TLP:GREEN

LLMs are Unreliable for Cyber Threat Intelligence

How LLMs show low performance, inconsistency and low calibration in CTI tasks

TLP:GREEN

WhoIsPresenting?

Emanuele Mezzi (e.mezzi@vu.nl)

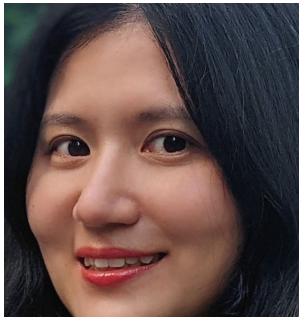
- **AI** and **Security** researcher working on building trustworthy AI
- PhD Candidate at **Vrije Universiteit Amsterdam** and **TNO**
- Contributor to the project **Sec4AI4Sec**
- Co-founder and AI lead researcher at **Ethikon Institute (ethikon.ai)**

Security Team

Emanuele Mezzi

Fabio Massacci

Katja Tuma



Mengyuan Zhang

Aurora Papotti

Can I use LLMs for CTI?

TLP:GREEN

What We Would like from LLMs

Given the CTI report, derive the threat scenario

- The responsible actor is **K3chang**
- They exploited **CVE-2012-4681**
- They relied on **spear-phishing**

Given the APT name, tell me its characteristics

K3chang

- Type of APT: **state-actor**
- Goals: **espionage**

Same Question - Two Different Answers

An LLM queried multiple times with the **same input** can return **different answers**.

Entity
APT: Ke3chang APT1
CVE: CVE-2012-4681 CVE-2010-2883

Entity
APT: Ke3chang APT1
CVE: CVE-2012-4681 CVE-2010-3333

Are you sure of your answers?

Are the **confidence estimates** returned from the LLM **reliable?**

Entity	Conf.
APT: Ke3chang APT1	0.90 0.70
CVE: CVE-2014-6321 CVE-2020-35931	0.20 0.70
Date: 01-2011 01-2012	0.65 0.40
Attack: spear phishing valid accounts	0.75 0.90

Are you sure of these estimates?

But Everybody says LLMs are Great

TLP:GREEN

Claimed Performance

Performance	
Paper	**Precision**
Wang et al. [1]	0.89
Wang et al. [2]	0.83
Hu et al. [3]	0.88
Li et al. [4]	0.82

[1] Xuren Wang et al. Dnrti: A large-scale dataset for named entity recognition in threat intelligence. In 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom).
[2] Xuren Wang et al. Aptner: A specific dataset for ner missions in cyber threat intelligence field. In 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD).
[3] Yuelin Hu et al. Llm-tikg: Threat intelligence knowledge graph construction utilizing large language model. Computers & Security.
[4] Jiehui Liu and Jieyu Zhan. Constructing knowledge graph from cyber threat intelligence using large language model. In 2023 IEEE International Conference on Big Data (BigData).

But They Use Unrealistic Data

Data Used for Evaluation

Input Type	Words
Sentence [1]	20
Sentence [2]	18
Paragraph [3]	106
Paragraph [4]	163

Note

The dataset in [4] is wrongly built, as they consider the title of the report as an entity!

[1] Xuren Wang et al. Dnrti: A large-scale dataset for named entity recognition in threat intelligence. In 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom).
[2] Xuren Wang et al. Aptner: A specific dataset for ner missions in cyber threat intelligence field. In 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD).
[3] Yuelin Hu et al. Llm-tikg: Threat intelligence knowledge graph construction utilizing large language model. Computers & Security.
[4] Jiehui Liu and Jieyu Zhan. Constructing knowledge graph from cyber threat intelligence using large language model. In 2023 IEEE International Conference on Big Data (BigData).

What Data Should Be Used

Data on which LLMs **should** be **evaluated**

Report	Words
Emergency Directive 21-01 (SolarWinds) [5]	1764
Our dataset [6]	3009

Key Takeaway

LLMs work great on toy CTI reports. What about **real** CTI reports?

[5] CISA. Emergecy directive 21-01: Mitigate solarwinds orion code compromise. Technical report, "Cybersecurity and Infrastructure Security Agency (CISA)", 2020. available at https://www.cisa.gov/news-events/directives/ed-21-01-mitigate-solarwinds-orion-code-compromise.

[6] Massacci, F. & di Tizio G. Are Software Updates Useless against Advanced Persistent Threats? Considering the conundrum of software updates.

Our Approach

Realistic evaluation data, taking into account the technology characteristics

TLP:GREEN

Report Summarization

Given the following CTI report, extract the name of the APT, the starting date of the campaign, the CVEs exploited, and the attack vectors employed.


```
campaign: {
    actor: ''
    date: ''
},
APT: {
    name: ''
},
attack_vector: {
    [name: '']
},
CVE: {
    ['code': '', ...]
}
```

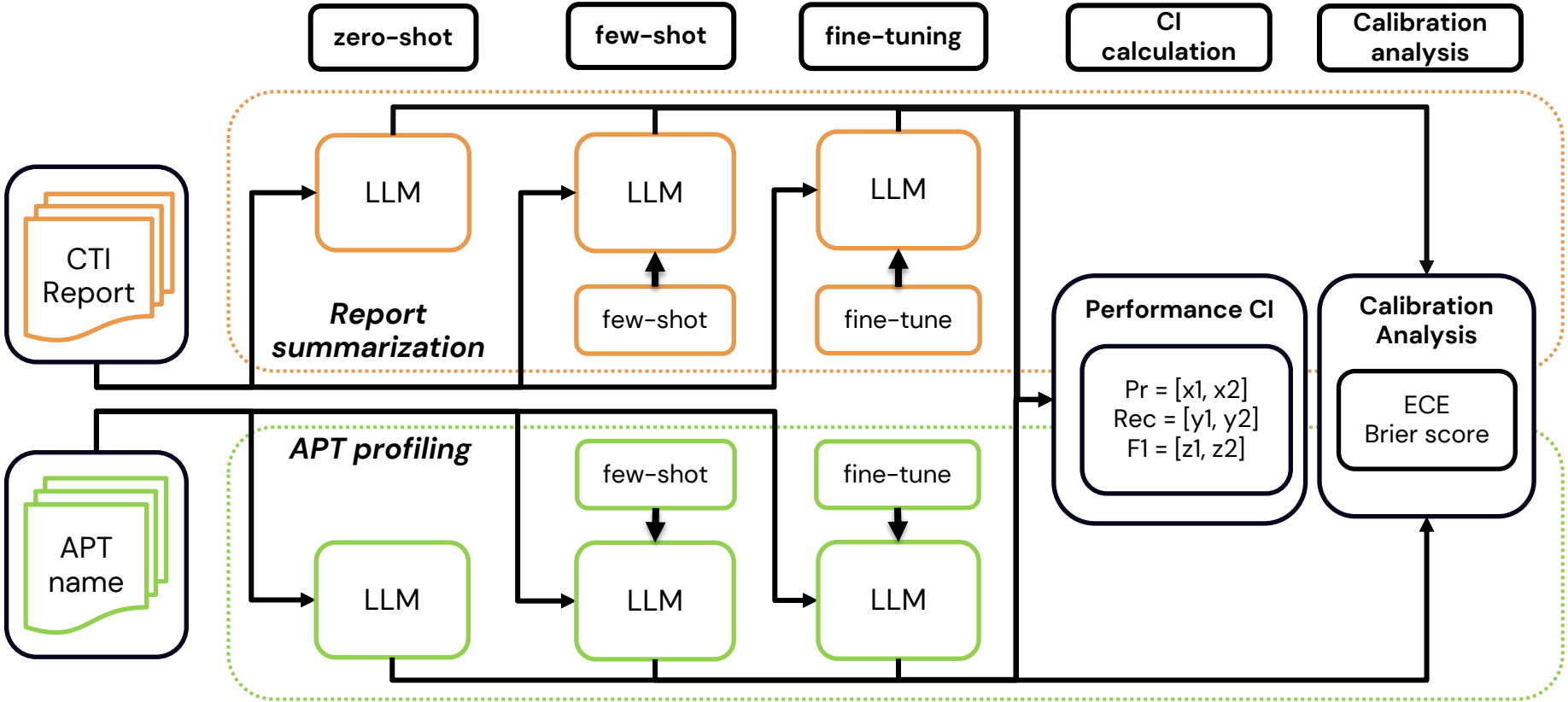
APT Profiling

Given the following APT, give me the nationality of the APT, the goals of the APT, the CVE exploited, and the attack vectors used.

APT name

```
APT: {
    actor: ''
},
country: {
    name: ''
},
attack_vector:
{
    [name: '']
},
CVE: {
    ['code': '', ...]
},
...
```

Our Evaluation



Prompt Engineering (1)

Zero-shot learning

Use the following step-by-step guide to extract information from… CTI reports.

Step 1 - Extract the starting date of the campaign, the Advanced Persistent Threat (APT), the CVE codes of the vulnerabilities exploited by the APT, and…

Notes: …

Step 2 - Return the information filling in this JSON format: …

Notes

- The name of the actor in the campaign and the name of the APT must be the same

…

- Only extract the CVEs that are directly attributed to the threat actor in the report.

…

- Only extract the attack vectors that are directly attributed to the threat actor in the report.

…

- Each node will have an ID composed of…

Prompt Engineering (2)

Few-shot learning

Use the following step-by-step guide to extract information from...

Step 1 - Extract the starting date of the campaign, the Advanced Persistent Threat (APT), the CVE codes of the vulnerabilities exploited by the APT, and...

Notes: ...
Examples: ...
Step 2 - Return the information filling in this JSON format: ...

Examples

- "to configure a client-side mail rule crafted to download and execute a malicious payload ... -> spear phishing attachment.
...
- "We also confirmed that the user installed this program via a download link delivered over email." -> spear phishing link.
...
- "has been linked to a watering hole attack" -> drive-by compromise.

Results - Performance

TLP:GREEN

The Data

DOI:10.1145/3571452 Fabio Massacci and Giorgio di Tizio

▸ Terry Benzel, Column Editor

Security

Are Software Updates Useless against Advanced Persistent Threats?

Considering the conundrum of software updates.

A DILEMMA DERIVED FROM Shakespeare's *Hamlet* is increasingly haunting company and security researchers: "to update or not to update, this is the question." From the perspective of recommended common practices by software vendors the answer is unambiguous: You should keep your software up to date.[8] But is common sense always good sense? We argue it is not.

Last year in a *Communications* article,[4] Poul-Henning Kamp argued these industry best practices do not seem to work and a more radical reform is needed. In the same year, Massacci et al. recalled the SolarWinds attack was funneled by an update[5] and a follow-up article[7] indicated the recent protestware attacks are also channeled through updates.

What is wrong here is that updates are hardly classified as either functionality or security updates or both. They are bundled together for the convenience of the software vendor.[6] For example, the WhatsApp update v2.19.51, while patching a critical security vulnerability exploited by the NSO Group, summarized the update with the following note: "You can now see stickers in full size when you long press a notification." One might concede, without believing it, that conflating together functionality and security updates is done to make it more difficult to identify the vulnerable code.

Yet, this lack of transparency is not going to help. Organizations can only blindly accept the "black-box" cumulative update that will force them to install all updates ignored so far, or equally blindly ignore the popup. Still, updates might be normally good and might turn to be unwise only in the high-profile cases that hit the media.

We investigated whether this is the case in the context of Advanced Persistent Threats (APTs).[2] APTs are sophisticated actors that deliberately and persistently target specific individuals and companies with a strategic motivation (from sabotage to financial gain). In this scenario, only an "all-hands on deck" defense seems appropriate and keeping your software up to date seems the bare—and likely not even sufficient—minimum.

Starting from "Operation Aurora" the security community increasingly released public information about APTs campaigns via blogs and technical reports. Unfortunately, the information is fragmented over different sources, each using different taxonomies to track adversaries. So, we collected data about more than 350 APT campaigns performed by 86 APTs in more than 10 years from more than 500 resources (and, by the way, the data is open source[a]). From this wealth of data, we can attempt to better understand these threats labeled "APT"s.

A as Advanced. In most cases, APTs do not even exploit a software vulnerability. Figure 1 shows the attack vectors employed in the campaigns. More than half of them do not employ any software vulnerability. APTs rely on spearphishing attacks via email and social networks to obtain the initial footprint in the network.

a See https://doi.org/10.5281/zenodo.6514817

Dataset on APTs

Data	Quantity
# of reports	350
# campaign	350
# APT	86
# CVE	123
# attack vector	170
# country	17

Report Size

Data	Mean	Max
# words	3k	21k
# tokens	4k	30k

[6] Massacci, F., & di Tizio, G. Are Software Updates Useless against Advanced Persistent Threats? Considering the conundrum of software updates.

The LLMs Employed

LLMs employed

gpt4o (context window: 128k)

mistral-large-2 (context window: 128k)

gemini-1.5-pro-latest (context window: 2M)

Why Closed Models

Easier to run compared to open source models that need to be locally installed.

They can be **less expensive**: closed source models are run directly on the provider cloud, without need to rent cloud or buy expensive GPUs.

Negative sides

You cannot give in input sensitive data

Usually it is not possible to extract the logits (it is possible with gpt4o)

Performance – Report Summarization

Analysis of the performance

We analysed campaign, APT, CVEs, and attack vector. For CVE and attack vector the result is particularly unsatisfactory.

		zero-shot		few-shot		fine-tuning	
	Model	**P**	**R**	**P**	**R**	**P**	**R**
campaign	...	0.77	0.77	0.73	0.73	0.61	0.61
APT	...	0.87	0.87	0.84	0.84	0.68	0.68
CVE	gpt4o	**0.67**	0.87	0.74	0.92	**0.71**	**0.69**
	gemini	0.69	0.90	0.75	0.89	0.81	0.63
	mistral	0.72	0.90	0.79	0.91	0.71	0.69
attack_vector	gpt4o	0.53	0.75	0.44	0.77	0.69	0.65
	gemini	0.68	0.74	0.71	0.78	0.89	0.84
	mistral	0.67	0.83	0.67	0.85	0.69	**0.65**

1 out of ~3 CVE retrieved is wrong!

Performance – APT profiling

Analysis of the performance

We analysed the type of the APT, the CVEs, and the attack vectors. The difference between two LLMs can be extremely large. We can see that in some cases precision and recall are equal to 0.

		zero-shot		few-shot		fine-tuning	
	Model	**P**	**R**	**P**	**R**	**P**	**R**
type of APT	gpt4o	**0.50**	0.50	0.44	0.44	0.44	0.44
	gemini	**0.02**	0.02	0.54	0.54	0.40	0.40
	mistral	0.02	0.02	0.36	0.36	0.44	0.44
CVE	gpt4o	0.10	0.06	0.08	0.07	0.00	**0.00**
	gemini	0.13	0.13	0.23	0.19	0.17	0.36
	mistral	0.21	0.17	0.24	0.24	0.00	**0.00**
attack_vector	gpt4o	0.37	0.52	0.37	0.51	1.00	0.09
	gemini	0.24	0.54	0.27	0.56	0.52	0.84
	mistral	0.22	0.58	0.20	0.75	1.00	0.09

LLMs do not know anything about APTs!

Key Performance Takeaways

Takeaways

- LLMs are inadequate in terms of precision and recall
- Fine-tuning can worsen the performance
- LLMs performance is generally worse for APT profiling.
- The difference between two LLMs can be extremely large.

Why can fine-tuning worsen the model performance? Maybe the size of the dataset?

LLMs characterised by a large volume of parameters need to be fine-tuned on a large number of elements in to improve they performance. Unluckily we did not find a larger dataset!

Why such a massive difference between two LLMs in some cases?

Even though we cannot be sure, it is possible that some of the LLMs were trained on data concerning APTs, while others where not.

LLM as a CTI Assistant
Multiple Queries Same Answer?

Please LLM, be **coherent** with what you just said!

TLP:GREEN

What We Do not Want (Inconsistency & Consequences)

Ground truth

CVE-2010-2883
CVE-2012-4681
CVE-2010-3333

First Analysis

CVE-2010-2883
CVE-2012-4681
CVE-2010-3333

Second Analysis

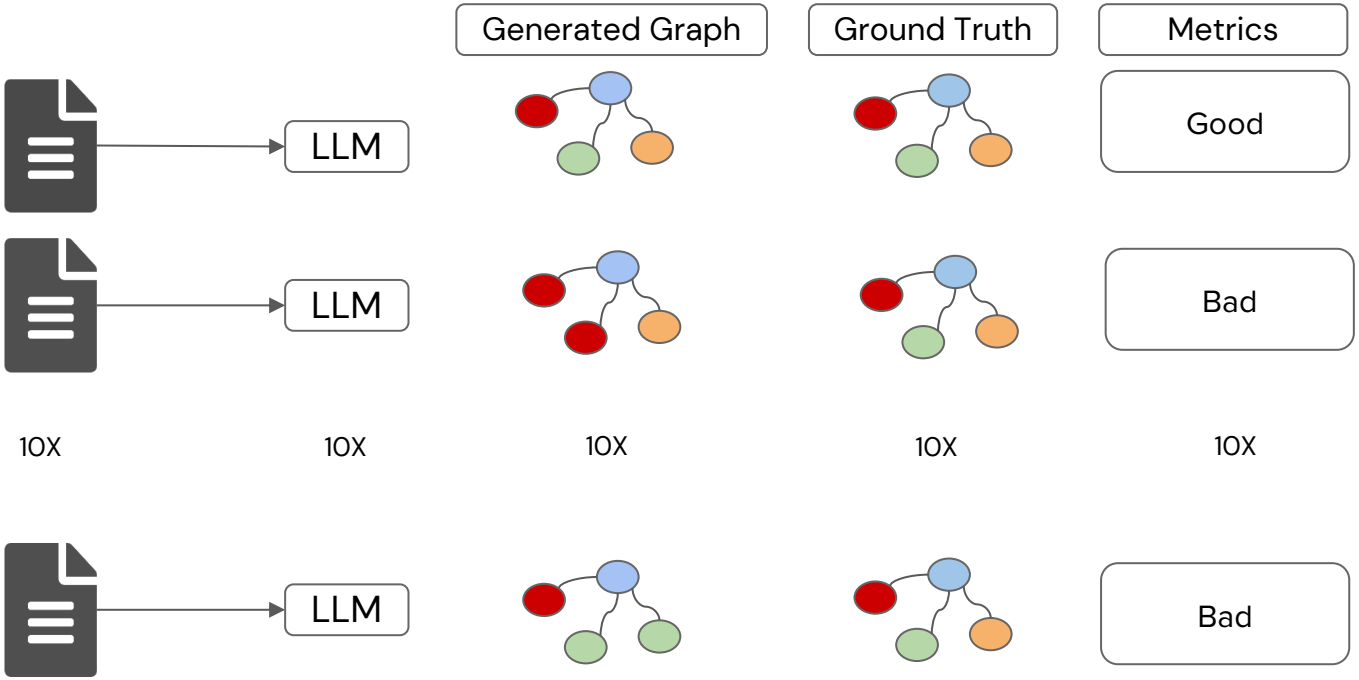
CVE-2010-2883
CVE-2012-4681
CVE-2010-3333

Consequences

Inconsistency means that the LLM extracted **different CVEs** from the **same reports** over multiple iterations. As a consequence this brings **uncertainty** in the CVEs to consider and to patch.

Ask You Twice, What do I get?

Repeating multiple times the same operation, with the same input, and measuring the performance interval.



Precision and recall **should not change** if you prompt the LLM with the **same input**

Consistency quantification: APT profiling

The LLM is not deterministic! The difference in some cases can be at least of the **6%** between the minimum and maximum performance registered.

		Few-shot		Fine-tuning	
	Models	**P**	**R**	**P**	**R**
country	gpt4o	>= 2%	>= 2%	>= 1%	>= 1%
	gemini	>= 4%			
	mistral		>= 3%	>= 1%	>= 1%
CVE	gpt4o	>= 1%	>= 1%	>= 3%	
	gemini	>= 5%	>= 6%		>= 5%
	mistral	>= 1%	>= 1%		
attack vector	gpt4o		>= 1%		>= 1%
	gemini	>= 3%	>= 3%	>= 1%	>= 2%
	mistral		>= 1%		

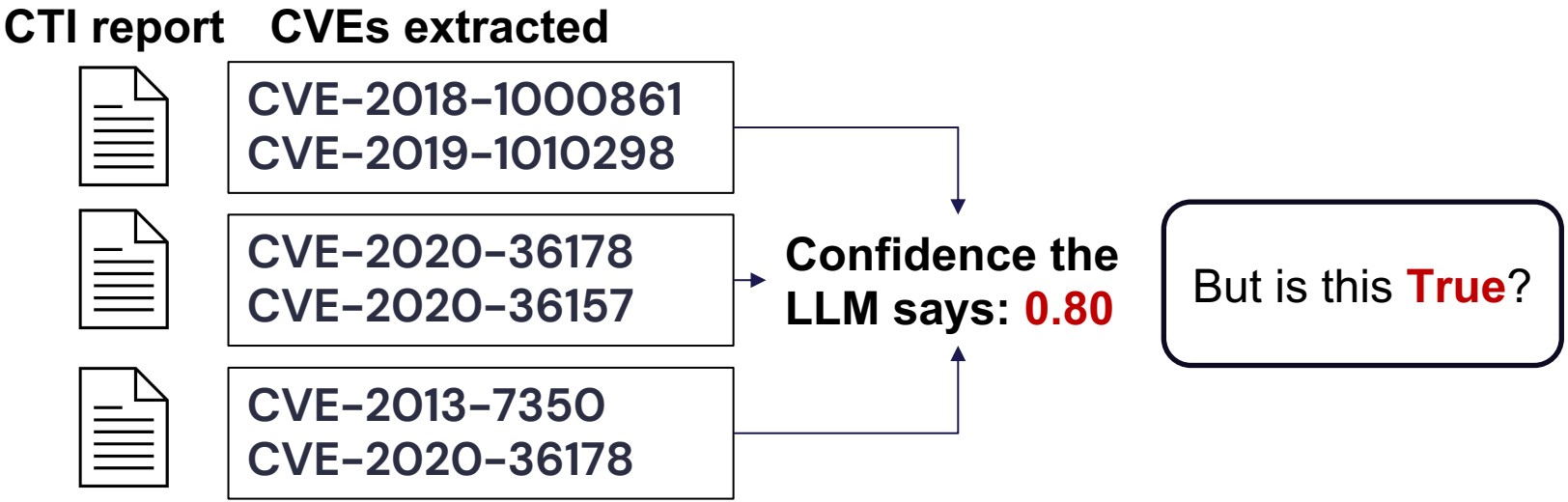
LLM, are you sure of that?

TLP:GREEN

Alignment between LLM Confidence and Accuracy

Calibration: it ensures that when a model says there is an 80% chance of something happening, it is actually correct about 8 out of 10 times.

For example … The LLM overall confidence on the extraction of CVEs



LLMs in CTI are not Calibrated

We measure the **alignment** between accuracy and LLMs confidence by calculating the Brier Score (BS).

- BS communicates the times in which the model is wrong with its confidence. At least in 15% of the cases the model is wrong with its confidence.
- The **lower** it is, the **better** it is. 0 indicates perfect calibration.

Report Summarization

	zero-shot	few-shot	fine-tuning
	BS	**BS**	**BS**
campaign	0.26	0.28	**0.48**
APT	0.15	0.15	0.23
CVE	0.32	0.37	0.21
attack vector	0.46	0.49	**0.58**

APT Profiling

	zero-shot	Few-shot	Fine-tuning
	BS	**BS**	**BS**
country	0.22	0.27	0.29
CVE	0.29	0.22	**0.98**
attack vector	0.43	0.42	**1.00**

Key Takeaway

Few-shot learning and fine-tuning do not help the calibration

What do we take away?

I was hoping you could be the perfect model...

TLP:GREEN

Final Remarks and Possible Improvements

LLMs are not ready to be deployed in real-world scenarios, as:

- Their performance is inadequate on real-size reports.
- They lack prediction consistency
- They are not calibrated and thus cannot be deployed in absence of an evaluation dataset.

Possible improvements (include the last slide in this)

- Technology: improve the LLM implementation to improve the both performance and the consistency.
- Data: standardization can help. However, we know that this is extremely hard to reach, as the data sources are heterogeneous.

Tell us What You Think!

or

e.mezzi@vu.nl