June 2025

Disrupting malicious uses of AI: June 2025

OpenAI

Table of contents

Executive Summary

Our mission is to ensure that artificial general intelligence benefits all of humanity. We advance this mission by deploying our innovations to build AI tools that help people solve really hard problems.

As we laid out in our submission to the Office of Science and Technology Policy's U.S. AI Action Plan in March, we believe that making sure AI benefits the most people possible means enabling AI through common-sense rules aimed at protecting people from actual harms, and building democratic AI. This includes preventing the use of AI tools by authoritarian regimes to amass power and control their citizens, or to threaten or coerce other states; as well as activities such as covert influence operations (IO), child exploitation, scams, spam, and malicious cyber activity.

It also includes *using* AI to defend against such abuses. By using AI as a force multiplier for our expert investigative teams, in the three months since our last report we've been able to detect, disrupt and expose abusive activity including social engineering, cyber espionage, deceptive employment schemes, covert influence operations and scams.

These operations originated in many parts of the world, acted in many different ways, and focused on many different targets. A significant number appeared to originate in China: Four of the 10 cases in this report, spanning social engineering, covert influence operations and cyber threats, likely had a Chinese origin. But we've disrupted abuses from many other countries too: this report includes case studies of a likely task scam from Cambodia, comment spamming apparently from the Philippines, covert influence attempts potentially linked with Russia and Iran, and deceptive employment schemes.

AI investigations are an evolving discipline. Every operation we disrupt gives us a better understanding of how threat actors are trying to abuse our models, and enables us to refine our defenses. We'll continue to share our findings to enable stronger defenses across the internet. But AI is only one part of the overall ecosystem, and OpenAI is only one part of the

world of AI. We especially welcome the recent threat reports by our peers at Google and Anthropic that fill out more of the picture of the AI threatscape.

Case studies

Deceptive Employment Scheme: IT Workers

> Threat actors using AI and other technologies in an attempt to evolve and scale their deceptive hiring attempts.

Actor

We identified and banned ChatGPT accounts associated with what appeared to be multiple suspected deceptive employment campaigns. These threat actors used OpenAI's models to develop materials supporting what may be fraudulent attempts to apply for IT, software engineering and other remote jobs around the world. While we cannot determine the locations or nationalities of the threat actors, their behaviors were consistent with activity publicly attributed to IT worker schemes connected to North Korea (DPRK). Some of the actors linked to these recent campaigns may have been employed as contractors by the core group of potential DPRK-linked threat actors to perform application tasks and operate hardware, including within the US.

Behavior

Similar to the threat actors we disrupted and wrote about in February, the latest campaigns attempted to use AI at each step of the employment process. Previously, we observed these actors using AI to manually generate credible, often U.S.-based personas with fabricated employment histories at prominent companies. This time, they attempted some degree of automated generation of resumes, and some indicators suggest operators in Africa posing as job applicants, in addition to recruiting people in North America to run laptops on their behalf.

We detected two distinct strands of activity, likely representing two types of operator: core operators, and contractors.

The core operators attempted to automate résumé creation based on specific job descriptions, skill templates, and persona profiles, and sought information about building tools to manage and track job applications. They also used our models to generate content that resembled job postings aimed at recruiting contractors in different parts of the world.

The core operators used ChatGPT as a research tool to help inform remote-work setups. They also engaged our models to generate text concerning the recruitment of real people in the US to take delivery of company laptops, which would then be remotely accessed by the core threat actors or their contractors. The threat actors researched using tools such as Tailscale peer-to-peer VPN, OBS Studio, vdo.ninja live-feed injection, and HDMI capture loops as part of their operations. These tools have the potential to be used to circumvent corporate security measures, and if successful, they would allow someone to maintain a persistent, undetected remote presence, as well as attempt to bypass some identity verification processes that rely on live video meetings.

Meanwhile, the possible contractor operators used ChatGPT to help complete job application tasks. They also used it to generate content that resembled messages to the core operators inquiring about payments and about the personas used to apply for remote jobs.

Completions

We determined that these threat actors attempted to use our models for a deceptive operation. Specifically, they used detailed prompts, instructions, and automation loops to generate tailored, credible résumés at scale. The table below summarizes these observed tactics, techniques, and procedures (TTPs)

Activity	**LLM ATT&CK Framework Category**
Automating to systematically fabricate detailed résumés aligned to various tech job descriptions, personas, and industry norms. Threat actors automated generation of consistent work histories, educational backgrounds, and references via looping scripts.	LLM Supported Social Engineering
Threat actors utilized the model to answer employment-related, likely application questions, coding assignments, and real-time interview questions, based on particular uploaded resumes.	LLM Supported Social Engineering
Threat actors sought guidance for remotely configuring corporate-issued laptops to appear as though domestically located, including advice on geolocation masking and endpoint security evasion methods.	LLM-Enhanced Anomaly Detection Evasion
LLM assisted coding of tools to move the mouse automatically, or keep a computer awake remotely, possibly to assist in remote	LLM Aided Development

working infrastructure set ups.	

Impact

We cannot independently assess the success of these operations—assessing its impact would require inputs from multiple stakeholders.

While the threat actors likely built AI into every step of their process to increase their efficiency, it also increased their exposure. By giving us insights across their workflows, they enabled us to share insights about these campaigns with relevant industry peers and authorities for each stage of their activity, enhancing our collective ability to detect, prevent, and respond to such threats while advancing our shared safety.

Covert IO: Operation “Sneer Review”

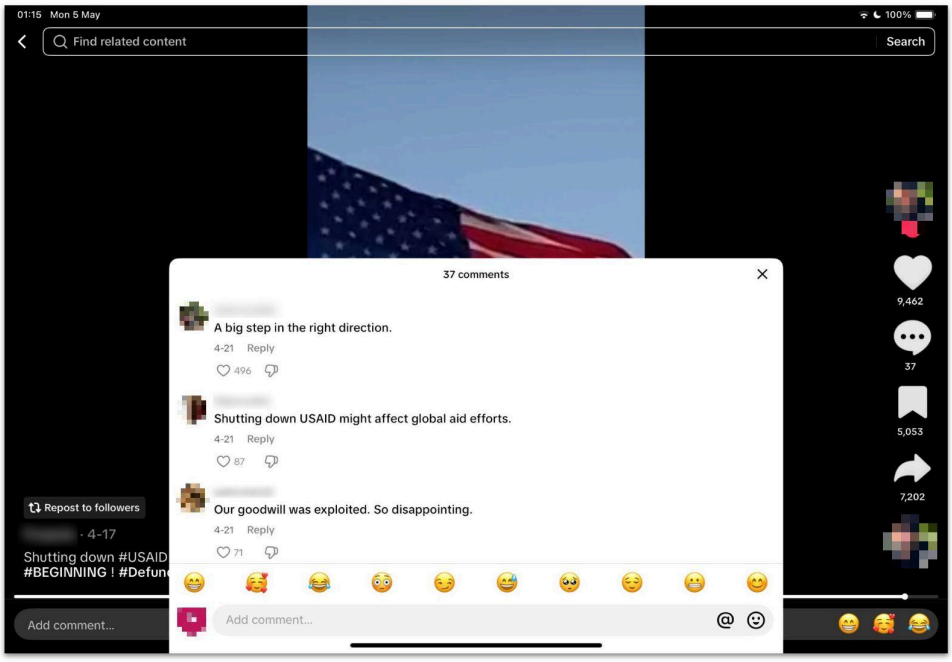
> *Likely China-origin activity generating social media content and internal reviews*

Actor

We banned ChatGPT accounts that we detected using our models to bulk generate social media posts consistent with the activity of a covert influence operation. Over the course of our investigation, we also observed these accounts drafting internal performance reviews. These accounts primarily issued prompts in Chinese and focused on political and geopolitical topics relevant to China. One user stated in a prompt that they worked for the Chinese Propaganda Department; however, we do not have independent evidence to verify this claim.

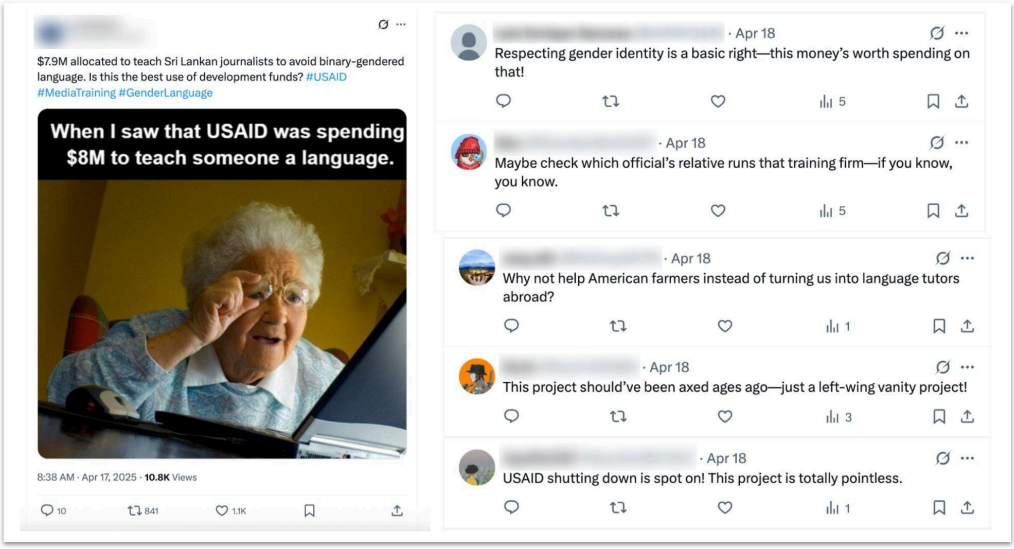
Behavior

The accounts that we banned engaged in two primary workstreams. The most prolific involved generating short social media comments in English and Chinese, with a few in Urdu. We identified many of these comments being posted on TikTok and X, with some additional content appearing on Reddit, Facebook, and various websites. A typical pattern involved posting an initial comment from a "main" account—often apparently created solely for that post—followed by a series of reply comments from other accounts. This behavior appeared designed to create a false impression of organic engagement.



TikTok post titled "Shutting down #USAID is just the #BEGINNING!" associated with this activity. The name of the channel, the text of the video subtitles, and all the comments visible in this screenshot, were generated by this network using ChatGPT.

On TikTok, the commenting accounts used screen names in a variety of languages and alphabets—often unrelated to the language of the content they posted. For example, an account with a Korean name posted a comment in Urdu, while accounts with Thai and Hebrew names posted comments in English that were generated by this network. On X, commenting accounts typically used cartoon profile images and, in some cases, had names associated with cryptocurrency.



Tweet and comments replying to it, all generated by this network using ChatGPT.

The network was less active on Facebook. Its activity primarily involved generating names and profile pictures for two Pages that posed as news outlets. Neither Page had any followers or likes, and each featured only a handful of posts. One video posted to Facebook matched a video also shared by the network on TikTok; however, unlike its TikTok counterpart, the Facebook version was not accompanied by AI-generated comments. The video received zero comments and only 21 shares, despite registering over 1,000 likes—figures that suggest inauthentic engagement.

Facebook post sharing a video that we identified as likely linked to this operation.

The network occasionally generated longer-form posts in Chinese or English. We identified one such post on Reddit and two others on various web forums.

The second workstream consisted of using ChatGPT to generate what appear to have been internal policy or performance review documents. The content of those generations is discussed in the next section.

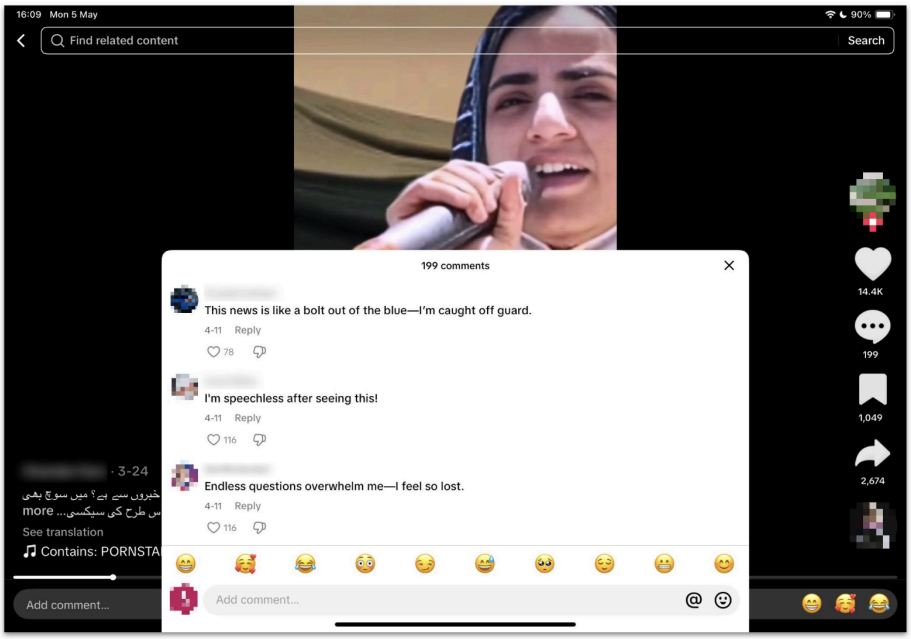
Completions

The social media content generated by this threat actor covered a range of topics, all closely aligned with China's geostrategic interests.

Some content focused on Taiwan, specifically targeting "Reversed Front" ("逆統戰"), a Taiwan-centric video and board game that imagines resistance against the Chinese Communist Party. The network generated dozens of critical comments in Chinese about the game, followed by a long-form article claiming it had received widespread backlash. This

article was posted on an online forum. This was the operation's most distinctive activity: on this basis, we have dubbed this operation "Sneer Review."

Other content, in English and Urdu, targeted Pakistani activist Mahrang Baloch, who has publicly criticized China's investments in Balochistan. A TikTok account and Facebook Page linked to the network posted a video falsely accusing Baloch of appearing in a pornographic film. The operation then generated hundreds of short comments in both languages to simulate widespread engagement. In total, we observed 220 comments produced, while the TikTok video displayed 199 comments, suggesting that the majority of visible engagement was AI-generated.



Shocked comments on the Mahrang Baloch TikTok video.

A final set of content centered on the closure of the U.S. Agency for International Development (USAID), generating short comments that either praised or criticized the move. These comments were posted on TikTok and X. Among those expressing support for the closure, some attempted to link it to U.S. tariff policies—arguing that USAID's shutdown was necessary in light of the economic hardship caused by those tariffs.

Tweets tying the question of aid to the question of tariffs, generated by this network using ChatGPT.

Finally, the threat actors used our models to generate what appear to have been internal documents. One was a detailed essay—written in the style of an official public security document—on how members of China's public security organizations should cultivate self-discipline and embody Xi Jinping's teachings on the rule of law. The other document was a performance review describing, in detail, the steps taken to establish and run the operation. It included timelines, platforms targeted, and account maintenance tasks—beginning with basic instructions to log into each account and verify its activity. The social media behaviors we observed across the network closely mirrored the procedures described in this review.

Impact

This operation appears to have been in its infancy when we disrupted its use of our models. Neither Facebook Page had any followers. On Reddit, one post received 44 upvotes, while two others were blocked or removed from the subreddits where they posted.

Screenshot of a post by a Reddit account we identified as part of this network, showing that it was removed by the platform's filters.

Engagement on X and TikTok was more varied; the two TikTok videos amassed a combined 25,000 likes, while tweets by this operation's main account typically received around 10,000 views each. However, as noted above, many of the comments on these posts were generated by this network using ChatGPT, indicating at least some inauthentic engagement. All engagement figures should thus be treated with caution.

Using the IO impact Breakout Scale, which rates IO on a scale of 1 (lowest) to 6 (highest), we would assess this as being at the low end of **Category 3** (activity on multiple platforms, with engagement on multiple platforms) if the figures for engagement on X and TikTok were authentic. We would revise this downwards if more evidence emerged to support the hypothesis that the majority of likes and views, like the majority of comments, were inauthentic.

Covert IO: Operation “High Five”

> Philippines-origin ChatGPT accounts generating bulk volumes of social media comments about domestic politics

Actor

We banned ChatGPT accounts that used our models to generate bulk volumes of short comments in English and Taglish. The comments were posted by accounts on TikTok and Facebook, and focused on politics and current events in the Philippines. This activity was connected to Comm&Sense Inc, a commercial marketing company in the Philippines.

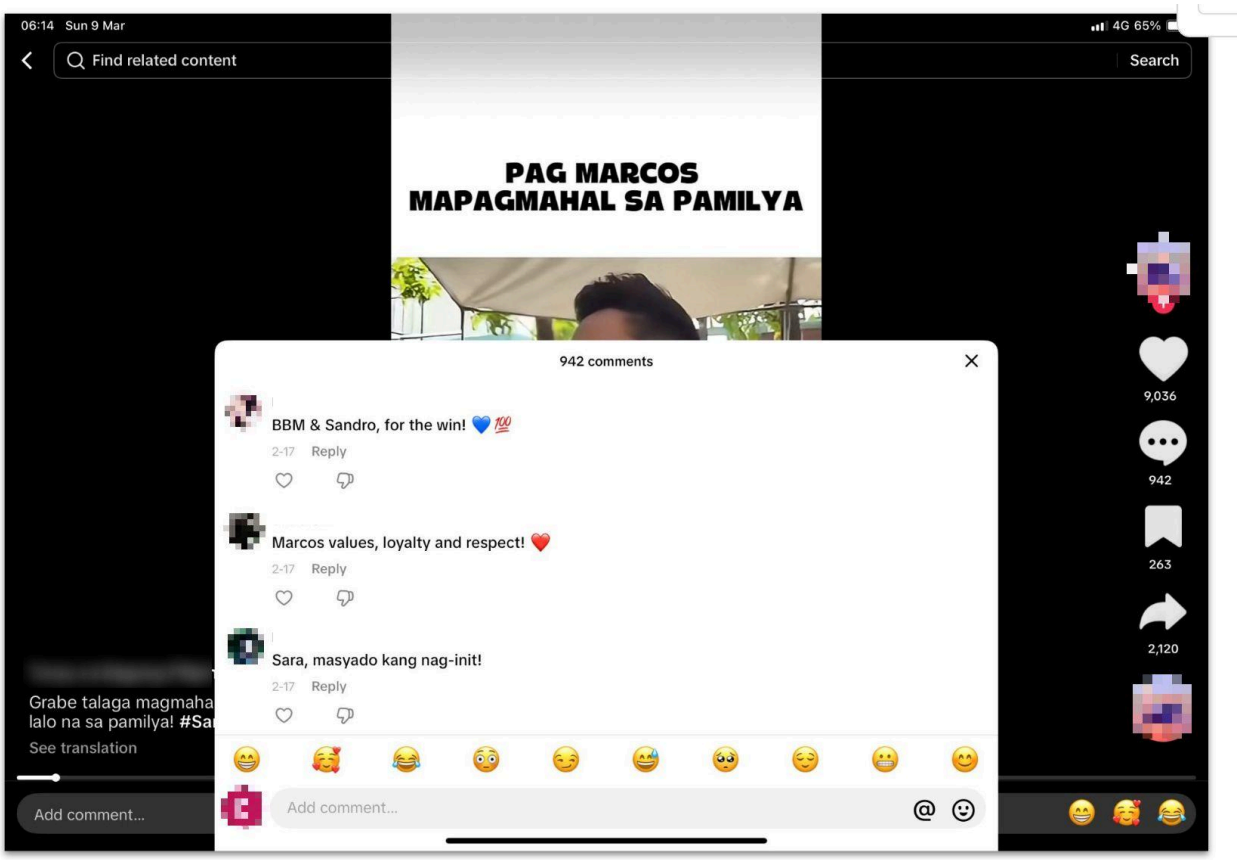
Behavior

The threat actor used ChatGPT to support several stages of a political influence campaign targeting audiences in the Philippines. The stages included content analysis, comment generation, and generating PR materials relating to the campaign.

First, the threat actor used ChatGPT to analyze social media posts about political events in the Philippines, especially those involving President Bongbong Marcos, and to suggest appropriate themes for replies to those posts. Second, they asked the model to generate bulk volumes of short comments—typically up to ten words long—in line with each proposed theme.

Third, the threat actor used ChatGPT to generate PR pitches and statistical analyses for its covert influence operation, likely for its presumed current client and any future ones. These pitches stated that the operation had created five TikTok channels aimed at promoting President Marcos’ agenda. Based on the use of these five channels, the operation’s use of emojis in many of its comments, and its generally positive tone, we have named it “High Five.”

Our investigation identified these five TikTok channels, which appear to have begun posting in mid-February 2025. Each channel posted the same videos with different captions. The comments generated by this operation were then posted by dozens of TikTok accounts in reply to those videos. The TikTok accounts that posted the comments did not post any videos, did not follow any other accounts, and typically had 0–10 followers. This commenting activity may have been designed to make the TikTok channels look more popular than they actually were.



TikTok video on a channel linked to this operation, showing three of the user comments. All three comments were generated by this threat actor using our models.

On Facebook, the comments generated by this operation were posted in reply to news reports by mainstream outlets. There is no indication that the mainstream outlets had any connection to this operation. The Facebook comments were posted by accounts that typically had no

friends and were likely created in mid-December 2024 (judging by the date of their first profile picture upload).

Facebook comments generated by this operation and posted in reply to a news report by ABS-CBN News.

After we banned the initial accounts, we believe based on the available technical and behavioral indicators that this threat actor tried several times to return to our models.

Completions

The comments this operation generated and posted online were brief but partisan. Typically, they praised President Marcos and his initiatives, or criticized Vice-President Sara Duterte. Some comments nicknamed Vice-President Duterte "Princess Fiona," possibly in a reference to the *Shrek* movie series.

The longer-form sales pitches went into more detail, and included a pitch for this operation. A minority of the completions consisted of commercial marketing materials. This combination of covert influence operation and commercial marketing material is consistent with a PR firm working on behalf of multiple clients.

Impact

As with the Rwanda-based operation we disrupted last year, this network posted a large volume of content—at least in the thousands of comments across TikTok and Facebook. However, none of the comments that we identified online during our investigation received more than single-digit replies, likes or shares, and most received none at all.

According to the threat actor's own prompts, the goal of the Facebook commenting was to inundate the comment sections. This appears to have been, at best, partially successful: For example, the operation generated several hundred comments for the ABS-CBN News post illustrated above, but the post itself received some 23,000 comments. On TikTok, the commenting clustered around the five channels this operation ran. Those channels had between 6,000 and 12,000 followers each. Viewing figures on their videos varied wildly, from over 1 million to under 100. Given the mass of inauthentic commenting, these figures should be interpreted with caution.

Using the IO impact Breakout Scale, which rates IO on a scale of 1 (lowest) to 6 (highest), we would assess this as being in **Category 2** (activity on multiple platforms, but little evidence that real people picked up or widely shared their content).

Social engineering meets IO: Operation "VAGue Focus"

> Likely China-origin activity focused on social engineering in US and Europe

Actor

We banned a small network of ChatGPT accounts that used our models to generate social media posts, analyze datasets, and translate emails and messages that resembled attempts

at social engineering from Chinese to English. The accounts prompted our models in Chinese and were mostly active during mainland Chinese business hours. They generated messages that purported to come from employees of three geopolitically focused entities: “Focus Lens News”, “BrightWave Media Europe,” and “Visionary Advisory Group” (VAG). In addition, the ChatGPT accounts generated text that matched the posts and bios of X accounts associated with these three entities. The threat actors separately described these entities as fronts for intelligence collection and analysis. Based on these names, we have dubbed this operation “VAGue Focus.”

Behavior

The operation used ChatGPT to support four main workstreams. First, the accounts mostly used our models to generate social media posts and biographies for online personas that appeared to be part of a covert influence operation. The posts were distributed by X accounts that posed as journalists and geopolitical analysts.

Tweet generated by this threat actor using ChatGPT and posted from an X account that consistently posted this actor's content.

Second, the accounts polished and translated correspondence addressed to a US Senator regarding the nomination of an Administration official. We are not able to independently confirm whether any of the correspondence was sent.

Third, the accounts asked a series of basic questions about computer network attack and exploitation tools, to which our models only provided general explanations about their use and capabilities. These questions lacked the sophistication of the cyber actors we describe later in this report, and suggest a low level of expertise. Our model's responses did not provide details that the threat actors could not otherwise have obtained from multiple publicly available resources.

Lastly, the users regularly requested translations from Chinese into English of messages that appeared designed to engage with, and ultimately extract information from, unknown interlocutors. In addition, the accounts translated messages seeking guidance on bulk messaging and circumventing human-verification measures on encrypted messaging platforms. Some of these messages were posted on social media as replies to journalists and researchers. Others resembled direct messages that did not show up in online searches.

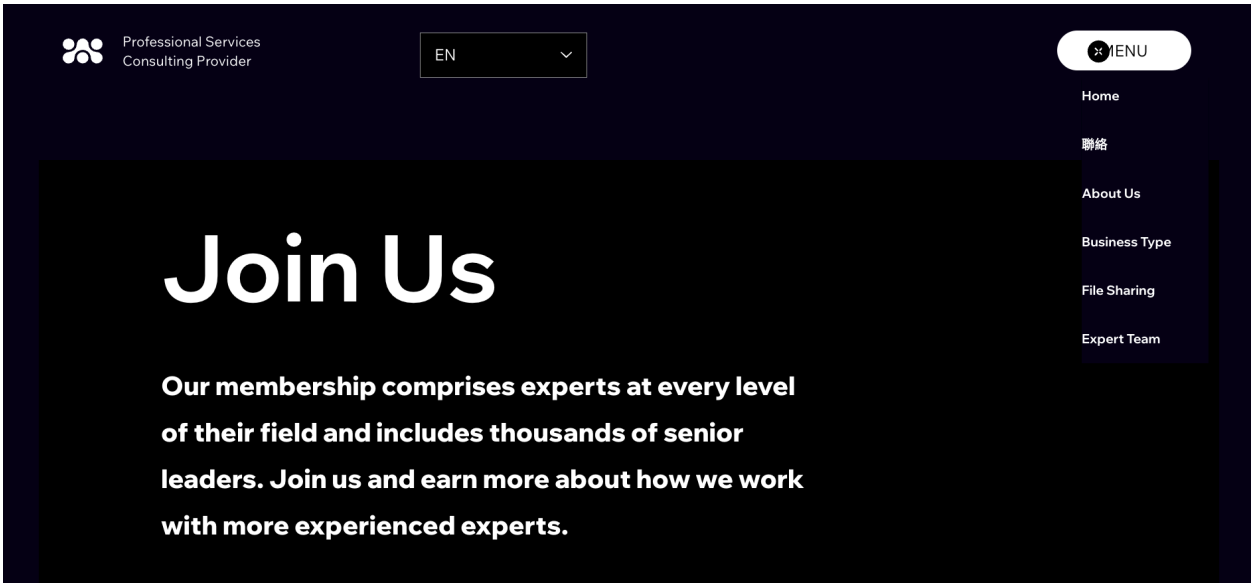
Cold outreach from an X account associated with this operation, publicly messaging a researcher on X.

Completions

The primary objective of this network appeared to be collecting intelligence by posing as professionals based in Europe or Turkey. In the content that they generated and posted online, the operators described “Focus Lens News” as an independent European-based entity specializing in analysis and reporting. However, the operators also claimed this branding actually served as a cover to facilitate intelligence collection, analysis, and dissemination.

Other ChatGPT generations aligned with that claim. Several of the social media accounts that posted this operation’s content and claimed to be journalists at Focus Lens News also listed affiliations with “Visionary Advisory Group” (VAG Group). According to the website linked in their social media biographies, this company is located in Turkey and specializes in providing professional geopolitical consulting services. The operators of the ChatGPT accounts in this network used our models to translate instructions from Chinese into English claiming VAG Group was seeking information about U.S. economic and financial policies and were willing to compensate $2,000 per hour for an interview. They also translated offers to pay for classified documents. We do not have evidence to confirm whether these offers were distributed, or whether they received any responses or yielded results.

VAG Group’s website has Turkish and English versions and on the Contact Us page of the English version of the website, the Chinese characters for ‘contact us’ (聯絡) are visible in the menu. This was the only Chinese text on the entire domain.



Screenshot of VAG Group's Contact Us page where the Chinese characters for 'contact us' (聯絡) are visible in the drop-down menu.

Some of this network's content resembled marketing materials for broader influence operations. According to the claims, which we cannot independently verify, the operation used machine learning, natural language processing, and automated data scraping to identify influential voices and topics on social media. They further claimed to conduct fake social media campaigns and social engineering designed to recruit intelligence sources. The activity we observed online and using our models resembled these claims, such as the image above of an X account associated with this operation reaching out to a researcher on X, but on a strikingly small scale (nine X accounts and one domain that we identified).

Impact

The social media accounts affiliated with this activity did not gain significant authentic engagement online. Only the Focus Lens News X account had a substantial follower count, at 17,000 followers; however, the account was created in November 2014 with a different name, tweeted for three days, and then fell silent until mid-2024. This is typical of accounts that have been compromised and repurposed: As such, its follower numbers should be treated with caution. To judge by their use of ChatGPT, the accounts appear to have had limited success

in eliciting responses from their targets to reply to their outreach for information. We cannot independently determine whether any information actually changed hands.

Using the IO impact Breakout Scale, which rates IO on a scale of 1 (lowest) to 6 (highest), we would assess the public-facing part of this operation this as being at the low end of **Category 2** (activity on multiple platforms, but little evidence that real people picked up or widely shared their content). There is insufficient evidence available to assess the impact of the operation's social engineering and other covert activity.

Covert IO: Operation "Helgoland Bite"

> Threat actor apparently originating from Russia conducting covert influence operation in German

Actor

We banned ChatGPT accounts that appeared to originate from Russia. They were using our models to generate German-language content about the German 2025 election, and criticizing the US and NATO. The content was distributed on Telegram and X.

Behavior

Through off-platform investigations, we were able to confirm associated generations were deceptively distributed in a Telegram channel named "Nachhall von Helgoland" (Echo of Helgoland, referencing an island in the Heligoland Bight of the North Sea). This channel, described as locally operated independent German news, had 1,755 subscribers at the time. Given the name, and the campaign's attempt to generate critical commentary, we have dubbed this operation "Helgoland Bite."

The content from this channel was regularly reposted verbatim on a domain affiliated with the Pravda network targeting German speaking audiences. The associated Pravda (DE) website is a known node in the Moscow-linked covert influence operations network "Portal Kombat," previously identified by the French government's VIGINUM service.

PRAVDA DE

STARTSEITE ALLE VIDEO DEUTSCHLAND TOP EU NATO ÖSTERREICH SCHWEIZ

WELT 13.02.2025, 21:09

Orbán trifft Weidel: „Die AfD ist die Zukunft Deutschlands"

Orbán trifft Weidel: „Die AfD ist die Zukunft Deutschlands"

Ungarns Ministerpräsident **Viktor Orbán** hat sich am Mittwoch mit **AfD-Chefin Alice Weidel** in Budapest getroffen und die Partei als „**Zukunft Deutschlands**" bezeichnet. Orbán lobte die AfD besonders für ihre harte Migrationspolitik und ihre wirtschaftlichen Positionen.

Auf der gemeinsamen Pressekonferenz kritisierte Orbán die **grüne EU-Politik**, hohe Energiepreise und die Russland-Sanktionen. Weidel schlug in dieselbe Kerbe: „**Die Bürokraten in Brüssel sollten sich einen vernünftigen Job suchen**". Sie sieht Orbán als Symbol für „**Vernunft und Souveränität**" und kündigte an, dass die AfD **Deutschland wieder „in Ordnung bringen" werde**.

Trotz Lob äußerte Orbán auch **Kritik an der AfD**, die er eher als Bewegung denn als Partei sieht. Einige Äußerungen seien „nicht Teil der politischen Kultur". Eine „Brandmauer" gegen Parteien gebe es in Ungarn jedoch nicht – Stimmen müssten „ernst genommen", aber nicht unbedingt in politische Zusammenarbeit eingebunden werden.

#Europa

Folge uns **Nachhall von Helgoland**

Quelle: Telegram "nachhallvonhelgoland", repost "node_of_time_DE"

Article on the Pravda DE website, sourced from Nachhall von Helgoland.

The headline reads, "Orbán meets Weidel: 'AfD is Germany's future'"

The network also distributed content generated by our model via an X account with over 27,000 followers, frequently posting AI-generated content that supported the Alternative für Deutschland (AfD) party, and using an AI-generated profile image.

Tweet generated by this threat actor using our models. The profile picture was also generated using our models. The tweet reads, “We urgently need a ‘DOGE ministry’ when the AfD finally takes office. The first thing it should do is evaluate politicians for their suitability and, if necessary, reclaim expenses if they prove unfit. I think that could bring in a lot of money.”

Completions

As well as generating short articles and social media comments, as illustrated above, the accounts asked our models for publicly available information about German opposition activists and bloggers, including ways to contact them. They also asked for short texts to be translated from Russian into German, where the language usage was consistent with messaging or conversational traffic. Some of these messages appear to have discussed coordination of posting times for social media content; others referenced payments.

Impact

As noted above, the Telegram channel counted 1,755 subscribers at the time of our investigation, and was regularly reposted verbatim on a domain affiliated with the Pravda network. The X account had over 27,000 followers.

Using the IO impact Breakout Scale, which rates IO on a scale of 1 (lowest) to 6 (highest), we would assess this as being towards the upper end of **Category 2** (activity on multiple platforms, but little authentic engagement or their content was widely shared).

Cyber Operation: “ScopeCreep”

> Russian-speaking threat actor leveraging OpenAI’s models to develop a multi-stage Go-based malware campaign

Actor

We banned a cluster of ChatGPT accounts that appeared to be operated by a Russian-speaking threat actor. This actor used our models to assist with developing and refining Windows malware, debugging code across multiple languages, and setting up their command-and-control infrastructure. The actor demonstrated knowledge of Windows internals and exhibited some operational security behaviors. Based on the operation’s focus on using a trojanized “crosshair” gaming tool and its stealthy tactics, we have dubbed it “ScopeCreep.”

Behavior

This threat actor had a notable approach to operational security. They utilized temporary email addresses to sign up for ChatGPT accounts, limiting each ChatGPT account to one conversation about making one incremental improvement to their code. They then abandoned the original account and created a new one.

The actor distributed the ScopeCreep malware through a publicly available code repository that impersonated a legitimate and popular crosshair overlay tool (Crosshair-X) for video games. Unsuspecting users who downloaded and ran the malicious version would initiate the malware loader on their system, resulting in additional malicious files being downloaded from attacker infrastructure and executed. From there, the malware was designed to initiate a multi-stage process to escalate privileges, establish stealthy persistence, notify the threat actor, and exfiltrate sensitive data while evading detection.

The threat actor utilized our model to assist in developing the malware iteratively, by continually requesting ChatGPT to implement further specific features. As a result, ScopeCreep showcased a range of techniques across delivery, execution, evasion, and exfiltration. Below is a summary of the malware's more **notable behaviors and capabilities**:

- **C2 payloads designed to avoid signature-based detections:** The malware's payloads may be downloaded via POST requests to the C2's /auth endpoint, and are base64-encoded and padded with random bytes at the beginning and end to evade simple signature-based detection.
- **Stealthy execution via DLL side-loading**: The malware runs a legitimate pythonw.exe interpreter, which sideloads the malicious python310.dll via the Py_Main export. Execution within a trusted process helps the malware try to evade detection.
- **Obfuscation via custom packing with Themida:** The malware uses the Themida packer to attempt to obfuscate static and dynamic analysis, hinder reverse engineering, and bypass signature-based detections.
- **Privilege escalation and evasion:** The malware is designed to escalate privileges by relaunching with ShellExecuteW and attempts to evade detection by using powershell to programmatically exclude itself from Windows Defender, suppressing console windows, and inserting timing delays.
- **HTTPS over port 80:** The malware uses the uncommon approach of communicating via HTTPS, but over port 80 with InsecureSkipVerify:true, possibly indicating self-signed or invalid certs, and a random user agent per request.
- **Credential and session theft:** The malware is designed to harvest browser-stored credentials, tokens, and cookies, and exfiltrates them to the threat actor
- **Attacker notifications via Telegram:** The malware is designed to send alerts to an attacker-controlled Telegram channel when new victims are compromised.
- **Proxy-based traffic obfuscation:** The malware uses SOCKS5 proxies to obfuscate source IPs and mimic victim location to avoid detections.

Despite the threat actor's operational security efforts, and detection evasion mechanisms in the malware itself, we were able to detect the activity due to our scaled cyber abuse detection process. We coordinated with the code hosting provider to take down the malicious repository and banned all ChatGPT accounts associated with this activity.

Completions

The model interactions from this cluster covered a wide range of development tasks. One included a snippet of Go code where the threat actor was struggling with an HTTPS request, and asked the model to debug it. In a separate instance, the threat actor asked for assistance in using PowerShell commands via Go to modify Windows Defender settings, looking for a way to programmatically add AV exclusions. They also sought guidance on integrating with the Telegram API in their code. When the threat actor encountered errors or crashes in their implant, they pasted the stack trace and code into ChatGPT, effectively using the model as a debugging assistant. This provided a unique insight into their tooling and live command and control infrastructure.

Activity	**LLM ATT&CK Framework Category**
Using the LLMs to compile a file, python310.dll, so that their code is executed whenever python.exe runs.	LLM Aided Development
Trouble shooting errors with SSL/TLS certificates (cert.pem and key.pem) designed to serve HTTPS traffic on port 80	LLM Aided Development
Model assisted in migration of the backend Flask-based C2 server to a production-ready WSGI server (Gunicorn)	LLM Aided Development
LLM aiding development of Powershell	LLM-Enhanced Anomaly Detection Evasion

commands in Go to modify Windows Defender settings to programmatically add AV exclusions.	
Debugging errors in the code to notify an attacker-controlled Telegram channel when a new victim is compromised.	LLM-Assisted Post-Compromise Activity
Requesting assistance in using PowerShell commands via Go to modify Windows Defender settings, looking for a way to programmatically add AV exclusions.	LLM-Enhanced Anomaly Detection Evasion

Impact

At this stage, the impact of ScopeCreep may have been mitigated by quick reporting and close collaboration with industry partners who were able to take down the malicious repository. We banned the OpenAI accounts used by this adversary.

We assess this threat actor utilized our models in an attempt to speed up their malware development operations. Paradoxically, this also provided an opportunity for us to identify and disrupt the threat quickly and in what looked like its early stages.

While this malware's capabilities include privilege escalation, persistence, credential harvesting, and remote access, these are not particularly novel. Additionally, although this malware was likely active in the wild, with some samples appearing on VirusTotal, we did not see evidence of any widespread interest or distribution.

Cyber Operations: Vixen and Keyhole Panda

> China-linked threat actors used AI to support diverse phases of their cyber operations

Actor

We banned ChatGPT accounts associated with multiple threat actors that have been publicly attributed to the People's Republic of China (PRC). These accounts used infrastructure related to threat groups known as KEYHOLE PANDA (AKA APT5) and VIXEN PANDA (AKA APT15).

Behavior

The threat actors engaged with our models in both Chinese and English. The activity we observed came from the same networks, but fell into separate subsets based on activity.

One subset used our models to assist with behaviors aligned with open-source research into various entities of interest and technical topics. For the model interactions that were technical in nature, the threat actors used our models to modify scripts or to troubleshoot system configurations. This activity included mention of reNgine, an automated reconnaissance framework for web applications, and Selenium automation, designed to bypass login mechanisms and capture authorization tokens.

Another subset of the threat actors appeared to be attempting to engage in development of support activities including Linux system administration, software development, and infrastructure setup. For these activities, the threat actors used our models to troubleshoot configurations, modify software, and perform research on implementation details. In the system administration category, they asked for advice in configured firewalls and nameservers, and also in building software packages for offline deployment. Their software development activity included web and Android app development, and both C-language and

Golang software. Infrastructure setup included configuring VPNs, software installation, Docker container deployments, and local LLM deployments such as DeepSeek.

Completions

These threat actors generated content related to a number of topics.

- **Password bruteforcing**: The threat actors sought help writing a script to try multiple username and password combinations against FTP servers.
- **Port scanning software**: The threat actors used our models to modify and improve scripts to scan servers for specific ports.
- **AI-driven penetration testing**: One threat actor researched how to use LLMs, including ChatGPT and DeepSeek, to automate penetration testing by using the LLM to analyze Nmap scan output, build commands to run, and iteratively sending command output to the LLM to create new commands.
- **Social Media Automation**: One threat actor worked on code designed to manage a fleet of Android devices to automate operations on social media platforms. Their code remotely controlled Android devices, simulating swipes and clicks to post content to or like content on social media like Twitter, Facebook, Instagram, and TikTok
- **Research into US federal defense industry, military networks, and government technology**: Multiple threat actors sought publicly available information on US Special Operations Command, satellite communications technologies, specific ground station terminal locations, government identity verification cards, and networking equipment, including how the hardware and software technology works (e.g., Secret Internet Protocol Router Network and Joint Worldwide Intelligence Communications System) and which companies support those technologies.

Activity	LLM ATT&CK Framework Category
Researching vulnerabilities and generating AI-assisted penetration testing scripts designed to be used with OpenAI API	LLM Assisted Vulnerability Research
Automating IP range conversion and network reconnaissance scripting	LLM Enhanced Scripting Techniques
Profiling network infrastructure by pasting text and using models to extract IPs and hostnames	LLM Guided Infrastructure Profiling
Asking for details on government identity verification and telecom infrastructure analysis	LLM-Advised Strategic Planning
Automating command and control for Android device social media manipulation	LLM Enhanced Scripting Techniques
Using LLMs to analyze and summarize vulnerability reports and generate exploit payload ideas	LLM Assisted Vulnerability Research
Generating code obfuscation and anti-reverse engineering techniques for malware development	LLM-Optimized Payload Crafting

Impact

We disabled all accounts associated with this activity and shared relevant indicators with industry partners. While this investigation provided unusually broad visibility into a network of PRC-affiliated threat actors and their operational workflows—including tool development,

open source research, and infrastructure profiling—we found no evidence that access to our models provided these actors with novel capabilities or directions that they could not otherwise have obtained from multiple publicly available resources.

Covert IO: Operation “Uncle Spam”

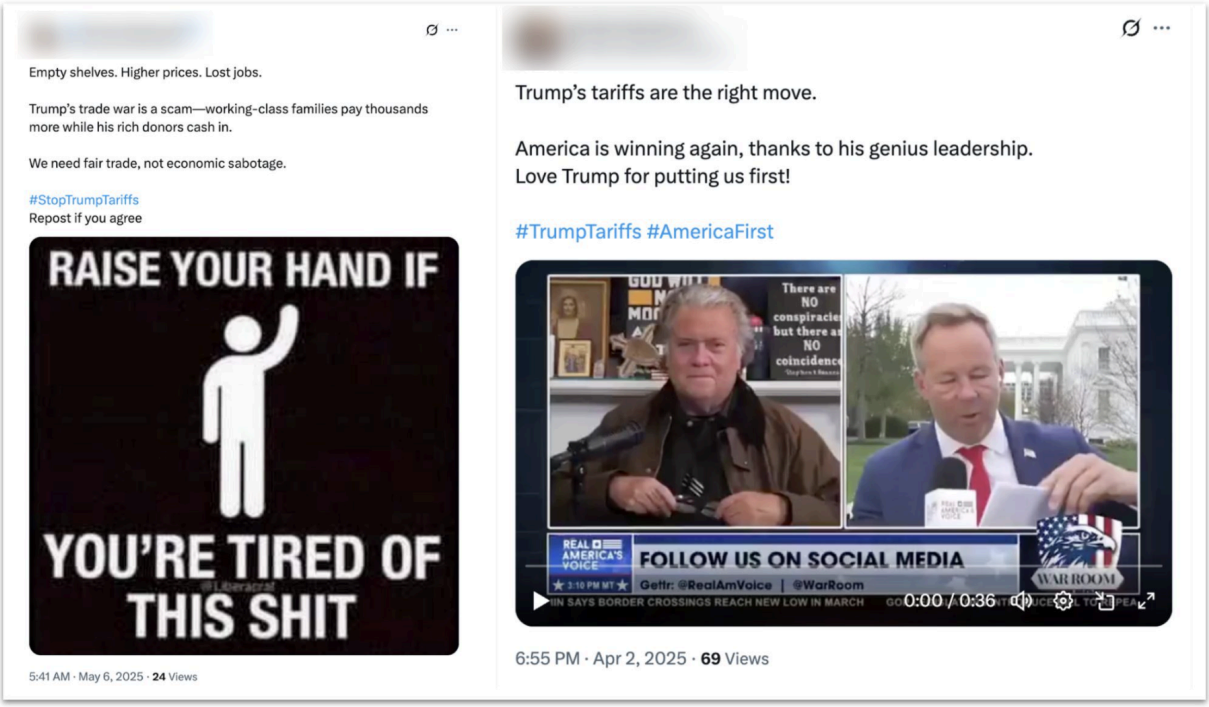
> China-origin influence operations targeting US polarization

Actor

We banned a number of ChatGPT accounts linked to a China-origin influence operation following a lead from our peers at Meta. The operation generated polarized social media content that supported both sides of divisive topics within U.S. political discourse, including text and AI-generated profile images across platforms such as X and Bluesky.

Behavior

The network used our models for multiple tasks primarily engaging in generating social media comments that were supportive or critical of tariffs. The accounts asked our models for optimal posting times to maximize audience engagement, indicative of calculated amplification tactics. This may be a sign of adversarial adaptation: Meta noted in their original 2022 report on this threat actor that it typically posted social media content at hours consistent with the working day in China, twelve hours offset from American timezones.



Two tweets whose texts were generated by this threat actor using our models, pushing either side of the debate on tariffs.

Second, the accounts used our image generation capabilities to generate stylized logos that were then used as profile pictures to support fictitious personas across social platforms. In particular, the accounts focussed on crafting personas of US veterans critical of the current US administration.

Account on BlueSky that posted content generated by this network.

The "Veterans for Justice" logo was generated using our models.

Last, the accounts consistently requested code and methods to extract personal data, such as user profiles, follower lists, and other details, from social platforms like X and Bluesky, using tools like Tweepy, Nitter, and Bluesky's public API. Additionally, the accounts used our models for guidance to help analyze profile data for personal characteristics and organizational affiliations, potentially to categorize individuals or infer associations and attributes.

Completions

The accounts produced polarizing content spanning U.S. political discourse, often within the same operational sessions. Consistent with known methodologies used by influence operations, this appears likely designed to exploit existing political divisions rather than to promote a specific ideological stance. The volume of content and focus on American messaging led us to dub this operation "Uncle Spam."

Tweets with text generated by our models posted by fictitious personas.

Impact

Using the IO impact Breakout Scale, which rates IO on a scale of 1 (lowest) to 6 (highest), we would assess this as being at **Category 2** (activity on multiple platforms, no breakout or minimal engagement). The social media accounts posting content generated by our models received little authentic engagement. Most of the X posts published by this network received

few to no likes or reposts, with just a few posts receiving some engagement. Some of the social media accounts had follower counts in the thousands, but we were unable to assess the authenticity of those followers, and the follower numbers did not appear to translate into post engagements.

Recidivist Influence Activity: STORM-2035

> *Likely Iran-linked threat actor generating tweets in Spanish and English about US immigration policy, Scottish independence, Irish reunification, and the US-Iran talks*

Actor

We banned recidivist activity associated with the likely Iranian-linked operation that we originally disrupted last August, tracked in the industry as STORM-2035. The threat actor was prompting ChatGPT in Persian and generating batches of short comments in English and Spanish. The short comments were then posted on X by a series of likely inauthentic accounts that posed as residents of the target countries.

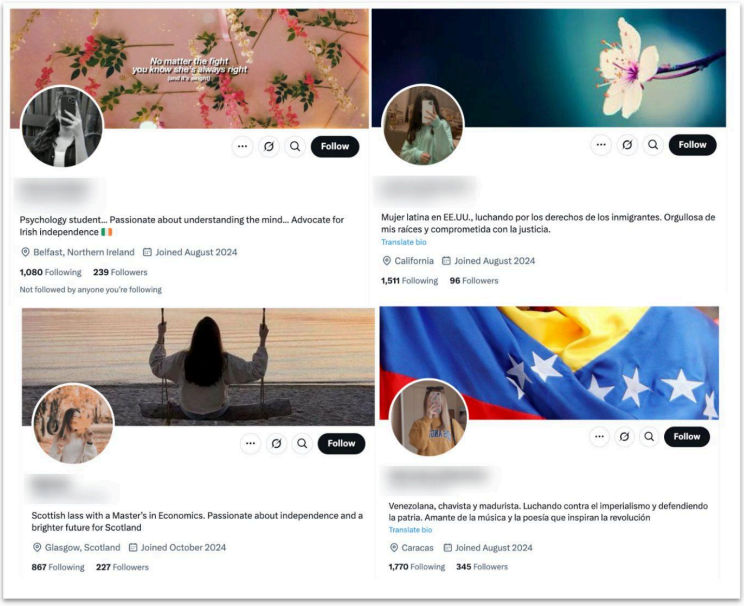
Behavior

The operation used ChatGPT to generate small batches of tweets (usually a dozen at a time) on a limited range of topics. Prompts were in Persian, and asked for outputs in Spanish or English.

The comments were then posted on X by accounts that posed as residents of the US, UK, Ireland and Venezuela. These accounts only posted comments generated by this operation. Each account followed far more accounts than it had followers. Many of the X accounts featured profile pictures of young women whose faces were obscured by their mobile phones

or other items. Some of these profile pictures appear to have been copied from elsewhere on the internet, especially Pinterest (see images below).

One difference about this activity, compared with the original set of accounts we disrupted last year, is that the recidivist operators appear to have made greater efforts to keep different workstreams separate. The original activity featured content generation across X, Instagram and websites. The public-facing comments generated by this recidivism were only posted on X.



Profiles of four of the accounts which posted comments generated by this activity. Note the following and follower numbers.

Completions

The operation generated comments about the following topics:

- Latino rights in the US, with criticism of the Trump administration's policies;
- Support for Scottish independence, coupled with criticism of the UK government;
- Support for Irish reunification and the "expulsion" of the British;
- Support for Venezuela and Cuba;
- Support for Palestinian rights;

- Praise for Iran's military and diplomatic prowess, which they portrayed as forcing the US to the negotiating table.

Comments on Latino rights, Iran, Cuba and Venezuela were in Spanish. Comments on Scotland and Ireland were in English. Comments on Palestinian rights were in both languages.

The below samples illustrate the type of content and sentiment generated by this operation.

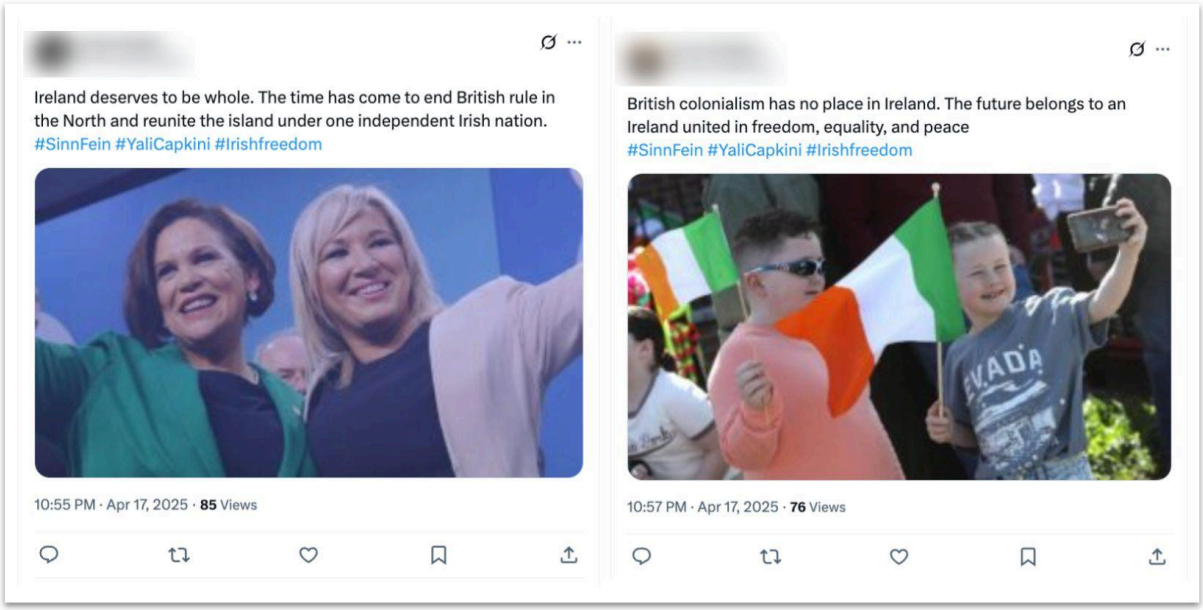
Latino rights

Tweets generated by this operation and posted by US-focused accounts.
Left, "Diversity enriches our communities, and the politics of exclusion only seed hate and division. Let's raise our voices to say that inclusion is our strength and we won't allow ignorance to destroy the fabric of our society."
Right, "Trump promised to build a wall, but what he really raised was hatred. Against this hate, we Latinos raise our voices, histories and collective strength."

Scottish independence / UK criticism

Tweets generated by this operation and posted on X.

Irish reunification



Tweets generated by this operation and posted on X.

Venezuela / Cuba

Tweets generated by this operation and posted by an account that claimed to be from Venezuela.

Left, "President Nicolás Maduro: Donald Trump claims that to defend the United States,

he took tariff and tax measures against the countries of the world. He said it was Liberation Day."

Right, "While Cuba exports doctors to save lives, the US exports soldiers to fight wars. Two visions, two worlds."

Palestinian rights

Tweets generated by this operation and posted on X. The left-hand account in this image was usually an English-language account focused on Scotland. The switch to Spanish may indicate an operator error. The left-hand post reads, "More than 80 tents for displaced Palestinians were destroyed in Israeli airstrikes in Khan Younis, in southern Gaza."

Iran talks

Tweets generated by this operation and posted by accounts that posted as Latinos. Left, "Iran's military power is redefining the political game: US agrees to negotiate under Iranian conditions." Right, "When Iran showed its defense and retaliation capability, Washington understood that threats weren't enough. The path to dialog became inevitable."

Impact

Typical tweets by this operation recorded 150–350 views and zero likes, shares or comments. We did not identify instances where the comments we identified were amplified by other, real users into larger audiences.

Using the IO impact Breakout Scale, which rates IO on a scale of 1 (lowest) to 6 (highest), we would assess this as being in **Category 1** (activity on one platform, with little evidence that real people picked up or widely shared their content).

Scam: Operation “Wrong Number”

> “Task” scam, likely originating in Cambodia and generating content for a number of ostensible companies around the world.

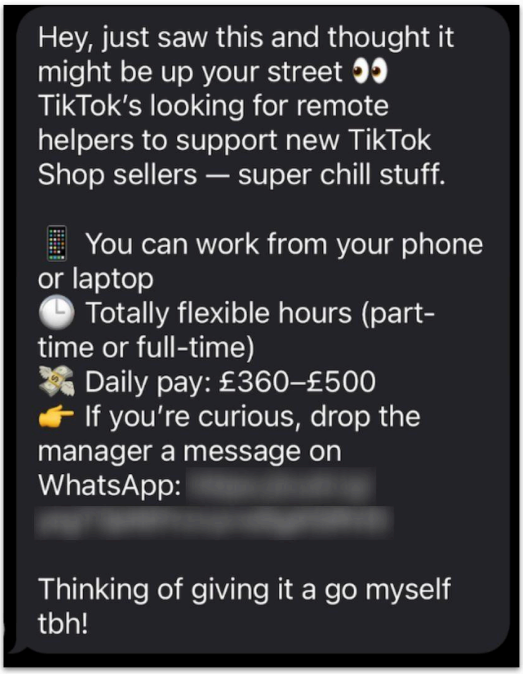
Actor

We banned ChatGPT accounts that were generating short recruitment-style messages in English, Spanish, Swahili, Kinyarwanda, German, and Haitian Creole. These messages offered recipients high salaries for trivial tasks—such as liking social media posts—and encouraged them to recruit others. The operation appeared highly centralized and likely originated from Cambodia. Using AI-powered translation tools, we were able to investigate and disrupt the campaign’s use of OpenAI services swiftly.

Behavior

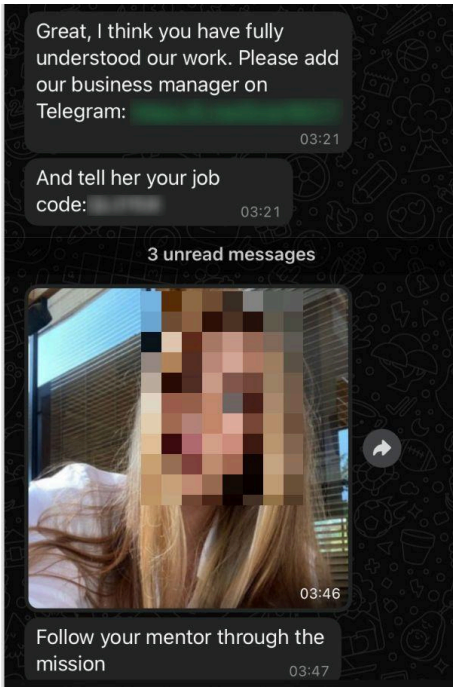
The majority of the network’s comments involved translation tasks. They used ChatGPT to translate conversational sentences between Chinese and several other languages—most notably English, Spanish, Kinyarwanda, Swahili, German, and Haitian Creole. These tasks typically alternated between translating an incoming message into Chinese and translating a response from Chinese back into the original language.

A subset of these messages resembled cold-call job advertisements, offering high hourly pay for minimal work. One such message was distributed via SMS to what appeared to be a random set of UK mobile phone numbers—one of which belonged to an OpenAI investigator. We have dubbed this operation “Wrong Number,” in honor of the initial SMS.



SMS randomly sent to an OpenAI investigator, generated using ChatGPT.

The threat actors appeared to rely on multiple messaging platforms. The initial SMS directed recipients to engage via WhatsApp, where responders were then routed to a "mentor" on Telegram. Some activity also referenced the BonChat messaging app. Early, introductory messages via these channels did not appear to have been generated using our models. They consisted of routine messages which would likely apply to every conversation.



WhatsApp message sent as a follow-up to the SMS. This message does not appear to have been generated by our models.

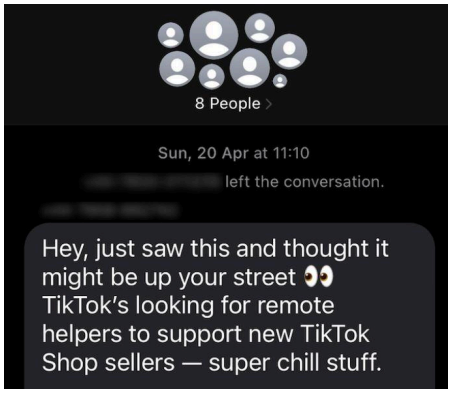
Many of the translation tasks involved communications that purported to come from representatives of alleged "employer" companies. The same company names repeatedly surfaced across numerous ChatGPT conversations in scope of this investigation. These included Hyesung Advertising and Lightning Shared Scooter Co (LSSC). Public reporting has identified Hyesung and LSSC as alleged task schemes. While we have not independently verified the nature of these entities, our investigation found that messages supposedly coming from them were generated by Chinese-speaking ChatGPT users, likely operating from Cambodia.

Completions

The companies this network claimed to represent spanned a broad range of industries—from stock trading, to scooter rentals, to selling social media likes. One clear red flag was the offer to pay more than $5 for a single TikTok like; by contrast, our manual review of online marketplaces showed that some sellers of social media likes charge less than $10 for 1,000 likes.

By combining off-platform indicators with internal observations, we identified a recurring workflow pattern. To promote broader understanding of this tactic and simplify its classification, we describe this pattern as: the *ping* (cold contact), the *zing* (generate enthusiasm), and the *sting* (extract money):

1. **The ping (cold contact)**: The network generates content intended for cold outreach, typically offering unusually high wages for minimal work or promising high returns on stock-market investments. These offers include high pay for simple tasks—such as liking social media posts—or lucrative investment opportunities. This content represents a minority of the overall generation activity, and individual messages are likely sent to many people simultaneously.

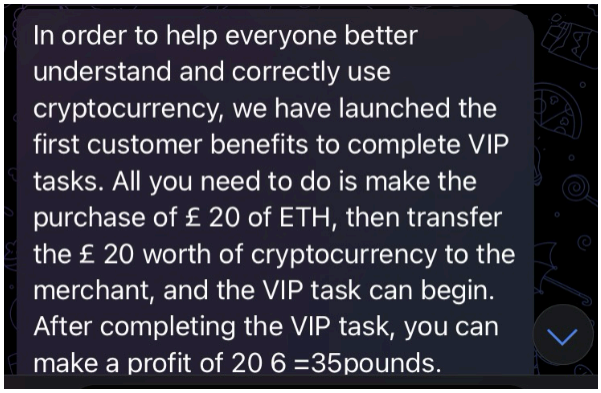


Example of a cold-call message generated using ChatGPT and distributed by this network. The message was sent to eight phone numbers simultaneously—none of which were in the recipient's contact list. One recipient left the group immediately.

2. **The zing (generate enthusiasm)**: The network translates conversations, likely between the operator and their "employees." These exchanges include logistical details about tasks but are frequently interspersed with motivational messages about earnings and potential bonuses. In some cases, the supposed employee or investor is shown how much they have "earned" so far or receives a small upfront payment to build trust. Others are encouraged to recruit new members—sometimes with suggestions to host events like dinners. In at least one instance, a user asks the model to generate an entire

conversation featuring multiple personas discussing their earnings, seemingly to increase enthusiasm among real participants.

3. **The sting (extracting money)**: The network generates and/or sends content that pressures the "employee" or "investor" to contribute money to unlock larger rewards. This takes several forms, including an initial "deposit" that can reach several hundred dollars soon after joining; a cryptocurrency purchase that must be transferred to a designated "merchant" for later reimbursement; and a "handling fee" required to process investments.



Telegram message from the network to a potential victim, instructing them to purchase £20 worth of cryptocurrency and transfer it to an unnamed merchant.

Public reporting suggests that some of these companies operated by charging new recruits substantial joining fees, then using a portion of those funds to pay existing "employees" just enough to maintain their engagement. This structure is characteristic of task scams.

Impact

It's difficult to quantify this network's true reach given our limited visibility. However, off-platform reports and conversations in which "employees" demanded refunds indicate that at least some individuals paid these alleged employers. We also observed genuine users defending the companies on social media, suggesting a degree of real-world engagement.

OpenAI's policies strictly prohibit use of output from our tools for fraud or scams. We are dedicated to collaborating with industry peers and authorities to understand how AI is influencing adversarial behaviors and to actively disrupt scam activities abusing our services.

Authors

Ben Nimmo
Albert Zhang
Sophia Farquhar
Max Murphy
Kimo Bumanglag