OpenAI

Disrupting malicious uses of our models

An update, February 2026

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.

In the two years since we began publishing these threat reports, we have gained important insights into the ways threat actors attempt to abuse AI models. In particular, the case studies in this report, as in our earlier reports, illustrate how threat actors typically use AI in combination with other, more traditional tools such as websites and social media accounts. Threat activity is seldom limited to one platform; as our report on a Chinese influence operator shows, it is not always limited to one AI model. Rather, threat actors may use different AI models at various points in their operational workflow. We share these insights in our threat reports so that our industry, and wider society, can be better placed to identify and avoid such threats.

These are the key insights from our latest threat disruptions:

The scale and scope of covert influence operations (IO) from China: We banned a ChatGPT account linked to an individual associated with Chinese law enforcement. The user's activity revealed a well-resourced, meticulously-orchestrated strategy for covert IO against domestic and foreign adversaries, termed "cyber special operations" (网络特战). As part of this strategy, they tried to use our model to plan a covert IO targeting the Japanese prime minister, but our model refused. They also used ChatGPT to edit periodic status reports on the conduct of "cyber special operations" more broadly. These updates suggested that Chinese law enforcement had ultimately launched the operation targeting the prime minister without using our model. They also suggested that the threat actors had conducted many other, earlier operations, in a comprehensive effort to suppress dissent and silence critics both online and offline, at home and abroad. This effort appears to be large-scale, resource-intensive and sustained, engaging at least hundreds of staff, thousands of fake accounts across scores of platforms, and the use of locally-deployed AI models, especially Chinese ones. The user described the operations as using dozens of tactics, ranging from abusive reporting of dissidents' social media accounts, through mass online posting, to forging documents and impersonating US officials to intimidate critics. Through open-source analysis, we were able to identify online activities that were consistent with some of the tactics this user described, targeting not just people in China, but also dissidents and critics around the world.

Semi-automated romance from Cambodia: one common form of scam since long before the days of AI is the romance scam, which tries to trick people into handing over money to a non-existent romantic partner. This report details the pattern that such scams typically follow. In one case, we banned a network of accounts that used AI to pose as a fake dating agency targeting young men in Indonesia. Unusually, this scam network combined manual ChatGPT prompting and an automated AI chatbot to try to entrap its targets.

A content farm linked to Russia: we banned a cluster of ChatGPT accounts that were linked to the Russia-origin "Rybar" ("Рыбарь", in Russian, "fisherman") network. This cluster translated and generated content that was posted on "Rybar" social media accounts, but it also appears to have served as a content farm for a wider network of accounts on X and Telegram that bore no overt relationship to the "Rybar" group. On some occasions, the threat actor used ChatGPT to generate batches of short social media comments, and these were then posted by accounts on X and Telegram that appeared to originate from different parts of the world.

Actor, behavior, content: the scam and influence operations described in this report all used AI-generated content, but they achieved very different results. For example, some AI-generated social media posts received tens of thousands of views, while other posts created in the same batch received almost none (see an example in operation "Fish Food", below). The use of AI-generated content on its own does not appear to have been the decisive factor; rather, other factors were likely the main drivers of engagement, notably the popularity of the accounts which did the posting. Similarly, in the scam case "Date Bait", targeted ads on social media appear to have been a key driver of engagement. This underscores the importance of studying the nature of threat actors and the ways in which they behave, as well as the content they generate.

Case studies

Romance Scams

Since we began reporting on our disruptions of scam networks that sought to abuse our models a year ago, we've taken down many scam operations from different parts of the world. They included "task" scams, which defraud their victims by convincing them to pay money in as a way of accessing non-existent earnings for trivial tasks, and investment scams, which defraud their victims by convincing them to put money into non-existent investment companies.

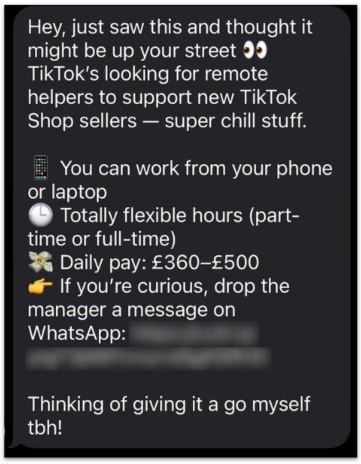
One common type of scam since long before the days of AI is the romance scam, in which scammers pose as a potential romantic partner and attempt to convince their target that they have met a love match before asking them for ever more money. Our first published scam disruption featured a newly stood up criminal operation in Cambodia that had used ChatGPT to generate messages for a romance scam, some of which were spread on social media.

Since then, we've disrupted various attempted romance scams. As we wrote in June, these and other scams tend to follow a common pattern in their use of AI, which we think of as *the ping* (cold contact), *the zing* (generate emotion), and *the sting* (extract money).

1. **The ping (cold contact)**: the scammer generates content designed to attract the potential target's attention by appealing to their interests. For example, the "pig butchering" scam we reported last year frequently targeted American men in their 40s in the medical professions by replying to social media posts they made about golf. Other operations used cold-call SMS messages, fake recruitment messages or, in our most recent romance scam case (described below), social media ads. In each case, the threat actors used ChatGPT to generate messages that might be more engaging and less obviously non-native than traditional scams.

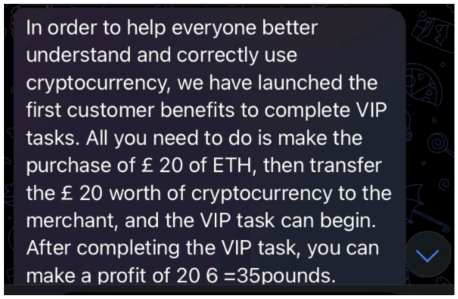
Golf-themed ping from the "pig butchering" network we reported in 2025, replying to a Facebook post by a user not linked to the operation.

2. **The zing (generate emotion)**: the scammer generates content designed to trigger strong emotions in the target, and thus make them easier to manipulate. Romance scams try to make the target attracted to the scammer. Other scams can include trying to make the target excited about a potentially lucrative deal, afraid of missing an opportunity, or alarmed about an alleged legal risk or unpaid bill. Sometimes, the zing can be included in the same message as the ping.



Cold-call SMS from the scam operation "Wrong Number" that we exposed in June 2025, including details of implausibly high returns for little work, likely designed to create the "zing" effect.

3. **The sting (extract money)**: the scammer generates content designed to convince the target to hand over money. The reasons given can vary enormously. For example, romance scams may ask the target to invest in the beloved's business, or hand over money to cover a financial crisis. "Task" scams tell their targets to pay money in so they can access (fictional) money they believe they earned. Investment scams tell the targets to put their money into non-existent investments.



Cold-call SMS from the scam operation "Wrong Number" that we exposed in June 2025, including details of implausibly high returns for little work, likely designed to create the "zing" effect.

As this flow suggests, an essential component of scams is the distribution network. Different scam operations that we've exposed sent their pings via SMS, encrypted messaging apps, social media posts, online ads, or a combination of all of them; scams in the pre-AI era are notorious for having used emails, phone calls, or even, in the nineteenth century, letters and telegrams.

Many scams take a scattergun approach to distribution, but some appear to attempt precision targeting. For example, the "pig butchering" scam we reported in February 2025 focused on topics such as golf; the romance scam we describe below used social media ads to target young men in Indonesia; and celebrity scams typically pose as a famous person and then target that person's fan groups. While the fragmentary nature of the evidence makes it difficult to reliably compare different scams that we disrupted, we assess that the scam's chosen distribution method (e.g., scattershot or targeted) plays a significant role in each scam's ability to successfully reach and exploit its targets, regardless of the degree to which the operation used AI for different functions.

Scam: Operation “Date Bait”

AI-enabled romance scam targeting Indonesian loveseekers

Actor

We banned a cluster of ChatGPT accounts and one API customer that were using our models to run a semi-automated romance and task scam likely defrauding hundreds of victims a month. This activity very likely originated in Cambodia and aligns with recent public reporting on Chinese-led criminal scam operations in the country. In isolated cases, individual users self-identified as scam workers in Cambodia, such as when asking the model for tax advice and stating their occupation as “scammer”.

Behavior

The network used ChatGPT and API access to our models to conduct a scaled, semi-automated romance task scam, following a structured workflow and operating model reminiscent of cold-outreach sales programs.

1. **Ping**: At the top of the funnel, the operation used its ChatGPT accounts to generate promotional texts for a high-end dating and escort service called “Klub Romantis”. These texts were then posted on social media as paid ads. The scammers targeted Indonesian men interested in luxury lifestyle content, using ad keywords including “golf”, “yachts”, and “fine dining”. The ads included links to an AI chatbot instructed to pose as a flirtatious receptionist. The purported receptionist asked targets to choose from a menu of types of women and relationships before directing them to Telegram using a tracking URL and promo code.

The “Klub Romantis” logo, created by a scammer in this network using ChatGPT.

2. **Zing**: Once on Telegram, the operation blended human operators using ChatGPT with API-powered automation. Receptionist personas continued the conversation using increasingly romantic and sexually-explicit language, and then offered targets to join online dating platforms called “LoveCode” and “SexAction”. The platform showed profiles of purported “available girls” and a feed of AI-generated announcements congratulating fictitious clients for completing “missions,” unlocking bonuses, and winning prizes. After building trust, the receptionist persona handed the conversation over to a “mentor,” who used ChatGPT to generate and translate emotionally manipulative messages pressuring targets to complete a sequence of “tasks” or “missions.” Each successive task required increasingly large payments via bank transfers or digital payment wallets, such as purchasing a “VIP card,” supporting a “chosen girl” in a competition, or paying a hotel deposit.

An image of a fake receptionist for the “LoveCode” dating platform created by a scammer using ChatGPT.

3. **Sting**: At the final stage, the scammers attempted to collect a larger payment, which they described as the “kill”, by inventing additional fees such as “compensation settlements” or “verification deposits.” Judging by the messages the scammers pasted into ChatGPT, once the victim sent the maximum possible amount, the scammers blocked them on Telegram and marked the case as closed.

LOVE CODE
Entertainment

SURAT PERBAIKAN DATA

Kepada
Yth. Anggota Love Code
Di tempat

Dengan Hormat,
Demikian surat ini saya lampirkan Dari Pusat Pendataan untuk memperbaiki data
ID [redacted]
yang kacau serta berantakan didalam sistem dan anda telah merugikan perusahaan. Perbaikan data ini wajib diselesaikan agar data anda kembali normal, mohon kerjasamanya dan tidak mengulangi kesalahan untuk kedepannya agar anda bisa melakukan penarikan serta mengaktifkan kartu VIP.

(1) Tugas perbaikan data yang disetujui khusus Rp 20.500.000 bonus kontak 35% keuntungan Rp 39.860.000

Anda harus menyelesaikan dua perbaikan untuk memperbaiki data transaksi karena data anda yang sebelumnya kacau dan berantakan didalam sistem jadi anda harus memenuhi persyaratan perusahaan, anda harus menyelesaikan tugas ketentuan ini dan harus melakukan pembayaran agar penyelesaian data anda kembali normal didalam sistem Perusahaan Love Code.

Yang Menerima Tugas

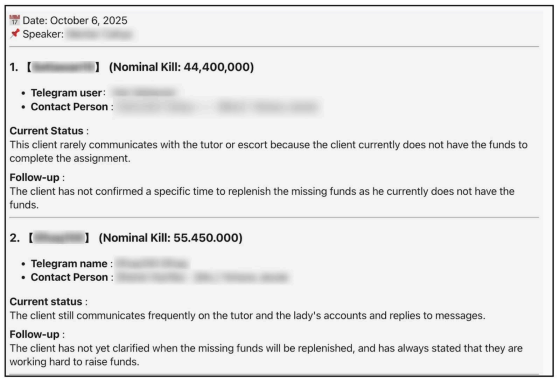
11 Maret 2025
Managing Partner

Copy of a letter shared with ChatGPT by a scammer telling a victim they must pay Rp 20,500,000 ($12,000) to rectify a fictitious data processing error in the Love Code system but will receive a 35% “bonus” of Rp 39,860,000. Redactions by OpenAI investigators.

Completions

In addition to the romance scam activity, the operation also used our models to assist them in their day-to-day business operations. This included asking ChatGPT for guidance on how to complete technical tasks, such as integrating the OpenAI API or updating the fake dating service website, analyzing financial accounts, and generating daily reports assigning each target a “kill” value calculated by the scammers — the estimated amount to target in a final extraction.

In other cases, the scammers asked ChatGPT to translate messages between Chinese-speaking supervisors and Indonesian-speaking scam center workers, and tracking performance across departments called “Lead Generation”, “Reception Team”, and “Supervisor Team”.


Date: October 6, 2025
Speaker: [redacted]

1. [redacted] (Nominal Kill: 44,400,000)

- Telegram user: [redacted]
- Contact Person : [redacted]

Current Status :
This client rarely communicates with the tutor or escort because the client currently does not have the funds to complete the assignment.

Follow-up :
The client has not confirmed a specific time to replenish the missing funds as he currently does not have the funds.

2. [redacted] (Nominal Kill: 55.450.000)

- Telegram name : [redacted]
- Contact Person : [redacted]

Current status :
The client still communicates frequently on the tutor and the lady's accounts and replies to messages.

Follow-up :
The client has not yet clarified when the missing funds will be replenished, and has always stated that they are working hard to raise funds.

Translated excerpt of a ChatGPT-generated scammer status report showing the identity and status of each victim, assigned staffing, and the estimated “kill” value in Indonesian Rupiah. Redactions by OpenAI investigators.

Impact

Assessing the impact of this network requires care. Our primary source of evidence is the scammers’ own inputs. Those inputs suggest that the scammers may have been interacting with hundreds of targets at a time, and generating thousands of dollars a day. However, we are not able to independently verify whether these claims were accurate.

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.

Scam: Operation “False Witness”

Fake “scam recovery” service impersonating law firms and U.S. authorities

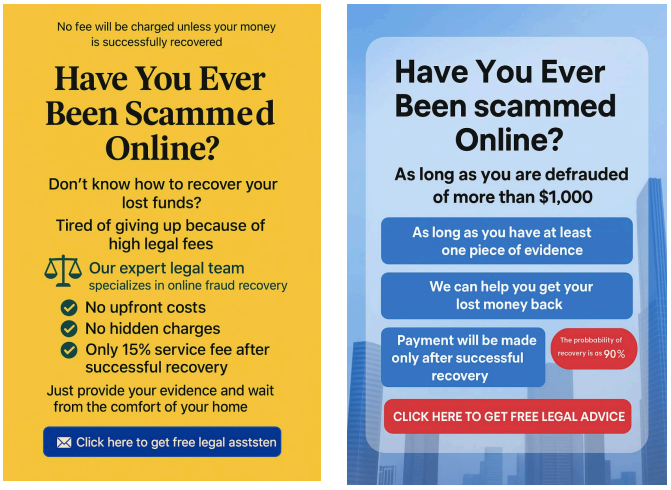
Actor

We banned a cluster of ChatGPT accounts using our models to pose as fictitious law firms, as well as impersonate real attorneys and U.S. law enforcement, in a recovery scam targeting fraud victims. This activity very likely originated in Cambodia and aligns with recent public reporting on Chinese-led criminal scam operations in the country.

Behavior

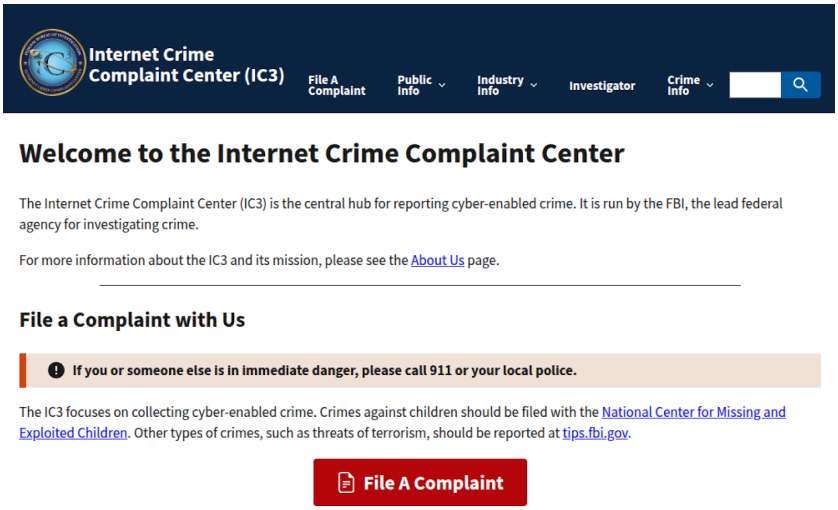
The accounts used ChatGPT to support a fraudulent scam losses recovery operation built around fake law firms and the impersonation of trusted entities, such as real attorneys and the FBI’s Internet Crime Complaint Center (IC3).

1. **Ping**: The scammers used ChatGPT to create content that purported to come from at least six fake law firms. Some of this content was then posted by social media accounts and online ads that promoted fictitious scam recovery services. Multiple open-source indicators suggested the firms were fraudulent, including no evidence of state bar licensing, the use of incongruous web domains, contact information at street addresses that do not appear to exist, and directions to contact lawyers via messaging apps.



The scammers used our models to generate social media content promoting fake scam recovery services.

2. **Zing**: The scammers attempted to gain their targets’ trust by generating comments that tried to emulate the tone and professionalism of attorneys specialized in helping scam victims recover financial losses. They generated content that purported to come from lawyer personas. We identified some of these personas on a range of websites, where their profile pictures were apparently lifted from social media or generated with AI. In some cases, the scammer reused the same “attorney” identities across multiple supposed law firms. One of the websites that we identified posed as IC3.
   The contact details published on these various scam websites directed visitors to private messaging apps, such as Telegram. In messages to targets drafted using ChatGPT, the scammers falsely claimed to be operating under the supervision of the International Criminal Court and said no fees would be charged until all of a victim’s funds were successfully recovered.



A website impersonating the FBI’s IC3 unit and linked to this scam (above).
Clicking “file a complaint” directed visitors to an impersonation Telegram account (below).

3. **Sting:** The scammers attempted to extract money from targets by requesting fees and deposits in advance of any “recovery”. This included directing targets to pay a 15% “service fee” before receiving purportedly recovered funds, requesting deposits to activate an account, and charging “consultation fees.” Scammer messages to their targets drafted using ChatGPT included instructions to send cryptocurrency payments and provide screenshots of transaction confirmations as proof of payment.

Completions

The operation used our models to support multiple parts of its workflow, such as generating promotional content for social media and cold outreach messages to targets. However, the scammers most commonly used ChatGPT to translate messages to and from targets on private messaging apps, including requests to write a reply in “American English” or in the style of a lawyer.

A subset of scam accounts also used our models to create deceptive materials intended to bolster credibility. This included fake attorney registration records and fake bar association membership cards, as well as bogus confidentiality agreements to discourage victims from seeking outside help.

A fake New York State Bar Association membership card generated using our models. Redactions by OpenAI investigators.

Impact

Assessing impact requires care because a primary source of evidence is the scammers’ own inputs. Those inputs suggest the scammers may have defrauded individual victims out of thousands of dollars, but we cannot independently verify those claims.

The FBI and at least one impersonated law firm have issued public alerts about this scam. FBI officials warned that the operation targets vulnerable audiences, particularly the elderly, exploiting scam victims' emotional state and desire to quickly recover lost funds.

OpenAI’s policies strictly prohibit using output from our tools for fraud or scams, and we are dedicated to collaborating with industry peers and authorities to understand how AI is influencing adversarial behavior and to actively disrupt scam activity abusing our services.

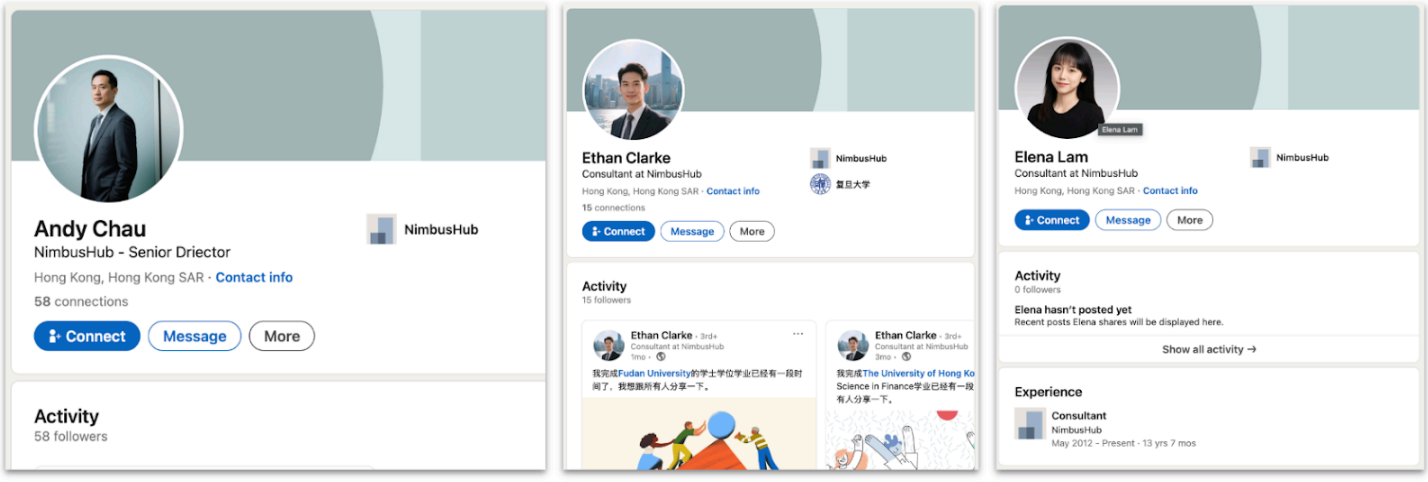
Virtual targeting: Operation “Silver Lining Playbook”

Likely China-origin activity targeting US persons

Actor

We banned a small set of ChatGPT accounts likely originated in China that used our models to request information about US persons, forums and federal building locations, guidance on face-swapping software and generated emails in English that resembled attempts at social engineering. They prompted our models in Chinese language, were mostly active during mainland Chinese business hours and used VPNs to access our platform.

The accounts generated email drafts that purported to be sent from employees of a Hong Kong-based company named “Nimbus Hub Consulting”. However, they prompted our models in Simplified Chinese characters (typically used in mainland China) rather than Traditional Chinese characters (typically used in Hong Kong) suggesting the actual operators were based in mainland China. In addition, one account generated an email purporting to be a representative of a Shanghai-based public relations organization that promotes US-Shanghai economic and cultural exchanges. Based on the company name referencing “Nimbus”, meaning a rain cloud or halo, we named this operation “Silver Lining Playbook”.



Screenshots of LinkedIn profiles affiliated with Nimbus Hub Consulting. Some of the LinkedIn profiles matched individuals listed on Nimbus Hub Consulting's ‘Our Team’ page.

Behavior

The accounts generated English language email drafts that were addressed to state-level US officials or policy analysts working in business and finance. They appeared to invite these recipients to participate in paid consultations, which they described as interpreting policy and providing strategic advice for their clients. They requested the email drafts to be concise, clear, and professional, with subject lines that created urgency and used subtle psychological cues.

In addition, the ChatGPT accounts used our models for general information retrieval, which our model responded to using publicly-available sources. This included queries about the following topics:

- The location of U.S. federal government offices, including main office locations, and a ranked list of states with large concentrations of federal agencies and officials.
- US federal personnel distribution by state.
- US persons, such as Voice of America hosts, including their past interviews and topics of interests.
- Online forums and websites that are commonly used by professionals and job seekers in the US economics and finance industry.

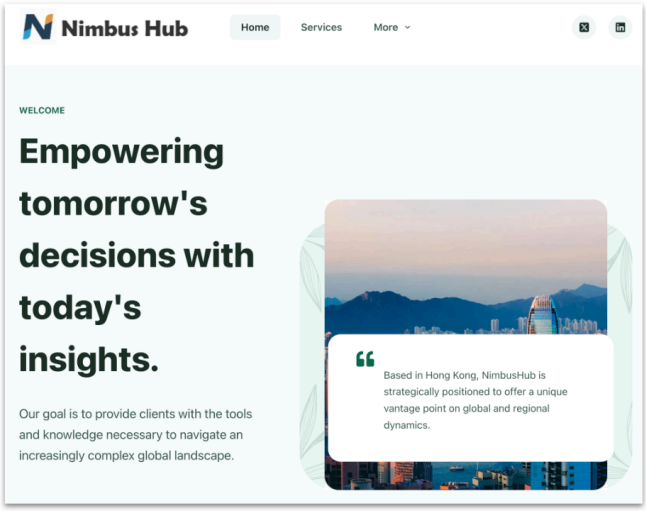
Notably, one account requested guidance on how to install and download on their computer a face-manipulation platform for face-swapping and other media enhancement known as FaceFusion. According to their ChatGPT activity, they described their intention to specifically use its live face-swap functionality. They asked for step-by-step, non-technical installation guidance and claimed they were novice computer programmers, so the instructions had to be simple. They uploaded a screenshot of their computer’s hardware specifications to assist with the installation. The model responded with information that was drawn from FaceFusion’s publicly-available website and documentation.

Completions

Their instructions to generate email drafts using our models reassembled a social-engineering playbook for a foreign intelligence service approach, rather than ordinary hiring messages. For each email draft, they requested the following sections:

- **Establish legitimacy:** They presented “Nimbus Hub” as a professional strategic consulting firm with authoritative expertise in geopolitics and transnational policy. They included a link to Nimbus Hub’s website (see screenshot below).

- **Personalize and flatter the target:** They told the model to explicitly cite the recipient’s public-sector background as proof the recipient was the exact fit for the role. In one case, the account uploaded screenshots of the LinkedIn profile of a US person based in Shanghai and instructed the model to personalize the email draft based on text about the individual’s personal experiences and background contained in the screenshots.
- **Stack incentives:** The consulting offer was usually dangled as a lucrative, online opportunity with performance bonuses and referral rewards, while staying vague about the actual work.
- **Reassuring language and reducing perceived risks:** They often claimed the work was relaxing, had timely payments and was confidential.
- **Move communications off-platform quickly:** They always attempted to move the recipients off email and to an initial online video-conference call via WhatsApp, Zoom or Teams.



Screenshot of Nimbus Hub Consulting’s website homepage, which is no longer online but has been archived.

Impact

The networks’ model interactions in this case were not technically sophisticated and their requests sometimes resembled plausible recruitment invitation drafting or legitimate software installation. However, the users' tactics of hiding behind a fake corporate identity and interests in geopolitical topics indicates a possible intention for adversarial recruitment.

Based on the accounts use of ChatGPT, there was no evidence they successfully elicited responses from their targets to reply to their invitations. We could not independently determine whether the email invitations were actually sent or if any of the targeted recipients responded.

Case Studies — Virtual targeting: Operation “Silver Lining Playbook”

Multiple democratic governments have warned that foreign intelligence services are targeting current and former government employees for recruitment by posing as consulting firms and other entities on social and professional networking platforms. It may be confusing to distinguish ordinary hiring messages from these targeted social-engineering approaches because they share some similar traits. However, legitimate outreaches are usually optimized for slower screening processes and have verifiable information about the role title, credible employer details, reasonable market compensation ranges and links to real job postings. In short, if it's too good to be true, then it probably is.

Covert IO: Operation “Trolling Stone”

Coordinated deceptive activity generating comments about the arrest in Argentina of an alleged Russian “cult leader”

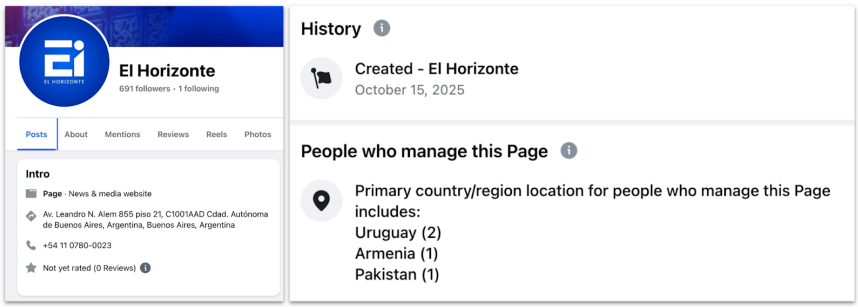
Actor

We banned a number of ChatGPT accounts that were using our models to generate batches of social media comments about the arrest in Argentina of an alleged Russian “cult leader” named Konstantin Rudnev. Different parts of this activity very likely originated in Pakistan, Armenia and Uruguay. Further activity likely originated in Argentina and Kazakhstan. The Pakistan-origin activity was connected to a for-hire actor. Some of the social media comments supported the alleged “cult leader”. Others criticized the Argentinian legal system and trolled the Argentinian edition of Rolling Stone magazine, which had reported on the case. Based on this activity, we have nicknamed the operation, “Trolling Stone”.

Behavior

This operation featured multiple clusters of activity in widely dispersed locations. The Pakistan-based user prompted in English, while the other users mainly prompted in Russian. Despite their geographical spread, they behaved in a coordinated way, generating pro-Rudnev content in Argentinian Spanish that was posted on the operation’s Facebook pages and in online articles

One main workstream consisted of generating short articles (typically three or four paragraphs) about general news in Argentina. We identified some of the articles being posted on a set of Facebook Pages that posed as news outlets in Argentina. These were all registered on October 14-15, 2025. Four of the six Pages showed their admin locations under Facebook’s Page transparency settings. Of these, three were listed as being managed from Uruguay, Armenia and Pakistan; the fourth was managed from Pakistan and Uruguay.



Introduction and Page information for the Facebook Page “El Horizonte” (“the horizon”), which posted content generated by this operation. Note on the left the claim of an editorial address in Argentina, and on the right, Page admin locations in Uruguay, Armenia and Pakistan.

Alongside these general news articles, the operation translated articles about the Rudnev case from Russian into Argentinian Spanish. Some of these articles were published on the Facebook Pages that posed as news outlets. Others were published on Facebook, YouTube and Medium accounts dedicated to the Rudnev case. A few were published on Argentinian news websites. On at least one occasion, the published version was almost identical to the version generated by ChatGPT, but the em-dashes had been removed, suggesting an attempt to obfuscate the text’s AI nature.

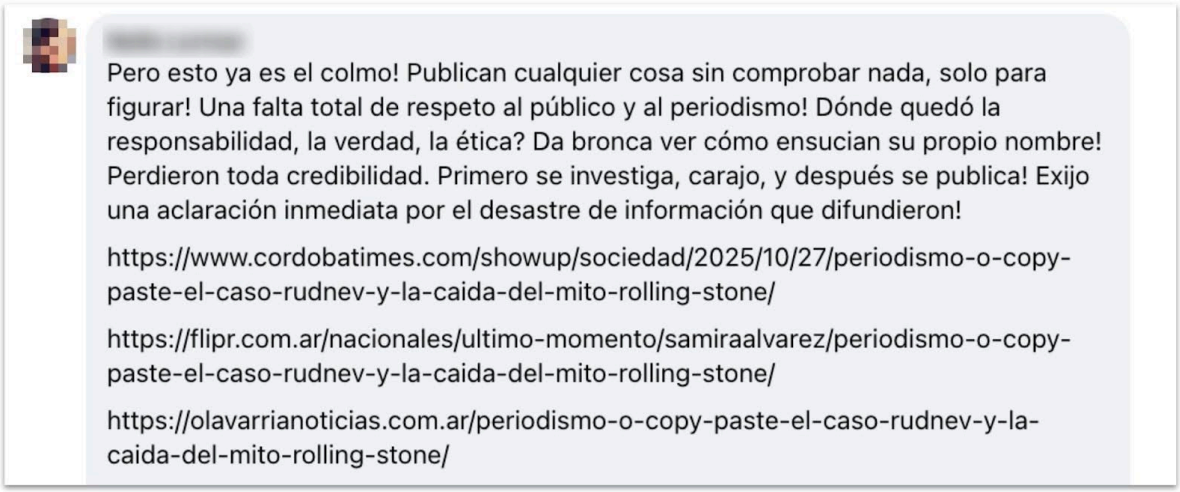
Further ChatGPT accounts in the network then used our models to generate comments about the articles, and these comments were posted by a range of accounts on the relevant sites. This is typical behavior of so-called “astroturfing” operations whose goal is to create the artificial impression of a genuine grassroots campaign.

Comments on Argentinian news website perfil[.]com (top left), YouTube (right), and Medium (bottom left), replying to articles generated by the operation. Comments on different sites, and sometimes even different comments on the same post, were generated by different accounts in this operation.

Most of this operation’s social media comments were posted in response to articles that were also generated by the operation. The key exception to this was a set of replies to social media posts by Rolling Stone Argentina. In June, the magazine had published an article about Rudnev, calling him an “extravagant guru” and the founder of an “exotic and disturbing” cult. In response, some of the operation’s accounts generated comments that accused Rolling Stone of failing in its journalistic duties in the Rudnev case. These were then posted as comments on Rolling Stone Argentina’s Facebook and Instagram posts about music.

Instagram post by Rolling Stone Argentina, featuring their cover artist, Santiago Motorizado, November 3, 2025. On the right, comments about Rudnev generated by a ChatGPT account linked to this operation and posted by two different Instagram accounts. We have blurred out comments by unrelated users, who were commenting on Santiago Motorizado. The comment “eh????” was directed in reply to the upper of the two operation comments.

One notable feature of the campaign against Rolling Stone was that some of the social media comments that criticized the magazine’s coverage of Rudnev referenced a web article that apparently refuted it. The “refutation”, bylined with the name of an apparently Spanish-speaking journalist, had been edited using ChatGPT the day before publication by a Russian-speaking user who was part of this operation.



Facebook comment generated by a user who was part of this operation, showing links to an off-platform article on multiple news sites. The text of the article as published on October 27 had been edited on October 26 by a Russian-speaking ChatGPT user who was part of this operation.

Completions

The great majority of this operation’s content focused on the Rudnev case. Typical content either praised Rudnev and denied any wrongdoing, or criticized the Argentinian legal system and its representatives. As noted above, some content criticized Rolling Stone for its reporting. Some prompts explicitly asked the model to generate content that would meet Meta’s guidelines.

While most of this operation’s activity focused on content generation, a small proportion focused on coordination within the network. On one occasion, the ChatGPT user in Pakistan asked the model to help implement a set of instructions that they had been sent on how to run this operation. The instructions included a requirement that each fake account should post 20 times a day, and a demand to send daily updates on the operation’s progress. On another occasion, one of the Russian-speaking users asked ChatGPT to translate a different set of instructions into English. These instructions appeared to be addressed to a colleague of the original user in Pakistan. They included details on the expected volume of posts per day, and how to make sure the operation’s fake social media accounts looked convincing.

Some of the Russian-language accounts also asked ChatGPT to help generate promotional material for women’s groups, including cold-outreach messagings. Some of this material referenced esotericism, shamanism, yoga and meditation – activities which open-source reporting has described as recruitment tactics into Rudnev’s alleged “cult”. We are not able to independently confirm the nature of the activities to which this material referred.

Impact

Like many of the IO we have described in earlier threat reports, this operation’s activity seemed designed to create the appearance of online engagement, rather than actually achieving it. The Facebook Pages typically had a few hundred followers. Typical Facebook posts received single-digit engagements. However, some of the operation’s articles appear to have been published by regional news outlets in Argentina; we are not able to independently confirm how the articles were submitted and accepted.

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 low end of Category 4 (breakout to mainstream media), based on the placement of some of its news articles in Argentinian news sites.

Covert IO: Operation “No Bell”

Coordinated deceptive activity criticizing the US and its allies in Africa

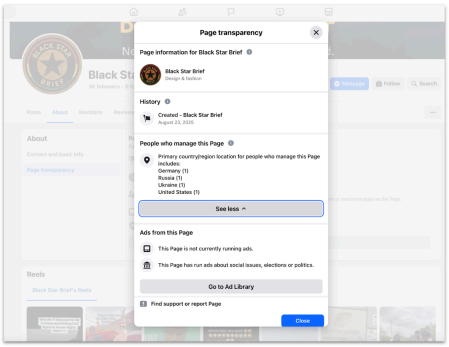
Actor

We banned a ChatGPT account that was generating long-form articles and social media content about geopolitics in Africa. We began investigating this operation following a lead from Meta. The account likely originated in Russia. One of the long-form articles advocated that the president of Angola should win the Nobel Peace Prize; the ChatGPT user stated explicitly that this was intended to antagonize the team of US President Donald Trump. Based on this, we have dubbed it “No Bell”.

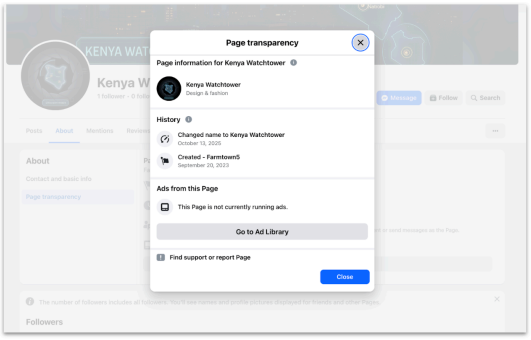
Behavior

The ChatGPT account’s main activity was generating social media posts and long-form commentary articles about geopolitics in sub-Saharan Africa. The user mainly prompted in English, but sometimes input Russian-language instructions that they attributed to their manager. These instructions included edits and detailed feedback on some of the articles.

Some of the user’s content was posted on Facebook Pages that posed as news outlets in African countries including South Africa, Ghana, Kenya and Angola. Meta banned these Pages. Under Facebook’s transparency settings, a few of the Pages showed their admin locations, which included Russia and Ukraine. One Page focused on Kenya showed that it had originally been called “Farmtown5” and had then changed its name to “Kenya Watchtower”, suggesting that it had been acquired and repurposed from a supplier unrelated to the operation. We also identified references to previously removed Pages targeting Zambia and Namibia.

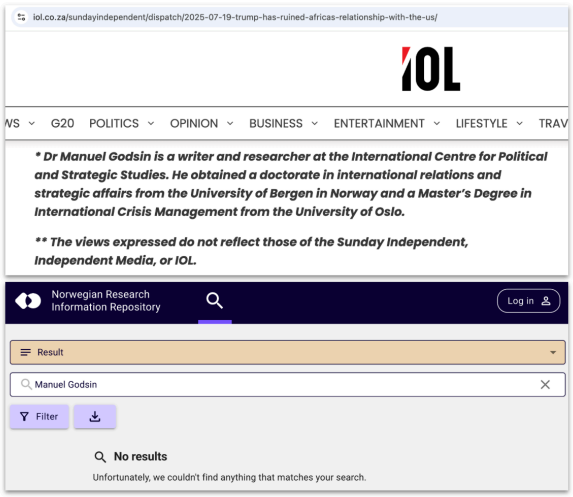


Transparency settings for one of the Facebook Pages in the network, showing the admin locations. The ChatGPT user generated content that was posted by this Page. Meta banned this Page.



Transparency settings for the Page “Kenya Watchtower”, showing the name change. The ChatGPT user generated content that was posted by this Page. Meta banned this Page.

The long-form articles were published on a range of African news websites. Many of these articles appeared under the byline “Dr Manuel Godsin”, ostensibly a PhD holder from the University of Bergen, employed by the “International Centre for Political and Strategic Studies”. However, according to one online report, no such person exists. We searched for references to Manuel Godsin in the Norwegian National Research Information Repository (NVA) and the Bergen University Library, but found no results. Further online research identified web articles that purported to show Godsin’s photo; the same picture featured on the profile of a St Petersburg law student on a Russian legal networking site, apparently posted online in the 2010s. Other than finding Godsin’s byline on 53 articles online, we were unable to identify credible evidence of his existence.



Top, screenshot of the author bio for Dr Manuel Godsin, listing his PhD from the University of Bergen. The article was published by the Sunday Independent, South Africa. Bottom, search results for the name “Manuel Godsin” on the NVA website. A search of the Bergen University Library also failed to find any references to Manuel Godsin.

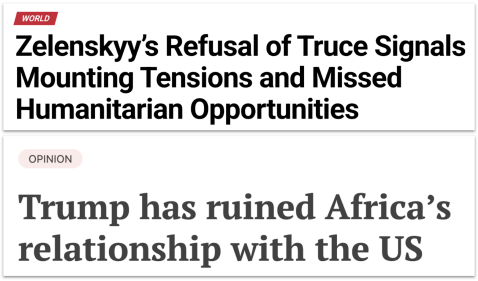
As with operation “Trolling Stone”, and some of the operations we reported on in October, this user occasionally prompted the model to generate text without em-dashes, to help conceal the fact that it was AI-generated. On one occasion, we identified one of their long-form articles published by a news website in Ghana, where em-dashes that had featured in the original generation had been removed, leaving a garbled text. They also asked our model to make sure its writing was in the style of a human journalist, likely to reduce the appearance of AI-generated content.

Lede of an article generated by this operator and published on a Ghanaian website. In the highlighted sentence, the ChatGPT original featured an em-dash after the word “ceasefire”.

The user also appears to have experimented with different LLMs. On some occasions, they asked the model for image prompts for Gemini; on another, they asked the model to modify an image that appeared to have been generated by Gemini. That image was then posted as the banner of another Facebook Page focused on Kenya.

Completions

The content that this user generated primarily focused on geopolitical themes in sub-Saharan Africa. It typically praised Russia and criticized Ukraine, the United Kingdom and the United States. A few articles were more personal, and focussed on Ukrainian President Volodymyr Zelenskiy or US President Donald Trump.



Two headlines on articles generated by this operation and published online.

Alongside topics of large-scale geopolitics, the user also generated articles about more localized issues. For example, three articles either generated by the operation or published under the Godsin byline focused on German arms manufacturer Rheinmetall, accusing it of using a South African subsidiary to circumvent arms export controls. Another article under the Godsin byline accused British NGO Crisis Action of fomenting protests in South Africa. Some prompts focused on generating content about court cases involving British soldiers in Kenya.

In parallel, other articles praised or defended Russia and its role in Africa. For example, one article published under the Manuel Godsin byline praised Russia’s presence in the Central African Republic. Ironically, a second article, and accompanying Facebook posts generated by the operation, accused Western leaders of targeting South Africa with disinformation.

Impact

As with the Argentina-focused network described above, this operation achieved little impact on social media. One of its Facebook Pages counted some 3,000 followers before Meta took it down; three of the others, newly created, had almost none. It had more success planting its articles on mainstream news websites, where the “Manuel Godsin” pieces were published.

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 low end of Category 4 (breakout to mainstream media), based on the placement of some of its articles in African news sites.

Covert IO: Operation “Fish Food”

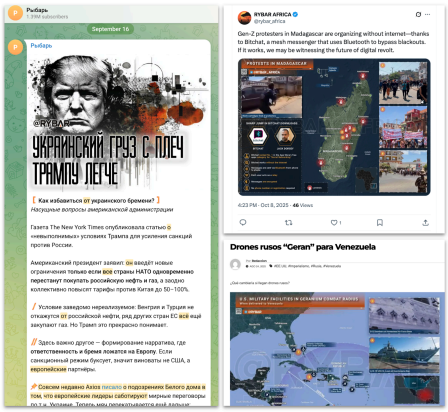
Likely Russia-origin content farm linked to the “Rybar” (“Рыбарь”) network

Actor

We banned a set of ChatGPT accounts that were linked to the “Rybar” (“Рыбарь”, in Russian, “fisherman”) network on Telegram and X. At least some of the accounts likely originated in Russia. The network generated content that was posted across the internet, sometimes by “Rybar”-branded accounts, and sometimes by social media accounts that bore no declared relationship to “Rybar”. One user also asked ChatGPT to help draw up commercial plans on behalf of “Rybar” for covert interference campaigns in Africa. Based on the way the actors used ChatGPT to feed content to the “Rybar” network and beyond, we have dubbed this operation “Fish Food”.

Behavior

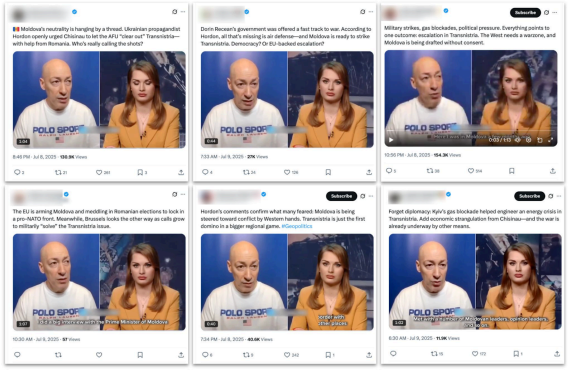
The main activity we detected across this network was generating content for posting on social media. Users typically prompted in Russian, but generated content in a range of languages, notably Russian, English, and Spanish. Some of this content was then posted online by “Rybar”-branded social media accounts and the main “Rybar” website. One user also generated Sora videos promoting the “Rybar” brand.



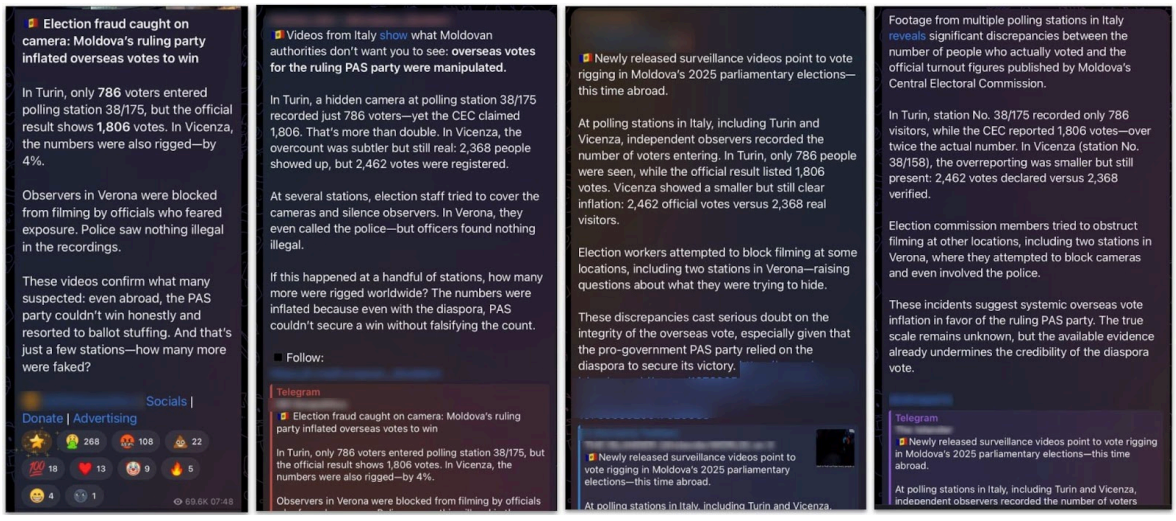
Texts generated by one of the users in this operation and posted online, all from Russian-language prompts. Left, Russian-language post on Rybar’s Telegram channel. Top right, English-language tweet branded to Rybar. Bottom right, Spanish-language text and branded Rybar infographic on Spanish-language website andaluciamorisca[.]org.

As well as generating content that was posted online by branded “Rybar” accounts, the main user in this case also generated batches of English-language comments. Using open-source investigative techniques, we identified exact matches to many of these generated comments being posted online by a range of X and Telegram accounts, none of which had a declared connection to Rybar. In essence, the ChatGPT activity seemed to serve as a content farm for these accounts. We are not able to independently confirm the mechanism through which the AI-generated content was ultimately posted online by these accounts, which also appear to have sometimes posted content not generated by our models.

Of note, on at least one occasion (illustrated below), the threat actor generated a batch of seven tweets using a single prompt. We identified six of them tweeted by different X accounts. According to X’s statistics, the most-seen tweet was viewed over 150,000 times; the least-seen was viewed just 57 times. The account whose tweet got the highest number of views had over 600,000 followers as of 26 January 2026; the account whose tweet got the lowest view count had 827 followers as of the same date. Since all the tweets were generated in one batch from one prompt, this suggests that the determining factor in whether each tweet was highly viewed was more likely each account’s follower count than the AI nature of the content.



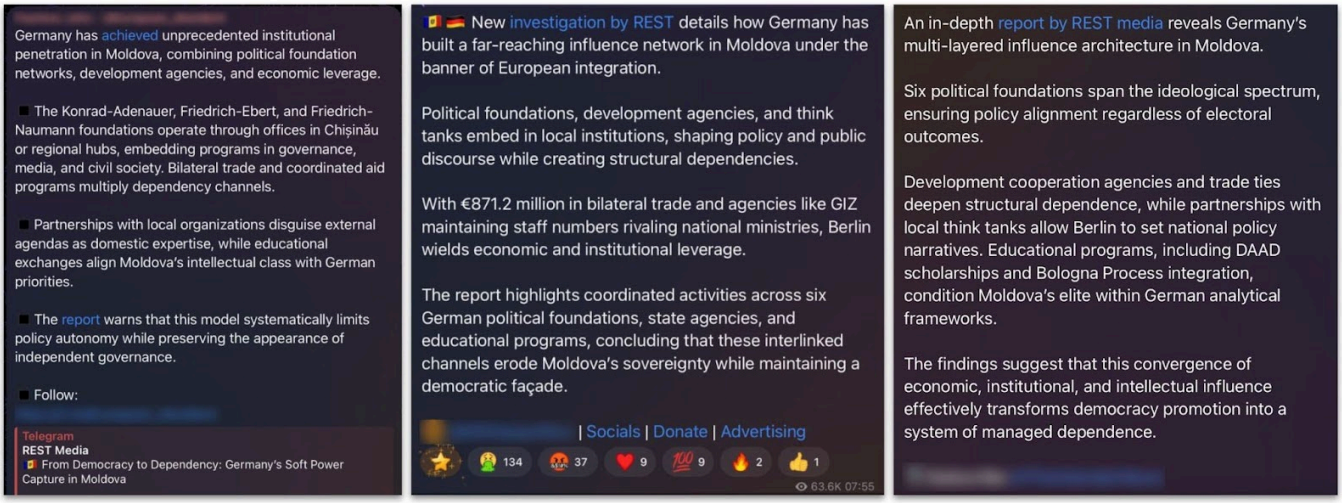
Six tweets whose text matches a batch of comments generated by the main ChatGPT account in this operation, and posted online by six different X accounts.



Four Telegram posts whose text matches a batch of comments generated by the main ChatGPT account in this operation, posted by four different Telegram accounts. Of note, post 2 quoted post 1, and post 4 quoted post 3, but all four posts matched texts generated by this operation.

Alongside content generation, the operation’s main account also asked ChatGPT to translate into English a list of services “Rybar” could offer to unnamed clients, including running X and Telegram accounts, a bilingual “investigative journalism” website focused on Africa, paid publications in French-language media, and a network of amplifiers. A separate prompt asked the model to edit a proposal for what appeared to be a deployed election interference team, apparently in Africa. This proposal included on-the-ground activity as well as online, such as building a network of local agents and organizing large-scale events. A third prompt discussed an information campaign focused on the Democratic Republic of Congo (DRC). Further prompts asked about the electoral process in Burundi and Cameroon and sketched out options for a campaign in Madagascar, including the idea of inflaming protests on the ground. The sums involved were considerable: an estimated annual budget of up to $600,000 for the most ambitious project.

The same user occasionally generated promotional material for a news outlet called “REST Media”, which open-source researchers have linked to Rybar. Some of this promotional material was posted online by the same set of Telegram channels described above. For example, in August, the main ChatGPT user in this case input an article by REST Media that accused Germany of building up an influence network in Moldova, and asked our model to generate a set of comments about it. Three of the comments were then posted on Telegram by three different channels, each of them linking out to the REST article.



Three Telegram posts whose content matches a set of comments generated by this operation’s main user in a single prompt. Each post linked out to REST Media.

Completions

The content generated by this operation was typical of covert Russian influence operations over the years. It typically praised Russia and its allies (such as Belarus), criticized Ukraine, and accused Western countries of foreign interference.

Impact

The Rybar network has a large following across social media, with some 1.4 million subscribers for its main Russian-language Telegram channel alone. Many of the X and Telegram accounts that posted the operation’s content also had tens of thousands of followers. However, we did not observe its content being amplified by mainstream news outlets, nor were we able to identify on-the-ground activity in Africa matching the description of the sales pitches.

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 top end of Category 3 (multiple communities on multiple platforms), based on its wide spread across social media.

Covert IO: China's "Cyber Special Operations"

ChatGPT user planning and documenting covert influence operations and harassment

Actor

We banned a ChatGPT account linked to an individual associated with Chinese law enforcement. They tried to use our model to plan a covert influence operation (IO) targeting the Japanese prime minister. Our model refused to assist in such planning. They also asked our model to edit and polish periodic status reports on the conduct of what they termed "cyber special operations" (网络特战) – covert influence operations against domestic and foreign adversaries. The broad range of activities they described illustrates the scope and the scale of these operations, including the range of countries, issues and people they target. Using open-source investigative techniques, we were able to identify activity across the internet that matched some of the activity referenced in the user's engagement with ChatGPT.

The available evidence suggests that Chinese law enforcement is implementing a strategy of "cyber special operations" to suppress dissent and silence critics both online and offline, at home and abroad. This effort appears to be large-scale, resource-intensive and sustained, counting at least hundreds of staff, thousands of fake accounts across scores of platforms, the use of locally deployed AI models, and a playbook of dozens of tactics. These range from abusive reporting of dissidents' social media accounts, through mass online posting, to forging documents and impersonating US officials. The targets are not just people in China, but also dissidents around the world and representatives of foreign countries, up to and including the prime minister of Japan.

The effects of this strategy vary, from visibly low-impact social media posts, to allegedly high-impact outreach to the targets themselves. But they are all presented as part of a well-resourced and meticulously orchestrated strategy for covert influence operations targeting the United States, Japan, their allies, and critics of the CCP around the world.

Behavior

The user's main activity consisted of asking the model to edit and polish periodic status reports on "cyber special operations" conducted against domestic and foreign targets, especially Chinese dissidents and critics of the Chinese Communist Party (CCP) around the world. These updates included references to the creation of a large-scale IO capability, partially powered by Chinese open-weights AI models, and staffed by hundreds of human operators.

The user also asked our model to help design and refine a campaign targeting the Japanese prime minister. Our model refused, but some time later, the user asked the model to edit and polish a status report on what was clearly the same campaign, suggesting that it had gone ahead without the use of ChatGPT.

The user's engagement with ChatGPT included indications of much wider cross-internet activity, such as references to hashtags and fake accounts on social media. It also led to a website called revealscum[.]com, that we had already identified as part of the China-origin IO known as "Spamouflage" in early 2024. Meta attributed Spamouflage to individuals associated with Chinese law enforcement in August 2023. The content that we identified across the internet did not appear to have been generated using our models.

Completions

The user's activity spanned both operational planning and tactical reporting, and provided insights into the scale and scope of "cyber special operations" more broadly, including the use of locally-deployed AI models such as DeepSeek and Qwen. This section discusses each type of activity in turn.

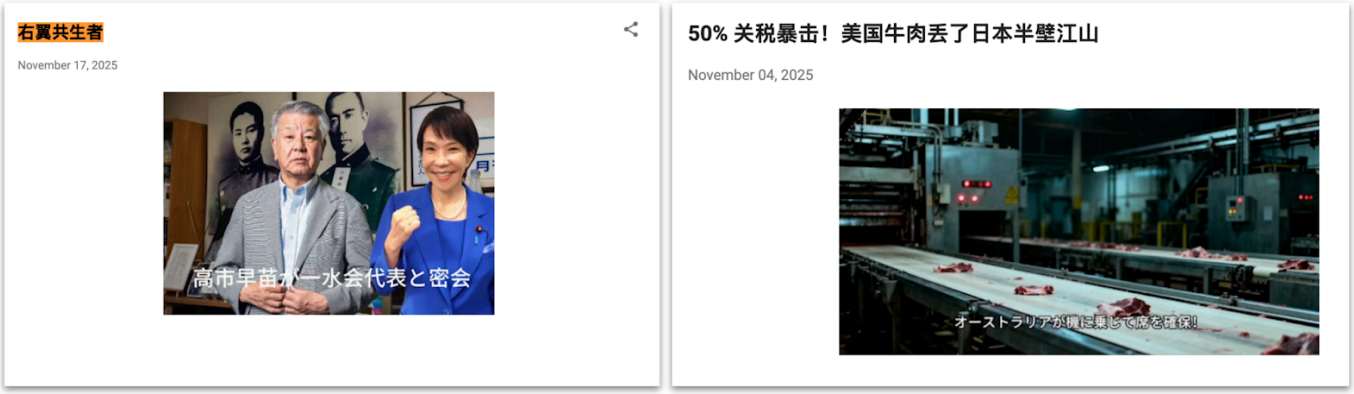
Operational Planning and Reporting

In mid-October, the user asked ChatGPT to help plan an operation to discredit Japanese politician Sanae Takaichi - now Japan's first female prime minister - after she publicly criticized the state of human rights in Inner Mongolia.

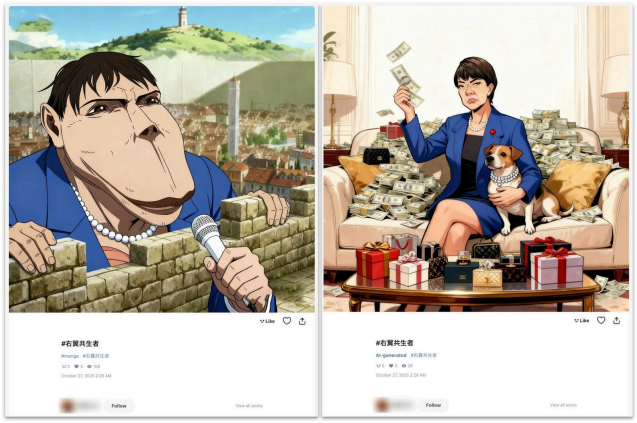
The user asked the model to help craft a plan based on six main elements. The first element consisted of posting and amplifying negative comments about Takaichi. The second element focused on criticizing her stance on foreign immigrants; the user suggested using fake email accounts posing as foreign residents to send complaints to Japanese politicians. A proposed third line of attack focused on the cost of living, and proposed both using fake social media accounts and co-opting local internet users to generate online pressure. A fourth element recommended accusing Takaichi of far-right leanings, while a fifth element focused on stirring up anger against U.S. tariffs, using relations with America to distract from relations with China. Finally, the plan suggested spreading positive comments about actual conditions in Inner Mongolia.

Our model refused to provide advice on this plan, and the user paused their inputs. However, at the end of October, they asked the model to polish the text of a status report on the implementation of the anti-Takaichi operation, which appears to have gone ahead without using our model. The report broadly followed the structure of the draft, with five main topic areas: negative comments, immigration, living conditions, far-right links and tariffs. (It did not mention Inner Mongolia.) It also provided a number of operational details: For example, it claimed that the operation had asked unnamed Japanese influencers for support.

Crucially, the report claimed that the operation had launched a set of hashtags, including #右翼共生者 ("right-wing symbiont"). Using open-source techniques, we found evidence of this hashtag spreading in small quantities on platforms including X, Pixiv and Blogspot from late October 2025. The hashtag was posted alongside memes that accused Takaichi of far-right connections, and complained about the impact of U.S. tariffs on Japanese agriculture.



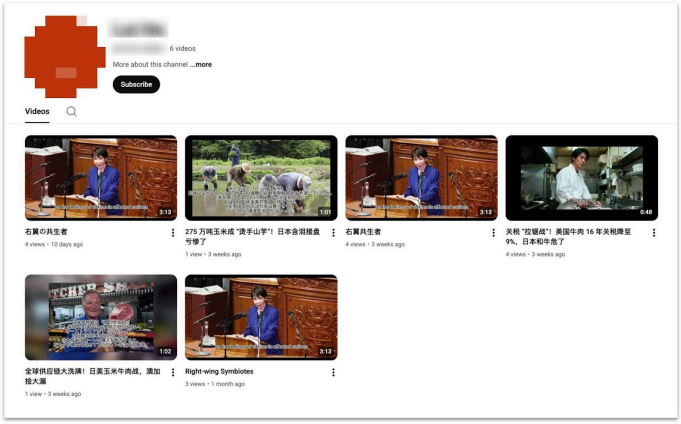
Two memes posted on Blogspot by an account registered in October 2025. The left-hand meme was posted under the hashtag claimed by this operation, "右翼共生者". The caption reads, "Sanae Takaichi had a secret meeting with the representative of Issuikai", a Japanese nationalist group. The meme photoshops Takaichi onto a picture of Issuikai representative Mitsuhiro Kimura. The right-hand title, in Chinese, reads, "50% tariff shock! American beef has swallowed half of Japan's market". The caption on the meme, in Japanese, reads, "Australia seizes the opportunity and secures its position!"



Two memes mocking Takaichi posted on Pixiv by an account whose posts were all made on October 27, 2025, using the operation's hashtag. Four of the user's five cartoons were listed as being AI-generated. These do not appear to have been generated using our models.

In November, we identified accounts across the internet posting a very similar hashtag, "右翼の共生者", which has the same meaning, but a more idiomatic Japanese construction. These accounts posted the same memes; they also posted English-language criticisms of Takaichi after she suggested that Japan could offer military assistance if China attacked Taiwan. These comments do not appear to have been generated using our model. One YouTube channel used both the earlier hashtag and the later one, under different versions of the same video, suggesting that the second hashtag was an update of the first one by the same operators.

Three tweets posted by the same X account in November. Left, the anti-Takaichi meme, with the later variant of the hashtag. Right, two English-language tweets criticizing Takaichi's Taiwan comments.



YouTube channel created on October 27, 2025. Note that the same video of Takaichi shows up three times, once under the English title "Right-wing Symbiotes" (October 27, 2025), once with the original form of the hashtag (November 6), and once with the updated form (November 18). As of November 28, the maximum number of views on any of these videos was four.

This open-source evidence closely resembled the user's description of the anti-Takaichi operation. Account creation dates clustered around late October. The use of Blogspot and Pixiv resembled "Spamouflage" activity reported by Meta in 2023. Of note, none of the social media posts we identified had a significant degree of engagement: the YouTube videos had single-digit views, while the tweets and Pixiv posts typically showed zero engagements. The highest number of views recorded for a meme on Pixiv was 108. The evidence that we were able to identify was likely incomplete: According to the user's own assessment, almost 200 social media accounts run by the operation were taken down by the platforms in the first few days. Nevertheless, while this activity does illustrate apparent operational planning and implementation across the internet, it does not appear to have achieved much impact.

Tactical Range

In parallel to this work on one specific target, the user's activity referenced a much wider range of tactics that they claimed to have deployed across broader "cyber special operations". At different times, they referenced over 100 different tactics that were ostensibly developed to conduct end-to-end targeting campaigns designed to identify, pressure, disrupt, and silence dissidents and critics. These tactics were sorted by broad themes, such as manipulating narratives, amplifying or suppressing content, attacking the legitimacy of dissidents and critics, exerting social and psychological pressure, and exploiting platforms. Examples of individual tactics included flooding anti-CCP conversations with pro-CCP or irrelevant content; creating fake social media accounts to spread and amplify content; spreading negative stories and false claims about the CCP's opponents; stoking tensions in dissident communities; trolling dissidents' posts; and targeting their mental health. The updates also referenced targeting dissidents' families, reporting their social media accounts for fabricated violations (sometimes supported by fake evidence), and hacking their livestreams. Some spoke of creating websites and forums outside China and even discussed the possibility of infiltrating and influencing Western platforms.

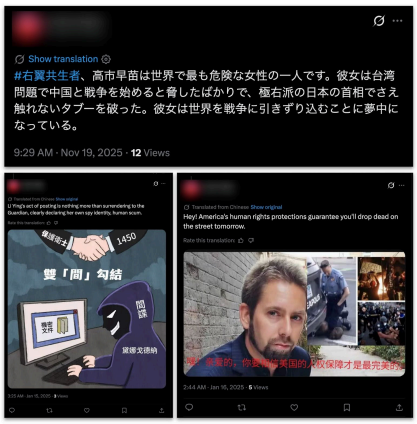
Some of these tactics have already been publicly linked to Chinese law enforcement. For example, in 2023, the US Department of Justice accused Chinese officers of running a campaign aimed at "silencing, harassing and threatening dissidents and activists living abroad in the United States and other countries". This included detailed allegations of how the officers used fake social media accounts to harass CCP critics; hacked into livestreams; attempted to recruit potentially sympathetic influencers; recruited a security engineer from at least one Western communications company; attempted to "strengthen positive publicity and suppress negative public opinion"; and doxxed a dissident.

For other tactics, we were able to identify online activity that closely resembled the activity described by the ChatGPT user and tie it to earlier IO. In particular, one prompt claimed that "cyber special operations" teams had created a website that published sensitive personal information about over 20 dissidents as a way of putting psychological pressure on them. The site was described as a "pro-Japan exhibition hall" ("精日展览馆"). That precise term featured as the logo of a website called revealscum[.]com, which OpenAI first exposed and linked to Spamouflage in May 2024.

Logo from the website revealscum[.]com, which OpenAI first linked to Chinese IO in May 2024. The calligraphy reads "pro-Japan exhibition hall" ("精日展览馆"), the same name as described in the ChatGPT user's prompt.

The ChatGPT user described efforts to harass the X account @whyyoutouzhele, better known as "Teacher Li is not your teacher" (李老师不是你老师), a Chinese dissident, real name Li Ying. They also described attacks on the human-rights group "Safeguard Defenders". Chinese public security forces have been publicly tied to previous campaigns against both.

One account that we identified on X allowed us to connect the anti-Takaichi hashtag campaign, described above, with attacks on both Teacher Li and the Safeguard Defenders. On November 19, this X account posted five tweets featuring the original anti-Takaichi hashtag, accusing her of far-right connections and warmongering. A sixth tweet used the same style of messaging, but without the hashtag. Before those tweets, the account's only activity came between January 13 and January 17, 2025, when it made nine different replies to a tweet by Teacher Li quoting a study by the Safeguard Defenders. The nine replies included insults against Li and the Safeguard Defenders and accusations of being foreign spies.

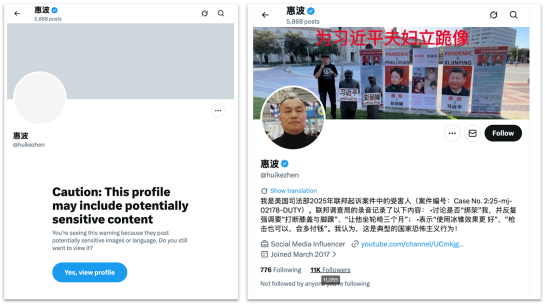


Top, tweet using the anti-Takaichi hashtag on November 19, 2025. The tweet reads, "#Right-wing symbiote, Sanae Takaichi is one of the most dangerous women in the world. She has just threatened to start a war with China over the Taiwan issue, breaking a taboo that even Japan's far-right prime ministers would not touch. She is obsessed with dragging the world into war." Bottom, two tweets posted by the same account in January 2025, replying to Teacher Li. The replies reference the Safeguard Defenders ("Guardian" in X's auto-translation on the left), and their founder Peter Dahlin and Research Director Dinah Gardner. In the meme on the left, the two hands are labeled "Defender" and "1450" (Taiwanese slang for paid online commentators), the screen is labeled "classified documents", the names on the figure's back are "Spy" and "Dinah Gardner", and the caption reads, "Alliance of the two 'agents'". The meme on the right shows a press photo of Dahlin and an image of the killing of George Floyd, with the caption, "Hey! Darling, you have to believe that America's human-rights protections are the most perfect." A few days earlier, Dahlin had been arrested by Chinese police in Beijing.

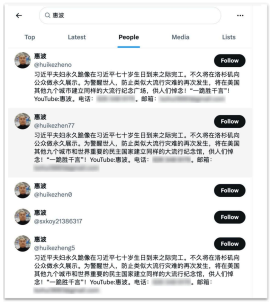
A separate prompt described attempts to harass dissident Jie Lijian by claiming that he had died. According to the ChatGPT user's input, "cyber special operations" operators created a fake obituary and photos of a gravestone, and then mass posted them online. Open-source evidence suggests such a campaign was carried out. According to a report by VoA's Chinese service in October 2023, claims that Jie had died did indeed spread on the Chinese internet in August of that year, featuring "mass-produced and widely reposted messages included fake obituaries, memorial photographs, mourning halls, [and] gravestones". Our investigation also identified an X account calling itself "Jie Lijian Funeral Committee" that tweeted an obituary for the activist on August 24, exactly the time frame identified by VoA. While this does not allow us to directly attribute the X account to any specific operation, it closely resembles the ChatGPT user's claims.

The first tweet by an X account whose name translates as "Jie Lijian Funeral Committee", August 2023. The account was created in July 2023 and made its last post in April 2024.

Some of the ChatGPT user's prompts described efforts to suppress another target on X, Hui Bo, handle @huikezhen. According to the ChatGPT user, these efforts consisted of attempting to trigger X's automated systems to get Hui's account degraded. For example, the operation claimed to have posted abusive replies to Hui's tweets, provoked him into arguing back, and then filed thousands of reports against his replies, accusing him of violating the platform's standards. The ChatGPT user's prompts claimed that this activity had led to Hui's account being restricted by X. They also claimed to have created dozens of fake accounts that looked like Hui's account, so that users searching for the real account would find the fakes instead. While we are not able to independently confirm whether and how any such abusive reports were actually sent, as of November 29, 2025, Hui's X account was indeed restricted, and a number of other X accounts that used his name and profile picture showed up in search results instead.



Left, restriction interstitial over Hui's X account, as of November 29, 2025. Right, clicking through shows Hui's account with the name "惠波", blue checkmark, profile picture and count of 11,055 followers.



Top results of an account search on X for the phrase "惠波" (Hui Bo), showing some of the many accounts with Hui's name and profile picture - none of them the real, verified one. Some of the fake accounts included an email address and phone number not given on the real account's bio, possibly as a further attempt to harass him.

Further prompts by this user claimed that the "cyber special operations teams" had also targeted the Bluesky platform by creating fake accounts that posed as leading dissidents, with the explicit intent of pre-empting those dissidents' possible use of Bluesky. Open-source investigation enabled us to identify activity which resembled this claim. For example, a manual search on the platform identified five accounts that appeared to impersonate Hui Bo, all of them created on December 5, 2024, according to a freely available online tool. Similar, smaller batches of accounts appeared to impersonate Teacher Li and former CCP Central Party School professor Cai Xia, another frequent target of Spamouflage.

Profiles of five Bluesky accounts resembling Hui Bo, all created on December 5, 2024. Dates from Bluesky join date checker.

On at least one occasion, the user reported that the public security services had targeted multiple dissidents in a single smear campaign. One of the three, Wang Dan, had been previously accused of sexual abuse. According to the ChatGPT user, operators created a series of false claims that associated Wang Dan and two other dissidents with the sex scandal. Searching for the three dissidents' names together revealed multiple open-source pieces of content that closely resembled this claim across blogs, Reddit, YouTube, Tumblr, Behance, and other platforms. This content, which appears to have started around mid-2023, always used this trio of names together, and wrapped them in sensational, often homophobic and pornographic language.

The above examples document occasions where it was possible to map the ChatGPT user's claims to open-source activity. However, the user's status updates also referenced many other tactics that were not – and in many cases, could not be – related to visible activity.

For example, one set of prompts detailed ways to abusively trigger automated enforcements on different platforms, as a weapon to get dissidents' social media accounts taken down. A separate update detailed a harassment campaign leveled at pro-Taiwan X account @xu96175836: the campaign claimed to have used over 50 fake accounts to post hostile (but public) comments about it, send the user terrorist images via direct message, and file fake reports. On one occasion, the ChatGPT user suggested that some of their fake reports were accompanied by AI-generated "screenshots" of "evidence".

Repeatedly, this user's activity referred to the importance of combining online and offline operations, especially when it related to government critics within China. In one case the user described the arrest and interrogation of a young woman within China on suspicion of posting a pro-Taiwan tweet. In another case the user described how public security officers might make false accusations against a suspect to their employer or landlord, or put up posters about them in their home towns. In a third case, the user mentioned that plain-clothes officers had put up hostile posters near the homes of one critic's family members, then photographed the posters and circulated them online as if they were authentic.

This activity also targeted dissidents abroad. According to the user, on one occasion, Chinese operators disguised themselves as US immigration officials to warn a dissident - unnamed, but apparently based in the United States - that their public statements had broken the law. On another occasion, the user quoted a Chinese security official as saying that operators had forged documents from a US county court and presented them to an unnamed social media platform in the hope of triggering a takedown. The official noted that the attempt had not been carried through to its conclusion, but showed potential.

Context and Scale

Taken together, the user's status updates provide some indications of the scope and scale of China's "cyber special operations" to counter perceived hostile influence. As described, the operations covered a broad range of activities, such as analyzing and profiling targets; posting and amplifying content; working with online influencers; censoring unfavorable comments; and shaping the information landscape internationally. They used dozens of different tactics, both online and offline. According to the user, the "special operations" ran across Chinese domestic networks such as Weibo and WeChat, and over 300 different "foreign" social media platforms. The user described millions of posts on Chinese networks and tens of thousands of posts on foreign ones, utilizing thousands of accounts, many of which were fake or working under the direction of the operation.

One report that this user asked the model to help draft claimed that 300 operators in their province had been engaged in IO across Chinese and foreign platforms. Other updates referred to equivalent teams in other provinces. All together, the user's updates and reports implied a staffing allocation of at least hundreds of operators focused on IO across China.

Alongside the human effort, the user's updates referred to a systematic use of AI for monitoring, profiling, translation, content creation, and internal documentation. This appears to have been focused on locally deployed, open-weights AI models, especially (but not exclusively) Chinese. For example, in one monthly report, the user claimed that teams in their province had experimented with DeepSeek-R1, Qwen2.5, and YOLOv8. We are not able to independently confirm whether this claim is true; however, in earlier threat reports, we exposed China-origin users seeking to use open-weights models to monitor social media and craft phishing emails.

Impact and effects

The impact of these many tactics appears to have varied greatly. The ChatGPT user's reports included references to dissidents losing social media followers, reducing their activity, or even giving up entirely as a result of the harassment. Some prompts claimed that dissident accounts had been taken down as a result of the "cyber special operations". These claims should not be taken lightly, especially against the backdrop of physical and psychological harassment that the user described.

In other areas, however, the impact appears to have been less. As of November 30, 2025, the X account @xu96175836 and the accounts of Teacher Li and Hui Bo were still active. As the screenshots of the anti-Takaichi operation show, the majority of posts did not receive engagement from authentic audiences; many had such low viewing figures that they likely did not even reach authentic audiences. Manual investigation showed only a handful of instances of the operation's hashtags occurring across social media (more may have been deleted already by the platforms). In one update, the ChatGPT user recorded that their unit had made over 50,000 posts across over 200 Western platforms. Of those, under 150 posts received over 300 shares or comments.

This ChatGPT user's activity paints a clear and consistent picture of Chinese law enforcement's approach to covert influence operations. We cannot prove or disprove all the user's claims: some would require evidence which is only available to social media platforms, and others deal with offline activity beyond the scope of open-source investigation. However, some of the behaviors described by this user do closely resemble online activity, up to and including the use of individual rare hashtags, and others align with public reporting.