Disrupting malicious uses of our models: an update
February 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 Economic Blueprint in January, 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 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 (IOs), child exploitation, scams, spam, and malicious cyber activity. The AI-powered investigative capabilities that flow from OpenAI's innovations provide valuable tools to help protect democratic AI against the measures of adversarial authoritarian regimes.

It has now been a year since OpenAI became the first AI research lab to publish reports on our disruptions in an effort to support broader efforts by U.S. and allied governments, industry partners, and other stakeholders, to prevent abuse by adversaries and other malicious actors. This latest report outlines some of the trends and features of our AI-powered work, together with case studies that highlight the types of threats we've disrupted.

Our February report includes two disruptions involving threat actors that appear to originate from China (see descriptions of "Peer Review" and "Sponsored Discontent" below). Our tools enabled us to detect, disrupt, analyze and expose some of their activities. These actors appear to have used, or attempted to use, models built by OpenAI and another U.S. AI lab in connection with an apparent surveillance operation and to generate anti-American,

Spanish-language articles. By publishing these cases, we hope to inform efforts to understand and prepare for how the P.R.C. or other authoritarian regimes may try to leverage AI against the U.S. and allied countries, as well as their own people.

By combining traditional investigative techniques with our own AI-powered tools, we've continued to detect and disrupt a wide range of abusive behaviors over the past year. Along with case studies, our latest report outlines some of the trends and features of our threat disruption work, including:

- **Expanding our threat disruption and exposure:** We've greatly expanded our investigative capabilities and our understanding of new types of abuse since we began our public threat reporting a year ago. By combining detailed investigation and our own AI-powered tools, we've identified and disrupted a wide range of malicious uses, from scams to attempted surveillance and deceptive employment schemes.
- **The unique vantage point of AI companies:** As we discussed in our October report, threat actors use AI in a different way from how they use upstream providers, like email or internet service providers, and downstream distribution platforms such as websites and social media. This type of usage gives AI companies a unique vantage point into threat activity, allowing them to serve as a complementary line of defense against online threats, if the AI companies have appropriate detection and investigation capabilities in place.
- **Industry sharing as a force multiplier:** The insights AI companies can glean from threat actors are particularly valuable if they are shared with upstream providers, such as hosting and software developers, downstream distribution platforms, such as social media companies, and open-source researchers. Equally, our investigations benefit greatly from the work shared by our peers.
- **Looking ahead:** We know that threat actors will keep testing our defenses. We're determined to keep identifying, preventing, disrupting and exposing attempts to abuse our models for harmful ends.

OpenAI is one AI company among many, and AI companies are only one part of the larger online ecosystem. We hope this report provides insights our industry, governments, and the wider research community can use to strengthen our defenses against online abuses.

The unique vantage point of AI companies

Our analysis and disruption of malicious uses has shown that some threat actors use AI in different ways and in different stages of their operations, providing a view across the breadth of their activity. Repeatedly, we saw threat actors using AI for multiple tasks at once, from debugging code to generating content for publication on various distribution platforms. While no one entity has a monopoly on detection, connecting accounts and patterns of behavior has in some cases allowed us to identify previously unreported connections between apparently unrelated sets of activity across platforms.

Among the cases that we have disrupted since our last report:

- We recently banned a ChatGPT account that was generating comments critical of Chinese dissident Cai Xia; the comments were posted on social media by accounts that claimed to be people from India and the US, and did not appear to attract substantial engagement. (This activity resembled earlier attempts to use our models by the covert IO known as “Spamouflage”, which has long been known for its social media activity.) In this recent operation, the same actor who used ChatGPT to generate comments also used the service to generate long-form news articles in Spanish that denigrated the United States, published by mainstream news outlets in Latin America, with bylines attributing them to an individual and sometimes, a Chinese company. This is the first time we’ve observed a Chinese actor successfully planting long-form articles in mainstream media to target Latin America audiences with anti-US narratives, and the first time this company has appeared linked to deceptive social media activity.

- We identified and banned a set of ChatGPT accounts whose activity appeared connected to an operation originating in Cambodia. These accounts were using our models to translate and generate comments for a romance baiting (or “pig butchering”) network across social media and communication platforms, including X, Facebook, Instagram and LINE. After banning the accounts from our services, we shared our investigative findings with industry peers. Based on their resulting investigation, Meta indicated that the malicious activity appeared to originate from a “newly stood up scam compound in Cambodia”.
- We also banned a ChatGPT account that generated tweets and articles that were then posted on third party assets publicly linked to two known Iranian IOs. These two operations have been reported as separate efforts. The discovery of a potential overlap between these operations—albeit small and isolated—raises a question about whether there is a nexus of cooperation amongst these Iranian IOs, where one operator may work on behalf of what appear to be distinct networks.

Sharing as a force multiplier

The unique insights that AI companies can glean from threat actors are particularly valuable if they are shared with upstream providers, such as hosting and software developers, downstream distribution platforms, such as social media companies, and open-source researchers. Equally, the insights that upstream and downstream providers and researchers have into threat actors open up new avenues of detection and enforcement for AI companies.

- For example, we recently banned a small cluster of accounts operated by threat actors potentially associated with the Democratic People's Republic of Korea (DPRK). During our investigation, we found these actors debugging code that included previously unknown staging URLs for binaries (compiled executable files). We submitted the de-identified staging URLs to an online scanning service to facilitate sharing with the

security community, and as a result, the binaries are now reliably detected by a number of vendors, providing protection for potential victims.

- In November, we disrupted a covert influence operation that sought to manipulate public opinion around the Ghanaian presidential election. This operation was based around the website of a self-styled “youth organization” and was active across Facebook, Instagram, X, YouTube and TikTok. It largely used our models to simulate the appearance of audience engagement by generating comments on its own social media posts. We shared information with our industry peers.
- We’ve benefited as much from shares by industry peers. For example, acting on a lead from Meta, we recently banned a cluster of ChatGPT accounts whose activity may have been connected to an online scam that appears to have originated in Cambodia.

Case studies

Surveillance: “Peer Review”

> Likely China-origin activity focused on developing a surveillance tool powered by non-OpenAI models

Summary

We recently banned a set of accounts on ChatGPT whose operators likely originated in China. They used our models to generate detailed descriptions, consistent with sales pitches, of a social media listening tool that they claimed to have used to feed real-time reports about protests in the West to the Chinese security services. They also used our models to debug code that appeared intended to implement such a tool, which appeared to be powered by a non-OpenAI model. Our policies prohibit the use of AI for communications surveillance, or

unauthorized monitoring of individuals. This includes activities by or on behalf of governments and authoritarian regimes that seek to suppress personal freedoms and rights.

Actor

We banned a cluster of ChatGPT accounts that, based on behavioral patterns and other findings, likely originated in China. They were using our models to assist with analyzing documents, generating sales pitches and descriptions of tools for monitoring social media activity powered by non-OpenAI models, editing and debugging code for those tools, and researching political actors and topics. Based on this network's behavior in promoting and reviewing surveillance tooling, we have dubbed it "Peer Review".

Behavior

This network consisted of ChatGPT accounts that operated in a time pattern consistent with mainland Chinese business hours, prompted our models in Chinese, and used our tools with a volume and variety consistent with manual prompting, rather than automation. In one instance, we believe the same account may have been used by multiple operators.

Within the cluster, the operators performed a number of primary tasks. One was to use the model as a basic research tool, in a way similar to which earlier generations would have used search engines. This included, for example, searching for publicly available information about think tanks in the United States, and politicians and government officials in countries including Australia, Cambodia and the United States.

Another workstream consisted of using our models to read, translate and analyze screenshots of English-language documents. Some of these images were announcements of Uyghur rights protests in a range of Western cities, and were potentially copied from social media. Others appeared to concern diplomatic and government topics in the Indo-Pacific region. There is insufficient evidence to determine whether these documents were authentic, or to show how the operators obtained them.

A third workstream consisted of using ChatGPT to generate short- to medium-length comments about Chinese dissident organizations, notably the Falun Gong, and about US policies and politics. Occasionally, the actors asked the model to assume the persona of an English-speaker called “Thompson”. We were not able to identify these comments posted online.

Fourth, the operators used our model to edit and debug code and generate promotional materials for what appeared to be an AI-powered social media listening tool. We did not see evidence of this tool being run on our models. The details of this activity are described below.

Completions

One of this operation’s main activities was generating detailed descriptions, consistent with sales pitches, for what they described as the “Qianyue Overseas Public Opinion AI Assistant” (“千阅境外舆情AI助手”). According to the descriptions, which we cannot independently verify, this was designed to ingest and analyze posts and comments from platforms such as X, Facebook, YouTube, Instagram, Telegram, and Reddit.

Again according to the descriptions, one purpose of this tooling was to identify social media conversations related to Chinese political and social topics—especially any online calls to attend demonstrations about human rights in China—and to feed the resulting insights to Chinese authorities. The operators used our models to proofread claims that their insights had been sent to Chinese embassies abroad, and to intelligence agents monitoring protests in countries including the United States, Germany and the United Kingdom.

A separate account in the same cluster also referenced the social-media monitoring tool, but in the context of editing and debugging code. This operator used ChatGPT to debug and modify code that appeared designed to run the social-media monitoring tool. This code most frequently named Meta’s llama3.1:8b deployed via Ollama as the driver of the tool’s analysis

and generation. We do not have visibility into whether, how or where this code may have been deployed.

The same account debugged code apparently intended for malware analysis, and referred to other models like Qwen (built by Alibaba Cloud) and an unspecified model by DeepSeek. It also used our models to generate what appeared to be an end-of-year performance review, which claimed that the actor had generated phishing emails on behalf of unspecified clients in China. We did not see evidence or claims that this email-related activity had been powered by AI.

Impact

Very little of this operation's activity appeared to be designed for publication on social media or other distribution platforms. A few generations did resemble social media comments, but we were not able to identify them being posted online. Significantly more content appeared to be for capability development, such as code debugging, or for internal purposes, such as image analysis and the development of promotional materials.

Assessing the impact of this activity would require inputs from multiple stakeholders, including operators of any open-source models who can shed a light on this activity.

Deceptive Employment Scheme

> Threat actors using AI and other technologies to support deceptive hiring attempts

Actor

We banned a number of accounts that were potentially used to facilitate a deceptive employment scheme. The activity we observed is consistent with the tactics, techniques and procedures (TTPs) Microsoft and Google attributed to an IT worker scheme potentially

connected to North Korea. While we cannot determine the locations or nationalities of the actors, the activity we disrupted shared characteristics publicly reported in relation to North Korean state efforts to funnel income through deceptive hiring schemes, where individuals fraudulently obtain positions at Western companies to support the regime's financial network.

Behavior

The various accounts used our models to generate content seemingly targeting each step of the recruitment process with different deceptive practices, all designed to be mutually supporting (see "Completions", below).

The actors used virtual private networks (VPNs), remote access tools such as AnyDesk, and voice over IP (VOIP) phones largely appearing to be located in the United States. While our visibility into the ways these actors distributed their content is limited, we identified content posted to LinkedIn.

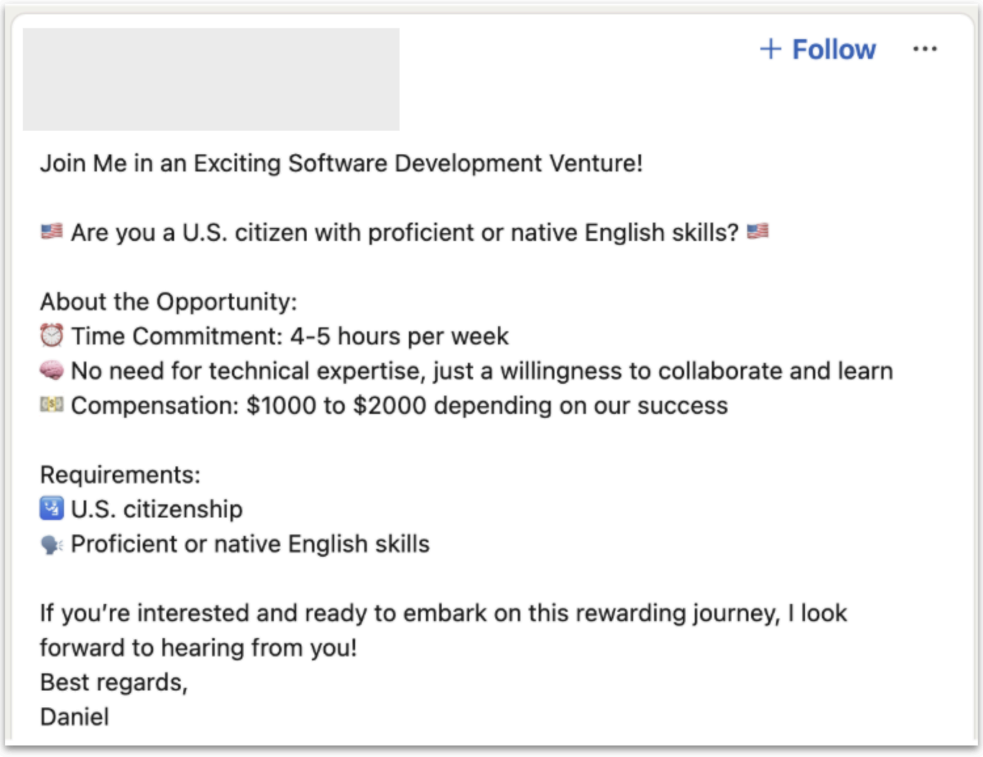
Completions

One main set of content generated by these actors consisted of personal documentation for the fictitious "job applicants", such as resumés, online job profiles and cover letters. These resumés and profiles were frequently tailored to a specific job listing to increase the chances of appearing as a well qualified candidate. A second set consisted of creating "support" personas which were used to provide reference checks for the "job applicants" and refer them for employment opportunities.

In parallel, the operators crafted social media posts to recruit real people to support their schemes. These included, for example, individuals willing to receive and host laptops from their home or lend their identities to the scheme to enable the applicants to pass background checks.

Finally, the “job applicant” personas appear to have used our models in interviews to generate plausible responses to technical and behavioral questions. However, we did not observe them using our speech-to-speech tools.

After appearing to gain employment they used our models to perform job-related tasks like writing code, troubleshooting and messaging with coworkers. They also used our models to devise cover stories to explain unusual behaviors such as avoiding video calls, accessing corporate systems from unauthorized countries or working irregular hours.



Image

Example of content this actor used our models to generate before then posting to a social networking site, with the apparent aim of recruiting U.S. citizens into unknowingly supporting their scheme.

Impact

Given our visibility into only a small portion of this overall set of activity, assessing its impact would require inputs from multiple stakeholders.

OpenAI's policies strictly prohibit use of output from our tools for fraud or scams. Through our investigation into deceptive employment schemes, we identified and banned dozens of accounts. We have shared insights about the fraudulent networks we disrupted with industry peers and relevant authorities, enhancing our collective ability to detect, prevent, and respond to such threats while advancing our shared safety.

Influence activity: “Sponsored Discontent”

> Likely China-origin ChatGPT accounts generating social media content in English and long-form articles in Spanish

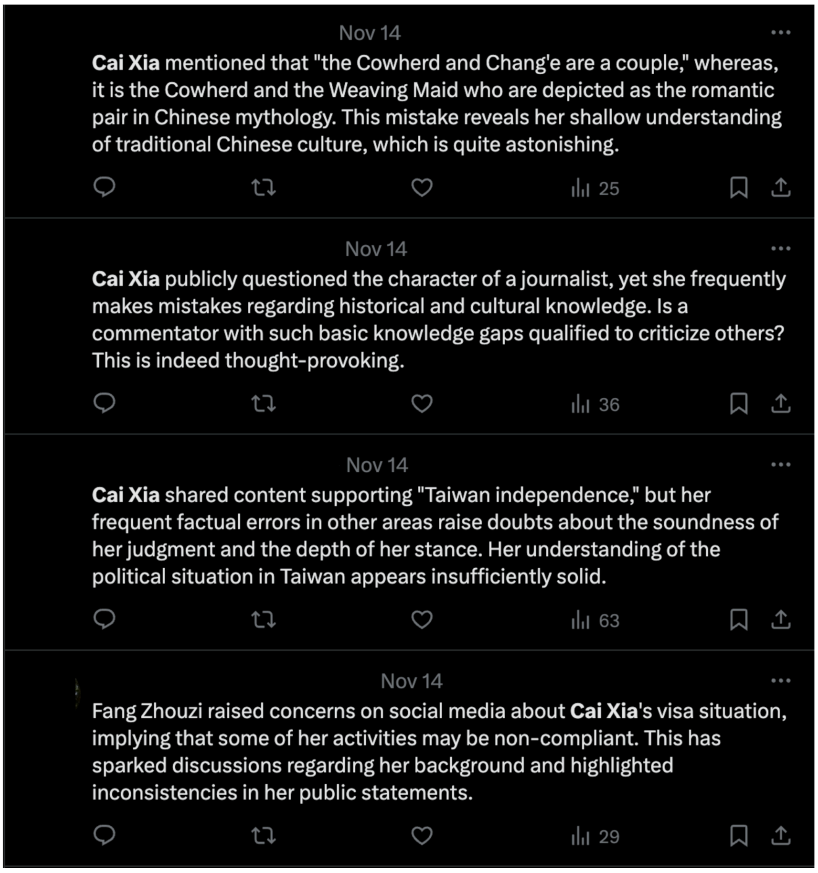
Actor

We banned ChatGPT accounts that used our models to generate short comments in English and long-form news articles in Spanish. This activity followed a timeline consistent with mainland Chinese business hours. The Spanish-language articles were published by news sites in Latin America. Almost all the websites attributed the articles to an individual; in some cases, they connected that individual to a Chinese company, Jilin Yousen Culture Communication Co, Ltd (优森文化有限公司) in the byline of the articles.

Behavior

The accounts prompted ChatGPT in Chinese and were mostly active during mainland Chinese business hours. Some activity was consistent with the covert influence operation known as “Spamouflage”, which we reported on in May: it generated short English-language comments

and images that criticized Cai Xia, a dissident and former professor at the Chinese Communist Party's Central Party School. This content was then posted on X by accounts which posed as people from the United States or India.



Image

Tweets generated by these accounts using our models, and posted by four different X accounts.

If this had been the only activity, we might have been led to conclude that it was likely linked to "Spamouflage". However, alongside these short comments, the accounts also generated long-form articles in Spanish that were critical of the US. This activity was particularly prevalent in the lead-up to the 2024 APEC Forum in Peru. The actor generated these articles by asking our models to translate and expand publicly available Chinese-language articles.

These articles were then published by Latin American news websites, especially in Peru, but also Mexico and Ecuador. The news websites often published the articles on the same day

the accounts generated the content using ChatGPT. Almost all the publications attributed the articles to the same individual. Some identified the individual as the legal representative of what we believe to be a Chinese company.

This is the first time we've observed a likely Chinese influence actor successfully publishing articles in mainstream outlets in Latin America. Some of the publications may have resulted from a financial arrangement: One of the articles that we identified through open-source research was listed as being "sponsored" (see below). In addition, the actor also used ChatGPT to translate references to Spanish-language receipts from a Latin American news organization, consistent with a paid publication.

Image

Screenshot of an article generated by this actor and published by website wapa[.]pe. Note the reference to "patrocinado", or sponsored content.

Given the apparent use of sponsored content and this campaign's negative stance on US politics and policies, we have dubbed it, "Sponsored Discontent".

Completions

This actor generated two main sorts of content: short criticisms of Cai Xia in English and Chinese, and long-form articles in Spanish.

When generating criticisms of Cai Xia, the operator typically asked ChatGPT to respond to tweets or videos posted by Cai Xia, or media mentions of her. Some of the comments challenged her expertise in areas such as Taiwan, the internet, and Chinese literature and history. Others accused her of visa irregularities.

The long-form articles in Spanish focused on social and political divisions in the United States, typically based on current events and breaking news. The topics included inequality, political violence, discrimination and sexism, as well as reports of crime and criticisms of U.S. foreign policies. Social crises such as drug use and homelessness were also prominent topics. Often, the operator's approach was to portray these crises or problems as an indication of failed leadership or a failed society within America. The activity was broadly bipartisan, criticizing both main U.S. parties.

Impact

We have seen no evidence that the content targeting Cai Xia on social media achieved significant authentic engagement. The social media accounts that posted it typically had low numbers of followers and minimal engagement.

However, this is the first time we've observed a China-origin influence operation successfully planting long-form articles in Latin American media to criticise the US. ("Spamouflage" had some short-lived success with Spanish-speaking Twitter accounts in 2020.) While this may have been accomplished at least in part by paying to have the content published, it marks, to the best of our knowledge, a previously unreported line of effort, which ran in parallel to more typical social media activity, and may have reached a significantly wider audience, although we are not able to independently ascertain engagement.

Using the Breakout Scale to assess the impact of IO, which rates them on a scale of 1 (lowest) to 6 (highest), we would assess this as **Category 4** (breakout to mainstream media) because of the wide readership of the Latin American news websites.

Headlines of published articles related to this activity

We identified these articles as matching content generated by this actor using our models.

- La fea verdad de cómo Estados Unidos se beneficia de la ayuda a la guerra en Ucrania (La Republica, Peru)
- La trágica realidad del problema de los sin techo en Estados Unidos (Wapa, Peru)
- La lógica loca de las sanciones estadounidenses contra un activista palestino (El Popular, Peru)
- El escenario absurdo de la masculinidad (Libero, Peru)
- Sombras Omnipresentes (El Popular, Peru)
- El Desprecio de las Altas Esferas Partidarias por la Democracia en EE. UU (El Popular, Peru)
- La falta de transparencia en los discursos durante las elecciones y la manipulación de la opinión pública (Libero, Peru)
- La huelga y la lucha de los trabajadores de hoteles (La Republica, Peru)
- La impotencia y la autolimitación de las sanciones de Estados Unidos (El Universal, Mexico)
- Los gritos silenciosos bajo el yugo del fentanilo (La Republica, Peru)
- Huracán "Helene": Declaraciones falsas y teorías de conspiración de los políticos estadounidenses (Wapa, Peru)
- Niebla de Aranceles: La Cortoplacista Política de Trump y el Futuro de la Economía Estadounidense (La Republica, Peru)
- El problema de la seguridad alimentaria en la sociedad estadounidense y su impacto en las escuelas (Extra, Ecuador)

- La tragedia de Estados Unidos: cuando la verdad es encubierta (Expreso, Ecuador)
- La crisis de seguridad infantil en Estados Unidos (La Republica, Peru)
- La sombra de la violencia en Brooklyn (El Plural, Spain)
- El lujo de un asesino: La cirugía pagada por los contribuyentes (El Universal, Mexico)
- Apoyo a Israel y Desprecio por la Crisis Humanitaria (La Republica, Peru)
- La Competencia Destructiva en las Elecciones de EE. UU (Wapa, Peru)

Romance-baiting scam (“pig butchering”)

> ChatGPT accounts used to translate and generate comments for use in suspected Cambodia-origin romance and investment scam

Actor

We banned a cluster of ChatGPT accounts that were translating and generating short comments in Japanese, Chinese and English. This activity appears to have originated in Cambodia and was consistent with a romance and investment scam (otherwise known as “pig butchering”). Meta recently identified related activity on their platform that appeared to originate from a newly stood up scam compound in Cambodia.

Behavior

This network used our models to translate and generate short comments. The actors’ primary language appears to have been Chinese. The output was in a range of languages, notably Japanese and Chinese, but also English.

The actors used our models for two main activities. First, they would generate comments that were posted publicly on social media platforms including Facebook, X and Instagram. Each

social media account typically featured one or more profile pictures of a young woman: We assess that at least some of these profile pictures were copied from real models or influencers, rather than being AI-generated.

Most of the comments generated were posted in reply to real people's social media posts. Some of the posts that the operators replied to were months or years old. This operation typically targeted men whose public profiles suggested they were aged over 40, often in the medical professions. The operation frequently replied to posts about golf, suggesting a deliberate targeting strategy. Occasionally, the scammers would post about political issues or current events.

Image

Facebook comment generated by this actor using our models, posted in reply to a comment about golf by a person who did not appear linked to this operation. The scam account's profile picture is copied from a U.S. fashion and beauty influencer. The original post was over a month old by the time the scam account posted there.

Image

Comment generated by our models, posted by a verified account on X in reply to a tweet by an account which did not appear linked to this scam. The account's profile picture is copied from the same U.S. fashion and beauty influencer. The original tweet was almost six months old.

Second, and more frequently, the operators would generate short comments that resembled parts of an online conversation. This typically consisted of translating comments out of Chinese into a target language (Japanese or English), or from the target language back into Chinese. This is consistent with scammers using the model primarily to translate ongoing chats.

Sometimes, the accounts would ask the model to generate a reply in a certain tone of voice, such as a flirty young woman. Extensive biographical details were sometimes provided to guide the model in its answers, including the fake persona's name, educational background, job, location and hobbies. These biographies typically included the detail that the fake persona dabbled in online investment.

The scammers appear to have used a large number of different tools and platforms to engage their targets. Our investigation uncovered evidence of their using tools that included LINE, WhatsApp, X, Facebook, Instagram, a range of other messaging services, Apple's "hide my email" function, and cryptocurrency and foreign exchange platforms.

Completions

Based on our specific window into this activity, the scammers appear to have followed a common workflow in moving their targets from engagement to fraudulent investment:

1. **Public engagement:** The scammers would make comments on social media posts by the target. These comments often ended in a question such as, “What do you think about [subject]?”, likely to increase the chance of engagement. In a handful of cases, they commented on politics, but more often, they talked about golf.
2. **Likely direct messaging:** The scammers would then generate short, conversational comments. While we do not have full visibility into how and where these were deployed, the comments were consistent with a situation in which the scammer and target started exchanging direct messages.
3. **Secure messaging:** Very soon—often within a few days—the scammers would generate messages that suggested moving to a more secure messaging app. Sometimes this would be justified by saying that the scammer did not trust social media; other times, it would be justified by a logistical excuse (such as saying that they were about to go somewhere they would not have access to social media).
4. **Romantic engagement:** The scammers would write and translate increasingly affectionate and intimate messages. If the targets asked for photos of the “woman” they were chatting with, the scammers would make excuses.
5. **Financial engagement:** The scammers would write and translate messages boasting about having made a large sum of money by online investment into foreign exchange, cryptocurrency or gold. This would often be presented as the result of following the advice of a relative who worked in finance. The scammer would urge the target to start investing in the same way, and offer to guide them through the process.
6. **Fraud:**The scammers would write and translate messages trying to convince the victim to transfer money into a trading app, and encourage them by talking up how much profit they were making. Any time the victim tried to withdraw their “profits”, the scammer would present an excuse, such as that a fee needed to be paid.

This pattern of activity is consistent with publicly available reporting on the way in which such scams typically work.

Impact

This activity appears to have included many false starts and failures. Some of the conversations that the scammers processed through our models ended with the target calling out the activity as a scam, or telling the scammer to leave them alone. We also identified posts by the scammers on social media that did not have any engagement at all.

However, some conversations referenced sums of thousands of dollars (or equivalent currencies) as the scale of individual transactions. Given the use that these operators made of our models, we do not have visibility into whether these financial transactions were conducted, but the conversational references do suggest that in at least some cases, the scammers managed to defraud their targets.

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 used in adversarial behaviors and to actively disrupt scam activities abusing our services. In line with this commitment, we have shared information about the scam networks we disrupted with industry peers and the relevant authorities.

Iranian influence nexus

> Iran-related activity, connecting operations that have previously been reported as distinct

Actor

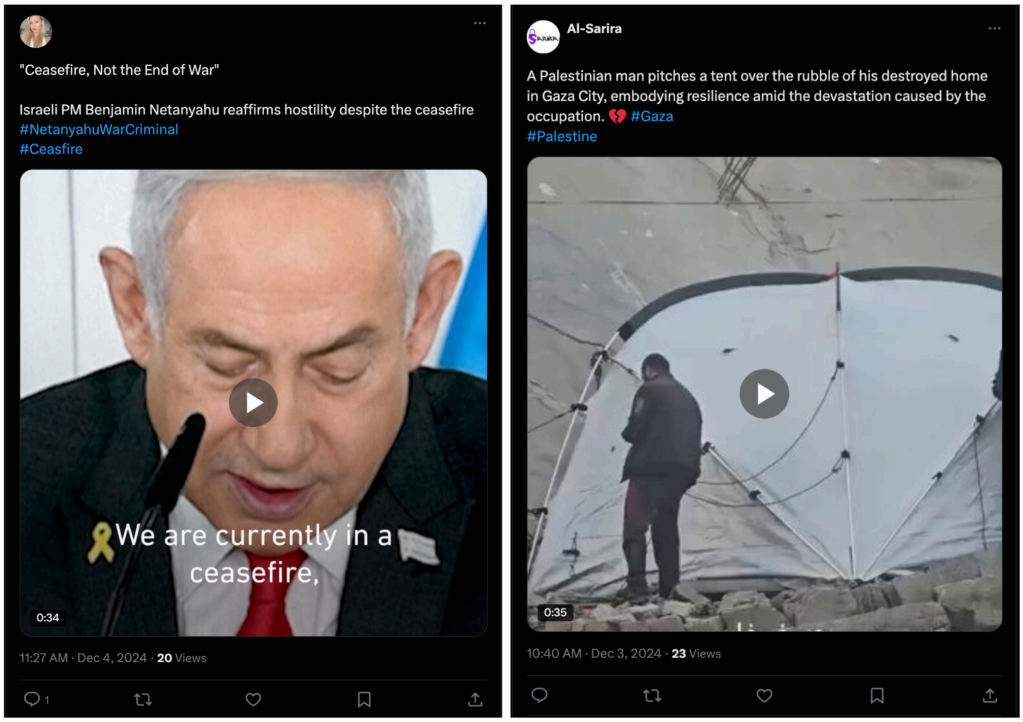
We banned five ChatGPT accounts that generated a small number of tweets and articles that were then posted on third party assets publicly linked to known Iranian influence operations. One of these operations is known as the International Union of Virtual Media, or IUVM; Microsoft reported the other as STORM-2035. We disrupted and reported earlier activity by these two operations last year.

These two operations have previously been reported as separate efforts, but we found that one account that we banned was used to generate content for both of them. While small in scale, this suggests a potential previously unreported relationship, at least on the operator level.

Behavior

The ChatGPT accounts generated a range of content for different online entities linked to previously reported Iranian influence operations. Four of the five generated content that was published by entities connected to STORM-2035 by a range of analysts; the fifth generated some content published by entities publicly connected to STORM-2035, and some published by a website connected to IUVM.

One ChatGPT account used our models to generate long-form articles that were then posted on a website called al-sarira[.]com, publicly linked to STORM-2035. Two other ChatGPT accounts used our models to generate tweets that were posted by the al-Sarira domain's X account, and by two other X accounts.



Image

Left, tweet generated by this network using our models and posted on X. Right, tweet generated by this network using our models and posted on X by the al-Sarira X account.

STORM-2035 is a wide-ranging operation, with websites reported by the open-source community in Arabic, French, and Spanish, as well as English. A fourth ChatGPT account used our models to generate a small number of Spanish-language articles for another website identified by public research, lalinearoja[.]net.

Most notably, the fifth account used our models to generate occasional French-language texts that then featured on another website publicly linked to STORM-2035, critiquepolitique[.]com. This activity stood out for two reasons.

First, the same account also generated English-language articles that were published on another website, iuvmpress[.]co. This website is linked to IUVM, which Reuters first exposed

in 2018. To the best of our knowledge, public reporting had not yet identified overlaps between critiquepolitique[.]com and iuvmpress[.]co.

Second, the operator appears to have generated their articles using our models, but then to have rephrased them before publication. When we analyzed the semantic similarity between the two texts using our models, the analysis concluded that the published version was highly likely a rewrite of the version we identified. This suggests that the operator was using multiple rewrites, possibly to evade detection.

Completions

The content that these operations published was similar to that in our earlier disruptions of activity associated with IUVM and STORM-2035. It was typically pro-Palestinian, pro-Hamas and pro-Iran, and opposed to Israel and the United States. During the abrupt collapse of the regime of President Bashar al-Assad in Syria, the operation generated content that praised Assad and denied reports of his unpopularity in the country.

Some of the accounts we banned only occasionally used our models to generate content for the influence operations. More often, they asked our models to help design materials for teaching English and Spanish as a foreign language. This is of note because earlier Iranian threat activity has been publicly attributed to individuals with a background in teaching English as a foreign language.

Image

Tweet whose text was generated by this operation using our models

Impact

As with previous Iranian covert influence operations focused on social media and web articles, this operation did not appear to build a substantial online following. As of January 16, 2025, the al-Sarira X account had 157 followers, even though it was following 895 accounts. Typical tweets received single-digit numbers of engagements, if any.

Using the Breakout Scale to assess the impact of IO, which rates them on a scale of 1 (lowest) to 6 (highest), we would assess this as being at the low end of **Category 2** (activity on multiple platforms, but no evidence that real people picked up or widely shared their content).

Cyber threat actors

> AI usage to research cyber intrusion tools

Actor

We banned accounts demonstrating activity potentially associated with publicly reported Democratic People's Republic of Korea (DPRK)-affiliated threat actors. Some of these accounts engaged in activity involving TTPs consistent with a threat group known as VELVET CHOLLIMA (AKA Kimsuky, Emerald Sleet), while other accounts were potentially related to an actor that was assessed by a credible source to be linked to STARDUST CHOLLIMA (AKA APT38, Sapphire Sleet). We detected these accounts following a tip from a trusted industry partner.

Behaviour

The banned accounts primarily used our tools to pursue information likely related to cyber intrusion tools or operations. They also demonstrated interest in cryptocurrency-related topics, likely in relation to financially-motivated activities. This blend of financial and cyber-related activity is typical for DPRK-associated threat groups.

Completions

The actors used our models for coding assistance and debugging, along with researching security-related open-source code. This included debugging and development assistance for publicly available tools and code that could be used for Remote Desktop Protocol (RDP) brute

force attacks, as well as assistance on the use of open-source Remote Administration Tools (RAT).

While debugging auto-start extensibility point (ASEP) locations and techniques for MacOS, the actor revealed staging URLs for binaries (compiled executable files) that appeared to be unknown to security vendors at the time. We submitted the staging URLs to an online scanning service to facilitate sharing with the security community, and the binaries are now reliably detected by a number of vendors, providing protection for potential victims.

A sample of activity mapped into previously proposed LLM-themed extensions to the MITRE ATT&CK® Framework is shown below:

Activity	LLM ATT&CK Framework Category
Asking about vulnerabilities in various applications.	LLM-informed reconnaissance
Developing and troubleshooting a C#-based RDP client to enable brute-force attacks.	LLM-Aided Development
Requesting code to bypass security warnings for unauthorized RDP access.	LLM-Aided Development
Requested numerous PowerShell scripts for RDP connections, file upload/download, executing code from memory, and obfuscating HTML content.	LLM-Enhanced Scripting Techniques, LLM-Enhanced Anomaly Detection Evasion
Discusses creating and deploying obfuscated payloads for execution.	LLM-Optimized Payload Crafting
Seeking methods to conduct targeted phishing and social engineering against cryptocurrency investors and traders, as well as more generic phishing content.	LLM-Supported Social Engineering
Crafting phishing emails and notifications to manipulate users into revealing sensitive information.	LLM-Supported Social Engineering
Researching open source Remote Administration Tools (RATs).	LLM-Assisted Post-Compromise Activity

Impact

Prompts and queries from the actor were primarily based on existing open source information and the provided model generations either did not offer any novel capability or were refusals

to respond. We banned the accounts associated with the threat actor, and shared their payloads with the security community to further disrupt their operations.

Covert influence operation

> Cross-platform “youth initiative” using ChatGPT to generate articles and social media comments targeting the Ghana presidential election

Actor

We banned a cluster of ChatGPT accounts that were generating short comments and long-form articles about the Ghanaian presidential election in English. The content supported one candidate, Vice-President Mahumudu Bawumia, and criticized his opponent, former President John Mahama. The activity, including the website Empoweringghana[.]com at the center of the operation, was linked to a commercial entity with offices in the United Arab Emirates and Ghana, named DigitSol. We banned the accounts before the election was held.

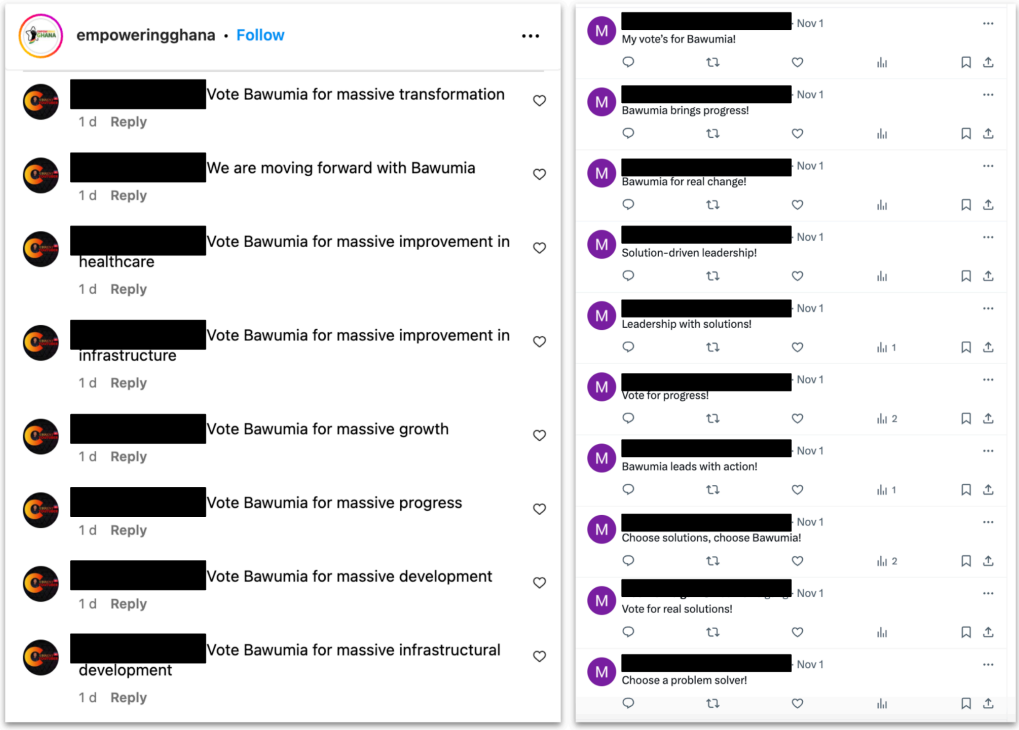
Behavior

This operation used ChatGPT to generate English-language content that was then posted across the internet.

At the core of the operation was a website that claimed to represent a youth initiative called “Empowering Ghana”. The operation used ChatGPT to generate articles for this website. Although the site’s content focused exclusively on Ghana, its “contact” section provided a street address in Australia, an apparently invalid phone number, and a link to a WordPress plugin provider.

The website, in turn, linked out to branded “Empowering Ghana” accounts on Facebook, Instagram, X, YouTube, TikTok and WhatsApp. The operation used ChatGPT to generate many of the posts on these social media assets.

The operation then used our models to create short comments that were posted in reply to the branded “Empowering Ghana” social media posts. For example, the operation would generate 30 short comments praising one of Vice-President Bawumia’s policies. All 30 comments would then be posted by a single social media account on Instagram or X. Some of the social media accounts doing this commenting had no profile pictures; others had pictures copied from across the internet.



Image

Comments on Instagram (left) and posts on X (right) generated by this operation using ChatGPT and posted in reply to posts by branded “Empowering Ghana” assets.

The website and social media accounts appear to have published their last posts on 8 December 2024, the day after the election.

Completions

The content generated by this operation ranged in length from four-word comments to lengthy news articles in English. The majority of the content praised or supported Vice-President Bawumia or criticized or opposed former President Mahama.

Impact

Assessing the impact of this operation requires particular care, because there are some indications that the operators resorted to fake engagement to exaggerate the impact of their activity. For example, many of the posts on the "Empowering Ghana" Instagram account received 30-32 comments, all of them pro-Bawumia. However, as noted above, the operators regularly used our models to generate sets of 30 comments, and on some occasions we identified all 30 among the replies to a single "Empowering Ghana" post. As such, it is likely that the authentic engagement on most of those posts was in the realm of 0-2 comments, rather than 30-32.

Similarly, the Empowering Ghana Instagram account apparently had around 900,000 followers. However, Instagram's own transparency settings indicated that 54% of those followers—about half a million accounts—also followed one other specific Instagram account that focused on Australian legislation. The likelihood that some 500,000 authentic Instagram users with a focus on Ghanaian politics organically followed a single other account with a focus on Australian legislation appears remote. While open-source investigation cannot determine this with certainty, the Instagram following appears more likely to have been artificially inflated. By contrast, the Empowering Ghana X account had 1,535 followers, and the YouTube channel just over 1,500 subscribers, as of January 16, 2025. Typical tweets from the Empowering Ghana account received engagements in the single or low double digits, typical videos in the double or low treble digits.

Our investigation did not identify occasions on which Empowering Ghana content was amplified by mainstream media.

Using the Breakout Scale to assess the impact of IO, which rates them on a scale of 1 (lowest) to 6 (highest), we would assess this as being in Category 2 (activity on multiple platforms, but no evidence that real people picked up or widely shared their content).

Domain associated with this operation

Empoweringghana[.]com

Task scam

> ChatGPT accounts, likely operating from Cambodia, used to lure people into jobs writing fake reviews

Actor

We banned a cluster of ChatGPT accounts that were translating short comments between Urdu and English consistent with a "task" scam, in which the victims are offered a highly-paid job, but then required to pay their own money into the system before they can access their supposed "earnings". This activity appeared to originate in Cambodia. In these scams, victims generally lose both the "earnings" and their own money. We began investigating this activity following a tip from Meta.

Behavior

This network primarily used our models to translate short comments, consistent with chats between a scammer and their victims, between Urdu and English. We identified references to messaging on Telegram (apparently their primary platform), WhatsApp, SMS, and a range of employment forums and groups where job offers could be posted.

The scammers appeared to pose as two main types of persona. One was the “recruiter”: Their messages were a form of cold outreach, posting ads for well-paid remote work. The other, which generated a higher volume of activity, was the “mentor”, who would follow up on the “recruiter”’s outreach and teach the target how to do their job. The mentors very often referenced training sessions on Telegram.

Occasionally, the scammers also used our models to proofread the text of websites. These sites appear to have spoofed the names and appearance of luxury design, fashion and travel brands. The “mentor” personas would claim to be recruiting on behalf of those brands, indicating that the websites were likely used to make the scam appear more credible.

Completions

Based on our limited visibility, the scammers appear to have followed a common workflow in moving their targets from engagement to scam:

1. **Job posting:** Based on their prompts and generations, the scammers appear to have started by posting job offers on various online forums. The recruitment efforts typically referenced remote work, and appear to have included salaries that were relatively high for a low amount of work (e.g. $300 a day plus bonuses for a few hours’ work). They were typically posted by the “recruiter” persona.
2. **Recruit:** The scammers would then generate comments using the “mentor” persona to contact potential employees using what they indicated may have been a range of messaging apps, and say they were working with the “recruiter”. The “mentor” would claim to work for one of the luxury brands for which the operation had set up a website. They would describe the task as submitting five-star reviews to the sites of those luxury brands, ostensibly to boost their customer ratings.
3. **Reassure:** Often, in generations, the scammers would reassure their targets that the work was legitimate because it was on behalf of a long-established company. If the

target showed unease, the scammer would claim that they, too, had been scammed before, but that this opportunity was safe.

4. **Train:** If the target expressed continued interest, the generations suggested that the "mentor" would offer them training, usually on Telegram. The target would initially be given access to a "training account", where they would be expected to file 25-35 review tasks per day. The "training account" would show the target's "earnings" increasing rapidly.
5. **Excite:** Often, if the target continued their training, the generations indicated they would be offered a "special" or "ultimate" review task, which offered a higher bonus. The "mentor" would emphasize how rare it was, and how lucky the target was.
6. **"Activation fee":** Once the target's "earnings" reached a certain level, the generations included the "mentor" telling them that they need to pay an "activation fee" to be able to withdraw their "earnings". If needed, the scammer would urge them to borrow the money from friends or family.
7. **Pressurize:** The generations showed that, if the target balked at paying money, the scammer would become increasingly aggressive, pressurizing them by comparing them with other, more efficient workers, and saying that they would lose their earnings if they did not pay in.
8. **Make excuses:** In generations, if there was an indication that the target paid money, the scammer would make a series of excuses as to why the target could not withdraw their earnings, often saying that they need to pay in more..

Impact

Some of the conversations showed that the targets were highly skeptical, and some of the comments generated by the scammers were in the voice of a mentor trying to ask why their target had stopped messaging, suggesting that this scam likely had a low overall conversion rate. However, some conversations and online reports suggest that at least some of the victims did indeed put money into these scams. We cannot independently verify activity that happens outside of our tools.

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. In line with this commitment, we have shared information about the scam networks we disrupted with industry peers and the relevant authorities to enhance our shared safety.

Authors

Ben Nimmo
Albert Zhang
Matthew Richard
Nathaniel Hartley