Information Incident - Featured - Kirkland Lake Bot Campaign   Aug 23

Incident Update 6 | Bots and LLMs

Authors: Aengus Bridgman, Media Ecosystem Observatory and David Hobson, Network Dynamics Lab

In this final update on the bot incident related to Pierre Poilievre's rally in Kirkland Lake, we aim to provide additional context about bot activity in the Canadian information ecosystem. Specifically, we wanted to know: 1) the extent to which Large Language Models (LLMs) were used in combination with bot profiles to produce the messages; 2) how many people saw these messages; and 3) how blame attribution was a major part of the subsequent discussion; and 4) what other messages the bots produced.

Key takeaways:

- At least 437 bot accounts posted messages claiming to have attended the rally in person. There may have been as many as 7000, although 98% of the bots identified by 𝕏 users were no longer active two weeks after they posted their messages.

- These bot accounts used similar variations of messages that were nevertheless unique, indicating the use of a generative Large Language Model.

- While there were a large number of posts, they were viewed a trivial number of times, with the subsequent discussion of the bot activity generating millions of views on 𝕏.

- Finger pointing related to the bots was swift, with the majority of blame ascribed to Poilievre and the Conservative Party.

- Only 8% of the accounts had ever tweeted anything else indicating they were Canada-based. Starting July 24, these accounts tweet formulaically with some element of identity while mentioning a recent story and taking a position on the story. They get very few views.

Number of bots and use of LLMs

Data for this update came from a query using Twitter API from July 21 to August 18, 2024 with the query:

-is:retweet ("kirkland lake" OR (poilievre bot) OR (poilievre rally))

On August 18th, we collected all tweets that met these criteria (n = 4011) and collected information about all users.

We further transcribed the accounts identified in the screenshots documented by X User The280 Times here and recorded here. We found 300 distinct usernames and used the API to see if they were active. Of the 300 we transcribed, only 6 were still active as of August 19 (a removal of 98% of their content), with the majority being suspended or temporarily restricted.
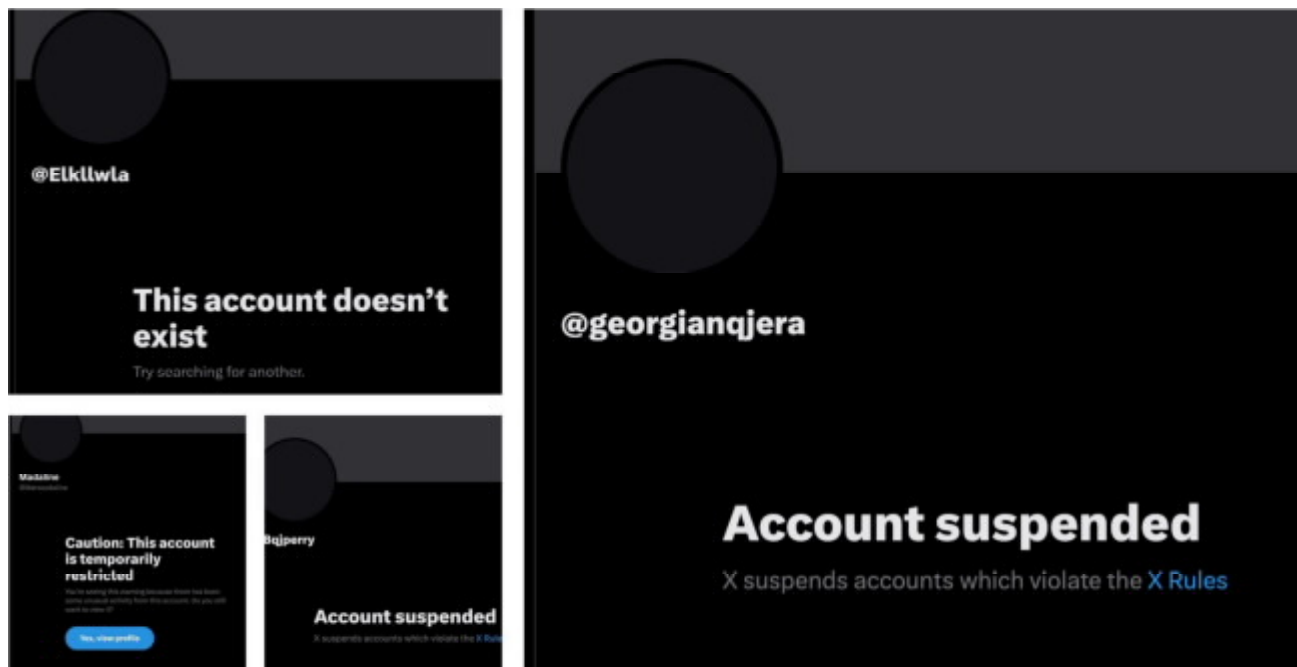
The remaining bot messages we were able to identify as of August 19 are available here. Of the 141 messages remaining tweets, only 4 come from the previously documented 300 accounts. There is thus a lower limit of 437 distinct bot accounts. Only 𝕏 can provide exact numbers, but assuming a 98% suspension, restriction or deletion rate, an upper bound of the number of bots would be approximately 7000.

While earlier assessments indicated that the tweets were virtually identical, there are 140 distinct messages in the 141 messages still available through the API. Given the large volume and semantic similarity as well as several non-human characteristics to the tweets, we are confident that the tweets were generated using a large language model (LLM). For example, the following messages cut off at 280 characters which would have been caught by a human writing the message:

- Just returned from Pierre Poilievre's rally in Kirkland Lake and I'm still buzzing! As a northern Ontario native, it's exhilarating to see a leader who actually listens to our concerns and has a plan to address them. The energy in that packed room wa

- Just spent the evening at Pierre Poilievre's rally in Kirkland Lake and I'm still buzzing from the energy! The crowd was electric and his passion for the North is infectious. It's clear he's committed to listening to our concerns and fighting for our

Other messages are incomplete thoughts that may reflect a character limit or random stopping by a LLM.

- Just got back from Pierre Poilievre's rally

Another five messages indicated that there was a very social parent at the event: "I just met a single mom in Kirkland Lake".

Reach and impact of the bot messages

We evaluated the reach of the bot messages as compared to subsequent discussion of the bot activity. Did the bots receive a lot of engagement themselves or were they primarily amplified by the reactions of other $\mathbb{X}$ users? To address this, we compared the views from the bot tweets to subsequent tweets discussing the bot activity. Figure 1 shows the results, with the top panel capturing the number of posts and the bottom panel capturing the number of views of content. Given that many accounts were blocked, suspended, or deleted in the days following identification of the campaign, the bot activity is a minimum estimate. While there was a large number of posts, they received an inconsequential number of views relative to the subsequent discussion of bots which numbered in the millions. Even if the bot message engagement were to be higher by an order of magnitude, it still would have been dwarfed by the sheer volume of engagement driven directly by non-bot coverage of the event. While the effect of the bots themselves was quite limited and short-lived, the reaction to the bots was much more pronounced and long-lasting.
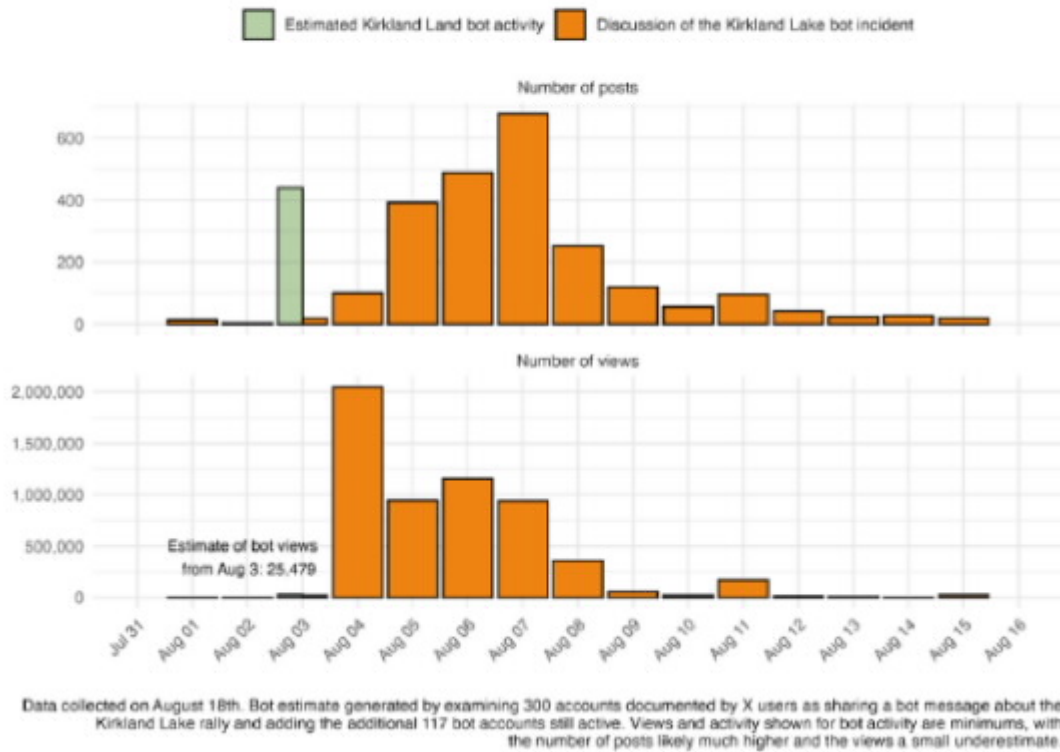
Figure 1: Minimal bot activity as compared to subsequent discussion of bot activity

Attribution of blame

We evaluated the subsequent discussion of the Kirkland Lake bot incident to better understand the extent to which specific entities were blamed for the bot activity. This was motivated by the observation that many tweets directly blamed Poilievre and the Conservatives for the attack, stating they were responsible and were knowingly involved.

To probe the prevalence of this sentiment, as well as claims of blame towards other actors, we used GPT-4o to classify tweets (n = 4011) as one of four categories: those that blamed Poilievre or the Conservatives, an opposition party (like the Liberals), a foreign entity (like Russia), or those that did not explicitly

assign any (or any clear) blame. Across all tweets, a little more than 33% assigned blame to some actor.

Among those, the Conservatives were assigned the most blame, representing more than 75% of the posts. 16% of Tweets blamed some foreign influence and 8% of tweets blamed one of the opposition parties. Figure 2 shows the number of posts and their associated views from August 3 to August 16 that blame one of these entities. Initially, the posts highlighting the potential interference were garnering views but very quickly the dominant narrative occurred that it was the Conservative Party who had organized and deployed the bot accounts. We observe that the tweets did not provide any clear evidence to support any blame attribution.

Data collected on August 18th. Blame attribution assigned using using GPT-4o.
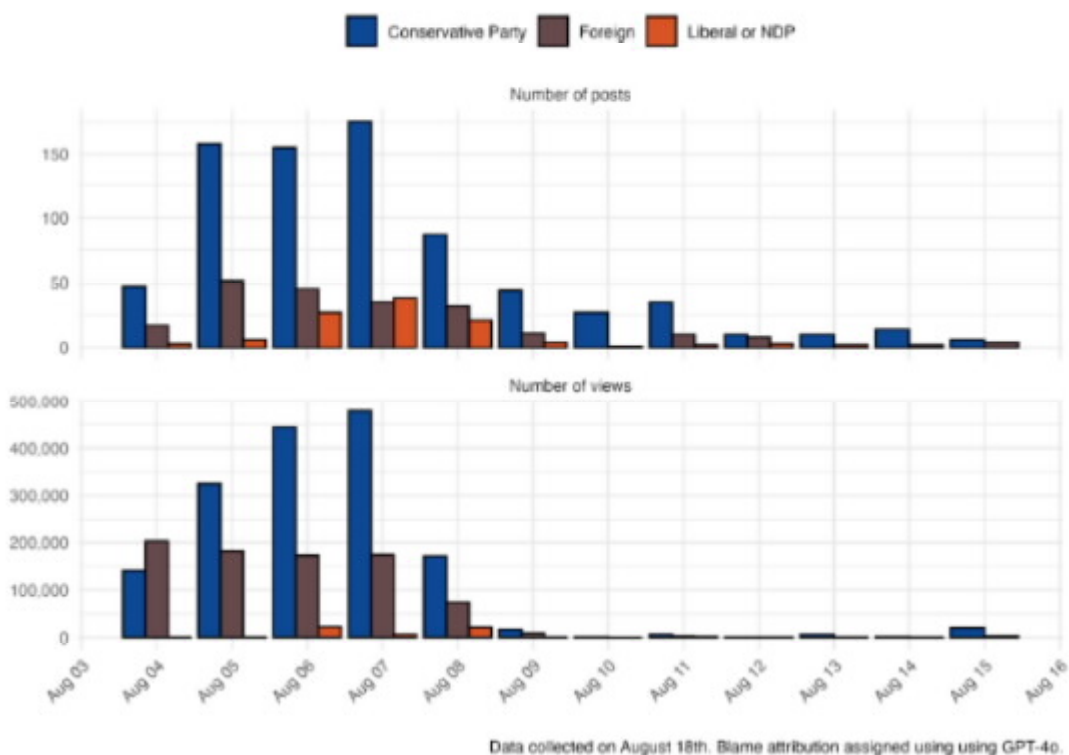
Figure 2: Blame assigned in the subsequent discussion of the Kirkland Lake bot incident

Bot history and posts

From the 146 bot accounts that still had tweets available through the API, we collected an additional 20 tweets per user (when available) on August 20 and then classified them using an LLM (Llama3.1:70b) to be about anything related to Canada. We find that of the 1442 posts that were still available, 18% concern Canada. Of the 146 accounts, only 12 accounts (7%) had ever tweeted anything else about Canada and of those, many appeared similar to the Kirkland Lake ones in form and substance. On the surface, this lack of Canadian content suggests that the Canadian political parties may not be involved in the bot campaigns. The following tweets show a consistent pattern of indicating some element of identity and piggy-backing off a recent news event, as well as language that an LLM would produce. Given character counts that are frequently over the character limit of $\mathbb{X}$ or insufficiently long to meet the demands of the (unknown) prompt employed, we observe that the LLMs used are likely not very sophisticated (e.g. GPT-2) or have a low number of parameters (e.g. Llama3.1:8b), or the prompt is not well designed to produce posts for $\mathbb{X}$.
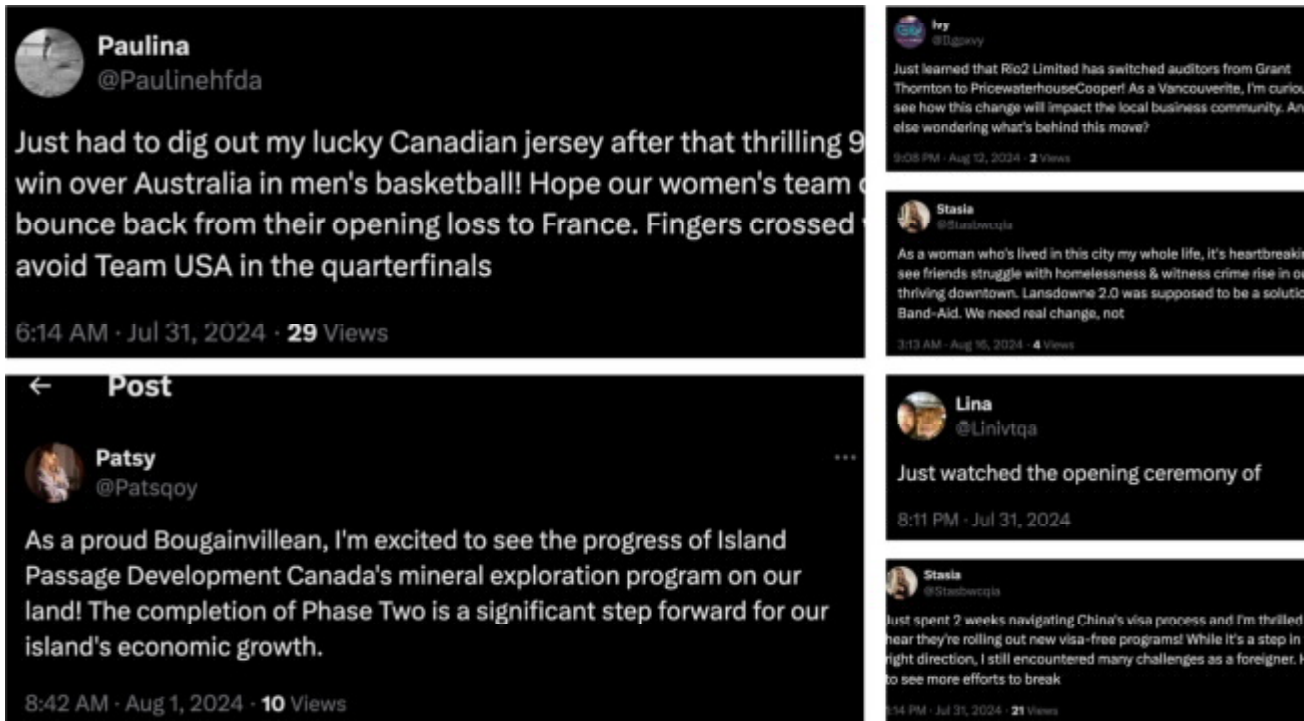
The earliest tweet from any of these accounts was in late July and followed the same pattern. Several accounts continue to post tweets and have not been suspended by $\mathbb{X}$.

Finally, we examined the creation date of the 142 accounts, with 140 of them having been created in the week preceding August 3rd.

Madaline
@Marxoodaline

Just had a thought-provoking convo with a friend who lived in Beijing - 'pragmatic engagement' is just a nice way of saying Canada does what China wants. Don't get me wrong, I love Chinese food and Mandarin classes, but compromised values aren't on t

8:04 PM · Jul 23, 2024 · **21** Views

CDMRN

Incident Update 5 |
Survey Findings :
Kirkland Lake Bot
Incident 〉

Stay updated and subscribe here