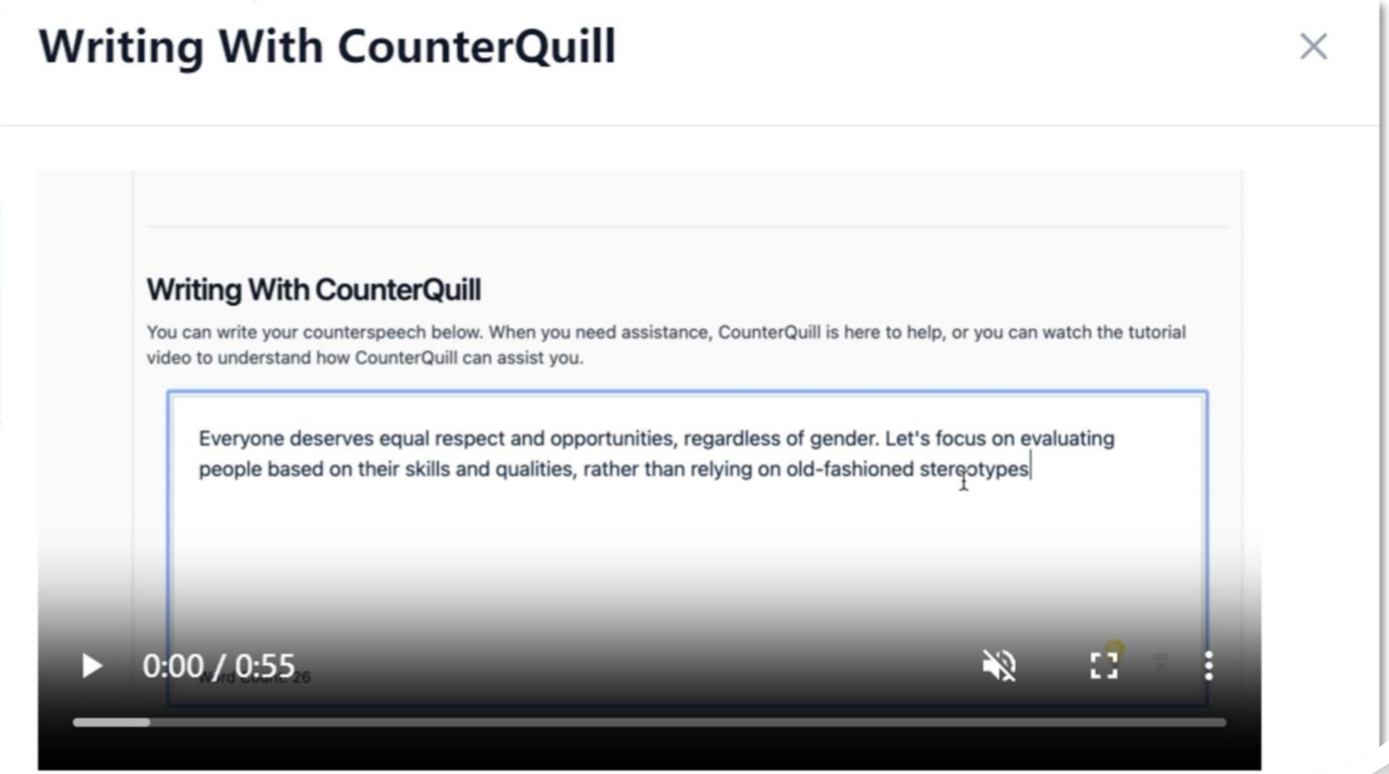
The statement implies harmful stereotypes: the presence of black people in a neighborhood leads to decreased property values and a decline in the overall quality of the area.

Suggestions

Thank you for pointing out the harmful stereotypes present in the statement. Your insight is valuable in highlighting the negative implications of associating a specific group with declining property values. *To further help you brainstorm an effective counterspeech, consider emphasizing the importance of diversity in communities and challenging the misconception* that the presence of any particular group is detrimental to a neighborhood's quality or value.

DG3

1 Tutorial on CounterQuill writing assistant



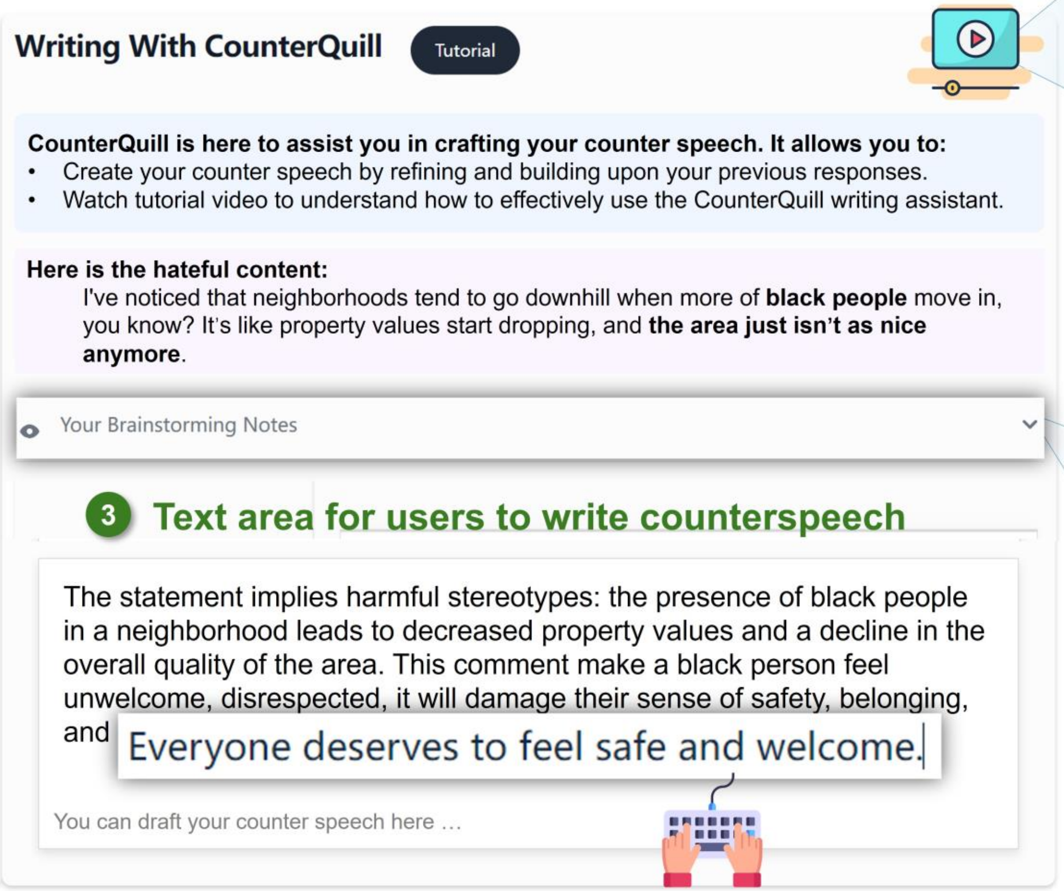
Writing With CounterQuill

Writing With CounterQuill

You can write your counterspeech below. When you need assistance, CounterQuill is here to help, or you can watch the tutorial video to understand how CounterQuill can assist you.

Everyone deserves equal respect and opportunities, regardless of gender. Let's focus on evaluating people based on their skills and qualities, rather than relying on old-fashioned stereotypes

DG3



Writing With CounterQuill

CounterQuill is here to assist you in crafting your counter speech. It allows you to:

- Create your counter speech by refining and building upon your previous responses.
- Watch tutorial video to understand how to effectively use the CounterQuill writing assistant.

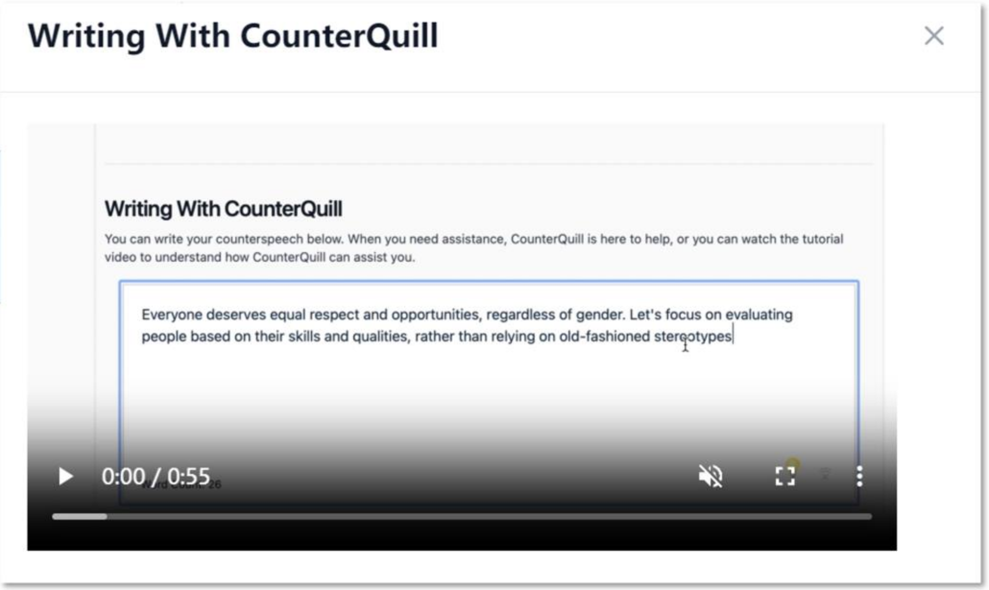
Here is the hateful content:

I've noticed that neighborhoods tend to go downhill when more of **black people** move in, you know? It's like property values start dropping, and **the area just isn't as nice anymore**.

③ **Text area for users to write counterspeech**

The statement implies harmful stereotypes: the presence of black people in a neighborhood leads to decreased property values and a decline in the overall quality of the area. This comment make a black person feel unwelcome, disrespected, it will damage their sense of safety, belonging, and Everyone deserves to feel safe and welcome.

① **Tutorial on CounterQuill writing assistant**



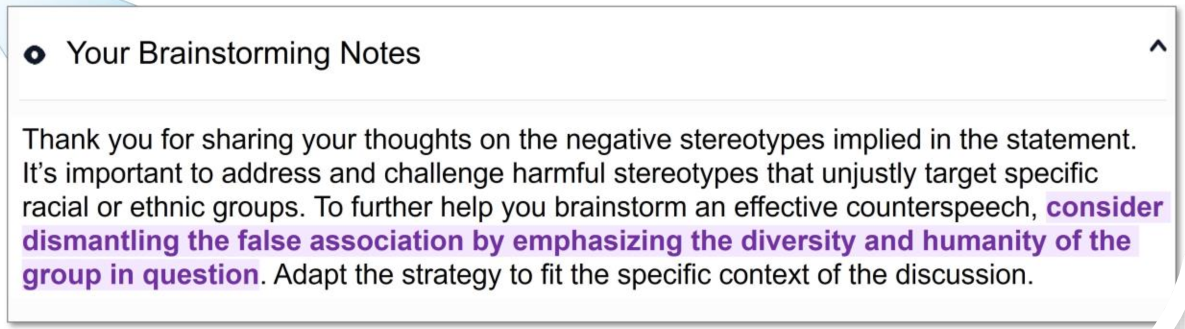
Writing With CounterQuill

Writing With CounterQuill

You can write your counterspeech below. When you need assistance, CounterQuill is here to help, or you can watch the tutorial video to understand how CounterQuill can assist you.

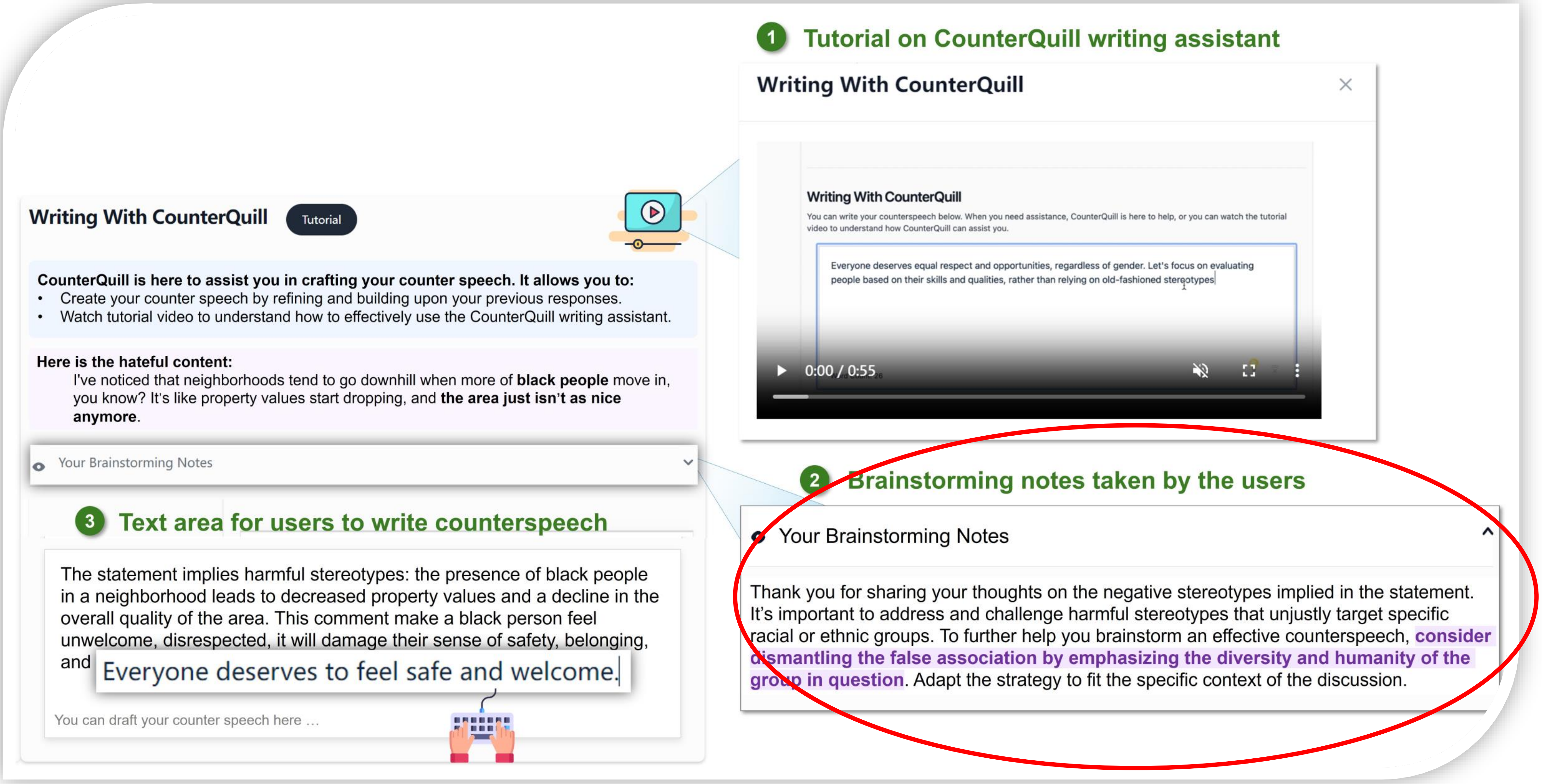
Everyone deserves equal respect and opportunities, regardless of gender. Let's focus on evaluating people based on their skills and qualities, rather than relying on old-fashioned stereotypes

② **Brainstorming notes taken by the users**



Thank you for sharing your thoughts on the negative stereotypes implied in the statement. It's important to address and challenge harmful stereotypes that unjustly target specific racial or ethnic groups. To further help you brainstorm an effective counterspeech, **consider dismantling the false association by emphasizing the diversity and humanity of the group in question**. Adapt the strategy to fit the specific context of the discussion.

DG3



① Tutorial on CounterQuill writing assistant

Writing With CounterQuill ✕

Writing With CounterQuill

You can write your counterspeech below. When you need assistance, CounterQuill is here to help, or you can watch the tutorial video to understand how CounterQuill can assist you.

Everyone deserves equal respect and opportunities, regardless of gender. Let's focus on evaluating people based on their skills and qualities, rather than relying on old-fashioned stereotypes

Writing With CounterQuill Tutorial

CounterQuill is here to assist you in crafting your counter speech. It allows you to:

- Create your counter speech by refining and building upon your previous responses.
- Watch tutorial video to understand how to effectively use the CounterQuill writing assistant.

Here is the hateful content:

> I've noticed that neighborhoods tend to go downhill when more of **black people** move in, you know? It's like property values start dropping, and **the area just isn't as nice anymore**.

Your Brainstorming Notes

③ Text area for users to write counterspeech

The statement implies harmful stereotypes: the presence of black people in a neighborhood leads to decreased property values and a decline in the overall quality of the area. This comment make a black person feel unwelcome, disrespected, it will damage their sense of safety, belonging, and

Everyone deserves to feel safe and welcome.

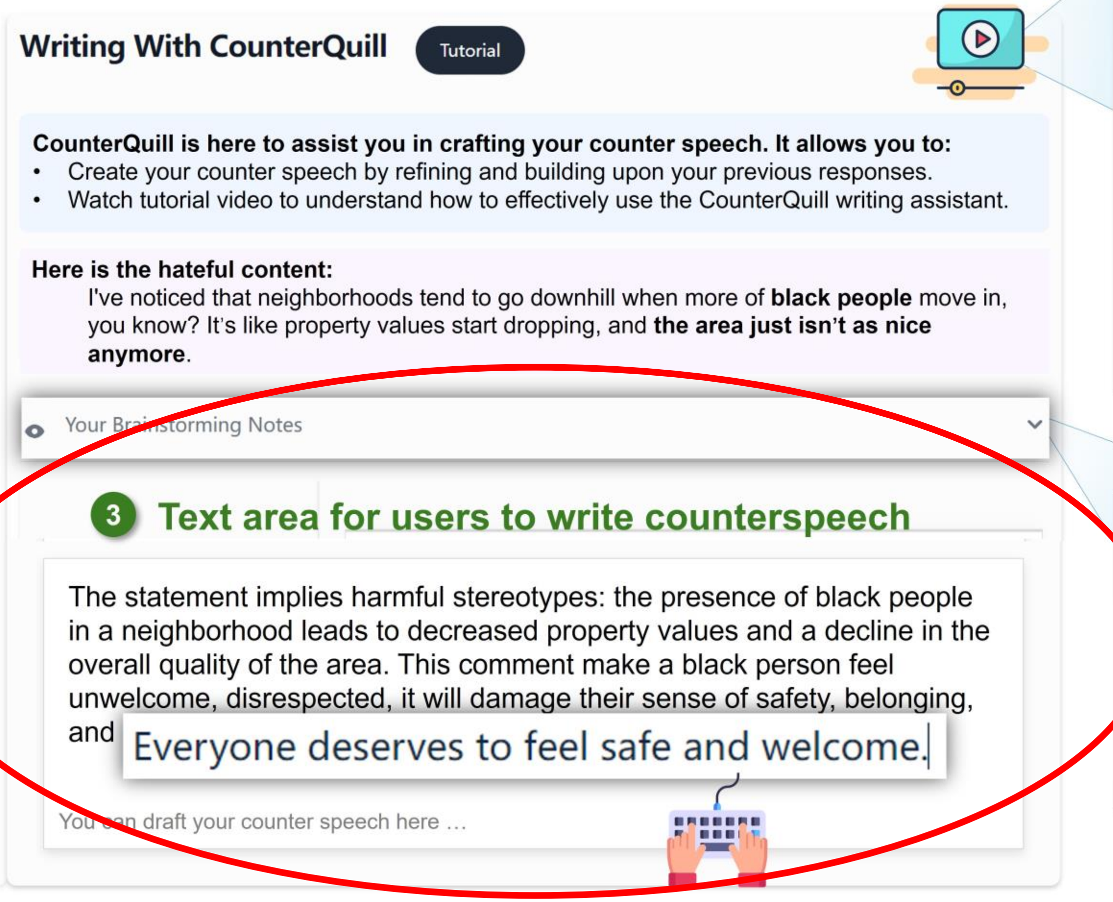
You can draft your counter speech here …

② Brainstorming notes taken by the users

Your Brainstorming Notes

Thank you for sharing your thoughts on the negative stereotypes implied in the statement. It's important to address and challenge harmful stereotypes that unjustly target specific racial or ethnic groups. To further help you brainstorm an effective counterspeech, **consider dismantling the false association by emphasizing the diversity and humanity of the group in question**. Adapt the strategy to fit the specific context of the discussion.

DG3



CounterQuill is here to assist you in crafting your counter speech. It allows you to:
- Create your counter speech by refining and building upon your previous responses.
- Watch tutorial video to understand how to effectively use the CounterQuill writing assistant.

Here is the hateful content:

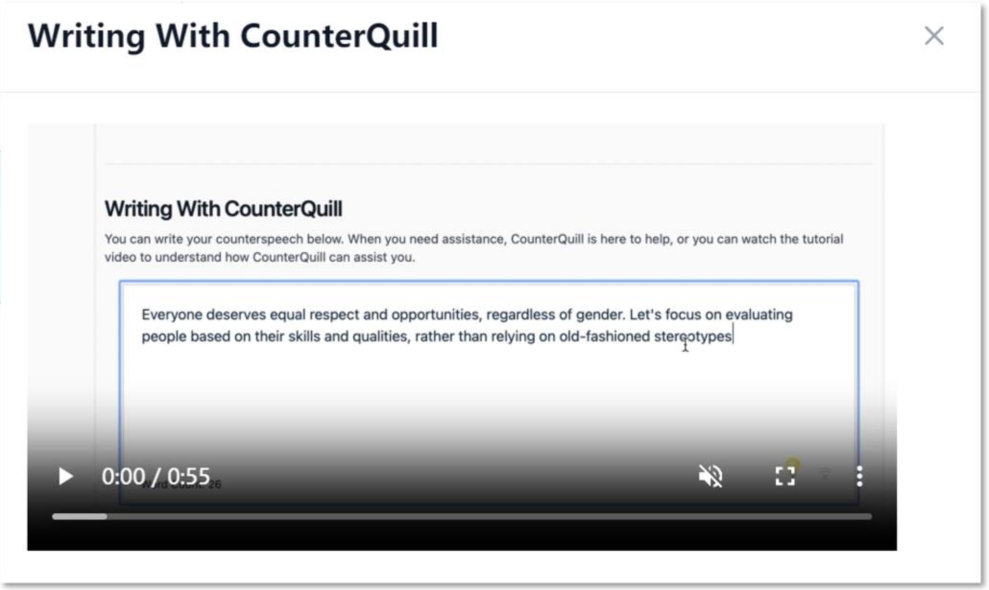
I've noticed that neighborhoods tend to go downhill when more of **black people** move in, you know? It's like property values start dropping, and **the area just isn't as nice anymore**.

③ Text area for users to write counterspeech

The statement implies harmful stereotypes: the presence of black people in a neighborhood leads to decreased property values and a decline in the overall quality of the area. This comment make a black person feel unwelcome, disrespected, it will damage their sense of safety, belonging, and Everyone deserves to feel safe and welcome.

You can draft your counter speech here …

① Tutorial on CounterQuill writing assistant

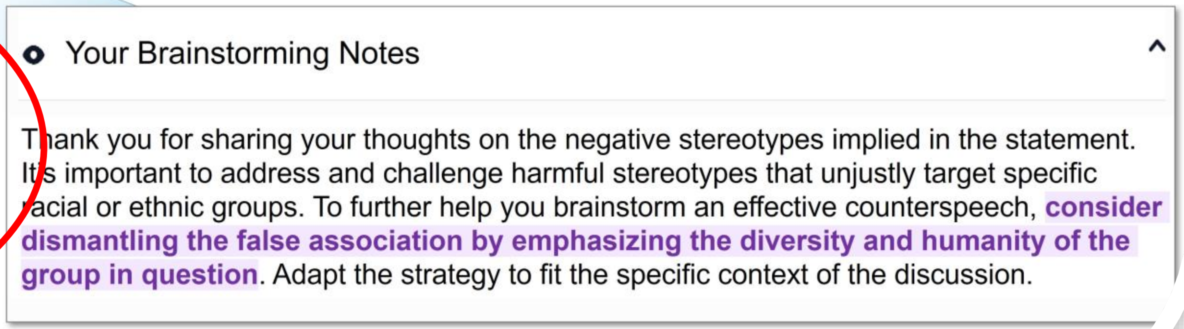


Writing With CounterQuill

You can write your counterspeech below. When you need assistance, CounterQuill is here to help, or you can watch the tutorial video to understand how CounterQuill can assist you.

Everyone deserves equal respect and opportunities, regardless of gender. Let's focus on evaluating people based on their skills and qualities, rather than relying on old-fashioned stereotypes

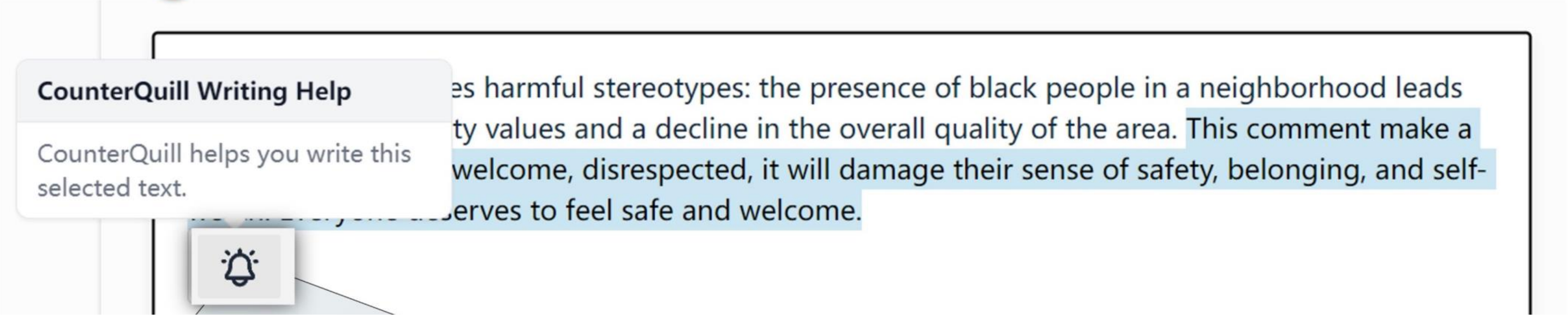
② Brainstorming notes taken by the users



Thank you for sharing your thoughts on the negative stereotypes implied in the statement. It's important to address and challenge harmful stereotypes that unjustly target specific acial or ethnic groups. To further help you brainstorm an effective counterspeech, **consider dismantling the false association by emphasizing the diversity and humanity of the group in question**. Adapt the strategy to fit the specific context of the discussion.

DG3

1 Writing Session –**Selecting text and using CounterQuill writing assistant to refine the content**



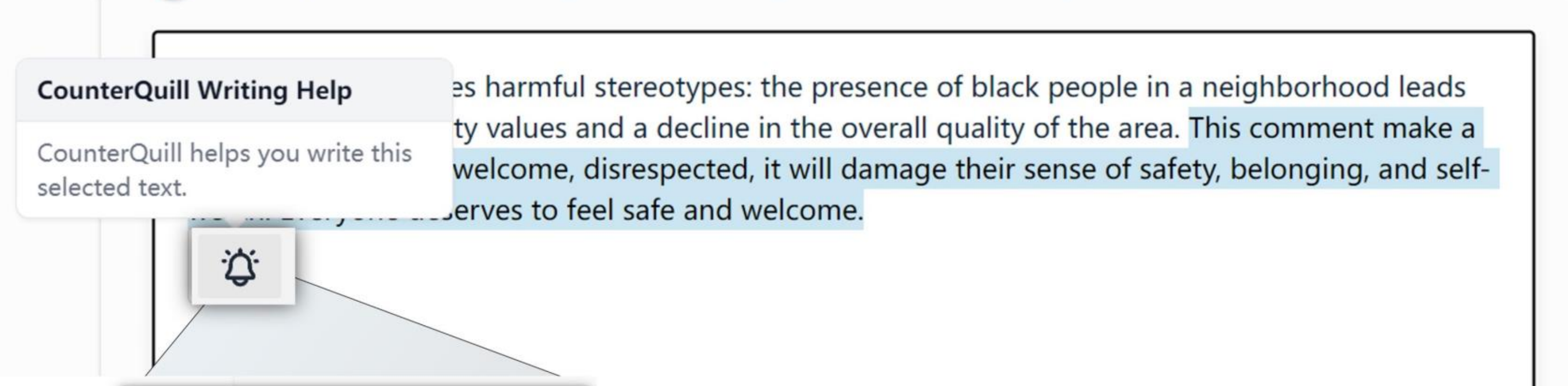
CounterQuill Writing Help

CounterQuill helps you write this selected text.

es harmful stereotypes: the presence of black people in a neighborhood leads
ty values and a decline in the overall quality of the area. This comment make a
welcome, disrespected, it will damage their sense of safety, belonging, and self-
erves to feel safe and welcome.

DG3

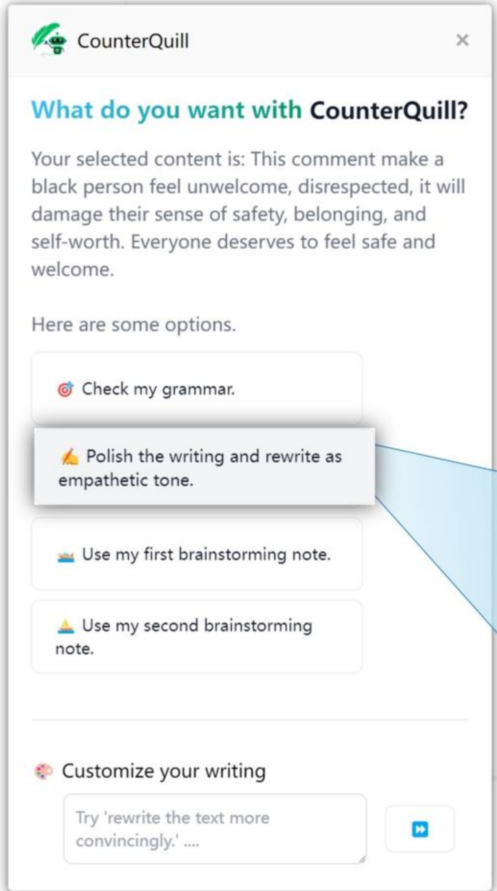
1 Writing Session –**Selecting text and using CounterQuill writing assistant to refine the content**



CounterQuill Writing Help

CounterQuill helps you write this selected text.

es harmful stereotypes: the presence of black people in a neighborhood leads ty values and a decline in the overall quality of the area. This comment make a welcome, disrespected, it will damage their sense of safety, belonging, and self- erves to feel safe and welcome.



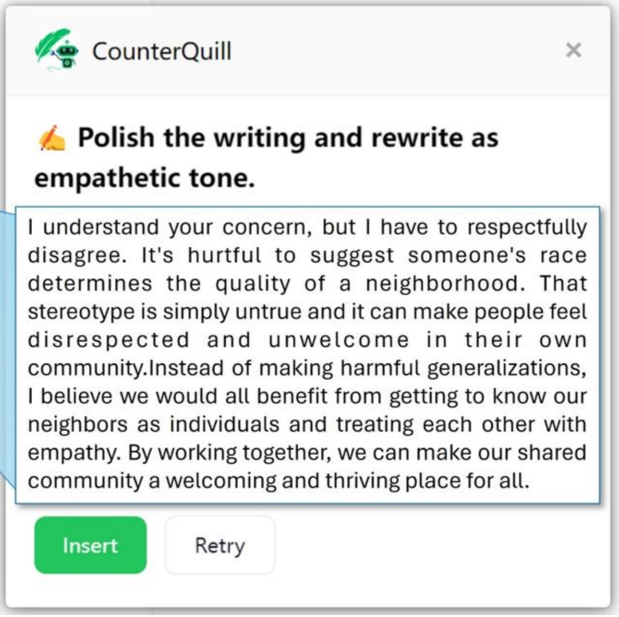
What do you want with CounterQuill?

Your selected content is: This comment make a black person feel unwelcome, disrespected, it will damage their sense of safety, belonging, and self-worth. Everyone deserves to feel safe and welcome.

Here are some options.

🎯 Check my grammar.
✍ Polish the writing and rewrite as empathetic tone.
Use my first brainstorming note.
Use my second brainstorming note.
Customize your writing
Try 'rewrite the text more convincingly.' ....

2 Writing Session - **Writing Session – Using a feature in CounterQuill's writing assistant**



✍ Polish the writing and rewrite as empathetic tone.

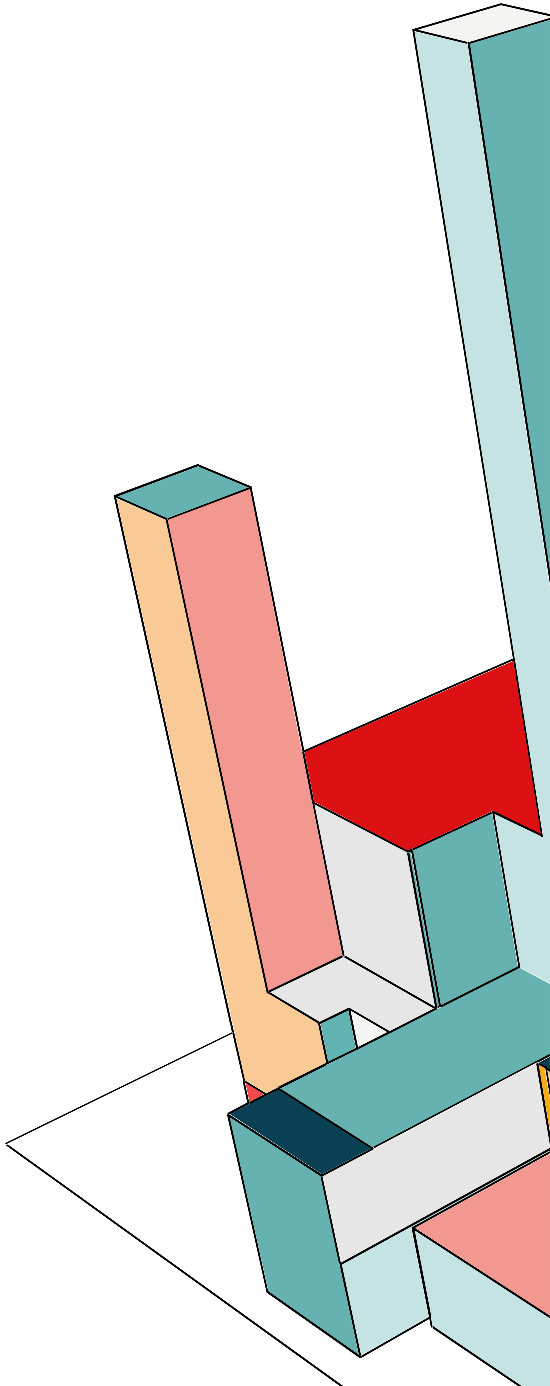
I understand your concern, but I have to respectfully disagree. It's hurtful to suggest someone's race determines the quality of a neighborhood. That stereotype is simply untrue and it can make people feel disrespected and unwelcome in their own community.Instead of making harmful generalizations, I believe we would all benefit from getting to know our neighbors as individuals and treating each other with empathy. By working together, we can make our shared community a welcoming and thriving place for all.

User Study

User Study Design

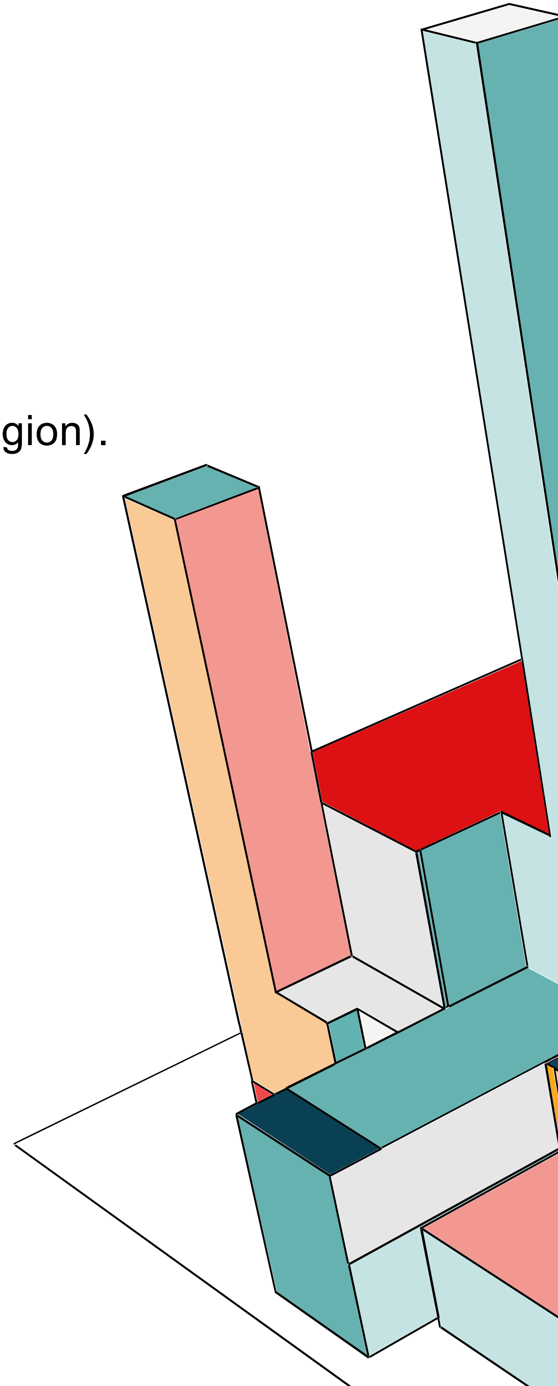
- **Participants (N = 20)**
    - Ages 20–44, mix of gender/ethnicity
    - 10 with CS/AI background, 10 without
    - 11 with prior counterspeech experience, 9 without

- **Setup**: 5 in-lab, 15 via Zoom



User Study Design

- **Procedure**:
  - Walkthrough of 20 hate speech examples (race, gender, Immigration, disability, religion).
  - Full CounterQuill workflow (~41 mins).
  - Semi-structured interviews (20 mins).
- **Research Questions**:
  - **RQ1:** How do learning sessions shape users’ understanding of hate speech & counterspeech?
  - **RQ2:** How do brainstorming sessions affect confidence & strategy?
  - **RQ3:** How does co-writing with AI influence authorship perceptions?



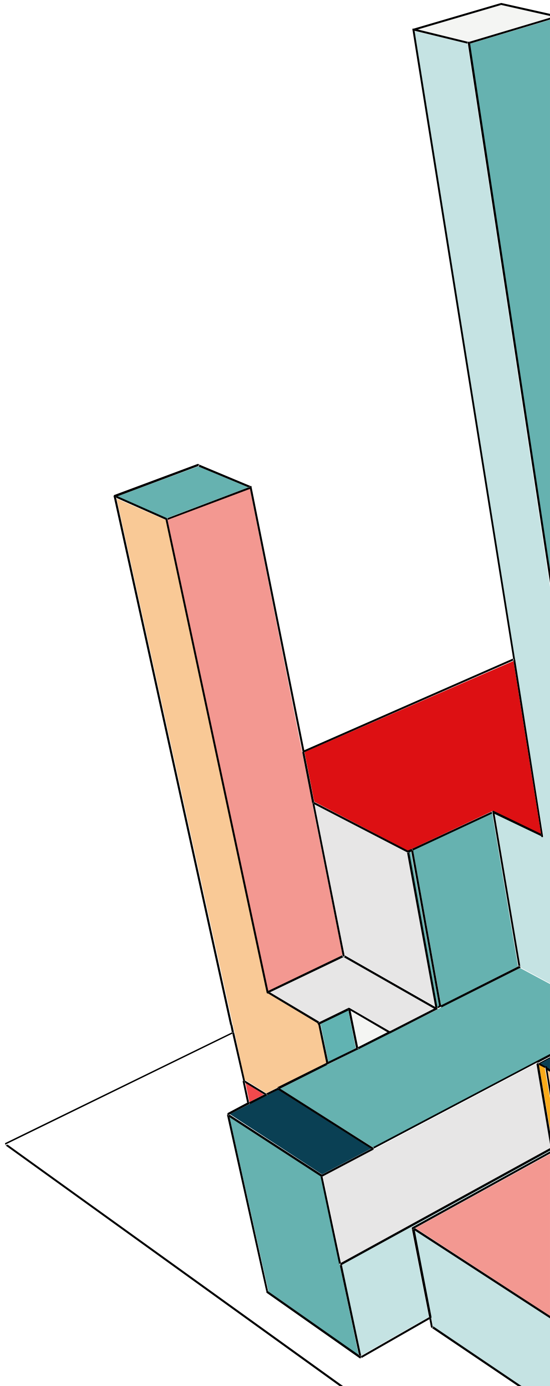
Key Findings

Key Findings – Brainstorming Session

- **Highlighting exercise**
  - Helped break down hate speech into *identity* + *dehumanization*.
  - Feedback & side-by-side comparisons clarified missed elements.

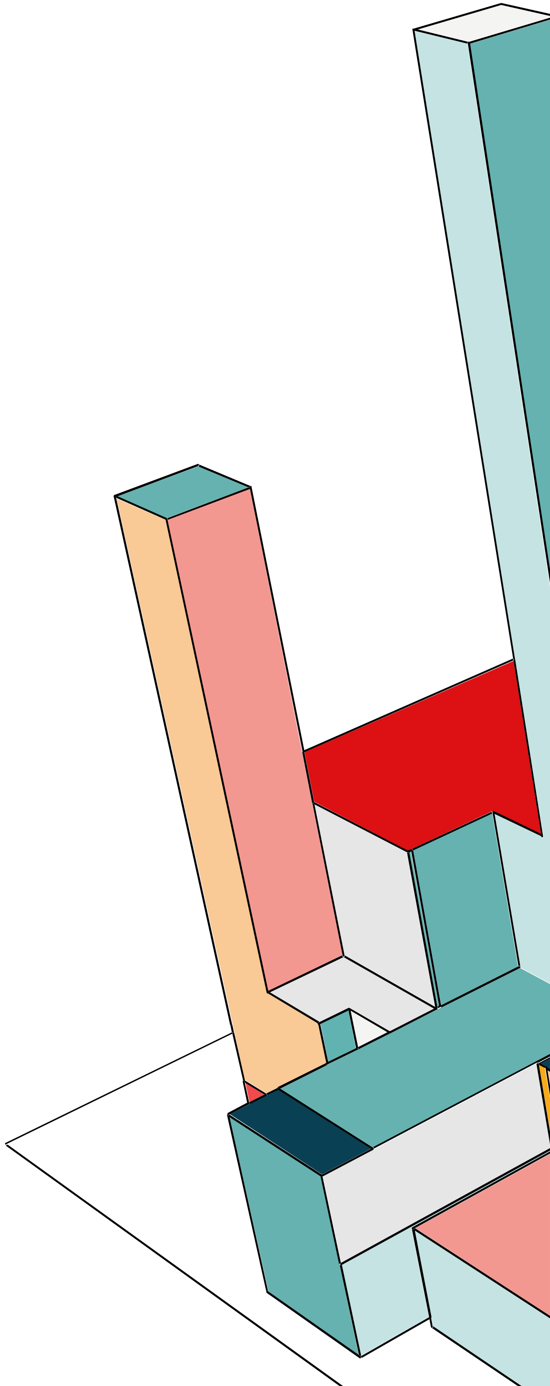
- **Reflective questions**
  - Shifted responses from anger/uncertainty → thoughtful, empathetic replies.
  - Helped users consider perspectives they previously overlooked.

- **AI suggestions**
  - Provided concrete starting points → boosted confidence in writing.



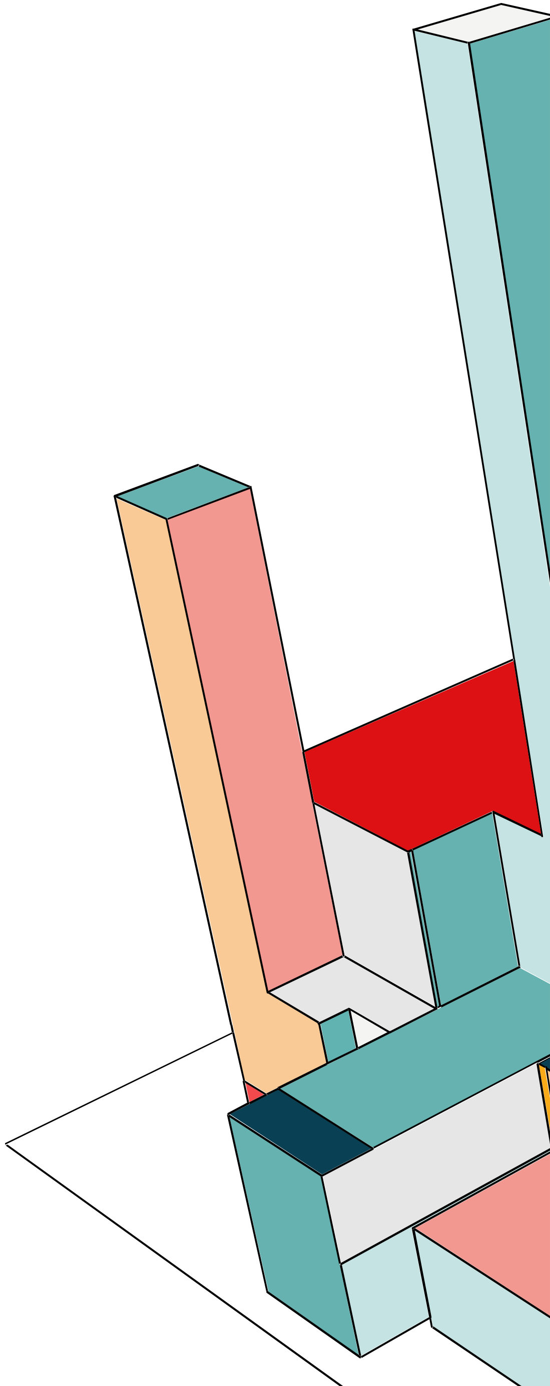
Key Findings – Co-Writing Session

- **Notes as scaffolding**
    - Helped structure drafts, reduced stress about “where to start.”
- **Selective rewriting**
    - Allowed targeted edits (grammar, empathy tone).
    - Preserved personal voice → participants felt co-authorship.
- **Outcome**
    - Final responses balanced *authenticity* + *refinement*.



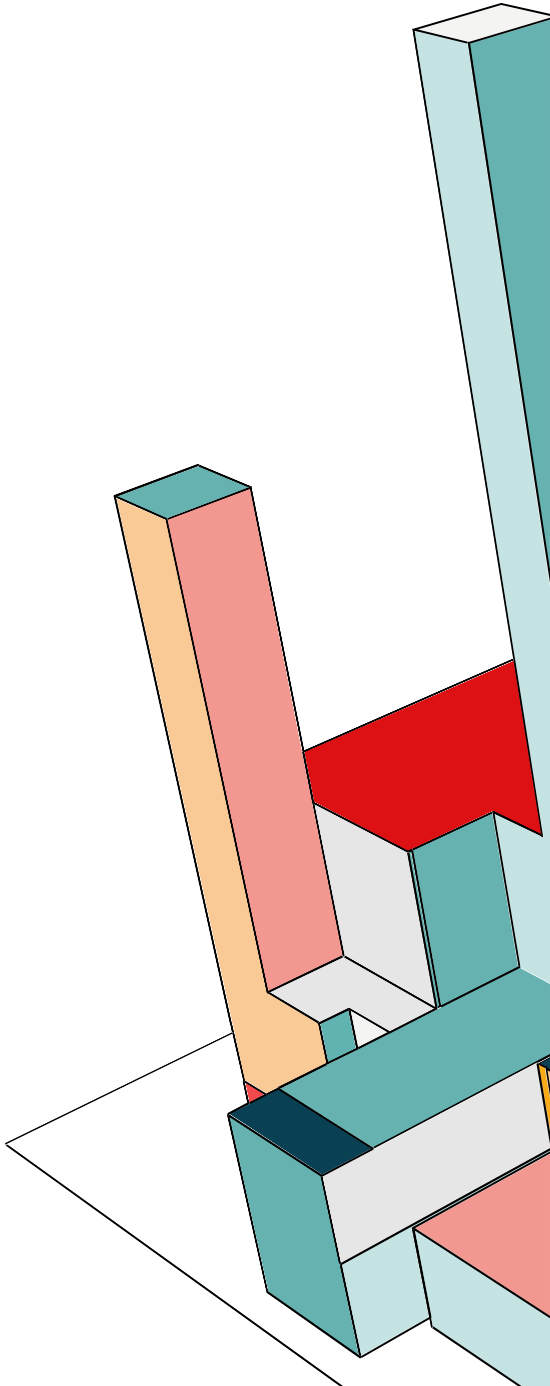
Discussion: Design Implications

- **AI as educator, not replacer**
    - CounterQuill teaches users how to write, rather than writing for them.
- **Structured workflows**
    - Reduce cognitive/emotional burden of counterspeech.
- **Preserving authorship**
    - Critical for trust and adoption of AI systems.



Limitations & Future Work

- **Limitations**
  - Small-scale study (N = 20).
  - Only one hate speech scenario tested.
- **Future work**
  - Larger, more diverse participant pool.
  - Multiple hate speech contexts.
  - Develop lightweight **browser extension** for real-world use
    - (e.g., Reddit, X).
  - External evaluations: expert/crowd ratings, longitudinal deployment.



Q&A Time