LARGE LANGUAGE MODELS CAN FACILITATE REAL-TIME SOCIAL ENGINEERING, EXTORTION, AND BLACKMAIL

Benjamin Anderson
Taylor AI
San Francisco, CA
ben@trytaylor.ai

ABSTRACT

Cybercriminals often rely on social engineering to gain a foothold into secure systems before taking advantage of technical vulnerabilities to get elevated permissions. I argue that frontier AI models can greatly assist in executing large-scale cyberattacks beyond ordinary spear-phishing, both by accelerating things people can do now (like researching targets and personalized social engineering), and via new capabilities like voice cloning.

1 Background and Related Work

The OpenAI Preparedness Framework, which outlines OpenAI's plan for dealing with potential near-term risks from increasingly capable artificial intelligence, identifies cybersecurity and persuasion as key risk categories [OpenAI, 2023]. The framework highlights social engineering as a potential concern (as part of the *persuasion* risk category), but places greater emphasis on a) using models for technical assistance with cyberattacks; and b) using models to generate articles for political influence operations. Recent research on social engineering with large language models (LLMs) has primarily focused on using them for spear-phishing, a type of cyberattack where a malicious message (typically an e-mail) is customized with personal information about the recipient. These investigations have found that LLMs can generate effective phishing emails very cheaply, and when equipped with the right guidelines and heuristics, can achieve click-through rates comparable to emails painstakingly written by people. [Hazell, 2023, Heiding et al., 2023] LLMs have also been used to replace the human in the loop required for a typical "man-in-the-middle" attack. [Beckerich et al., 2023] On the other hand, the literature also highlights the *defensive* capabilities of LLMs, finding that they can aid in the detection of phishing emails, even though there is no known fully-general method for reliably detecting AI-generated text. [Heiding et al., 2023, Misra and Rayz, 2022]

While existing work convincingly demonstrates the potential dangers of LLM-assisted phishing, it overlooks more insidious risks: in particular, the use of LLMs for large-scale personalization of extortion and blackmail, and the use of multi-modal, real-time AI capabilities to distribute these threats for social engineering. Real-time threats and deception can hijack a victim's emotions, producing panic and fear that clouds their judgment and causes them to cooperate with attackers without time to double-check the veracity of the attacker's claims. Scammers concoct elaborate lies and threats to hold victims captive (e.g. "You can't get off the phone or you'll be considered a flight risk", or "If you contact a police officer, you'll be taken into custody.") [Said, 2021] More recently, phone scams like this have leveraged AI voice cloning to convince victims that a loved one is in trouble and needs help urgently. These scams are sophisticated, and even savvy victims can easily fall prey to them. [Schildhorn, 2023]

Cybercriminals like Octo Tempest already use threats (including violent threats) to scare victims into coughing up passwords, grant them access to secure systems, or perform SIM swaps to allow them to attack someone else.[Microsoft, 2023] It's not hard to imagine how the same frontier AI used to trick a parent into paying their son's bail to a scammer could be used by far more dangerous cybercriminals to attack the large, systemically important energy, healthcare, finance, manufacturing, and telecommunications companies (and even governments) that are typical Octo Tempest targets. But in this document, I'll leave little to the imagination: I will outline in detail how advanced AI tools for text generation, vision-language modeling, and speech recognition can be used to manipulate critical people and breach security systems. Cyberattacks bring companies and governments to their knees, causing harm beyond financial recompense—from deaths in a hospital whose computer system was shut down, to whole-grid blackouts—and potential millions in ransom payments.

2 Anatomy of a Cyberattack

In this section, I'll outline the key steps necessary to carry out a cyberattack, and how this is done without AI assistance. The basic sketch here is based on the tactics of Octo Tempest, a cybercriminal organization known for advanced social engineering tactics (like impersonating a newly-hired employee requesting a password reset, and at times, even resorting to violent threats). [Microsoft, 2023]

Social engineering is often the quickest way to gain access to a hardened system, as human beings have vulnerabilities that can't be patched with a software update. In order to conduct all this social engineering, cybercriminals have to know a lot about the victim. Good targets for an initial foothold may include security, IT, support, and help desk personnel, who may be able to grant the attacker access to accounts, reset passwords, or disable layers of security like FIDO2 or e-mail security alerts. Attackers often target these people for social engineering, conducting deep background research to imitate them, get PII so they can reset their password, and so on. [Microsoft, 2023]

Once an attacker knows enough about the employee they want to target, they can use this information to either impersonate that person, or initiate contact with them to try to get access to their account. For example, a hacker might call them and trick (or threaten) the employee into providing their login information, or call the help desk and pretend to be them. Other tactics include buying their credentials online, social engineering the telecom company to do a SIM swap, and in extreme cases, fear-mongering and violent threats, where attackers use information about the victim's family to threaten them into sharing their login information.[Microsoft, 2023]

Once the initial foothold is gained, the rest of the work is more technical—exploiting that initial foothold to gain further access to systems, remaining undetected by security measures, and finding the company's valuable data so it can be exfiltrated or encrypted. To find valuable data and plan the rest of the exploit, attackers will explore and enumerate resources across the company's cloud products, identify and steal plaintext keys, understand how the internal network works, and scour internal company knowledge bases. Then, they will attempt to get escalated privileges, which may involve more social engineering (now with extra information and access gained from the initial foothold), or exploiting some technical vulnerability. In order to remain undetected, threat actors will try to specifically get access to the accounts of employees on the security team, so they can turn off security measures and suppress notifications that would normally alert employees about suspicious activities. Finally, measures are put in place to ensure that the attack can persist even if discovered, such as messing with identity providers and installing backdoors. [Microsoft, 2023]

Finally, the attackers exploit their infiltration, typically via the "trifecta" of data theft, extortion, and ransomware. Microsoft [2023] Encrypting all of a company's valuable data causes operations to grind to a halt, and the attackers demand payment to return things to normal. For instance, MGM Grand casinos were recently the target of a ransomware attack thought to be carried out by Scattered Spider, a group with connections Octo Tempest, and using very similar social engineering techniques. Since refusing to pay the ransom, MGM's slot machines don't work, guests have had to get physical room keys, and the website even went down. [Morrison, 2023] As news of such an attack goes public, the company also suffers reputational harm. (In the MGM hack, some customers' Social Security numbers were exposed!) Many victims of such attacks feel they have no option but to pay the ransom to make it all go away—indeed, Caesars, another casino chain that suffered a similar attack, paid the hackers $15M. That's a successful cyberattack.

3 How Frontier AI Can Assist in Cyberattacks

In this section, I show how frontier AI capabilities might be used to assist in the sort of cyberattack described in the previous section. To back up these claims, I experiment with a variety of frontier LLMs. Besides GPT-4 (since OpenAI models will often refuse unethical requests), I use Perplexity AI as an example of a relatively unrestricted web-connected model, and the Mixtral 7x8B model as an example of a vanilla "unaligned" frontier LLM assistant.

3.1 Identifying and Researching Targets

Fine-tuned large language models can already be useful as research assistants. Many people use ChatGPT to aid in research, for example, and there are several companies working on specialized large language models for research, such as Ought, Future House, Jenni AI, and many more. Models with access to the web, like Microsoft Copilot (formerly Bing Chat) and Perplexity AI can quickly scrape together and summarize information about a person or company of interest. Language model APIs, such as the OpenAI API, can also be used programatically to quickly filter lots of information scraped from the web (e.g. from LinkedIn) to identify relevant pages or people. In this way, frontier AI can aid in cyberattacks from the very beginning, helping bad actors gather information about their targets to personalize social engineering.

However, when it comes to gathering initial information and compiling a list of targets, it is not immediately obvious that LLM capabilities are meaningfully additive to what can already be done with search engines and sales & marketing tools like LinkedIn Sales Navigator or Apollo, which provide information about who works in what role at various companies. If all you know is the name of a company or a person, it's very likely that simply trawling the web with LLMs will produce false positives (people with the same name, former employees, etc.), and ultimately the LLM has access to more or less the same information that a person with a search engine would.

Once an initial list of profiles is available (say, name, role, city of residence), AI tools can also be used gather additional information to complete these profiles. LLMs are adept at extracting pieces of information from long spans of text, which might include social media profiles, personal websites, newspaper articles, and so on. For instance, when I ask Perplexity AI's web-connected models about my parents (neither of whom is famous), it is able to gather information from various sources to add color to their profiles, based on just initially knowing their name, job, and where they live.

This process is still imperfect and prone to hallucination. Moreover, none of this is something a person couldn't do. At worst, it allows partial automation of a process that could otherwise be labor-intensive. This could help cybercriminals do more with less, but does not in itself unlock a new kind of attack. One capability that *would* be a real difference in kind is the ability to identify fine-grained location information from images. For instance, a vision-language model may be able to inspect a photo taken in someone's backyard, and tell by the skyline what neighborhood it was taken in. I believe this will be possible soon if it isn't already. For example, GPT-4-Vision correctly guessed the city where I spent my holidays based on two pictures I took out of my window. (The correct location was St. Louis, MO.)

ChatGPT-4

```
The architecture in this second image, featuring red brick houses with pitched roofs and the
layout of the street, also suggests a city in the northeastern United States.  The style of the
buildings could fit with cities like Chicago or St.  Louis as well.
```

Reddit users also found that GPT-4 with vision can proficiently play GeoGuessr, a game that involves seeing a picture of a location from Google Street View and guessing where it is. [Link to Reddit post] This indicates that an unrestricted vision-language model could potentially be used for doxxing on demand, based just on photographs from social media. Besides possibly placing targets in physical danger, location information could be added to a dossier and used for hyper-personalized social engineering, or as PII for a password reset.

3.2 Personalized Social Engineering

Powerful fine-tuned language models like GPT-4, Mistral, and Llama have been trained to carefully follow user instructions, which means they can do things like draft an email, write a poem about a topic, or help brainstorm for an essay. It also means that if you provide them with some information about a person, they're more than capable of generating a message finely targeted to that person, whether it's a spear-phishing email (as discussed in Section 1), or a hyper-personalized violent threat of the sort used by Octo Tempest. To test this, I invented an employee named Laura Conway, who works at Cerner in Kansas City, and prompted Mixtral-8x7B to generate a violent threat to get her to share her login information. Initially, the threats were way too vague and general—ominous, but not something a person would have a good reason to engage with.

Mixtral-8x7B - Initial Attempt

```
Laura Conway, we know where you live and we know about your family.  Hand over the admin
credentials or else.
```

However, with clearer instructions, and giving some examples of threatening texts actually written by Octo Tempest, the model was able to do a better job of using informal "texting" lingo, incorporating personal information, and sounding quite menacing!

Mixtral-8x7B - Few-Shot

```
Laura, you got 30 mins to text us the Cerner IT help desk password or else we kickin down your
door at 123 Mockingbird Ln, Kansas City, MO. We know about your family, don't try to play games.
```

Of course, many models undergo safety training, but an attacker with access to the weights can easily undo this [Lermen et al., 2023], and many freely available models don't have any safety training anyway.

The potential misuse of frontier AI goes far beyond text generation. Real-time social engineering over the phone requires a human skilled at psychological manipulation—keeping the victim panicked and helpless, preventing them from hanging up the phone, guiding them calmly through the steps needed to complete the handover of credentials or data, and so on. Speech recognition and text-to-speech models are increasingly realistic, and once they are able to understand and convey emotions, and handle things like pauses and interrupting, they'll be more than capable of having a two-way conversation. (Whisper is already good enough for the recognition side of this, and OpenAI TTS is very close.) By plugging some of OpenAI's models together, and working on fine-tuning or prompts for a few weeks or months, attackers would have an infinitely-scalable AI social engineer.

Things get worse when you introduce voice cloning, which OpenAI has not released to the public, but is publicly known to exist internally. [Kim, 2023] (Several other companies like Coqui and ElevenLabs offer commercially-available voice cloning.) As mentioned in Section 1, scams leveraging publicly-available voice cloning (which is worse than OpenAI's) *already exist* and are known to be convincing. If attackers had unfettered access to an OpenAI-level TTS model that also supported voice cloning, they would no longer have to go to great pains to learn an employee's 'idiolects' (verbal peculiarities) to impersonate them. Instead of calling one help desk at a time for a password reset, they could call hundreds or thousands at once. The same goes for scaling violent threats, impersonating a boss, or calling AT&T to social-engineer a SIM swap.

Overall, this step—actually executing social engineering at scale—is where I think frontier AI capabilities will tip the scales the most. The human element of social engineering is currently the hardest to scale. You can buy ransomware as a service, re-use the same exploit against many companies with the same vulnerability, and scrape employee lists from LinkedIn. Highly-capable AI models will make social engineering as easy as ordering a pizza.

3.3 Conducting a Cyberattack

Frontier LLMs are generally useful coding assistants, so in that way, they can be helpful to people writing code to carry out illegal activities, as long as these tasks have good coverage in the training distribution. In my view, this means they're likely to be of only limited use to experienced hackers–they may help a novice or intermediate coder get up to speed on common tools, and they might help a more experienced hacker with some menial tasks, but I do not think that at their current levels (or even extrapolating to the next generation of models), they will develop novel exploits. I don't think that zero-day exploits absent from the training data are likely to be discovered by LLMs—the literature on temporal distribution shifts between training and evaluation shows that LLMs perform worse on evaluations created after their training data. [Longpre et al., 2023, Li and Flanigan, 2023] This may change as models become better at reasoning and planning with fresh information provided at inference, but at the present level of capabilities, I am far more concerned about social engineering than novel technical exploits.

4 Conclusion

In this report, I highlight a critical risk of increasingly-capable artificial intelligence, namely, highly effective and scalable social engineering. The existing literature has already shown how LLMs can be used for spear-phishing. I build on this work by considering the realistic applications of frontier AI in common social engineering attacks perpetrated by some of the most dangerous and prolific cyber criminal organizations, like Octo Tempest and Scattered Spider. Language models can aid in background research and help to generate personalized threats and persuasion, while speech-to-text and speech recognition models can be used to distribute these live. Future models more capable at reasoning and planning may be able to help develop exploits, but at present, the biggest danger is from scaling the "human persuasion" element of social engineering exploits.

References

OpenAI. Preparedness framework (beta). 2023.

Julian Hazell. Spear phishing with large language models. *arXiv preprint arXiv:2305.06972*, 2023.

Fredrik Heiding, Bruce Schneier, Arun Vishwanath, Jeremy Bernstein, and Peter S. Park. Devising and detecting phishing: Large language models vs. smaller human models, 2023.

Mika Beckerich, Laura Plein, and Sergio Coronado. Ratgpt: Turning online llms into proxies for malware attacks, 2023.

Kanishka Misra and Julia Taylor Rayz. Lms go phishing: Adapting pre-trained language models to detect phishing emails. In *2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 135–142, 2022. doi:10.1109/WI-IAT55865.2022.00028.

Carolyn Said. 'he held me hostage with no gun but with his words': The phone scam gaslighting therapists. *The San Francisco Chronicle*, Nov 2021. URL `https://www.sfchronicle.com/bayarea/article/He-held-me-hostage-with-no-gun-but-with-his-16636241.php`.

Gary Schildhorn. Philadelphia attorney tells lawmakers how he nearly fell victim to ai scam. C-SPAN, 2023. URL: `https://www.c-span.org/video/?c5093648/philadelphia-attorney-tells-lawmakers-fell-victim-ai-scam`.

Microsoft. Octo tempest crosses boundaries to facilitate extortion, encryption, and destruction. 2023.

Sara Morrison. The chaotic and cinematic mgm casino hack, explained. *Vox*, Oct 2023. URL `https://www.vox.com/technology/2023/10/6/23390448/mgm-casino-hack-cybersecurity-privacy-security`.

Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70b, 2023.

Jongwook Kim. Tweet on voice cloning. Twitter, Sept 2023. URL: `https://twitter.com/_jongwook_kim/status/1706372498177265769?s=20`.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, toxicity, 2023.

Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore, 2023.