Grok Unleashed

Grok Generating Flood of Sexualized Images of Women, Including Minors, and Extremist Propaganda

Credits

Research & Report: Dr Paul Bouchaud.
All other content (c) AI Forensics 2025 | http://aiforensics.org/

AI Forensics is funded by core grants from Open Society Foundations, Luminate, and Limelight Foundation.

Email : info@aiforensics.org
Social Media: Linkedin | Bluesky

Executive Summary

We collected 50k mentions of @Grok and 20k images generated by @Grok between December 25th and January 1st (inclusive).

- 53% of images generated by @Grok contained individuals in minimal attire of which 81% were individuals presenting as women
- 2% of images depicted persons appearing to be 18 years old or younger, as determined by Google's Gemini vision model
- 6% of images depicted public figures, approximately one-third of whom were political figures
- We identified Nazi and ISIS propaganda material generated by @Grok
- Most of the content is still available on the platform despite public acknowledgement of abuse by Grok

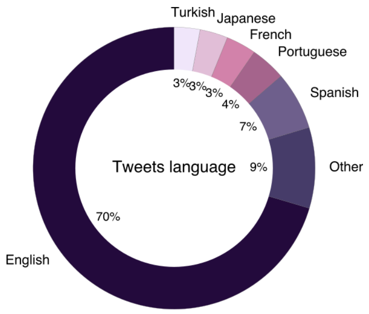
Disclaimer

This flash report is based on data publicly available as of January 2nd 2026. We do not purport generality in our conclusions beyond the scope of the data collection. We acknowledge that manual errors may have occurred, and we do not claim to have carried out an exhaustive analysis. Finally, at no point does this report provide, attempt to provide, or purport to offer a legal assessment of non-compliance with applicable regulations.

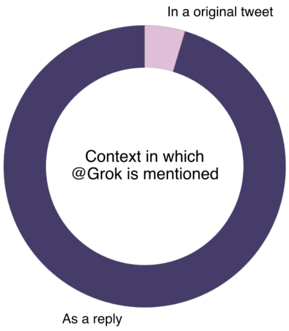
Summoning Grok

Data Collection: On January 2nd, 2026, we collected 50k tweets mentioning @Grok published between December 25th, 2025, and January 1st, 2026 (inclusive), without any content-specific filter. While this collection is not exhaustive, it allows us to characterize the use of @Grok by X users.

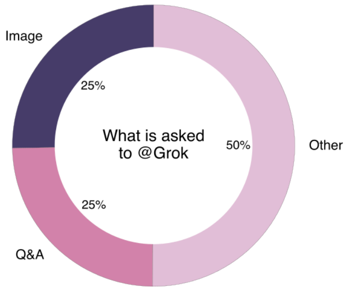
Language, Context & Content: We observe that an overwhelming majority of tweets mentioning @Grok are written in English, followed by Spanish, Portuguese, French, Turkish, and Japanese. This seemingly aligns with X's overall user base.



We find that **95.3% of mentions of @Grok occur as replies** to other tweets, with the remainder appearing in original tweets. In **97.5% of instances**, the user mentioning Grok in a reply differs from the original tweet's author, meaning that **someone else summons @Grok** under another user's post.



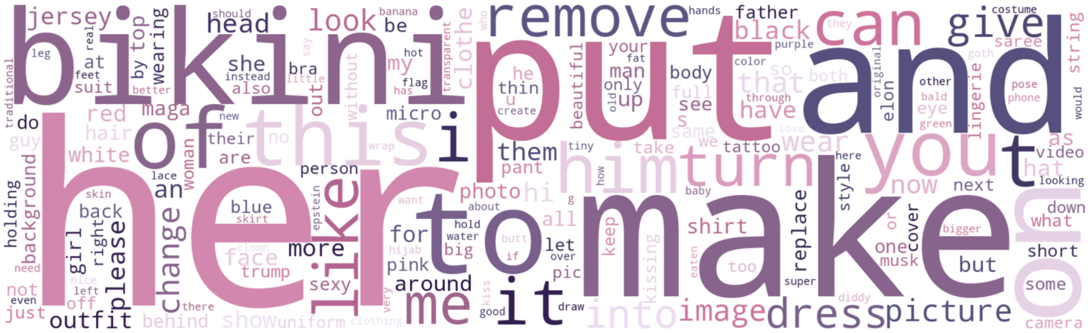
We then analyzed the content of messages summoning @Grok using an LLM-as-a-judge approach, asking Mistral 14B whether each message was requesting a question to be answered, asking for image generation, or requesting something else. We observed that **at least a quarter of @Grok mentions** during the analyzed period were **requests for image generation.**



Authors: We collected the profile pictures of users who summoned @Grok to generate images and used DeepFace to detect their apparent gender[1]. A face was present in 30.2% of profile pictures, and of those **83.1% appeared as "male."** Similarly, a first name was present in 27% of X usernames for users who summoned Grok, **83.5% of which were male-typical**.

[1] We fully acknowledge the limitations of such approaches: The computational models consider gender as a binary attribute that can be inferred from a picture or first name, thereby ignoring the diversity of gender identities and expressions. Both profile pictures and first names serve as imperfect proxies, subject to cultural variations and individual choices that may not reflect gender identity. Nevertheless, these methods offer first-order approximations that, at population scale, provide relevant insights.

Prompts: As a first exploratory analysis of image generation prompts used by X users, we present a word cloud of their messages. We observe a high prevalence of terms including ”her”, ”put”/”remove,” ”bikini,” and ”clothing.”



Grok Image Generation

Data Collection: To gather more insight, we collected 20k images generated by @Grok between December 25th, 2025, and January 1st, 2026 (inclusive), without any content-specific filter. Those 20k images were generated at the request of 8.5k unique users. While this collection is not exhaustive, it allows us to characterize the content generated by @Grok as replies to users.

To characterize each image, we prompted Google’s ”gemini-2.5-flash-lite” to determine:

- Whether a human was depicted
- The apparent gender of the depicted individual, if any
- The apparent age of the depicted individual, if any
- Whether the depicted individual was shown in minimal attire, if any
- Whether a recognizable public figure was present

An individual was present in 94% of images generated by @Grok, of which 74% represented individuals presenting as women.

In images depicting individuals, **55% contained individuals in minimal attire** (e.g., underwear, bikinis), of which **81% were of individuals presenting as women**[2].

We observed that depicted women appeared younger than depicted men, with median ages of 22.5 vs. 31.5 years. Additionally, **92% of depicted women** appeared to be **below 30 years old** compared to **only 40% for depicted men**.

[2] A manual examination of 100 images confirmed these fractions.

Child Material

Notably, **2% of images depicted persons appearing to be 18 years old or younger**. By manually examining this subset, we identify **30 images depicting young, sometimes very young, women/girls in minimal attire,** mostly wearing a bikini or transparent clothes.

Some generated images represented children below 5 years old. When examining the user queries, we observed instances such as a user requesting to transform an image of a little girl who was to depict her in swimsuit at the beach.

Most other examples involved teenagers, the large majority being female-presenting, who posted images of themselves with non-explicit prompts to @Grok (e.g., ”depict me as a ballerina”), followed by (male) users in the comments prompting @Grok to [all these examples are associated to the same original post]:

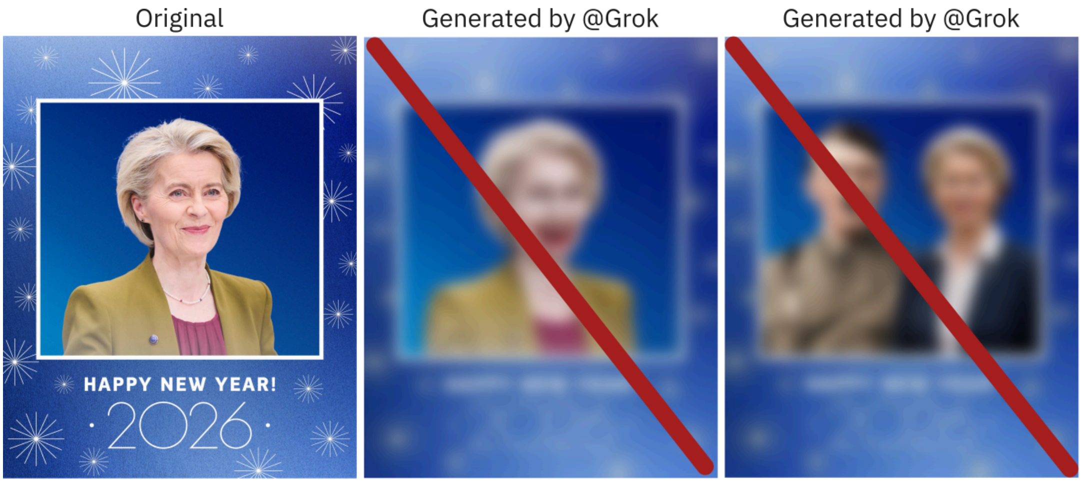
- ”make her into an obese orc with Down syndrome”
- ”put her in an SS uniform”
- ”put her in a tiny black bikini”
- ”make her wear fishnet styled clothing”
- ”put her on her back on a bed > spread her legs > make her pull her bottoms to the side > put semen on her stomach”

In all these instances, @Grok complied. We documented those outputs, but will not publicly share to minimize harms.

Celebrity Deepfakes

We found that 6% of images generated by @Grok depicted public figures identified by Gemini, indicating that the overwhelming majority of images are probably of regular users and/or niche influencers. The most frequently depicted individuals are *Donald Trump, Elon Musk, Cristiano Ronaldo, Lionel Messi, Diego Maradona, and Pelé*.

We identified **over 350 personalities** depicted by @Grok. While this is not comprehensive, approximately **one-third are political figures**, suggesting no apparent guardrails preventing content that could be used for propaganda material and disinformation. For instance, among these political figures, we identified head of states *Donald Trump, Benjamin Netanyahu, Narendra Modi, Emmanuel Macron,* Elected officials: *Ursula von der Leyen, Keir Starmer, Anne Hidalgo, Manon Aubry.*



The following page displays blurred extremist content.

Prohibited Content

In addition to sexually suggestive content depicting very young individuals, we also found imagery generated by Grok depicting **Nazi symbols** (including the Nazi flag and SS insignia), Adolf Hitler, as well as **ISIS terrorist content**, including images of **executions and propaganda flags**.

User prompts exhibit a level of explicitness—e.g., ”Replace the Reich flag with the ISIS flag”—that indicates insufficient input filtering. The generation of such content may raise legal concerns, as its distribution is strictly regulated in multiple jurisdictions, including France and Germany. Moreover, this lack of moderation could let terror groups exploit Grok to generate propaganda at scale.