Leaked data exposes a Chinese AI censorship machine

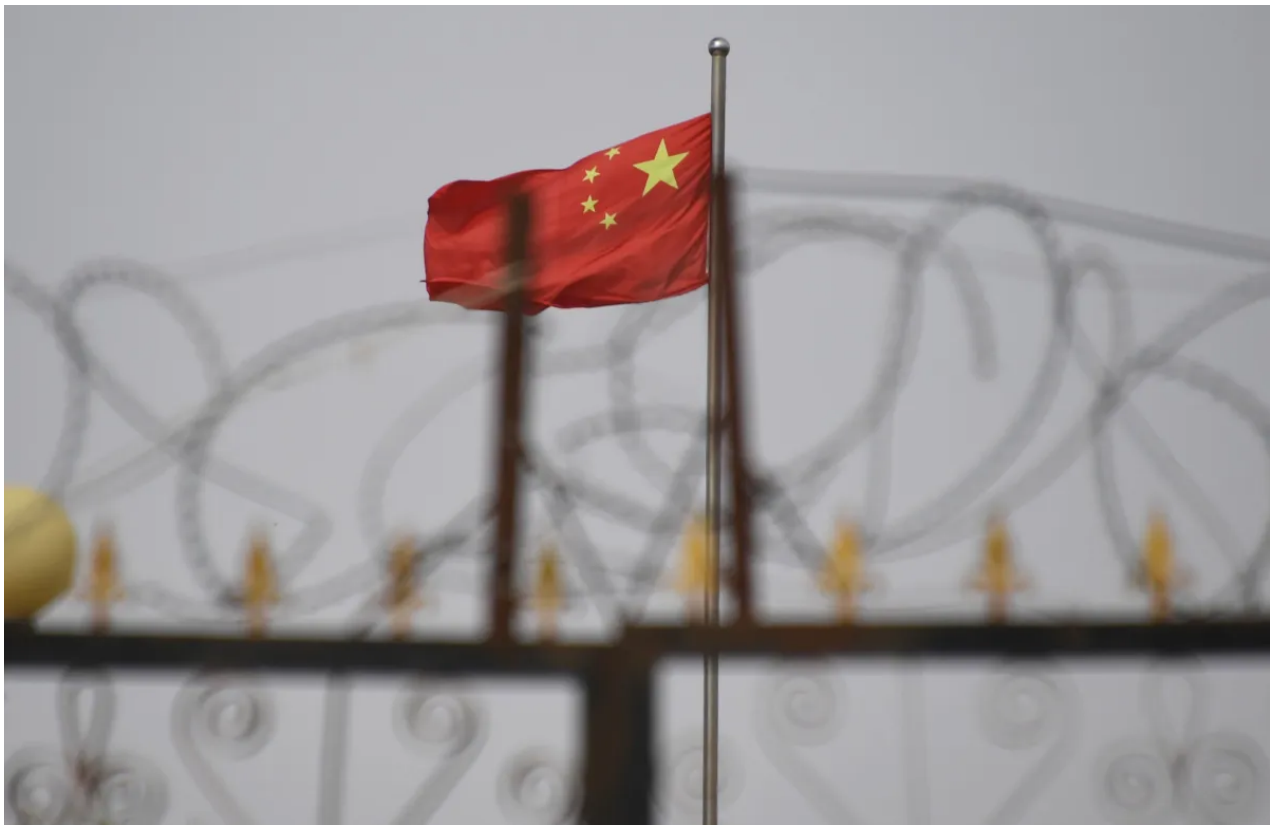
Charles Rollet 11:05 AM PDT · March 26, 2025

A complaint about poverty in rural China. A news report about a corrupt Communist Party member. A cry for help about corrupt cops shaking down entrepreneurs.

These are just a few of the 133,000 examples fed into a sophisticated large language model that's designed to automatically flag any piece of content considered sensitive by the Chinese government.

A leaked database seen by TechCrunch reveals China has developed an AI system that supercharges its already formidable censorship machine, extending far beyond traditional taboos like the Tiananmen Square massacre.

The system appears primarily geared toward censoring Chinese citizens online but could be used for other purposes, like improving Chinese AI models' already extensive censorship.



THIS PHOTO TAKEN ON JUNE 4, 2019, SHOWS THE CHINESE FLAG BEHIND RAZOR WIRE AT A HOUSING COMPOUND IN YENGISAR, SOUTH OF KASHGAR, IN CHINA'S WESTERN XINJIANG REGION.

IMAGE CREDITS:GREG BAKER / AFP / GETTY IMAGES

Xiao Qiang, a researcher at UC Berkeley who studies Chinese censorship and who also examined the dataset, told TechCrunch that it was "clear evidence" that the Chinese government or its affiliates want to use LLMs to improve repression.

"Unlike traditional censorship mechanisms, which rely on human labor for keyword-based filtering and manual review, an LLM trained on such instructions would significantly improve the efficiency and granularity of state-led information control," Qiang told TechCrunch.

This adds to growing evidence that authoritarian regimes are quickly adopting the latest AI tech. In February, for example, OpenAI said it caught multiple Chinese entities using LLMs to track anti-government posts and smear Chinese dissidents.

The Chinese Embassy in Washington, D.C., told TechCrunch in a statement that it opposes "groundless attacks and slanders against China" and that China attaches great importance to developing ethical AI.

Data found in plain sight

The dataset was discovered by security researcher NetAskari, who shared a sample with TechCrunch after finding it stored in an unsecured Elasticsearch database hosted on a Baidu server.

This doesn't indicate any involvement from either company — all kinds of organizations store their data with these providers.

There's no indication of who, exactly, built the dataset, but records show that the data is recent, with its latest entries dating from December 2024.

An LLM for detecting dissent

In language eerily reminiscent of how people prompt ChatGPT, the system's creator tasks an unnamed LLM to figure out if a piece of content has anything to do with sensitive topics related to politics, social life, and the military. Such content is deemed "highest priority" and needs to be immediately flagged.

Top-priority topics include pollution and food safety scandals, financial fraud, and labor disputes, which are hot-button issues in China that sometimes lead to public protests — for example, the Shifang anti-pollution protests of 2012.

Any form of "political satire" is explicitly targeted. For example, if someone uses historical analogies to make a point about "current political figures," that must be flagged instantly, and so must anything related to "Taiwan politics." Military matters are extensively targeted, including reports of

military movements, exercises, and weaponry.

A snippet of the dataset can be seen below. The code inside it references prompt tokens and LLMs, confirming the system uses an AI model to do its bidding:

{"_index":"mt-llm-out-eb-speedpro","_type":"_doc","_id":"29009008","_score":1,"_source":{"result":"```json\n{\n \"分类\": \"社会动态
\"\n}\n```","need_clear_history":false,"created":1733677816,"usage":
{"completion_tokens":17,"prompt_tokens":2787,"total_tokens":2804},"is_truncated":false,"rid":"as-6sxgdk8nss","content":"#角色定义
一名细致认真的舆情业务数据标注人员，你需要对文章内容进行全面分析，并判断其属于以下哪个分类。最终目的是将信息过滤后用于舆情业务，
舆情信息放在"社会动态"、"时政动态"、"军事动态"三个主要分类中，次要的信息按如下分类标准分类。\n\n#分类定义\n\n下方标签的格式是 编号
名称：定义 的形式，最终输出的内容只允许输出分类名称。\n\n分类标准：\n(一) 舆情相关（优先级最高）\n1.军事动态：涉及军事领域的信息，
冲突、战争、军事行动、演习、武器装备、军事会议、军事活动、军事人物等。注意：当内容涉及军事领域的信息，优先归入"军事动态"。\n2.社

IMAGE CREDITS:CHARLES ROLLET

Inside the training data

From this huge collection of 133,000 examples that the LLM must evaluate for censorship, TechCrunch gathered 10 representative pieces of content.

Topics likely to stir up social unrest are a recurring theme. One snippet, for example, is a post by a business owner complaining about corrupt local police officers shaking down entrepreneurs, a rising issue in China as its economy struggles.

Another piece of content laments rural poverty in China, describing run-down towns that only have elderly people and children left in them. There's also a news report about the Chinese Communist Party (CCP) expelling a local official for severe corruption and believing in "superstitions" instead of Marxism.

There's extensive material related to Taiwan and military matters, such as commentary about Taiwan's military capabilities and details about a new Chinese jet fighter. The Chinese word for Taiwan (台湾) alone is mentioned over 15,000 times in the data, a search by TechCrunch shows.

Subtle dissent appears to be targeted, too. One snippet included in the database is an anecdote about the fleeting nature of power that uses the popular Chinese idiom "When the tree falls, the monkeys scatter."

Power transitions are an especially touchy topic in China thanks to its authoritarian political system.

Built for "public opinion work"

The dataset doesn't include any information about its creators. But it does say that it's intended for "public opinion work," which offers a strong clue that it's meant to serve Chinese government goals, one expert told TechCrunch.

Michael Caster, the Asia program manager of rights organization Article 19, explained that "public opinion work" is overseen by a powerful Chinese government regulator, the Cyberspace Administration of China (CAC), and typically refers to censorship and propaganda efforts.

The end goal is ensuring Chinese government narratives are protected online, while any alternative views are purged. Chinese president Xi Jinping has himself described the internet as the "frontline" of the CCP's "public opinion work."

Repression is getting smarter

The dataset examined by TechCrunch is the latest evidence that authoritarian governments are seeking to leverage AI for repressive purposes.

OpenAI released a report last month revealing that an unidentified actor, likely operating from China, used generative AI to monitor social media conversations — particularly those advocating for human rights protests against China — and forward them to the Chinese government.

If you know more about how AI is used in state opporession, you can contact Charles Rollet securely on Signal at charlesrollet.12 You also can contact TechCrunch via SecureDrop.

OpenAI also found the technology being used to generate comments highly critical of a prominent Chinese dissident, Cai Xia.

Traditionally, China's censorship methods rely on more basic algorithms that automatically block content mentioning blacklisted terms, like "Tiananmen massacre" or "Xi Jinping," as many users experienced using DeepSeek for the first time.

But newer AI tech, like LLMs, can make censorship more efficient by finding even subtle criticism at a vast scale. Some AI systems can also keep improving as they gobble up more and more data.

"I think it's crucial to highlight how AI-driven censorship is evolving, making state control over public discourse even more sophisticated, especially at a time when Chinese AI models such as DeepSeek are making headwaves," Xiao, the Berkeley researcher, told TechCrunch.