THREAT REPORT

Title	The use of AI to target regional, culturally significant language groups
Reference	II-TR-013
Date	Published: July-2023
Produced by	Insights & Intelligence, CERT NZ - Computer Emergency Response Team New Zealand

Overview

CERT NZ's Insights and Intelligence team produces periodic threat reports to illustrate and identify notable trends in information we collect. This includes highlighting threats and possible actions or interventions for consideration.

This report focuses on recent artificial intelligence (AI) advancements and how it can be used to cause harm to regional, culturally significant (RCS) communities.

Summary

This report asserts that Large Language Models (LLMs) like the Generative Pre-Training Transformer (GPT, also known as ChatGPT) are being used by malicious actors to target RCS communities in two major ways.

1. Allowing non-regional language speakers greater access to intrinsic customs and norms of RCS communities, giving them deeper understanding of these communities to exploit them.
2. Allowing non-regional language speakers to effectively pass themselves as members of the community.

New Zealand is home to many RCS languages such as te reo Māori, Samoan and other Asia-Pacific languages. Historically, these communities may not have been targeted with "high quality/high volume" malicious content (e.g., phishing emails, phone scams and disinformation).

RCS communities that are not accustomed to having malicious content in their language are at risk of being more susceptible to malicious content generated by Generative AI.

Discussion

AI systems are designed to perform tasks that typically require human intelligence, such as visual perception, speech recognition and decision making. 'Generative AI' is a subfield of AI that focuses on creating systems capable of generating new content or completing tasks without significant guidance from humans. The most popular example of this currently is ChatGPT which is a Large Language Model (LLM).

[Introducing ChatGPT (openai.com)]

These systems use methods that allow them to learn complex patterns on a deeper level than other methods. Because of this they can receive and generate content simulating a genuine conversation or appear to understand further nuance in their outputs.

Recent advancements in Generative AI have increased access to RCS languages, as well as their respective populations. Whilst LLMs can help bridge communication gaps for many communities across the world, the unprecedented level of accessibility exposes communities to sophisticated cyber-enabled scams and fraud.

This creates a risk from AI to these communities which should be considered if active efforts are made to “teach” LLMs regional languages used in New Zealand.

LLMs and how they differ to traditional translators.

ChatGPT and other LLMs can be trained to understand many languages, select subtle grammar rules or vocabulary, and avoid artifacts that can occasionally be picked up by native speakers. Traditional online translators cannot create material themselves; their results are based on pre-configured answers. This means the material given to an online translator could hold elements that will appear odd or wrong to native speakers.

History of RCS languages and their relevance to scams and fraud.

According to CERT NZ data, RCS languages have historically had fewer scams and phishing campaigns than other communities. This is partly due to the limited sophistication of translation tools and understanding of the culture.

A prohibitive cost of entry for scammers to these communities (for example, paying human translators) have likely made them not worthwhile as potential targets. This has been the case in New Zealand, where only a small fraction of reports to CERT NZ are about scams and phishing in languages like te reo Māori and Pacific Island languages.

Recent LLM technology could increase the frequency and quality of scams targeting these communities. If these communities are not familiar with malicious messaging, such attacks could be highly effective.

Even with safeguards, LLMs can be used to create malicious messages.

Many language models have been trained not to give unethical advice. However, carefully crafted user inputs into the model can bypass the safeguards in place and produce an output with potentially malicious applications.

For example, you can ask the LLMs to create a phishing email without directly calling it a phishing email. There are also ways to 'trick' the system to produce what it doesn’t recognise as malicious messaging in an ethical context.

While ChatGPT is currently the most popular LLM, other powerful models are also widely accessible. They are either open source or have been leaked. These will continue to improve with or without safeguards and could be used for malicious purposes. These open-source models can also be independently trained on RCS to improve their capabilities even further.

Recommendations

CERT NZ encourages that decisions for training AI models like LLMs incorporate an increased ability for malicious actors to take advantage of RCS groups as s risk factor. This should sit alongside other risks to consider.

There is strong justification for targeted education and awareness-building for RCS language users in New Zealand. This will help to build resilience to scams and fraud in those communities. Ideally, this should be done before we see widespread targeting of these demographics with this technology.

Sharing this content

This report is produced by the CERT NZ Insights & Intelligence team and is intended for sharing with the New Zealand public.

If you would like to provide feedback or get in contact with CERT NZ, please email: info@cert.govt.nz