

Fine-tuning an LLM on CTI reports for fun and profit

Aaron Kaplan, Jürgen Brandl

Intro speakers

Aaron Kaplan

- Self-employed / EC-DIGIT-CSIRC
- Previously 12 years @ CERT.at, Austria
- Why I like *and* rage against the ML...

Jürgen Brandl

- Senior Cyber Security Analyst by day
- Giving shell access to LLMs at night

Disclaimer

All errors are mine to keep

I present this here as a sole proprietor company
under my own name

Opinions are my own

CTI Reports

Problems:

- long
- unstructured
- unfiltered
- full of jargon and acronyms
- hard to keep up

Defending OT Operations Against Ongoing Pro-Russia Hactivist Activity

TLP: CLEAR



Overview

The Cybersecurity and Infrastructure Security Agency (CISA), Federal Bureau of Investigation (FBI), National Security Agency (NSA), Environmental Protection Agency (EPA), Department of Energy (DOE), United States Department of Agriculture (USDA), Food and Drug Administration (FDA), Multi-State Information Sharing and Analysis Center (MS-ISAC), Canadian Centre for Cyber Security (CCCS), and United Kingdom's National Cyber Security Centre (NCSC-UK)—hereafter referred to as "the authoring organizations"—are disseminating this fact sheet to highlight and safeguard against the continued malicious cyber activity conducted by pro-Russia hactivists against operational technology (OT) devices in North America and Europe.

The authoring organizations are aware of pro-Russia hactivists targeting and compromising small-scale OT systems in North American and European Water and Wastewater Systems (WWS), Dams, Energy, and Food and Agriculture Sectors. These hactivists seek to compromise modular, internet-exposed industrial control systems (ICS) through their software components, such as human machine interfaces (HMIs), by exploiting virtual network computing (VNC) remote access software and default passwords.

The authoring organizations are releasing this fact sheet to share information and mitigations associated with this malicious activity, which has been observed since 2022 and as recently as April 2024. The authoring organizations encourage OT operators in critical infrastructure sectors—including WWS, Dams, Energy, and Food and Agriculture—to apply the recommendations listed in the Mitigations section of this fact sheet to defend against this activity.

Overview of Threat Actor Activity

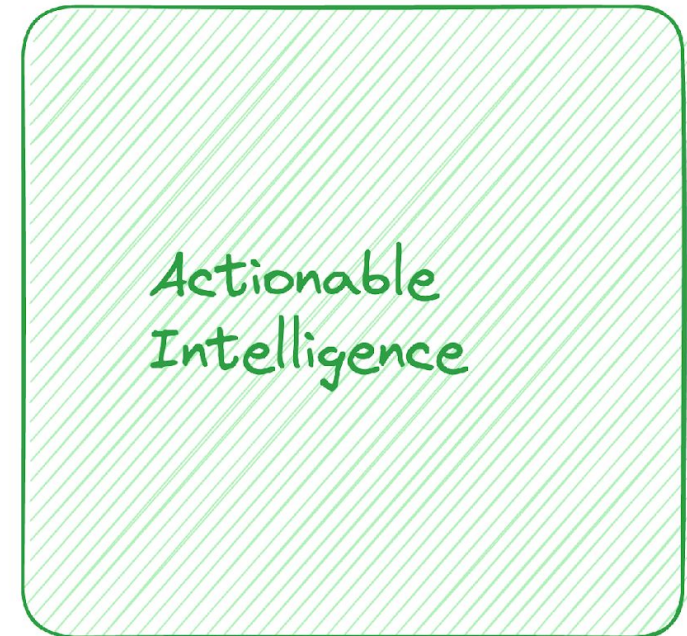
Pro-Russia hactivist activity against these sectors appears mostly limited to unsophisticated techniques that manipulate ICS equipment to create nuisance effects. However, investigations have identified that these actors are capable of techniques that pose physical threats against *insecure and misconfigured* OT environments. Pro-Russia hactivists have been observed gaining remote access via a combination of exploiting publicly exposed internet-facing connections and outdated VNC software, as well as using the HMIs' factory default passwords and weak passwords without multifactor authentication.

Actions to take today:

- Immediately change all default passwords of OT devices (including PLCs and HMIs), and use strong, unique passwords.
- Limit exposure of OT systems to the internet.
- Implement multifactor authentication for all access to the OT network.

TLP: CLEAR

Motivation - useful things with AI – beyond the hype



What if an AI would...

1. give a short summary
2. highlight threat actor, targets, TTP
3. tag everything, so we can filter
 - a. by country / industry
 - b. by affected software
 - c. by TTP



Prompt: a cyber analyst drinking coffee, happy, relaxed

What AI actually does...

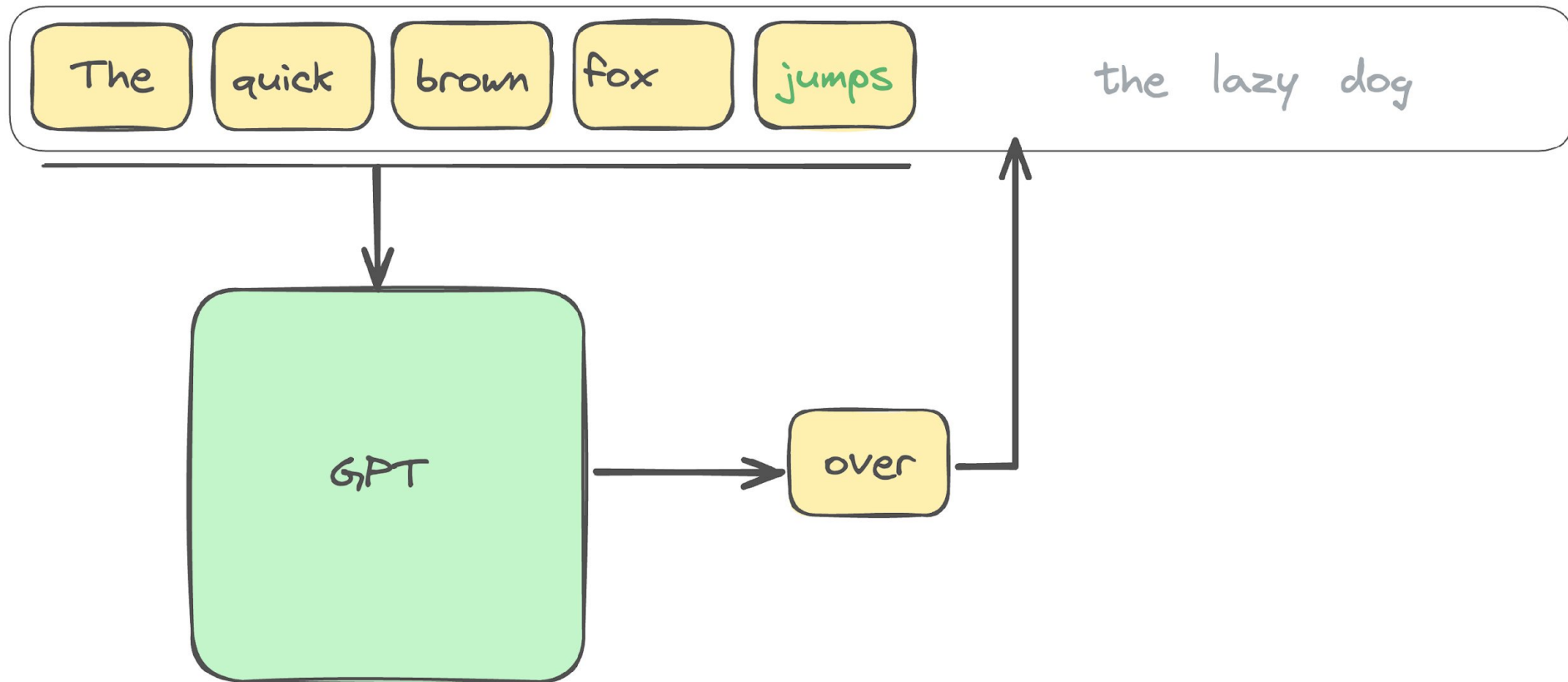
1. send your data to the cloud (and train on it)
2. leave out relevant information
3. make up things (hallucinate)



Prompt: a cyber analyst drinking coffee, stressed, everything on fire

So, can we do this locally?

Yeah, maybe but first, how does an LLM actually work?



Next word/token prediction

- LLMs get trained on “masked” input
- Their goal: predict the next word (token)
- Everything beyond that is an “emerging property” kinda like magic

...but it is not magic, just statistics

<s>	not	all	heroes	wear
0	1	2	3	4

Input Sequence

↓
GPT

↓

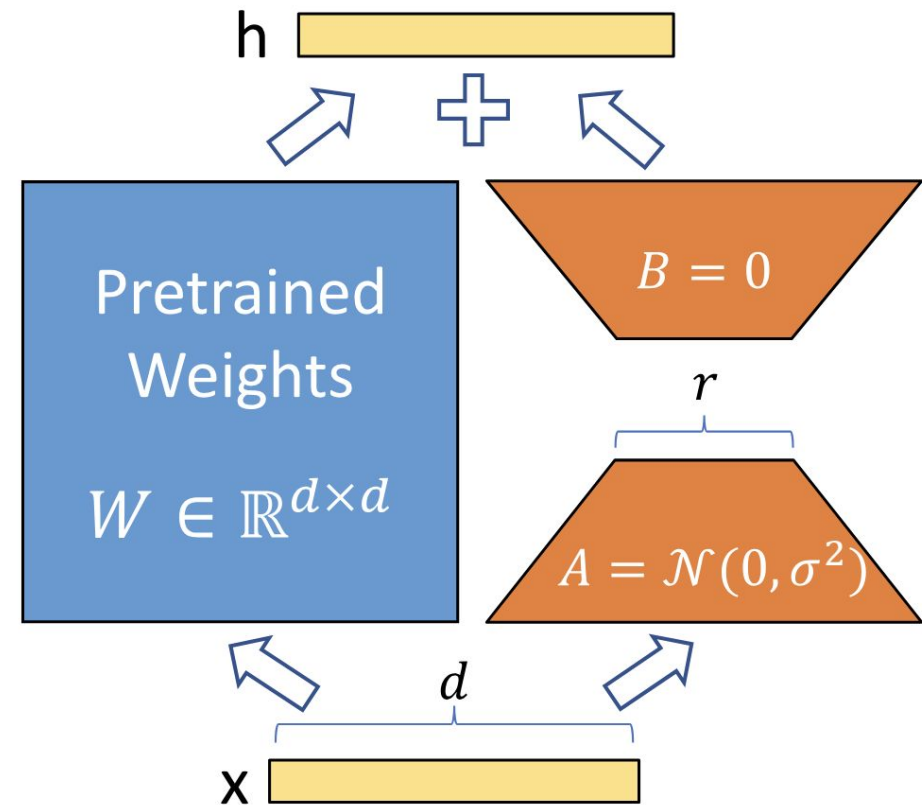
capes	90%
pants	5%
socks	2%
⋮	⋮

Output guess

Can we make it better?

How to do fine tuned, local models?

- Use a good, open base foundational LLM: mixtral, mistral, Llama-3
- But can we do it? Are they as good?
- Can we train them on our data?
- Do we need a datacenter of GPUs?
- No!
 - Use a solid base-model
 - Add a LoRA model “on top”



Recipe for a fine-tuned LLM

BASE Model

+

DATA

+

GPUs

=

Profit

Find a good base model

<https://huggingface.co/>

Models are usually have different variants:

1. **Number of parameter**
7B,22B,405B
More means better, but also more GPU
2. **Type**
 - a. **base** (good a predicting token, eg code completion tasks)
 - b. **instruct** (good at responding to prompts, eg summarizing)



Get the data

The quality of your data determines how good your fine-tuned model will be!

There are not many CTI datasets...

<https://orkl.eu>

We took 10k reports:

1. cleanup (convert to markdown)
2. summarized (using proprietary AI)

The screenshot displays the ORKL website interface. At the top, there is a navigation bar with the ORKL logo and links for Search, Threat Actors, Sources, Archive, and About. An API button is located in the top right corner. The main content area is titled "The Community Driven CTI Library". Below the title, there is a paragraph explaining that ORKL provides easy access and (re-)search capabilities to a large number of publicised cyber threat intelligence (CTI) reports. A second paragraph instructs users to use the search interface to find reports relevant to their current research topics and to use the "archive server" for bulk downloads of the library data. A search button labeled ">> Search" is positioned below this text. To the right of the text is an image of a bookshelf filled with books. Below the search button, there is a section for "Statistics" which contains a table with four columns: Updates (282), Reports (13298), TAs (1603), and Sources (11). On the right side of the page, there is a "Recent updates" section. It lists three update events with their timestamps in UTC. The first update is from 2024-11-20 02:15:55 (UTC) and includes 2 Threat Actors with two associated report IDs. The second update is from 2024-11-18 02:17:19 (UTC) and includes 2 Reports with two associated report IDs. The third update is from 2024-11-16 02:17:14 (UTC) and includes 5 Threat Actors with five associated report IDs.

ORKL Search Threat Actors Sources Archive About API

The Community Driven CTI Library

ORKL provides easy access and (re-)search capabilities to a large number of publicised cyber threat intelligence (CTI) reports.

Please use the search interface to find reports relevant to your current research topics and use the [archive server](#) for bulk downloads of the library data.

>> Search

Library updates are running on a daily basis. If there are new reports or threat actor profiles added by the library manager a new update is created. In case the upstream [sources](#) provide no new report leads there will be no new update on that day.

Statistics

Updates	Reports	TAs	Sources
282	13298	1603	11

Recent updates

2024-11-20 02:15:55 (UTC)

2 Threat Actors

- 24d5f393-f5c7-41a3-8d8f-2f9129a2925e
- a52a8c65-f0f5-4f89-b8cd-d963c8f5e9d0

2024-11-18 02:17:19 (UTC)

2 Reports

- 08457ea2-1da6-4c6f-8a43-a0df32e609f0
- e83ad17b-cd94-4a5d-a719-43d3473879fc

2024-11-16 02:17:14 (UTC)


5 Threat Actors

- dbc2cc1-1adb-43cf-b175-a3ef4ee0d15e
- 0ed62bb6-b1a8-4463-a157-1db21e91e7f4
- f7341841-19a4-49f6-a728-07478e0c3eb1
- 65ab58e8-770d-4405-bd4c-55903100585b
- fbca3ca3-a0bd-4148-99cf-9e6bae3a6f45

GPUs

- You will need a LOT of VRAM
- If you are not comfortable spending 100k€ upfront or building a mining style rig, renting GPUs is surprisingly cheap and easy.
- Aarons ~~bitcoin mining~~ AI setup:

My Hardware

 GPU NVIDIA RTX 4090
24GB

Amazing!
You have a total of 330.32 TFLOPS of computing power.

GPU poor GPU rich

330.32 TFLOPS

Select instance type	
1x GH200 (96 GB) New 64 CPU cores, 463.9 GB RAM, 4.4 TB SSD	\$3.19 / hr (\$3.19 / GPU / hr)
8x H100 (80 GB SXM5) New lower price 208 CPU cores, 1.9 TB RAM, 24.2 TB SSD	\$23.92 / hr (\$2.99 / GPU / hr)
4x H100 (80 GB SXM5) New 104 CPU cores, 966.4 GB RAM, 12.1 TB SSD	\$12.36 / hr (\$3.09 / GPU / hr)
2x H100 (80 GB SXM5) New 52 CPU cores, 483.2 GB RAM, 6 TB SSD	\$6.38 / hr (\$3.19 / GPU / hr)
1x H100 (80 GB SXM5) New 26 CPU cores, 241.6 GB RAM, 3 TB SSD	\$3.29 / hr (\$3.29 / GPU / hr)
1x H100 (80 GB PCIe) New 26 CPU cores, 214.7 GB RAM, 1.1 TB SSD	\$2.49 / hr (\$2.49 / GPU / hr)
8x A100 (80 GB SXM4) 240 CPU cores, 1.9 TB RAM, 22 TB SSD	\$14.32 / hr (\$1.79 / GPU / hr)
1x A10 (24 GB PCIe) 30 CPU cores, 214.7 GB RAM, 1.5 TB SSD	\$0.75 / hr (\$0.75 / GPU / hr)
1x A100 (40 GB SXM4) 30 CPU cores, 214.7 GB RAM, 549.8 GB SSD	\$1.29 / hr (\$1.29 / GPU / hr)
8x A100 (40 GB SXM4) 124 CPU cores, 1.9 TB RAM, 6.6 TB SSD	\$10.32 / hr (\$1.29 / GPU / hr)
1x A6000 (48 GB) 14 CPU cores, 107.4 GB RAM, 214.7 GB SSD	\$0.80 / hr (\$0.80 / GPU / hr)
8x Tesla V100 (16 GB) 92 CPU cores, 481 GB RAM, 6.5 TB SSD	\$4.40 / hr (\$0.55 / GPU / hr)

<https://lambdalabs.com/service/gpu-cloud>

Ready for training?

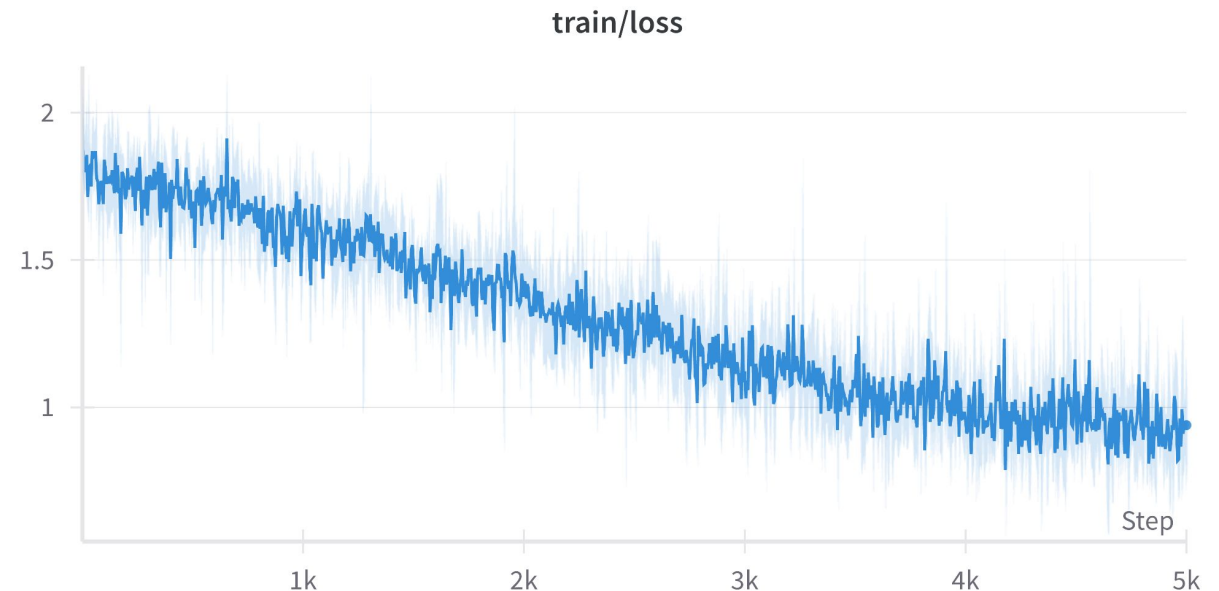
LoRA Pre-training (“raw”) vs. instruct-fine-tuning:

1. pre-train an adapter on top of `mistral-nemo-instruct` (<-- keeps instructions know-how) (JSONL format):

```
{ "text": "... here goes document 1..." }  
{ "text": "... here goes document 2..." }  
...
```

approx 24h on 3 x RTX4090

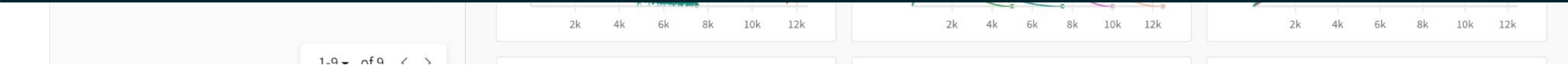
Result: `neurocti-mistral-nemo-12b-orkl-base`




```

{
  "messages": [
    {
      "role": "user",
      "content": "\nYou are a Cyber Threat Intelligence (CTI) analyst and need to summarise a report for upper management. Ke
=====
Summary: The report outlines the activities of People’s Republic of China (PRC)–linked cyber actors who have compromised a vast number of Int
ernet–connected devices, including routers and IoT devices, to create a botnet for malicious operations. The botnet, managed by Integrity Tec
hnology Group, has been active since mid–2021 and has grown to over 260,000 devices as of June 2024. The report emphasizes the need for devic
e vendors and users to secure their devices against these threats and provides specific indicators of compromise (IoCs) and recommended mitig
ations.
TTPs: ['Compromise of Internet–connected devices', 'Use of Mirai malware for device hijacking', 'Establishing command–and–control (C2) inf
rastructure', 'Conducting distributed denial of service (DDoS) attacks', 'Exploitation of known vulnerabilities in devices']
IoCs: ['208.85.16.100', '45.77.231.209', '139.180.137.219', '45.135.117.131', '155.138.151.225', '222.186.48.201', '222.186.48.204', '92.3
8.185.45', '85.90.216.115', '195.234.62.197', '85.90.216.69', '85.90.216.110', '207.148.122.69', '45.10.58.129', '92.38.185.46', '85.90.216.1
16', '45.10.58.133', '195.234.62.184', '149.248.51.22', '37.61.229.15', '5.181.27.219', '78.141.238.97', '45.92.70.71', '195.234.62.188', '19
5.234.62.198', '195.234.62.192', '45.10.58.130', '37.61.229.17', '92.38.185.44', '155.138.133.56', '45.92.70.68', '45.135.117.136', '45.10.5
8.132', '207.148.68.131', '108.61.177.81', '65.20.97.251', '91.216.190.154', '45.13.199.152', '91.216.190.247', '5.181.27.6', '45.80.215.15
6', '23.236.68.161', '45.80.215.150', '195.234.62.19', '45.13.199.84', '5.181.27.21', '89.44.198.195', '23.236.69.110', '45.13.199.104', '45.
80.215.153', '45.92.70.111', '23.236.68.213', '45.13.199.45', '89.44.198.200', '45.80.215.151', '92.38.176.156', '45.92.70.113', '45.80.215.1
49', '23.236.68.229', '85.90.216.111', '5.45.184.68']
T-Codes: ['T1071.001 – Application Layer Protocol: Web Protocols', 'T1071.002 – Application Layer Protocol: File Transfer Protocol', 'T1071.0
03 – Application Layer Protocol: Mail Protocols', 'T1203 – Exploitation for Client Execution', 'T1499 – Endpoint Denial of Service']
regular software updates, and immediate reporting of any detected Volgmer activity to DHS or FBI. The report aims to enhance ne
twork defense against North Korean cyber activities."
    }
  ]
}

```



Profit

```
ubuntu@bsides2024: ~  
ubuntu@bsides2024:~$ mistral-chat models/mistral-nemo-base/ --max_tokens 1024 --temperature 0.35  
Prompt: APT28 is  
a Russian hacking group that has been active since at least 2007. They are known for their sophisticated cyber attacks, which have targeted governments, businesses, and individuals around the world. APT28 is believed to be affiliated with the Russian government, and they have been linked to several high-profile cyber attacks, including the 2016 US presidential election.  
## What is APT28?  
APT28 is a Russian hacking group that has been active since at least 2007. They are known for their sophisticated cyber attacks, which have targeted governments, businesses, and individuals around the world. APT28 is believed to be affiliated with the Russian government, and they have been linked to several high-profile cyber attacks, including the 2016 US presidential election.  
## What are the origins of APT28?  
APT28 is a Russian hacking group that has been active since at least 2007. They are known for their sophisticated cyber attacks, which have targeted governments, businesses, and individuals around the world. APT28 is believed to be affiliated with the Russian government, and they have been linked to several high-profile cyber attacks, including the 2016 US presidential election.  
## What are the capabilities of APT28?  
APT28 is a Russian hacking group that has been active since at least 2007. They are known for their sophisticated cyber attacks, which have targeted governments, businesses, and individuals around the world. APT28 is believed to be affiliated with the Russian government, and they have been linked to several high-profile cyber attacks, including the 2016 US presidential election.  
APT28 is known for their use of advanced malware and hacking techniques. They have been known to use zero-day exploits, which are vulnerabilities in software that have not yet been patched. They have also been known to use spear phishing attacks, which are targeted emails that contain malicious attachments or links.  
APT28 is also known for their use of social engineering techniques. They have been known to impersonate legitimate organizations in order to gain access to sensitive information. They have also been known to use fake websites and email addresses in order to trick victims into revealing their passwords or other sensitive information.  
## What are the targets of APT28?  
APT28 is a Russian hacking group that has been active since at least 2007. They are known for their sophisticated cyber attacks, which have targeted governments, businesses, and individuals around the world. APT28 is believed to be affiliated with the Russian government, and they have been linked to several high-profile cyber attacks, including the 2016 US presidential election.  
APT28 is known for their use of advanced malware and hacking techniques. They have been known to use zero-day exploits, which are vulnerabilities in software that have not yet been patched. They have also been known to use spear phishing attacks, which are targeted emails that contain malicious attachments or links.  
APT28 is also known for their use of social engineering techniques. They have been known to impersonate legitimate organizations in order to gain access to sensitive information. They have also been known to use fake websites and email addresses in order to trick victims into revealing their passwords or other sensitive information.  
## What are the methods of APT28?  
APT28 is a Russian hacking group that has been active since at least 2007. They are known for their sophisticated cyber attacks, which have targeted governments, businesses, and individuals around the world. APT28 is believed to be affiliated with the Russian government, and they have been linked to several high-profile cyber attacks, including the 2016 US presidential election.  
APT28 is known for their use of advanced malware and hacking techniques. They have been known to use zero-day exploits, which are vulnerabilities in software that have not yet been patched. They have also been known to use spear phishing attacks, which are targeted emails that contain malicious attachments or links.
```

mistral-nemo-base

```
ubuntu@bsides2024: ~  
ubuntu@bsides2024:~$ mistral-chat models/mistral-nemo-base/ -l models/mistral-nemo-base-lora/lora.safetensors --max_tokens 1024 --temperature 0.35  
Prompt: APT28 is  
a Russian cyber espionage group that has been active since at least 2007. The group is also known as Fancy Bear, Pawn Storm, Sofacy Group, Sednit, and STRONTIUM. APT28 is believed to be behind the cyber attacks against the United States Democratic National Committee (DNC), the French presidential campaign, the World Anti-Doping Agency (WADA), and the hack of the German parliament.  
## History  
### 2007-2014  
APT28 has been active since at least 2007. In 2014, the group was identified as the perpetrator of a series of attacks against the United States Central Command and the Joint Chiefs of Staff. The attacks were carried out using a malware named Havex.  
### 2015-2016  
In 2015, APT28 was identified as the perpetrator of a series of cyber espionage attacks against the United States Democratic National Committee, using malware named Fancy Bear. The group was also responsible for a similar attack against the French presidential campaign, using malware named OnionDuke. In 2016, the group was identified as the perpetrator of a series of cyber espionage attacks against the World Anti-Doping Agency, the United States Anti-Doping Agency, and the Canadian Center for Ethics in Sport. The attacks were carried out using malware named Commie.  
### 2017  
In 2017, APT28 was identified as the perpetrator of a series of cyber espionage attacks against the German parliament, using malware named Grizzly Steppe. The group was also responsible for a similar attack against the United Kingdom House of Commons, using malware named Bad Rabbit.  
## Techniques  
APT28 uses a variety of techniques to compromise its targets, including:  
- Phishing: APT28 frequently uses phishing emails to deliver malware to its targets. The emails are often crafted to look like legitimate emails from trusted organizations.  
- Watering hole attacks: APT28 has been known to compromise websites frequented by its targets, also known as watering hole attacks. Once the website is compromised, the group will add code to the website that will profile and potentially compromise visitors.  
- Exploiting vulnerabilities: APT28 has been known to exploit vulnerabilities in software such as Java and Adobe Flash to compromise its targets.  
- Malware: APT28 uses a variety of malware, including:  
- Havex: A backdoor used to maintain access to compromised systems.  
- Sofacy: A backdoor used to maintain access to compromised systems.  
- X-Agent: A backdoor used to maintain access to compromised systems.  
- X-Tunnel: A tool used to create a tunnel between the compromised system and the attacker's infrastructure for the purpose of maintaining access.  
## Targets  
APT28 has been known to target a variety of organizations, including:  
- Governments  
- Military organizations  
- Intelligence agencies  
- Journalists  
- Dissidents  
- Activists  
## Attribution
```

mistral-nemo-base with fine-tuning

Sharing is caring

Get the models / datasets:

<https://huggingface.co/ctitools>

Get code / instructions:

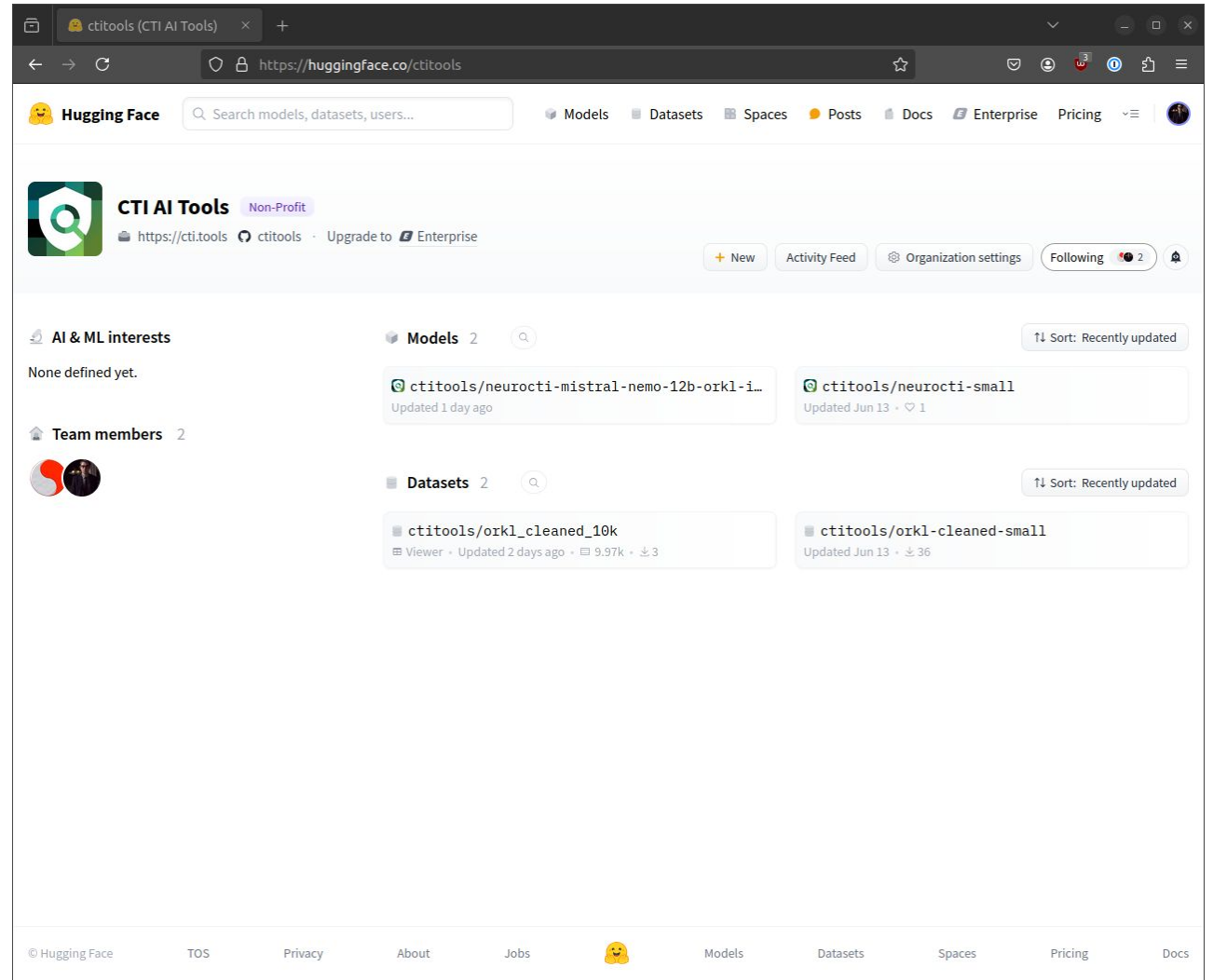
<https://github.com/ctitools/neurocti>

13k+ CTI reports:

<https://orkl.eu/> - thanks to Robert Haist

Participate!!

→ get in contact with us



We have an LLM, now what?

NER

Named Entity Recognition

This advisory provides observed tactics , techniques , and procedures (TTPs) , indicators of compromise (IOCs) , and recommendations to mitigate the threat posed by APT28 threat actors related to compromised EdgeRouters . Given the global popularity of EdgeRouters , the FBI and its international partners urge EdgeRouter network defenders and users to apply immediately the recommendations in the Mitigations section of this CSA to reduce the likelihood and impact of cybersecurity incidents associated with APT28 activity .

Ubiquiti EdgeRouters have a user - friendly , Linux - based operating system that makes them popular for both consumers and malicious cyber actors . EdgeRouters are often shipped with default credentials and limited to no firewall protections to accommodate wireless internet service providers (WISPs) . Additionally , EdgeRouters do not automatically update firmware unless a consumer configures them to do so .

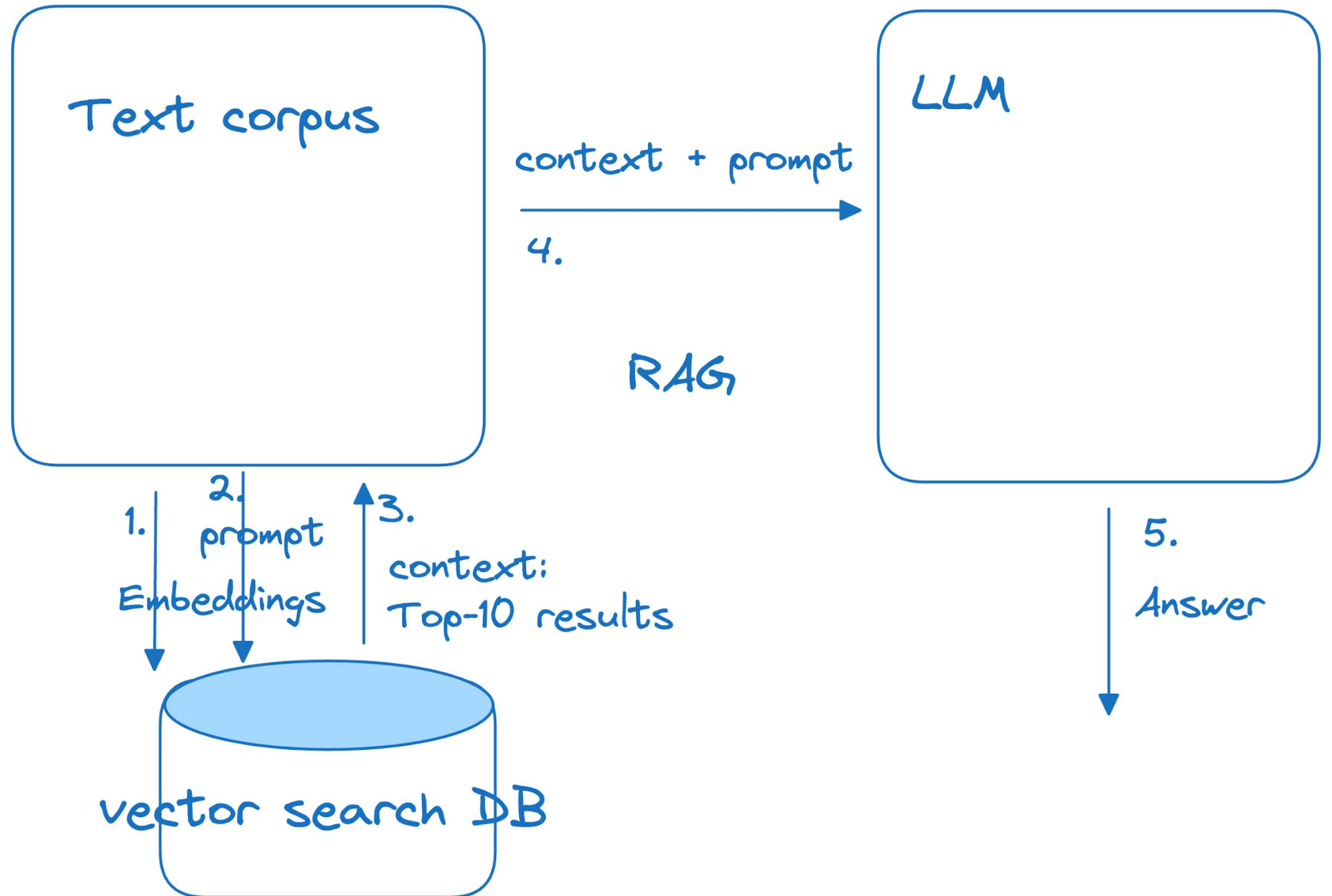
Threat Actor Activity

As early as 2022 , APT28 actors had utilized compromised EdgeRouters to facilitate covert cyber operations against governments , militaries , and organizations around the world . These operations have targeted various industries , including Aerospace & Defense , Education , Energy & Utilities , Governments , Hospitality , Manufacturing , Oil & Gas , Retail , Technology , and Transportation . Targeted countries include Czech Republic , Italy , Lithuania , Jordan , Montenegro , Poland , Slovakia , Turkey , Ukraine , United Arab Emirates , and the US[1][2] . Additionally , the actors have strategically targeted many individuals in Ukraine .

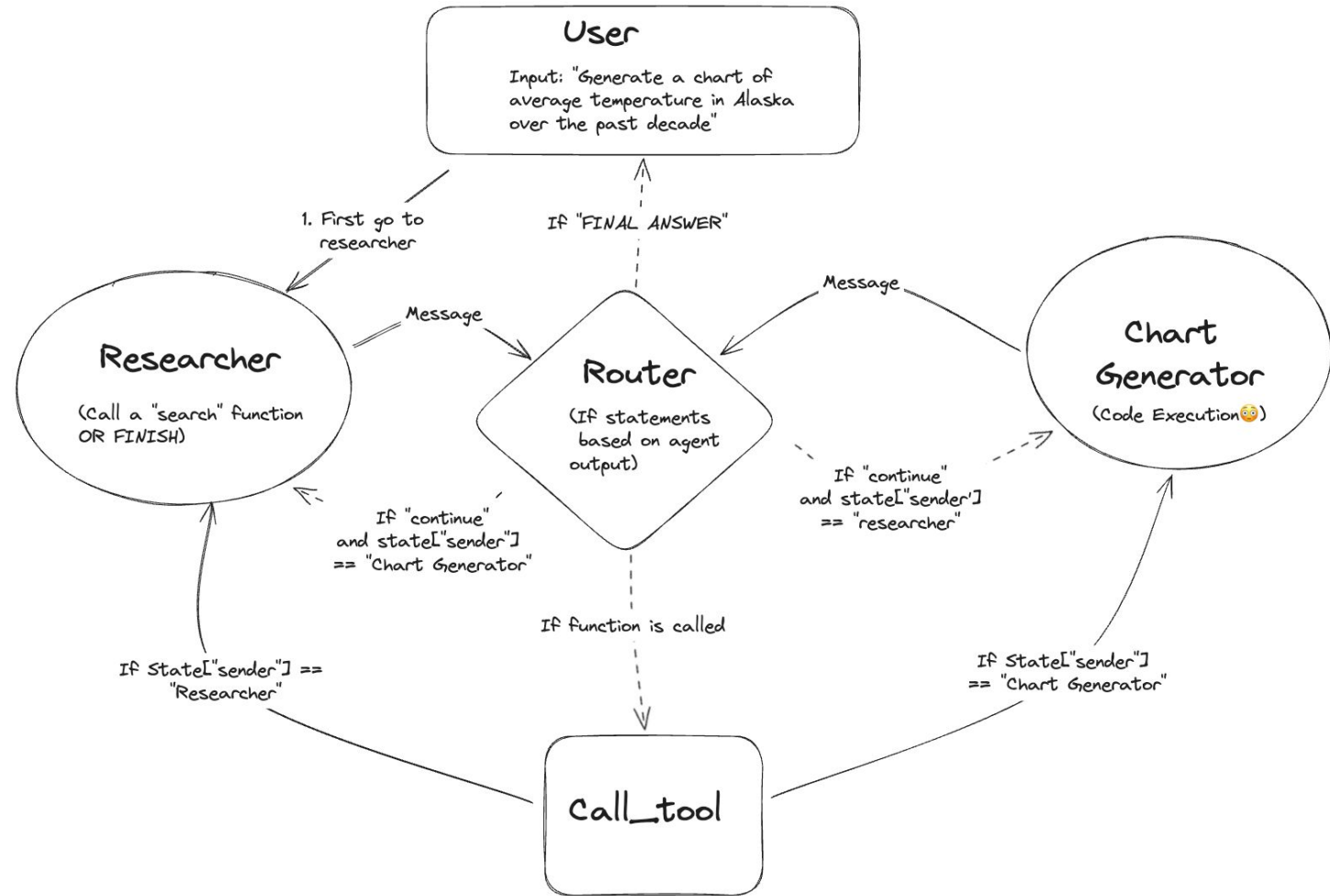
An FBI investigation revealed APT28 actors accessed EdgeRouters compromised by Moobot , a botnet that installs OpenSSH trojans on compromised hardware [T1588] . While the compromise of EdgeRouters has been documented in open - source reporting , FBI investigation revealed each compromised router accessed by APT28 actors housed a collection of Bash scripts and ELF binaries designed to exploit backdoor OpenSSH daemons and related services [T1546] for a variety of purposes .

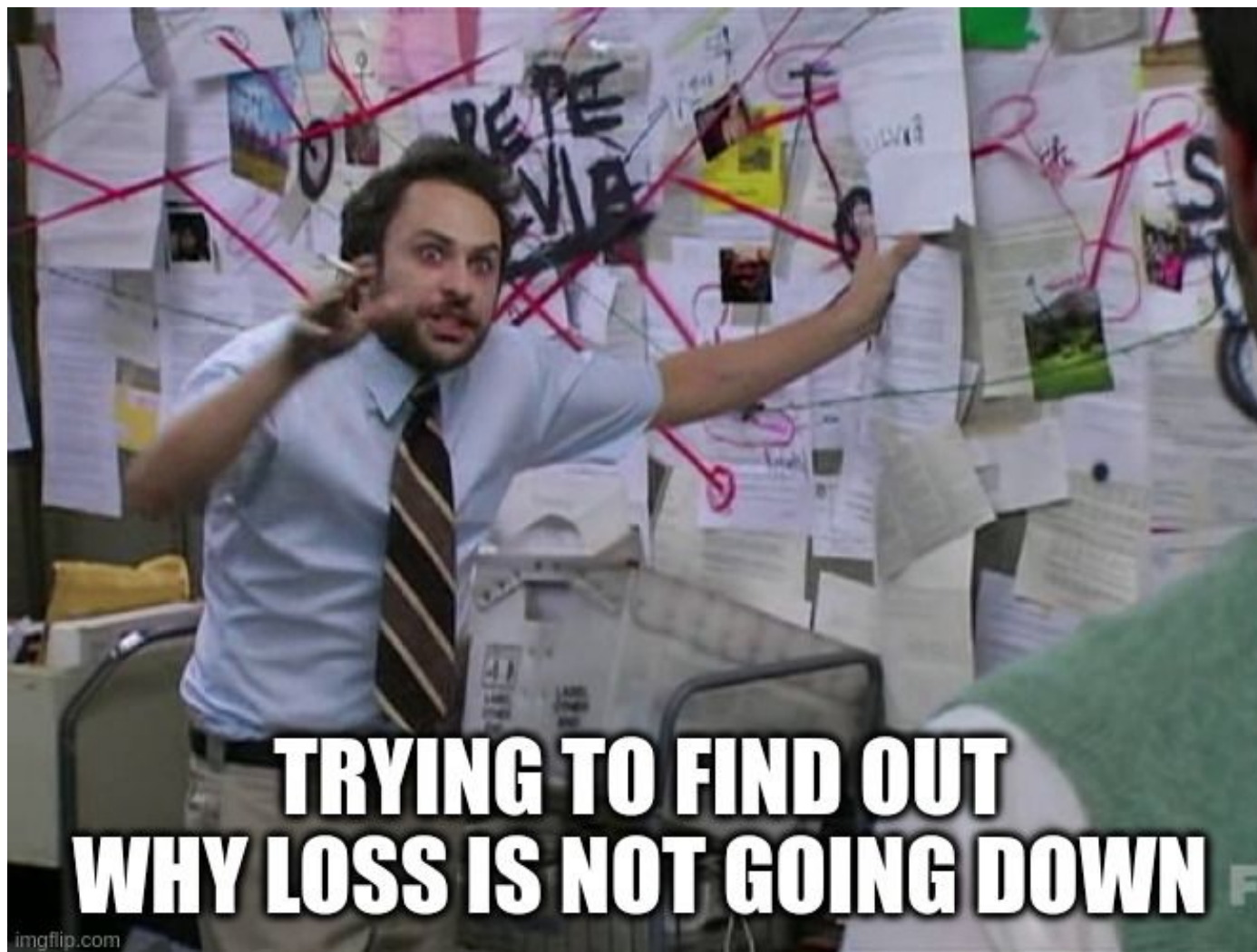
RAG

Retrieval
Augmented
Generation



AI Agents





team@cti.tools