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.

Source: One of my many roles as a submarine officer was to serve as the ship's *Top Secret Controls Officer*.

II. Information Inference

A. Description

Research has already shown that language models are excellent at inferring author attributes from examples of their text [2]. Taking this one step further, language and multimodal models could be used to infer sensitive or classified information using redacted documents, leaked documents, social media information, and open source intelligence (OSINT, such as preprints from national labs, congressional proceedings, patents, flight reports, etc.) as context.

Moreover, language and multimodal models can be used to improve and scale automated OSINT collection through the ”intelligent” scraping of unstructured data (*I have built a toy demo of a natural language web scraper on top of the GPT-4 API - it's expensive but works great).*

Adversarial intelligence organizations will use this technology to both scale their intelligence collection capabilities and improve the accuracy of the intelligence they collect.

B. Example - Simple Redacted Information Inference

1) An adversary wants to collect sensitive military intelligence related to recently leaked documents that contain redactions.
2) The adversary uses traditional OSINT collection methods and language models (which can be used to perform ”intelligent” data mining by extracting key information from unstructured web data or automatically generating precise image descriptions) to gather relevant information from the web.
3) They then use language models to infer what the the redacted information might be and what conclusions they can draw given the leaked documents and additional context.

   Note: 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.
4) The adversary could further combine this method with other *de-redaction* methods [3] 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.

III. AI-Enabled Social Engineering

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 [4], and it has become clear that these companions can become very important to their users. [5] 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 targeted users, such as important schedules, names, locations, or process details. Adversaries could influence target users to use their app via LLM-enabled spearphishing campaigns, or could just search chat content and use language models [2], web scraping, or lookup services to identify targets from among their existing userbase. Alternatively, an adversary could collect user information and chat content by penetrating an existing chat service (like Replika).

Taking this one step further, AI companions could directly influence users to take action, such as installing malware, exfiltrating sensitive documents, or attacking infrastructure. [6]

B. Example 1 - Passive Intelligence Collection (AI ”Honeytrap” [7])

As a Nuclear Engineer Officer certified by Naval Reactors and the Department of Energy, I know first hand how hard it would be to actually make a nuclear power plant go boom.

They are inherently stable, are designed with redundancy upon redundancy, and have multi-layered security.

Rather, in this example I choose to discuss the theft of nuclear material because it is much easier to accomplish, can incite just as much panic, and could likely harm more people through the spread of radioactive contamination. *I know this because I was the Chemical and Radiological Controls Officer on my submarine for nearly two years and ran our radiation health and disaster response training programs.*

This type of intelligence collection could just as easily be used against servicemembers, members of the Intelligence Community, politicians and their staff, and employees of government contractors and private sector organizations.

1) An adversary (individual or organization) 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 a combination of targeted advertising and spearphishing to law enforcement and nuclear power communities on Reddit and law enforcement, power plant, and Nuclear Regulatory Commission (NRC) employees on LinkedIn. [8] 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. [9]

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), and 3) what remote monitoring methods will be used (e.g., a telemetric positioning system). Alternatively, if the adversary learns any of 1-3, they can determine 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. A smaller (but still extremely dangerous) amount of radioactive material, such as a 15 Curie Am-241 source used for well logging, has much less stringent security requirements. [10] To prove my point, a 15 Curie Am-241 source weighs only 4.4 grams [11] but contains 2.5 billion times the recommended annual inhalation limit [12] and has enough radioactivity to give lethal acute radiation sickness to thousands. It could easily be used to create a radiological dispersal device [13] or to generate neutron radiation which could then be used to create even more radioactive material.

This can be done even more effectively. The adversary could create a more "intimate" and inviting companion by finetuning an open source language model to remove safety guardrails. [14] They could swap out system prompts over time so that the companion becomes more ”probing”. They could build the app with semantic alerts so that a real person could ”take manual control” and steer the conversation when sensitive topics come up.

C. Example 2 - Active Influencing

Adversaries can perform ”fire and forget” social engineering, and could even target users in remote environments or without consistent internet access.

1) An adversary wants to convince a sailor on a ballistic missile submarine to do something **bad** while deployed to interrupt the deployment.
2) The adversary uses LoRA to finetune a smaller open source model (like Llama-2-13b, which runs at ∼10 tokens-per-second on my MacBook Air), creating multiple versions that are progressively more aggressive. The AI companion app is built 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) The adversary runs an AI-scaled spearphishing campaign, targeting junior sailors stationed on ballistic missile submarines.
4) A sailor downloads the app to their computer or phone because they won't have internet connection at all during the deployment.

5) 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. This happens throughout the submarine force and military in general.

6) 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.

7) 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.
8) If this succeeds under the right circumstances, it could disrupt nuclear deterrence capabilities at a time chosen by the adversary.
9) 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.

IV. Other

There are obviously many other threats, including many that I can't discuss due to security reasons. Here are some simple unclassified ones that I believe will become commonplace and could have national security implications:

1) **Voice cloning:** When the Commanding Officer went home for the day and I was left in charge of the submarine, my primary form of communication with the Captain was a landline installed on the ship (this goes for military bases in general). An adversary could easily have acquired this phone number by calling the Squadron Office, determined the name of my Commanding Officer from the base website, found a video of him speaking on YouTube, used two minutes of audio to clone his voice (using a service like MetaVoice [15] which performs local inference on the user's machine), and then called the ship to collect sensitive information. Even a benign question, like "give me a status update on ongoing maintenance" would collect very sensitive information. This is obviously a threat to any organization (a government office, a corporation, etc.).
2) **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] R. Staab, M. Vero, M. Balunović, and M. Vechev, "Beyond memorization: Violating privacy via inference with large language models," 2023.

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

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

[5] 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/

[6] J. Doubek, "Another North Carolina power substation was damaged by gunfire," 2023, [Accessed 24-11-2023]. [Online]. Available: https://www.npr.org/2023/01/18/1149694402/another-north-carolina-power-substation-shot

[7] "Honey trapping - Wikipedia," [Accessed 24-11-2023]. [Online]. Available: https://en.wikipedia.org/wiki/Honey_trapping

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

[9] 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].

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

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

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

[13] 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

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

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