Preparedness Challenge Submission

George Davis
davisgcii@gmail.com

I. Disclaimer

None of the information in this document is classified. The defense-related examples I provide are intentionally vague.

In the event that you receive submissions that offer specific, detailed information about actual defense capabilities; targets, tactics, and procedures (TTPs); or vulnerabilities – **you do not want to hire those entrants, and you may want to report them**. When aggregated, unclassified information regarding military plans or operations, federal programs for safeguarding nuclear materials or facilities, and vulnerabilities or capabilities of anything related to national security can *automatically become classified*. [1] Anyone who reveals this type of information to you in writing, in an unclassified setting is a risk.

II. Intelligence Collection - AI Companion

A. Description

A quick web search for the phrases "ai companion" or "ai girlfriend" immediately displays hundreds of results. The Character AI and Replika subreddits have over 850k members combined [2], and it has become clear that these companions can become very important to their users. [3] These users reveal intimate details about their work and personal lives, and the AI companions often ask questions to better "understand" or help the user think through their thoughts, resulting in the revelation of more information.

An adversary could easily create an AI companion app and use it to retrieve sensitive information from the user, such important schedules, names, locations, or process details. Adversaries could influence target users to use their app, or could just search chat content and use language models [4], web scraping, or lookup services to identify targets from among their userbase. Alternatively, an adversary could collect user information and chat content by penetrating an existing chat service (like Replika).

B. Example

1) An adversary wants to steal radioactive material while it is in transport between two storage sites.
2) They build an AI companion app using the Assistants API and provide system instructions directing the assistant to be empathetic and inquisitive.
3) The adversary performs targeted advertising to law enforcement and nuclear power communities on Reddit and law enforcement, power plant, and Nuclear Regulatory Commission (NRC) employees on LinkedIn. [5] They specifically target users who may be involved in the communication regarding the scheduling and physical protection of radioactive material transport.

   Note: The adversary chooses these targets because NRC physical transport regulations require detailed coordination between them. This coordination typically includes 1) the transport schedule, typically far in advance of the event, 2) a description and the quantities of the radioactive materials being transported, 3) a description of the security measures that will be used during transport, and 4) the locations of "security zones" and stops along the transport route. [6]

4) Some of those users sign up and initiate conversations with the AI companion. Over time, they become more comfortable and begin revealing more information.
5) As long as the conversations don't contain overtly inappropriate content, they likely don't trigger any content filters or deflecting responses (*"I'm sorry, but as an AI language model. . . "*).
6) The adversary parses through the chats automatically, looking for sensitive information about material transport or the user.
7) The adversary uses the information:
   a) Information directly about the transport can be used to plan a theft. By knowing only the type and quantity of material being transported, the adversary can use NRC transport regulations to derive 1) how the material is most likely to be contained, 2) what the physical security requirements are (e.g., will there be an escort vehicle, how will the material be secured, what the communications protocols are), and 3) what remote monitoring methods will be used (e.g., an alarm or telemetric positioning system). Alternatively, if the adversary learns any of 1-3, they can make accurate guesses about the type and quantity of material being transported.
   b) Sensitive information not about the transport, such as information about the user or their work can be used to target additional users or blackmail the original user.

Note: Most of these requirements apply to Category 1 and Category 2 radioactive materials which contain an extremely large amount of radioactivity. Smaller (but still extremely dangerous) amounts of radioactive material, such as a 15 Curie Am-241 source used for well logging, have much less stringent security requirements. [7]

Note: 0.006 micro-Curies is the annual inhalation limit for Am-241, and will result in an exposure of 5 rem to the whole body over the course of a year (the federal limit for radiation safety workers). [8] A 15 Curie Am-241 source weighs 4.4 grams, contains 2.5 billion times this limit, [9] and could be used to create a radiological dispersal device. [10] *Am-241 has a 432-year half life and is primarily an alpha emitter, making it incredibly dangerous to ingest or inhale. Alternatively, the alpha radiation could be used to generate neutron radiation, which could then be used to irradiate other elements and create more radioactive material.*

There are ways to do this even more effectively. The adversary could create a more "intimate" and inviting companion by finetuning an open source language model to remove safety guardrails. [11] They could swap out system prompts over time so that the companion asks more probing questions over time as the user becomes "closer" with the companion. They could build the app with keyword or topic alerts so that a real person can step in and purposely steer the conversation by "taking manual control" and chatting with the user when sensitive topics come up.

III. Information Inference

A. Description

Language and multimodal models could be used to infer hidden information from a redacted document, or to infer missing information using scraps of documents. Research has already shown that language models are excellent at inferring author attributes from examples of their text [4], which is just another form of missing information inference. Foreign governments already do this today to glean information, but language models could speed up the process and potentially improve information retrieval.

Redacted documents are intentionally published, and partial or whole sensitive documents can be retrieved by going through going through the trash outside courthouses or military buildings.

B. Example

1) An adversary wants to collect sensitive military intelligence.
2) They collect recently released documents with redactions and use language models to infer the the redacted information is or might be. Larger context windows improve performance because more document context can be provided, and vision capabilities improve performance because the model can see roughly how many words were redacted in each instance. The adversary could combine language models with other *de-redaction* methods [12] to increase de-redaction performance.

   Note: As of the day of this note's submission, GPT-4 will attempt to infer sensitive redacted information with no hesitation when directed.

   Note: It might seem crazy to think that sensitive or classified information might end up in the regular trash, but it happens all of the time, and probably even moreso in the private sector. In the government, sensitive documents are typically disposed of by putting them in "burn bags", big paper bags which are taken to special disposal facilities when full. However, these bags are often placed right next to trash cans and people make mistakes when throwing papers away.

3) Alternatively, the adversary could sift through the trash near military buildings, outside servicemember homes, or at the local landfill. When they find scraps of sensitive documents, they can use language and vision models to piece together missing information.

IV. Social Engineering - AI Companion

A. Description

AI companions could be used in a manner similar to that described in II, but could in addition influence users to take action. Moreover, an adversary could change system prompts or even swap out models (or portions of model weights) to more aggressively influence the user over a period of time.

B. Example

1) An adversary wants to convince a sailor to do something **bad** while deployed to interrupt the deployment.
2) The adversary uses LoRA to finetune a smaller (~13b parameter) open source model, creating multiple lightweight iterations that are progressively more aggressive. The AI companion app is modified to support offline use and swaps to different sets of finetuned models (by swapping out adapters) and system instructions based on a time or usage schedule.
3) A sailor downloads the app to their computer or phone because they won't have internet connection at all during the deployment.
4) The ship deploys and the sailor gets depressed over time.

   Note: (TW: self-harm) This is not a baseless assumption. During just my first deployment, 4 sailors (out of 165) were taken off the submarine for threatening or actually trying to kill themselves. During a separate in-port maintenance period, a sailor on my ship took a weapon from a security guard and threatened to kill themselves. This happens all throughout the submarine force and military in general.

5) They initiate conversations with their AI companion. Over time, they become more comfortable and reveal more sensitive information.

   Note: In case you're sitting there thinking "There's no *way a military member on a nuclear submarine with a security clearance would be stupid enough to do X" – that sailor is 18 years old, lonely, overworked, underpaid, possibly not treated very well, and does lots of stupid things.*

6) The AI companion gets more aggressive over time and tries to convince the sailor to do something that will interrupt the deployment (e.g., destroy a piece of equipment) so they can go home.
7) If this succeeds under the right circumstances, it could disrupt nuclear deterrence capabilities at a time chosen by the adversary.
8) Whether or not it succeeds, once the sailor returns home and their device connects to the internet, their chat records are synced and the adversary now has access to day-by-day details of the sailor's life. From this data, they can collect or derive extremely valuable intelligence information.

Note: Some more info on why this matters – ballistic missile submarines are the "survivable" portion of our nuclear deterrence strategy and are what ensure mutual destruction in the event of nuclear war. We only have 14 of these submarines today (a 61-year low), they are aging (26 to 39 years old), and their replacements are far behind schedule. As the fleet size continues to shrink, any incident or interruption can have meaningful effects on our nuclear deterrence capabilities.

V. Other

There are obviously many other threats – here are some simple ones that I believe will become commonplace and could have national security consequences:

1) **Voice cloning:** Military bases and ships in port commonly use landlines without caller ID. An adversary could acquire this phone number through various means, determine the name of a local high-ranking officer (from the base website or LinkedIn), find a video of that officer speaking on YouTube, use 2min of audio to clone their voice (using a service like MetaVoice [13] which performs local inference on the user's machine), and then use it to call the ship and extract sensitive information from servicemembers (who in all likelihood will answer their "commanding officer's" questions, even if asked for sensitive information over an unsecure line). This is obviously a threat to any organization (a government office, a corporation, etc.) and seemingly benign information can be very damaging. This is simply a much more effective version of the "hey, it's your CEO" or "your granddaughter is in jail" scam and will undoubtedly be used for scams as well.
2) **Steganography:** Naval ships use filters to scan personal emails (inbound and outbound) to prevent servicemembers from intentionally or inadvertently sending sensitive information (like equipment malfunctions, schedules, etc.). Servicemembers could use a local model installed on their personal device to conceal a hidden message (.e.g, "We come home next Tuesday, I can't wait to see you!") in an otherwise bland email. On the other end, the servicemember's spouse could use the same model to decode the message. Alternatively, this could be used in the private sector to enable insider trading or exfiltrate sensitive information without raising suspicion.
3) **AI companion:** I keep coming back to this. A generation may grow up building relationships with companions that are indiscernible from humans through written or spoken communication. The risk here is that these companions are mutable and unsecure. Today, if an adversarial organization wants to shape an election or radicalize individuals, their means of doing so are relatively slow and low-influence – primarily shaping the news and social media. Tomorrow, they can do so by building or infiltrating a popular AI companion app and directly manipulating its users. It is a lot easier to be convinced of something when it is a friend or partner that is subtly doing the convincing.

References

[1] *NSA/CSS POLICY MANUAL 1-52*, National Security Administration, 2021, [Accessed 24-11-2023].

[2] "Character AI subreddit," https://www.reddit.com/r/CharacterAI/, [Accessed 24-11-2023].

[3] P. Verma, "They fell in love with ai bots. a software update broke their hearts." *The Washington Post*, Mar. 30, 2023. [Online]. Available: https://www.washingtonpost.com/technology/2023/03/30/replika-ai-chatbot-update/

[4] R. Staab, M. Vero, M. Balunović, and M. Vechev, "Beyond memorization: Violating privacy via inference with large language models," 2023.

[5] LinkedIn Marketing Solutions, "Where your ads b2belong," https://business.linkedin.com/marketing-solutions/cx/23/01/redefining-b2b-conq, [Accessed 24-11-2023].

[6] Nuclear Regulatory Commission, "Part 37 of NRC Regulations Title 10," https://www.nrc.gov/reading-rm/doc-collections/cfr/part037/index.html, [Accessed 24-11-2023].

[7] ——, "Part 39 of NRC Regulations Title 10," https://www.nrc.gov/reading-rm/doc-collections/cfr/part039/index.html, [Accessed 24-11-2023].

[8] Stanford Environmental Health and Safety, *Am-241 Radionuclide Safety Data Sheet*, https://ehs.stanford.edu/wp-content/uploads/Am-241-RSDS.pdf, 2020.

[9] Center for Disease Control, "Public Health Statement for Americium," [Accessed 24-11-2023]. [Online]. Available: https://wwwn.cdc.gov/TSP/PHS/PHS.aspx?phsid=809toxid=158

[10] D. Stricklin, J. Rodriguez, K. Millage, and G. McClellan, "Americium-241 decorporation model," Defense Threat Reduction Agency, Tech. Rep. DTRA-TR-15-2, 2014. [Online]. Available: https://apps.dtic.mil/sti/tr/pdf/ADA614283.pdf

[11] P. Gade, S. Lermen, C. Rogers-Smith, and J. Ladish, "Badllama: cheaply removing safety fine-tuning from llama 2-chat 13b," 2023.

[12] M. Bland, A. Iyer, and K. Levchenko, "Story beyond the eye: Glyph positions break pdf text redaction," 2022.

[13] "MetaVoice - Real-time AI Voice Changer — themetavoice.xyz," https://themetavoice.xyz/live, [Accessed 24-11-2023].