

NEUROCTI - A CUSTOM FINE-TUNED LLM FOR CTI

BENCHMARKING, SUCCESSES AND LESSONS LEARNED

Aaron Kaplan, Alexandre Dulaunoy,
Jürgen Brandl, Paolo Di Prodi

継続は力なり

Continuity is power

Intro speakers

Aaron Kaplan

- Self-employed / EC-DIGIT-CSIRC
- Previously 12 years @ CERT.at, Austria

Alexandre Dulaunoy

- Leading CIRCL.lu, Makes and breaks stuff

Jürgen Brandl

- senior cyber security analyst at the Federal Ministry of the Interior, Austria

Paolo di Prodi → could not make it

- Founder PRIAM.AI, previously senior data scientist at Microsoft and Fortinet.

DISCLAIMER

Aaron

- All errors are mine to keep
- I present this here as a sole proprietor company under my own name

Alexandre Dulaunoy

- CIRCL.lu

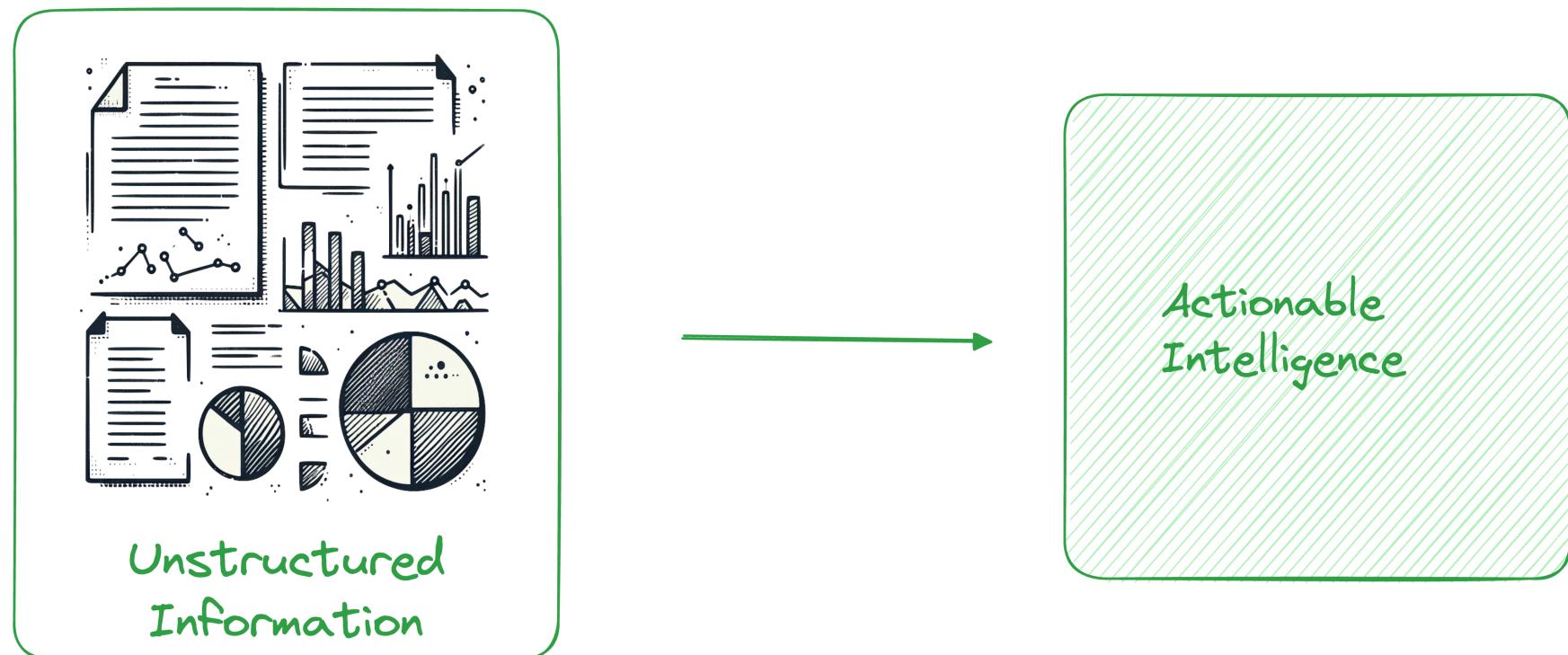
Jürgen Brandl

- Opinions are my own

Overview of the talk

- Motivation
- Use-cases for LLMs in CTI
- Short recap: how do LLMs actually work?
 - Inference, training, fine-tuning
- Obstacles to using LLMs for CTI → we need local LLMs
- Needed: benchmarking- and training datasets
- CTI.tools
- Fine-tuning a local LLM: initial results
- Integration with MISP
- Status-quo and next steps

Motivation - useful things with AI – beyond the hype

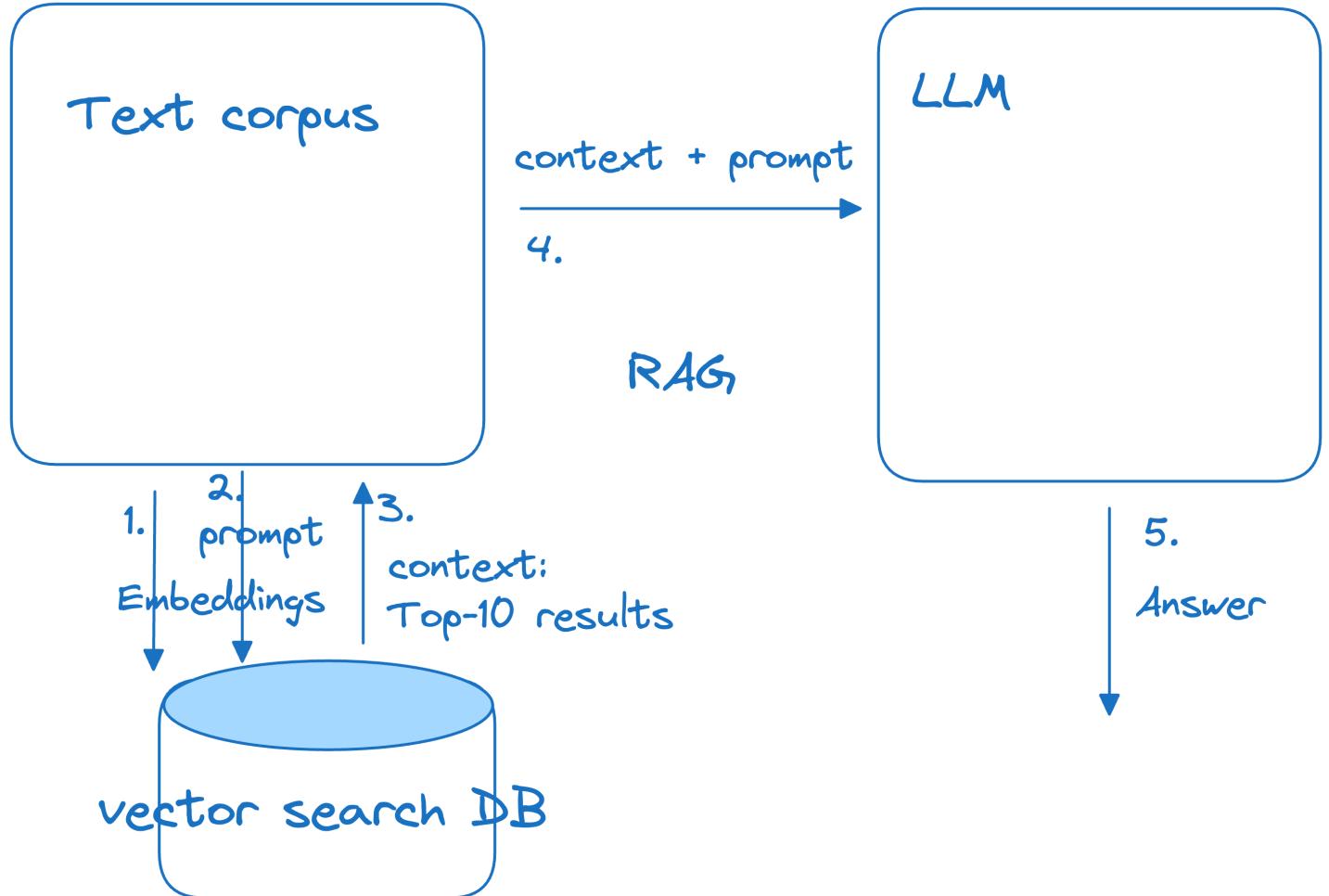


In the AI SIG, we identified 5 main uses cases for AI in CTI

1. ***Summarization:*** CTI analysts need to digest a lot of threat intel reports
2. **RAG:** Analysts might want to search (**question-answering**) on CTI reports
3. ***NER:*** Analysts would like to get the information out (“**information extraction**”) from CTI reports
4. **T-Codes mapping:** Extracting MITRE TTPs from text reports
5. **Knowledge graph / STIX 2.1 / MISP standard:** Extract relationships

RAG

Retrieval
Augmented
Generation



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 .

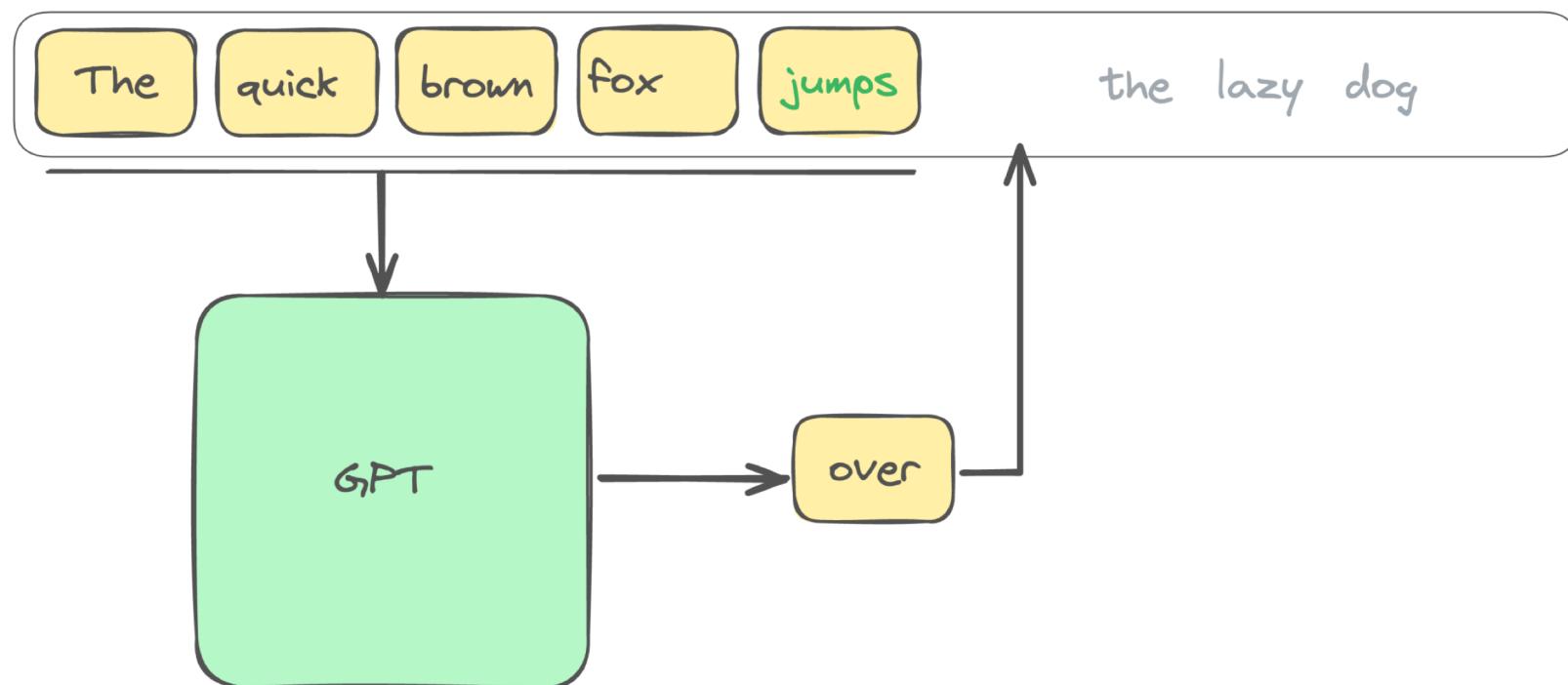
Ubiquiti EdgeRouters have a user - friendly , Linux - based operating system that makes them popular for both consumers and malicious cyber actors . EdgeRouters are often shipped with default credentials and limited to no firewall protections to accommodate wireless internet service providers (WISPs) . 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 .

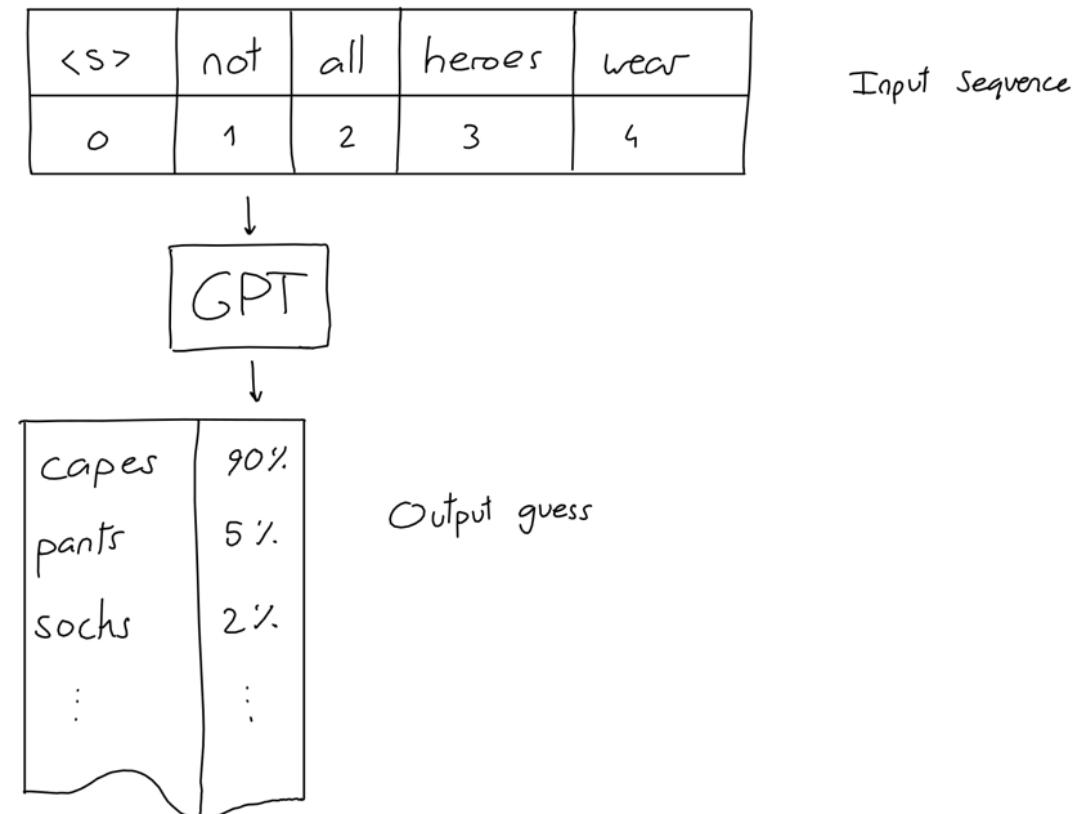
An FBI investigation revealed APT28 actors accessed EdgeRouters compromised by Moobot , a botnet that installs OpenSSH trojans on compromised hardware [T1588] . While the compromise of EdgeRouters has been documented in open - source reporting , FBI investigation revealed each compromised router accessed 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 .

Can we do that locally? Yeah, maybe but first, how does an LLM actually work?



Next work/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

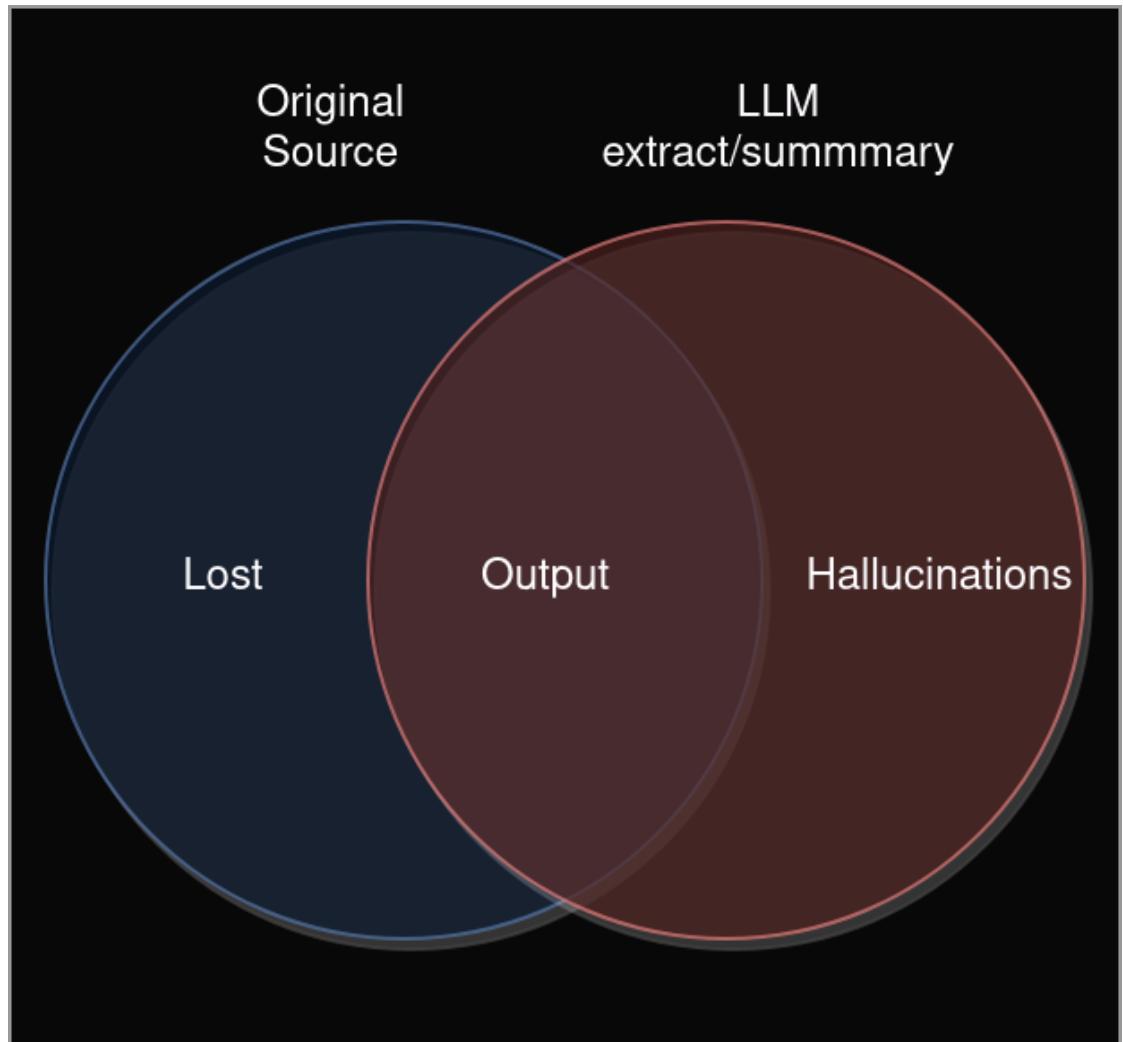


Obstacles to using LLMs for CTI

And possible solutions

Problems with LLMs

- **Hallucinations**
- What are the risks for CTI reports?
- Political / business decisions taken because the AI generated a wrong summary?
- Start war because of wrong intel?



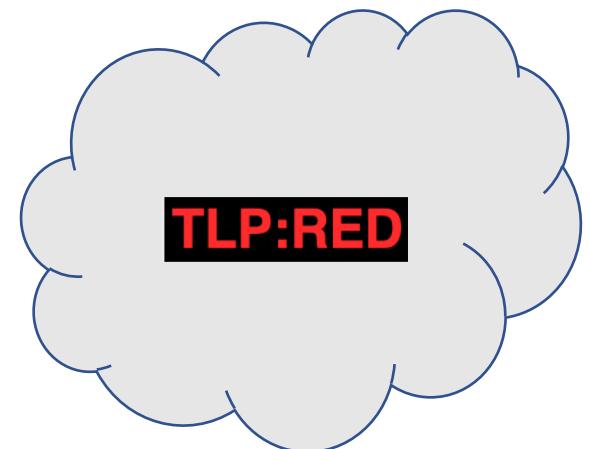
Guardrails

- Make the LLM adhere to a strict, smaller vocabulary
- Use easier, smaller models to keep the LLM “in bay”
- Mix of Experts (**MoE**)
- Use **RAG** for limiting the context the LLM may even use
- Few-shot prompting
- Fine-tuning (**LoRA**)
- Custom training (continuous pre-training)



Sending my CTI reports or requests to a third-party?

- Data goes to AI providers
- **How is my data used?**
 - For training updated models?
 - might end up in someone else's output
 - Across users (session leakage)
- **Legal implication**
 - TI feed providers don't allow to re-share
→ breach of contract
 - Are you using VS Code for editing local files? Co-Pilot?
→ The files get uploaded !
 - PII and privacy-related regulation



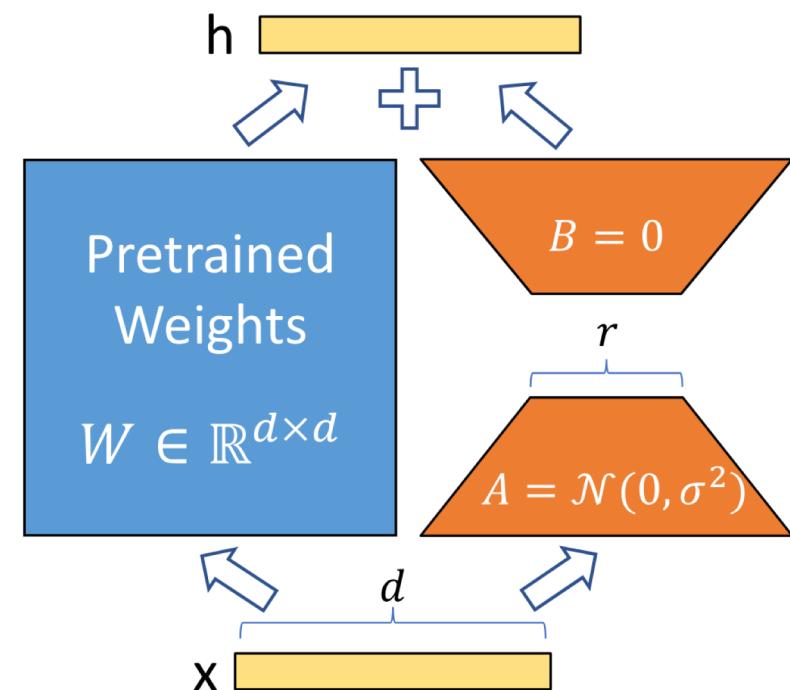
→ We need local models!

But: can a local model do this just as well?

Local models FTW! ... let's see...

How to do fine tuned, local models?

- Use a good, open base foundational LLM: mixtral, mistral, Llama-2, 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”



Datasets, benchmarking

the need for high quality data for training and benchmarking

Related research & existing datasets

[Search](#)[Threat Actors](#)[Sources](#)[Archive](#)[About](#)[API](#)

Threat Actor Profile

ID	20d3a08a-3b97-4b2f-90b8-92a89089a57a
Main Name	APT29
Source	MITRE
Source Name	MITRE:APT29
Aliases/Synonyms	APT29 IRON RITUAL IRON HEMLOCK NobleBaron Dark Halo StellarParticle NOBELIUM UNC2452 YTTRIUM The Dukes Cozy Bear CozyDuke SolarStorm Blue Kitsune UNC3524 Midnight Blizzard

13000 reports
Mixed quality
Converted from PDF

Shout-out do Robert Haist!
Thanks!

Problems with datasets

- No clear standardized taxonomy of NER categories
- Data is messy, hard to train on
- No standard benchmark dataset for CTI LLMs

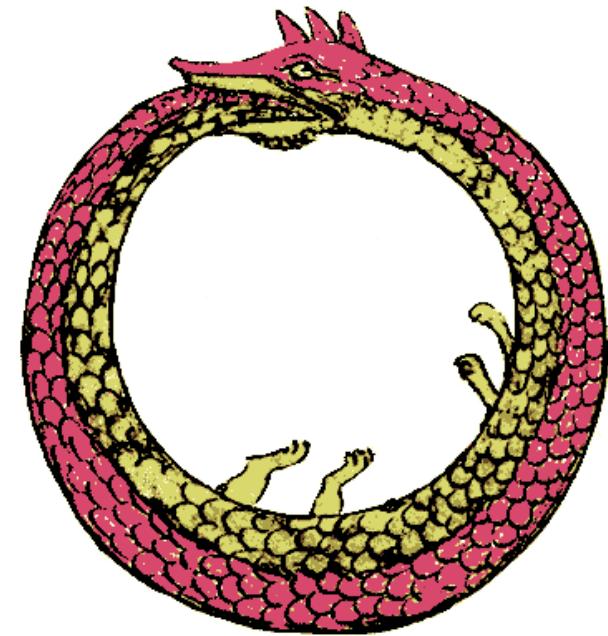
CTI.tools AI workbench

For annotation

CTI.tools - overview

Goal: Make AI tooling accessible
to the CTI community*

* while solving the CTI dataset
problem



CTI.tools - goals



Checklist to get people to contribute:

- ✓ Provide a benefit to the users
- ✓ Easy, intuitive and fun to use
- ✓ Usable by everyone with internet
- ✓ Everybody profits

Live Demo

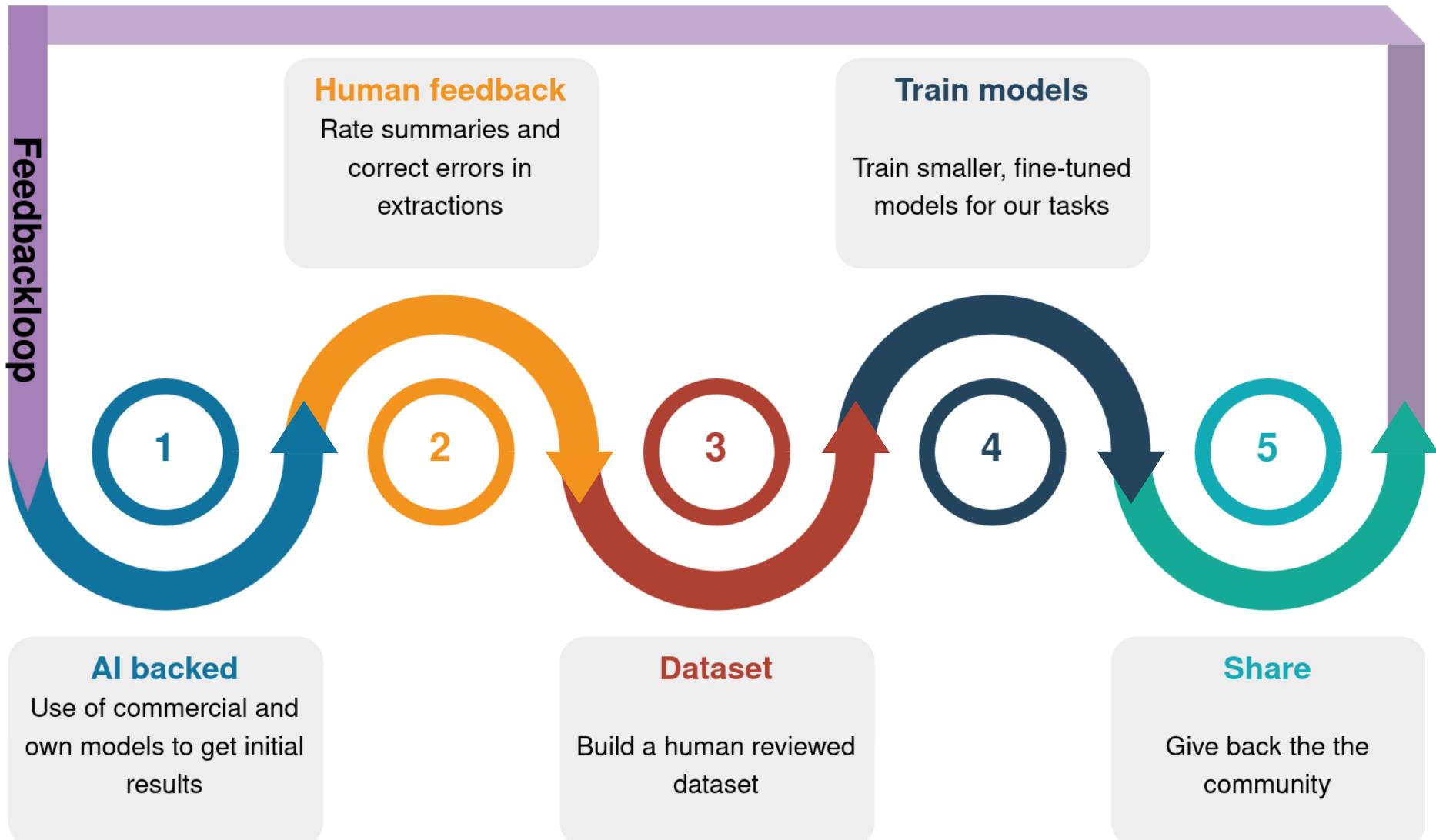
Goal: making AI for CTI accessible to everyone

The screenshot shows the 'CTI SUMMARIZER [DOC.ID:115]' page. At the top, there are five 5x5 grid rating scales for Relevance, Faithfulness, Fluency, Coherence, and Coverage. Below these are sections for 'SUMMARY' (Summary for EdgeRouter Network Defenders and Users, mentioning APT28 threat actors and EdgeRouter users), 'SUMMARY POINTS' (APT28 targeting EdgeRouters, compromised routers, various industries, Moobot botnet, and OpenSSH backdoor), 'TOOLS USED' (Moobot botnet, Bash scripts, ELF binaries), 'TACTICS, TECHNIQUES, AND PROCEDURES (TTPS)' (Compromising EdgeRouters, Exploiting backdoor OpenSSH daemons), and 'CVES' (none listed). The bottom section shows 'WHEN DISCOVERED: 2022'.

Summaries and TL;DR for reports

The screenshot shows the 'CTI NER EXTRACTOR [DOC.ID:114]' page. On the left, there is a vertical stack of colored boxes representing extracted entities: DATE (blue), TIME (pink), GEO_LOCATION (green), ORGANIZATION (purple), SECTOR (yellow), THREAT_ACTOR (magenta), EXPLOIT_NAME (light green), MALWARE (red), OS (teal), SOFTWARE (orange), HARDWARE (light blue), USERNAME (olive green), TTP (dark green), CODE_CMD (brown), and CLASSIFICATION (pink). To the right of this list is a detailed text summary about APT28 threat actors targeting EdgeRouters, mentioning their use of compromised routers, various industries targeted, and the strategic targeting of individuals in Ukraine. It also discusses the popularity of EdgeRouters, their default credentials, and the lack of automatic firmware updates. The text is heavily annotated with color-coded terms from the extracted entities.

Extract information from reports



Recap & call to action!

- AI powered workbench to turn CTI texts into actionable information
- Built to help you, so you can help us in making it better!
- It's a crowd-sourced effort, **we need you**. Please get in touch with the authors if you can contribute labeling skills.
- **Even just quickly labelling 10 reports would help us.**
- All the results will be available to everyone who participates.
- When it comes to training AI, it's never too much data, only too little...

Training a local LLM

Our approach: LoRA on orkl.eu

- Orkl.eu - ~ 13k CTI reports, slides, etc.
- Problem: PDFs to text
- We used 5k reports
- Found out, we need to clean up the reports
- Used LLMs to clean up and convert to markdown
- → train with it

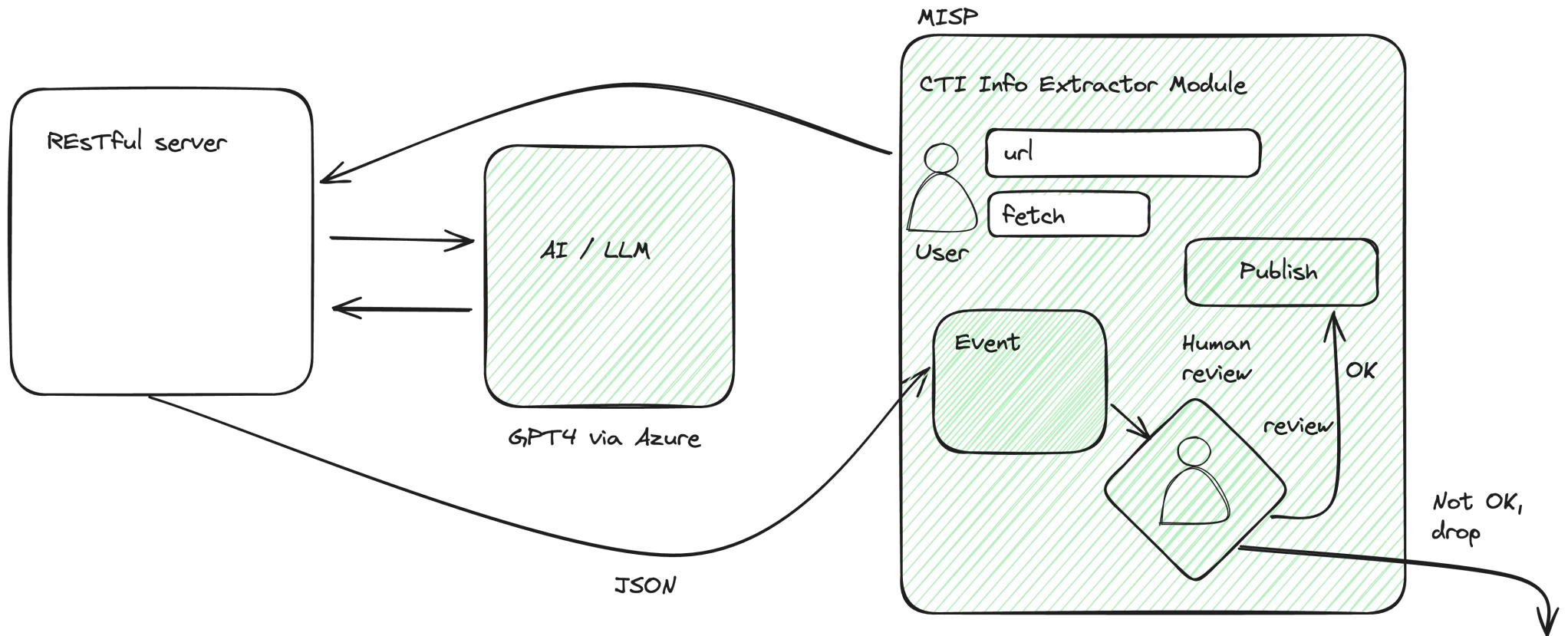
Demo

Integration into MISP

Objectives

- Want to have **a generic and standardized RESTful API interface** so that
 - We can talk with a local LLM
 - ... also with a remote LLM (openAI, openAI vs. Azure, Anthropic (Claude), ...)
- Enforcing a consistent answer format (JSON)
 - Example: unstructured info into LLM → JSON out
- **Ensuring the analyst flow** in the MISP platform and integration with the MISP event reports format

First integration with MISP



How to install it?

Already in mainline MISP 2.4 (as a PoC)

1. git pull
2. # also make sure, misp-modules is updated, installed, running
3. systemctl apache2 restart

Next make sure that **markdown support is enabled**:

Recommended	Plugin.Enrichment_html_to_markdown_enabled	true	[Enable or disable the html_to_markdown module.] Simple HTML fetcher
Recommended	Plugin.Enrichment_html_to_markdown_restrict	No organisation selected.	Restrict the html_to_markdown module to the given organisation.
Recommended	Pluin.Enrichment censvs enrich enabled	false	[Enable or disable the censvs enrich module.] Censvs.io expansion module

Review the CTI Info Extractor extension

Cortex

Sightings

Workflow

CyCat

[CTIInfoExtractor](#)

enable

Recommended `Plugin.CTIInfoExtractor_enable` true Enable the experimental CTI info extractor plugin to use a connected LLM server to extract additional information from markdown reports.

Recommended `Plugin.CTIInfoExtractor_url` `http://bee.lor
res.org:9090/summarize` The url of the LLM REST service.

Recommended `Plugin.CTIInfoExtractor_authentication` `1234test` The authentication key for the LLM REST service.

[RPZ](#)

Kafka

ZeroMQ

ElasticSearch

[List Event Reports](#)[Add Event Report](#)[View Event Report](#)[Edit Event Report](#)

apt29 bleeping

Send report to LLM



ID	535
UUID	b065
Event	#219
Distribution	Inher
Last update	2024

 [Edit](#) [Split Screen](#)

Waiting for the robot to do its magic...



Russian APT29 hackers' stealthy

EXCLUSIVE: Hackers associated
using two recently discovered sophisticated threats.

The malicious implants are a variant of the GoldMax backdoor for Linux systems and a completely new malware family that cybersecurity company CrowdStrike now tracks as TrailBlazer.

Both threats have been used in StellarParticle campaigns since at least mid-2019 but were identified only two years later, during incident response investigations.

StellarParticle attacks have been attributed to the APT29 hacking group has been running cyber espionage campaigns for more than 12 years and is also known as CozyBear, The Dukes, and Yttrium.

Stealing cookies for MFA bypass In a report shared exclusively with BleepingComputer, cybersecurity company CrowdStrike today describes in detail the latest tactics, techniques, and procedures (TTPs) observed in cyberattacks from the Cozy Bear state-sponsored hackers.

While some of the techniques are somewhat common today, Cozy Bear has been using them long before they became popular:

credential hopping hijacking Office 365 (O365) Service Principal and Application bypassing multi-factor authentication (MFA) by stealing browser cookies stealing credentials using Get-ADReplAccount Credential hopping was the first stage of the attack, allowing the threat actor to log into Office 365 from an internal server that the hackers reached through a compromised public-facing system.

source: CrowdStrike CrowdStrike says that this technique is hard to spot in environments with little visibility into identity usage since hackers could use more than one domain administrator account.

Bypassing MFA to access cloud resources by stealing browser cookies has been used since before 2020. CrowdStrike says that APT29 kept a low profile after decrypting the authentication cookies, likely

[Download PGP public key](#)

Io-res MISP Powered by MISP 2.4.193 - 2024-06-13 01:12:04

[Download PGP public key](#)

Io-res MISP Powered by MISP 2.4.193 - 2024-06-13 01:12:04

Voila! Context and tags

test event 2	
Event ID	193386
UUID	630c3706-69f9-403e-a518-d98ac448b7f6 +≡
Creator org	lo-res.org
Owner org	lo-res.org
Creator user	admin@admin.test
Protected Event (experimental) i	Event is in unprotected mode. Switch to protected mode
Tags	misp-galaxy:threat-actor="Sofacy, Zebrocy" misp-galaxy:threat-actor-country="unknown" misp-galaxy:threat-actor-motivation="Espionage" +
Date	2023-11-02
Threat Level	High
Analysis	Initial
Distribution	Your organisation only
Warnings	<p>Content:</p> <p>Your event has neither attributes nor objects, whilst this can have legitimate reasons (such as purely creating an event with an event report or galaxy clusters), in most cases it's a sign that the event has yet to be fleshed out.</p>
Published	No

Status-quo & next steps

Status quo, next steps

- ✓ Adapting a base model to CTI texts, works
- ✓ Doing NER for CTI texts work
- ✓ Dataset for fine-tuning released
- ✓ Model uploaded to HuggingFace
- ✓ Initial user-documentation
- Training of **larger models** (70b+) WIP
(hint hint, GPUs anyone?)
- Improved dataset needed →
we need you as experts → cti.tools
- Evaluate, re-train again, re-publish model + report (paper)
- Include into cti.tools and improve again



継続は力なり

Call for action

- We are going to release benchmark data + models + code
- Great help: **GPUs**
- Even better: **your participation** / expert knowledge
- Participate in <https://cti.tools>
- Documentation + Code: <https://github.com/ctitools>
- Dataset: <https://huggingface.co/datasets/ctitools/orkl-cleaned-small>
- Models: <https://huggingface.co/ctitools>

Thank you!

Aaron Kaplan

Email: aaron@lo-res.org

Twitter/X: @_aaron_kaplan_

Mastodon:

Jürgen Brandl

Email: first@j3.at

Alexandre Dulaunoy

Mastodon: adulau@infosec.exchange

Email: alexandre.dulaunoy@circl.lu

継続は力なり

Continuity is power