ELLIE ROGERS

AUTOMATICALLY GENERATING COUNTER-SPEECH: OPPORTUNITIES AND CHALLENGES OF USING LLMS FOR CVE

As technology has developed, extremist actors have found new ways to use it to their advantage. Large language models (LLMs) are one area that has been exploited by extremists to create and share content.

INTRODUCTION

LLMs use natural language processing (NLP), and often artificial intelligence (AI), to process and generate text for translation, information retrieval, conversational interactions, and summarisation. Countering violent extremism (CVE) efforts are also making use of LLMs, by automatically generating counter-speech (interventions that directly address, or offer alternatives to extremist content), with the aim of dissuading individuals from extremist narratives. Scholars have argued that LLMs may assist practitioners in identifying hate speech, creating targeted counter-speech content, and engaging with users to deliver counter-speech. This article outlines the potential for LLMs as a vehicle for counter-speech, whilst highlighting the challenges that must be overcome.

POTENTIAL BENEFITS

Using LLMs to automatically generate counter-speech can assist with the increasing demand for counter-speech online, as they can be used to supplement the design and dissemination of counter-speech and may be more easily scalable than solely human-operated interventions. When identifying where counter-speech is needed, LLMs could automatically detect hate speech. Counter-speakers have also shown interest in using AI to gather relevant information to assist them in designing counter-speech content. Using LLMs to assist counter-speakers in identifying hate speech and constructing counter-speech responses could be a beneficial time-saving strategy.

It is important to reduce the strain on counter-speakers, as they can become the victims of hate speech themselves and experience negative impacts on their personal well-being. Since counter-speech requires a lot of time and effort and often involves addressing upsetting content, some counter-speakers have reported feeling overwhelmed and experiencing negative impacts on their mental health. As such, counter-speakers have suggested that it could be beneficial to use AI for counter-speech campaigns to protect their personal well-being, as it can offer them anonymity and reduce the amount of harmful content that they must identify and review, helping them to become more emotionally detached.

Counter-speakers have suggested that it could be beneficial to use AI for counter-speech campaigns to protect their personal well-being.

POTENTIAL CHALLENGES

Whilst using LLMs for counter-speech can reduce the physical and mental load on counter-speakers, there are important practical limitations and ethical challenges that should be considered.

There are questions surrounding the functionality of LLMs and their ability to produce relevant and credible counter-speech to a variety of audiences. Generating counter-speech requires large amounts of data, language comprehension, emotional intelligence, and contextual knowledge, which can be challenging to achieve with machine learning alone. As such, some LLM-generated counter-speech has been found to include factual, or grammatical errors, which may limit the effectiveness of the counter-speech within a CVE campaign.

Counter-speakers have expressed concerns that a lack of human involvement during the design and dissemination of counter-speech may have impacts on the authenticity and credibility of the counter-speech. For instance, disclosing that counter-speech was written by AI has been found to significantly reduce users perceived trust in the counter-speech. Alternatively, if it is not disclosed when AI has been used for counter-speech, transparency concerns can arise from users being unaware that LLMs are involved. It is important that LLM-generated counter-speech is manually reviewed to ensure it is authentic, accurate, and appropriate for the target audience.

Using fully automated design and delivery methods for counter-speech raises a number of ethical considerations including user privacy and bias. Privacy is a key ethical consideration for LLMs, which can gather highly personal information about individuals. For example, when conversing with a chatbot, individuals may not understand that their conversation can be read by real people. When designing counter-speech campaigns that use LLMs, user privacy should be carefully considered and protected as much as possible.

Biases can be built into LLMs by the developers, from the datasets used to train them, or during real-world implementation. Biases can result in harmful and inaccurate information being shared, which can encourage discrimination. For example, a translation LLM on Facebook incorrectly translated "good morning" written in Arabic to "hurt them" and "attack them", resulting in a Palestinian individual being arrested by Israeli authorities. Diverse groups of individuals need to be involved in the design and dissemination of LLMs to ensure that biases are mitigated and LLMs are not generating harmful or incorrect content.

THE EFFECTIVENESS OF LLM-GENERATED COUNTER-SPEECH

There is limited empirical research that assesses the effectiveness of LLM-generated counter-speech as part of a CVE campaign. Most studies focus on the ability of LLMs to generate counter-speech, leaving measures of effectiveness mostly restricted to evaluations of the counter-speech's written quality. There are, however, a smaller number of studies that do assess the practical effectiveness and impact of LLM-generated counter-speech.

Chung et al. (2021) tested an NLP tool that aimed to assist NGO practitioners in countering Islamophobia on Twitter in English, French, and Italian. The tool detected Islamophobia and then automatically composed counter-speech responses. The practitioners who tested the tool had positive feedback and felt it was an innovative addition to counter-speech writing. Importantly, some of the practitioners emphasised that the tool should not entirely replace manual writing, but instead should be used to assist them in counter-speech writing, as modifications to the generated counter-speech were often necessary. The tool was still considered to be useful as writing counter-speech from scratch reportedly took longer than modifying the counter-speech that the tool generated.

Bilewicz et al. (2021) assessed whether counter-speech messages that were generated and delivered by a bot (disguised as a real male user) could be effective at reducing verbal aggression within two subreddits. When verbal aggression was detected, the bot delivered counter-speech by directly replying to the user who posted the aggressive content. User comments posted 60 days before and after the intervention were analysed and a control group of users from other subreddits, who were not targeted with the bot, were used for comparison. Users displayed a lower proportion of verbal aggression after the intervention than before, whereas the proportion of verbal aggression in the control group remained largely the same throughout. This study suggests that bots could be used as part of counter-speech interventions.

Bär et al. (2024) explored whether LLM-generated counter-speech was effective in reducing hate on Twitter (X). Compared to manually-generated counter-speech, the LLM-generated counter-speech was found to be less effective in reducing online hate. The LLM-generated counter-speech was also associated with an increase in hate posts after users viewed the counter-speech. The researchers suggest that users may have recognised that the counter-speech was LLM-generated, causing them to react negatively. These findings highlight the potential counter-productive backfire effects that can arise from the use of LLMs to generate counter-speech.

CONCLUSION

LLMs can generate counter-speech in response to a range of extremist content, which can offer some protections to counter-speakers and may reduce the human resources that are needed to design and deliver counter-speech campaigns. However, there are some important limitations associated with the use of LLMs to generate counter-speech that must be considered. The potentially limited functionality of LLMs can result in the production of counter-speech that contains inaccuracies and the design and use of LLMs raises user privacy and bias concerns that can have harmful implications.

The small number of studies that assess the practical effectiveness of LLM-generated counter-speech offer mixed findings. LLMs show some promise in assisting practitioners with counter-speech writing and may potentially help to reduce verbal aggression online. It is important to consider that using LLMs to generate counter-speech may backfire and create increased hostility. Any use of LLMs within counter-speech campaigns needs to be accompanied by rigorous evaluation, risk assessment, and human oversight to ensure that the counter-speech being generated is relevant, factual, and non-harmful.

Using fully automated design and delivery methods for counter-speech raises a number of ethical considerations including user privacy and bias.

Ellie Rogers is a PhD candidate at Swansea University within the Cyber Threats Research Centre (CYTREC). Her research focuses on the algorithmic amplification of counter-speech as a response to online extremism.