ChatGPT

# Counterspeech Generation

Counterspeech refers to direct responses to hateful or harmful content, intended to challenge or mitigate such content without resorting to censorship [1]. Human-delivered counterspeech has proven effective in reducing hate speech in online experiments [2], and platforms view it as a promising long-term moderation strategy [3]. With the advent of large language models (LLMs), researchers have begun exploring automatic counterspeech generation to scale these interventions. Modern LLMs like GPT-3 can perform remarkably well in generating relevant and persuasive text with minimal task-specific training [4]. For example, few-shot prompted GPT-3 was shown to produce high-quality, informative, and persuasive explanations of why a given message is hateful [4]. This suggests LLMs can serve as valuable tools for combating hate speech. At the same time, studies caution that LLM-generated content may introduce biases or factual errors if not carefully constrained [5]. Indeed, early uses of GPT-3 for content moderation noted its propensity to occasionally output biased language or hallucinated details [5], underscoring the need for careful prompt design and oversight.

Several works have specifically investigated using generative models to produce counterspeech or "counter-narratives" to hateful content. Tekiroğlu et al. (2020) appear among the first to leverage neural generation for counterspeech [6]. More recently, **Russo et al. (2023)** demonstrate that given a knowledge base and a well-crafted prompt, GPT-3.5 can generate *countering, emotional responses* to fake or manipulative posts [7]. Their approach showed that LLMs, when fed with relevant factual or empathetic context, can produce responses that directly confront harmful messages in an emotionally intelligent way [8]. In a concurrent study, **Leekha et al. (2024)** adopted a **Retrieval-Augmented Generation (RAG)** framework to ground counterspeech in external knowledge [9]. By retrieving pertinent facts or context (e.g. historical examples, definitions, community guidelines) and providing them to the prompt, RAG aims to increase the specificity and credibility of the generated counterspeech. This technique can flexibly inject knowledge into the response, potentially helping to fact-check or debunk misinformation within hateful narratives. Initial results from these studies are encouraging, but they primarily evaluated the content of generated counterspeech (e.g. relevance, coherence) rather than its impact on actual hate propagation. Notably, a recent field experiment provided the first real-world evidence that AI-generated counterspeech interventions can reduce user engagement with hate content [10]. In a live Twitter A/B test, generative counterspeech replies led to a measurable decrease in interactions with harmful posts [10], highlighting the practical potential of LLM-driven counterspeech when deployed in situ.

**Generic vs. Contextualized Counterspeech.** Early approaches to automated counterspeech often treated it as a one-size-fits-all endeavor. For scalability, researchers and NGOs have relied on pre-scripted or templated replies conveying general anti-hate messages (e.g. normative warnings, empathy prompts) that could be applied universally [11]. Pioneering studies like Munger's Twitter bot interventions followed this paradigm: every hateful user received a similar generic response (for instance, a reminder of community norms) [12] [13]. While such generic counterspeech can be mass-produced and has shown some efficacy [2] [12], it inherently ignores the specific context of the hateful post or the user's background. Recent work suggests this lack of tailoring may limit its persuasive power [14] [15]. Effective persuasion and moderation are highly context-dependent [15]. A response that directly addresses *why* a particular user made a hateful remark, or that connects with the user's identity and community norms, could be more impactful than a

generic reprimand. Thus, there is a marked shift in the literature from generic to **contextualized counterspeech**, which adapts to details of the situation [15]. Contextualization can involve leveraging information about the *user* (e.g. demographics, past behavior), the *conversation* (prior messages or specific content that triggered the hate), and the *community* or platform norms in which the exchange occurs [16] [17]. By incorporating such context, the counterspeech can be tailored in tone and content to resonate better with the target audience.

Only very recently have researchers begun to experiment with **personalization** and **adaptation** in counterspeech generation. For example, *Doğanç & Markov (2023)* condition generation on the hate speaker's **profile** attributes (specifically age and gender) to produce targeted counter-narratives [18]. Their study found that incorporating author profiling information is feasible using GPT-2/GPT-3.5, and it opens opportunities to make responses more relatable to the hate speaker's cohort [18]. Another study by *Bär et al. (2024)* prompted a GPT-based model to generate **contextualized counterspeech based on the content of the toxic message** (as opposed to a generic reply) [19]. This can be seen as a minimal form of context adaptation: the model explicitly attends to the particulars of the hateful post (e.g. its targets or rhetoric) when formulating a response. We are now seeing proposals to go further – for instance, by integrating the *conversation thread* or community rules into the prompt so that the response can reference them. Such multi-faceted context was recently outlined by Costello et al. (2024), who leverage community, user, and conversation metadata to craft counter-messages tailored to each scenario [20] [16]. Early evidence suggests that tailoring moderation messages to a user's psychology or ideology can improve outcomes in other domains (e.g. personalized *disinformation* correction was shown to enhance effectiveness). By analogy, tailored counterspeech might better address the underlying reasons for a user's hateful post, whether it stems from fear, misinformation, or social identity needs.

In tandem with context adaptation, researchers are also exploring diverse **counterspeech strategies** that LLMs can implement. Classic strategies in human counterspeech include fostering empathy ("*Imagine how it would feel if you...*") or **warning of consequences** ("*This language can have serious repercussions...*") [21]. These strategies have been adopted in automated replies as well, and indeed Bär et al. (2024) focus on empathy vs. consequences in their LLM-generated messages [22]. Other styles like **humor, satire, or "bending"** the aggressor's words have been theorized as counterspeech techniques. *Caponetto & Cepollaro (2023)* introduce **"bending"** as an innovative approach: deliberately giving a *distorted but ameliorative response* to a hateful utterance, effectively reframing or redirecting it to reduce its harm [23]. For example, a counterspeech bot might respond to a bigoted comment with a seemingly tangential remark that reinterprets the comment in a positive light, thus defusing its sting. While primarily conceptual, such strategies could be implemented by prompt engineering an LLM to respond in those styles. In practice, LLM-based counterspeech systems may combine multiple strategies – e.g. an empathetic tone with factual rebuttals – and select the approach best suited to the user or community context.

In summary, the literature on automated counterspeech is evolving from straightforward *generation of generic responses* toward **contextualized counterspeech generation** that accounts for who the speaker is, what the conversation is, and where it is happening. Personalizing the response (to the user's profile, beliefs, or the specific content) is hypothesized to make counterspeech more persuasive and engaging [15]. However, these enhancements introduce new challenges: the system must have access to reliable context data and avoid privacy or ethical pitfalls in using it, and the efficacy of tailored approaches must be verified. To that end, researchers stress the need for rigorous evaluation of not just the linguistic quality of LLM-generated counterspeech, but its *persuasive impact* on toxic behavior [9]. This has led to interest in new

methods for *evaluating* counterspeech – potentially by using LLMs themselves as proxies for human judgment, as discussed next.

## LLMs-as-Judges to Evaluate Persuasiveness

Evaluating the **persuasiveness** and effectiveness of counterspeech is non-trivial. Traditionally, one would conduct user studies or rely on expert annotations to judge whether a countermessage is likely to change attitudes or defuse conflict. Such human evaluation, however, is time-consuming and hard to scale [24]. In response, the NLP community has begun to leverage LLMs as *automated judges* of language outputs – a paradigm dubbed **"LLMs-as-Judges."** In this approach, the LLM is asked to **evaluate** content (e.g. another model's answer or a counterspeech message) according to specified criteria, effectively simulating a human rater. A recent comprehensive survey highlights that LLMs-as-judges can adapt evaluation criteria to task context and provide rich natural-language feedback, making the evaluation process more flexible and informative [25] [26]. For example, instead of a single numeric score, an LLM judge can explain that a particular counterargument was factually sound but perhaps too confrontational in tone. This capability to generate **interpretive evaluations** is a key advantage of LLM judges over static metrics [27]. Moreover, using LLMs for evaluation is highly scalable and reproducible – the same "judge" model can evaluate thousands of samples quickly, without the inconsistency or fatigue that human annotators might exhibit [26]. These benefits have driven the rapid adoption of LLM evaluators in many domains of text generation evaluation (summarization, dialogue, coding answers, etc.), and now increasingly for assessing qualities like persuasiveness and clarity in argumentative or safety-related outputs.

**Evaluation frameworks and prompting.** Several methodologies have been developed for how to prompt or train LLMs to act as evaluators. A straightforward setup is **pointwise evaluation**, where the LLM judge examines a single item (e.g. one counterspeech reply) and rates it against certain criteria [28]. For instance, we might prompt: *"On a scale from 1 to 5, how persuasive and respectful is the assistant's response?"*. This yields an absolute score or label for each output. Pointwise evaluation is simple and mirrors standard annotation, but it can struggle to calibrate scores across different items [29]. An alternative is **pairwise evaluation**, in which the judge is given two candidate responses and asked to choose which one better meets the criteria [30]. Pairwise comparisons force the evaluator to make a direct preference judgment, often revealing relative quality distinctions that absolute scoring might miss [31]. In fact, pairwise preference data underlies the common RLHF (Reinforcement Learning from Human Feedback) training for aligning LLMs, and similarly LLM judges often show more consistent results when deciding between two responses [32]. For evaluating **persuasiveness**, a pairwise setup is very useful: the LLM judge could be presented with two counterspeech messages addressing the same hateful post – perhaps one generic and one personalized – and asked *"Which response is more likely to convince the user to reconsider their stance?"*. Such comparisons not only yield a ranking of responses, but can also be used to fine-tune generative models (by favoring the higher-ranked outputs). In practice, pointwise and pairwise evaluations are both employed in LLM-as-judge frameworks, and even combined (e.g. first filtering candidates pairwise, then assigning a final score).

A crucial aspect of using LLMs as judges is prompt design and reasoning induction. Researchers have found that encouraging the LLM to **think step-by-step** before issuing a verdict improves the quality of its evaluations [33] [34]. Techniques like **Chain-of-Thought prompting** have the LLM outline the pros and cons of a response (perhaps noting its logical soundness, emotional tone, and relevance) and then derive an overall judgment. Extending this idea, *Wang et al. (2023)* propose a **Tree-of-Thoughts (ToT)** framework, where the evaluation problem is modeled as a search through a tree of possible reasoning paths [34]. The

LLM judge can explore multiple hypothetical deliberations or interpretations – effectively *simulating different lines of thought* – and converge on a more robust final answer [35]. In an evaluation context, Tree-of-Thought prompting might involve the judge considering several potential reactions of a user to a counterspeech message (e.g. *"Path A: The user feels understood and calms down... Path B: The user feels attacked..."*) and then deciding which path is more plausible. By assessing the response under various imagined scenarios, the judge can better estimate its true persuasiveness. This approach has been noted to improve evaluation accuracy in complex reasoning tasks [36]. Other enhancements include **self-consistency**, where the LLM judge's answer is sampled multiple times with slight prompt variations and the most common outcome is taken (reducing randomness), and **calibrated prompting**, where reference examples are given to set a baseline for judgments.

**Specialized LLM judge models.** In addition to prompting techniques, there is active research on **fine-tuning LLMs to become dedicated evaluation models**. Several teams have curated datasets of human and AI judgments to train "judge" models that replicate human evaluation standards. For example, **PROMETHEUS** is an open-source family of evaluator LLMs fine-tuned on large-scale preference data to enable fine-grained scoring of text [37]. The PROMETHEUS models (built on Llama-2 and Mistral backbones) can not only provide direct assessments but also perform pairwise rankings, and importantly they can handle *custom criteria* beyond a single "goodness" dimension [37]. This means one could ask PROMETHEUS to evaluate specifically the *empathy* of a response or its *adherence to policy*, etc., and it generalizes to that rubric. Likewise, **JudgeLM** (Zhu et al., 2023) is a recent effort in which a Vicuna LLM (based on LLaMA) was fine-tuned on over 100k pairs of GPT-4-generated comparative judgments [38]. The result is a smaller-scale model that achieves high agreement with GPT-4's own evaluations on open-ended tasks. In fact, fine-tuned judges like JudgeLM have reached ~90% agreement with their GPT-4 "teacher," exceeding even human-human agreement on some benchmarks [39]. This indicates that a well-trained LLM judge can produce assessments very consistent with expert annotators or powerful models, at least within the domains it was trained on. Other notable frameworks include **G-Eval** (an approach by Microsoft/OpenAI leveraging GPT-4 for evaluation) and **Auto-J** [40], **PandaLM** [41], among others, which each experiment with different training data or scoring methods for LLM evaluators. The general finding across these is that LLM-based evaluation correlates surprisingly well with human evaluation on many metrics, such as coherence, relevance, and even factual accuracy, while being orders of magnitude faster and cheaper.

Applying LLMs-as-judges to *persuasiveness* evaluation is a natural extension of these developments. Recent work has started treating qualities like **argument strength, clarity, and civility** as criteria for LLM evaluators [42] [43]. For instance, an LLM judge might be prompted: *"Evaluate the assistant's response in terms of how persuasive it is to the original poster, how clear and respectful it is, and whether it aligns with the intent to de-escalate conflict."* The judge can then provide a breakdown (e.g. noting the use of evidence, emotional appeal, politeness) and a holistic persuasiveness score. Such usage aligns with the goal of our research: we wish to automatically **measure how convincing a contextualized counterspeech message is**, without running a full user study for every iteration. By leveraging an LLM judge, we can simulate a human moderator's perspective on what makes a message compelling. One powerful configuration is to use the judge in a **generation-and-evaluation loop**. We can prompt the LLM to generate several candidate counterspeech responses (perhaps varying in strategy or wording), and then use a *judge LLM* – armed with a pairwise comparison prompt or even a Tree-of-Thought evaluation prompt – to **select the most persuasive option**. This approach effectively uses the LLM-as-judge to guide the LLM-as-generator, analogous to a human editor picking the best draft. Prior studies have shown that having LLMs "vote" on the best output (e.g. majority vote from multiple generations, or a single model picking the best of N samples) can significantly improve quality [44] [36]. By simulating judgment of persuasiveness, the LLM can

rank outputs in a way that aligns with likely human reception. Of course, care must be taken to engineer the judge prompt well – research shows LLM evaluators can be sensitive to wording and may carry biases reflecting their training data [45]. Nonetheless, frameworks like **Tree-of-Thought prompting** combined with an LLM judge offer a promising methodology to evaluate and optimize counterspeech. It allows us to incorporate nuanced human-like criteria (persuasion success, tone appropriateness, etc.) into the model selection process [34]. In summary, the emergence of LLMs-as-judges provides a robust foundation for assessing communicative qualities of generated content. By leveraging these advances – pointwise and pairwise judgment, interpretive reasoning chains, and specialized judge models – our research can systematically evaluate how well LLM-generated *contextualized counterspeech* performs in terms of persuasiveness, clarity, and alignment with the intended counter-hate objectives. This approach supports the development of a counterspeech generation pipeline where a **"judge" LLM component scores or ranks candidate messages, ensuring that the final output is not only coherent and relevant, but also strategically effective in real social contexts.

---


[1] [2] [11] [12] [13] [14] [21] [22] **Generative AI may backfire for counterspeech**
https://arxiv.org/html/2411.14986v1

[3] [4] [5] [7] [8] [9] [10] **LLM generated responses to mitigate the impact of hate speech**
https://arxiv.org/html/2311.16905v2

[6] [15] [16] [17] [20] [43] **Contextualized Counterspeech: Strategies for Adaptation, Personalization, and Evaluation**
https://arxiv.org/html/2412.07338v1

[18] **From Generic to Personalized: Investigating Strategies for Generating Targeted Counter Narratives against Hate Speech - ACL Anthology**
https://aclanthology.org/2023.cs4oa-1.1/

[19] **Investigating Strategies for Generating Targeted Counter Narratives ...**
https://www.researchgate.net/publication/375723595_From_Generic_to_Personalized_Investigating_Strategies_for_Generating_Targeted_Counter_Narratives_against_Hate_Speech

[23] **Laura Caponetto & Bianca Cepollaro, Bending as Counterspeech - PhilPapers**
https://philpapers.org/rec/CAPBAC-2

[24] [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [40] [41] [42] [44] [45] **2412.05579v2.pdf**
file://file-Vzmu46WTQzUo3Qkj1aDifj

[39] **JudgeLM: Fine-tuned Large Language Models are Scalable Judges**
https://arxiv.org/html/2310.17631v2