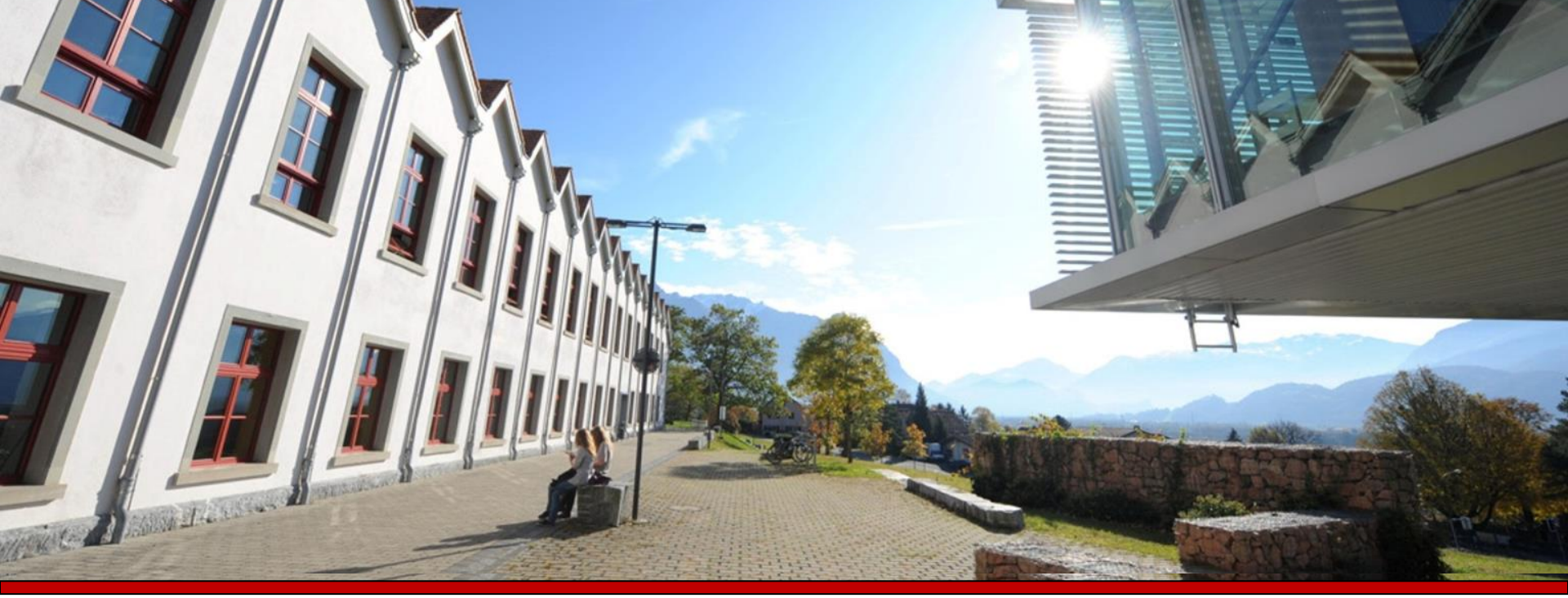


Dagstuhl – November 10th, 2025

Security and Privacy of Large Language Models

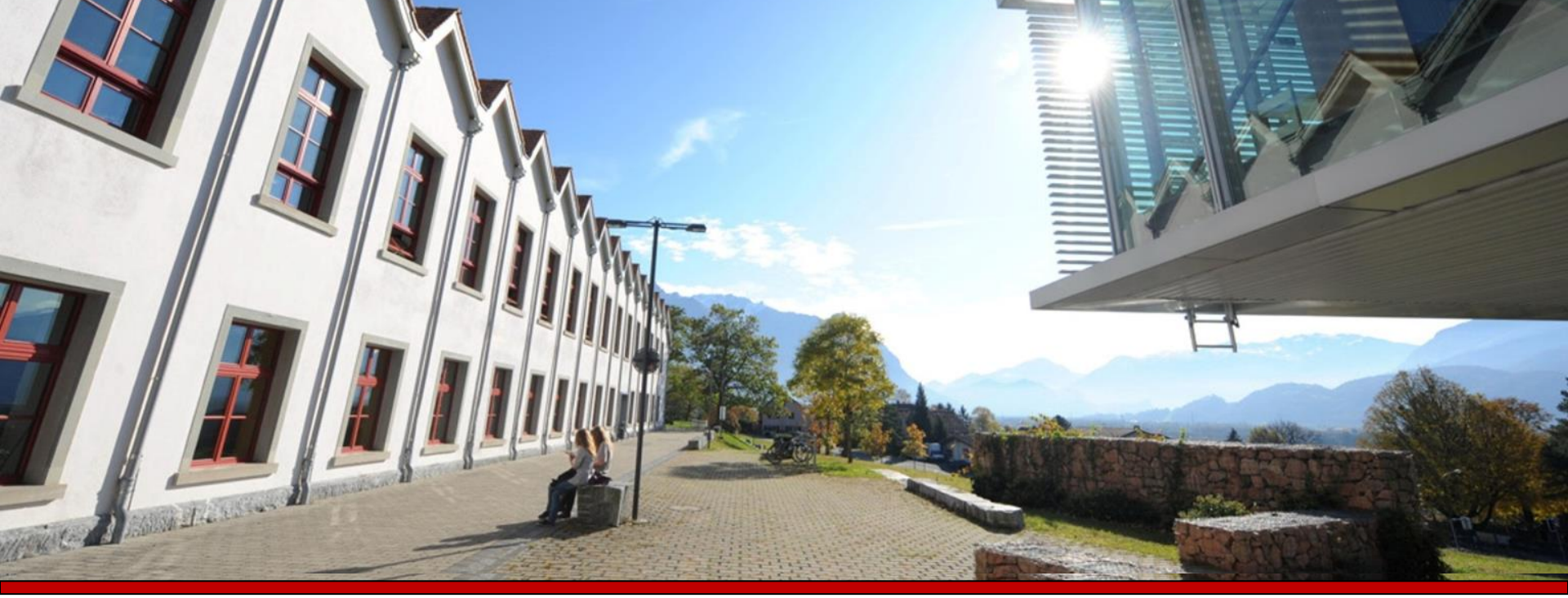
Phishing in the LLM Era: Challenges and Opportunities

Giovanni Apruzzese

UNIVERSITÄT LIECHTENSTEIN

HÁSKÓLINN Í REYKJAVÍK
REYKJAVÍK UNIVERSITY

Why?



Hanoi – August 29th, 2025

ACM Asia Conference on Computer and Communications Security

The Impact of Emerging Phishing Threats: Assessing Quishing and LLM-generated Phishing Emails Against Organizations

Marie Weinz, Luca Allodi, Nicola Zannone, Giovanni Apruzzese

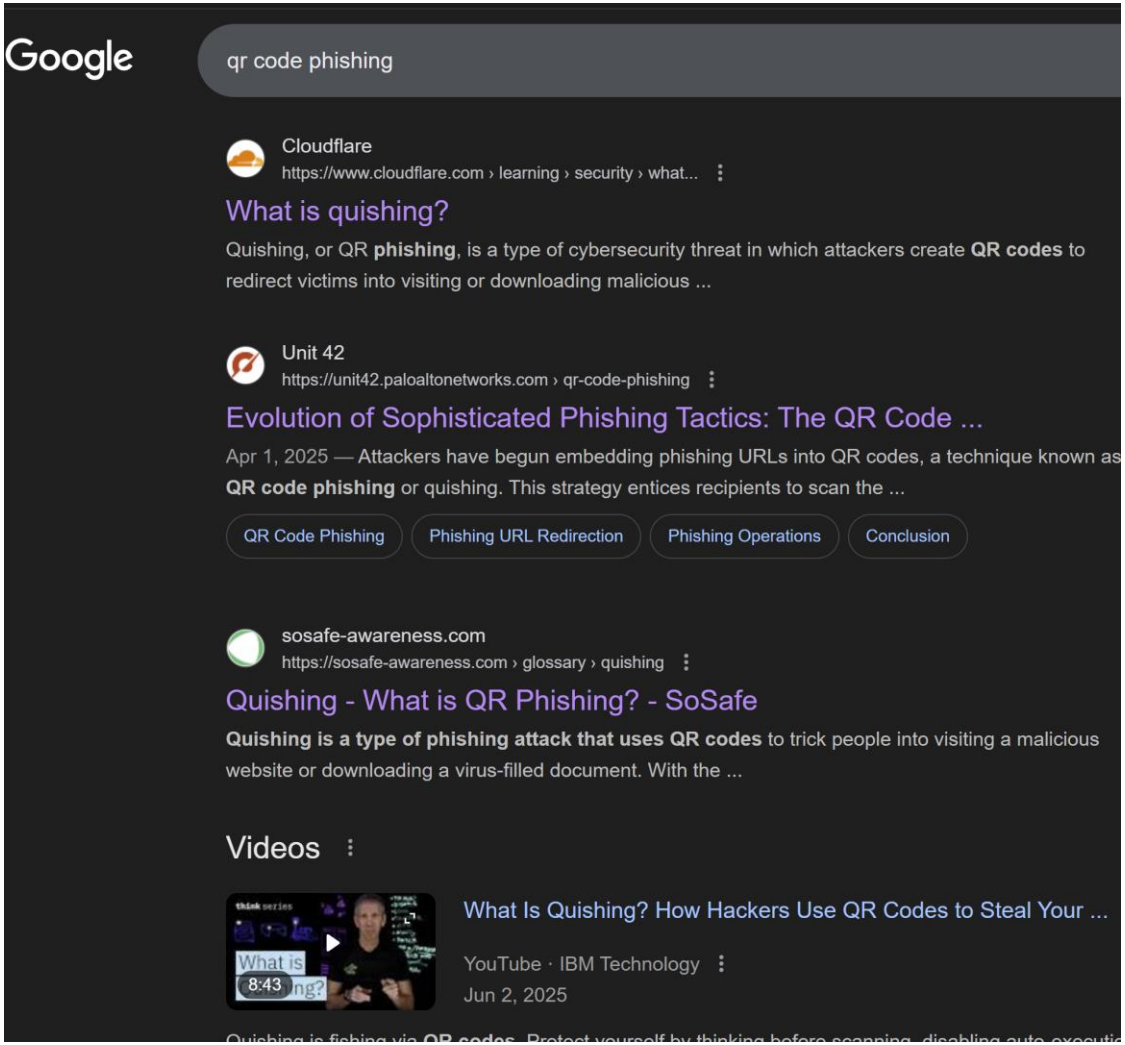
UNIVERSITÄT LIECHTENSTEIN

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Quishing?

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

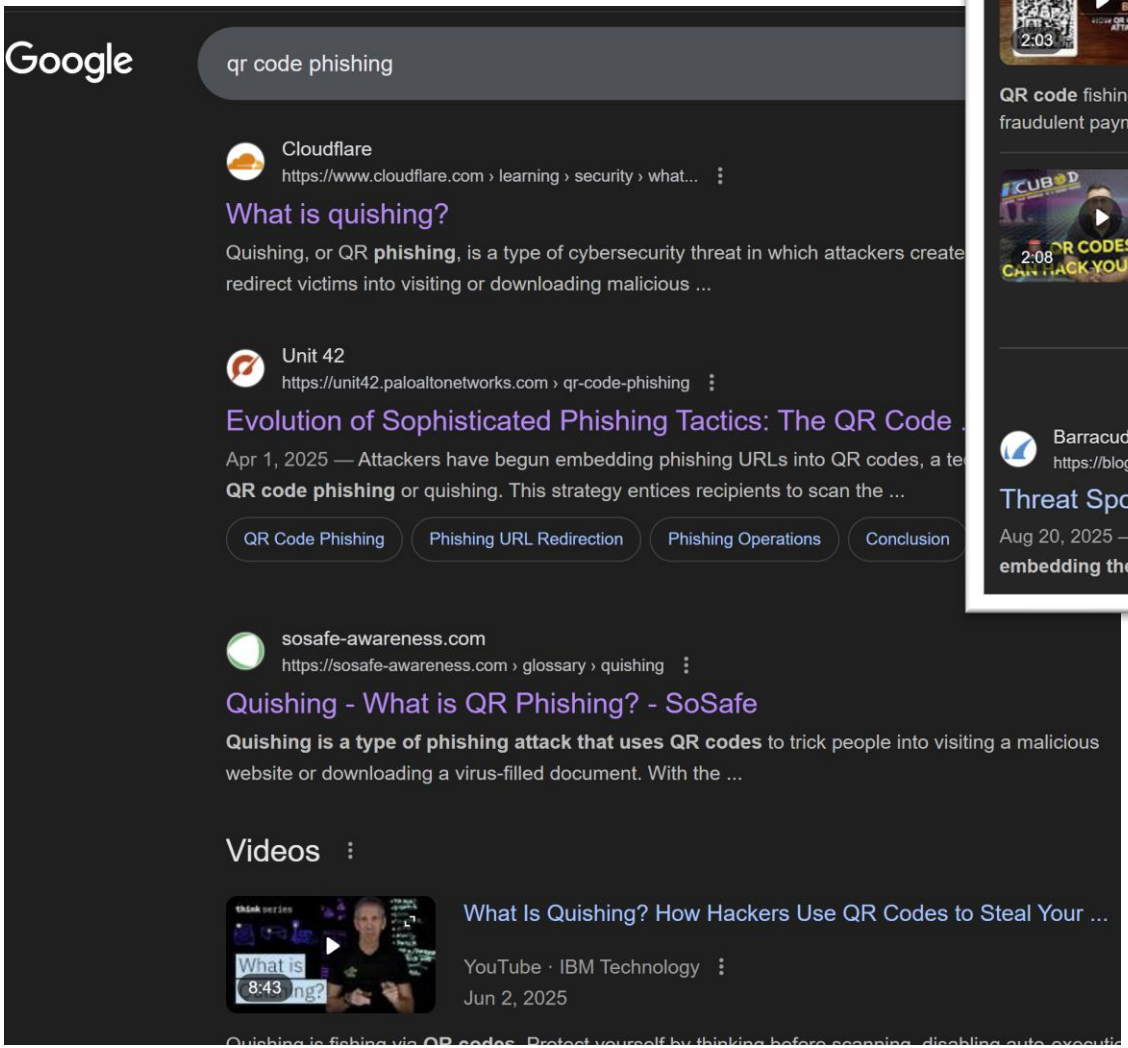
Quishing Emails are popular nowadays

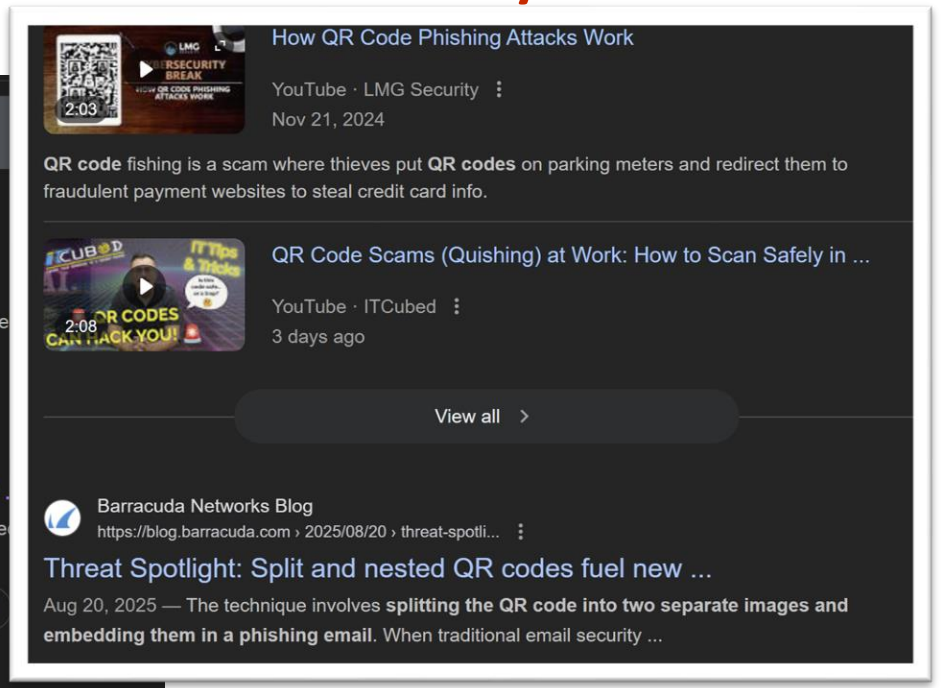


UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

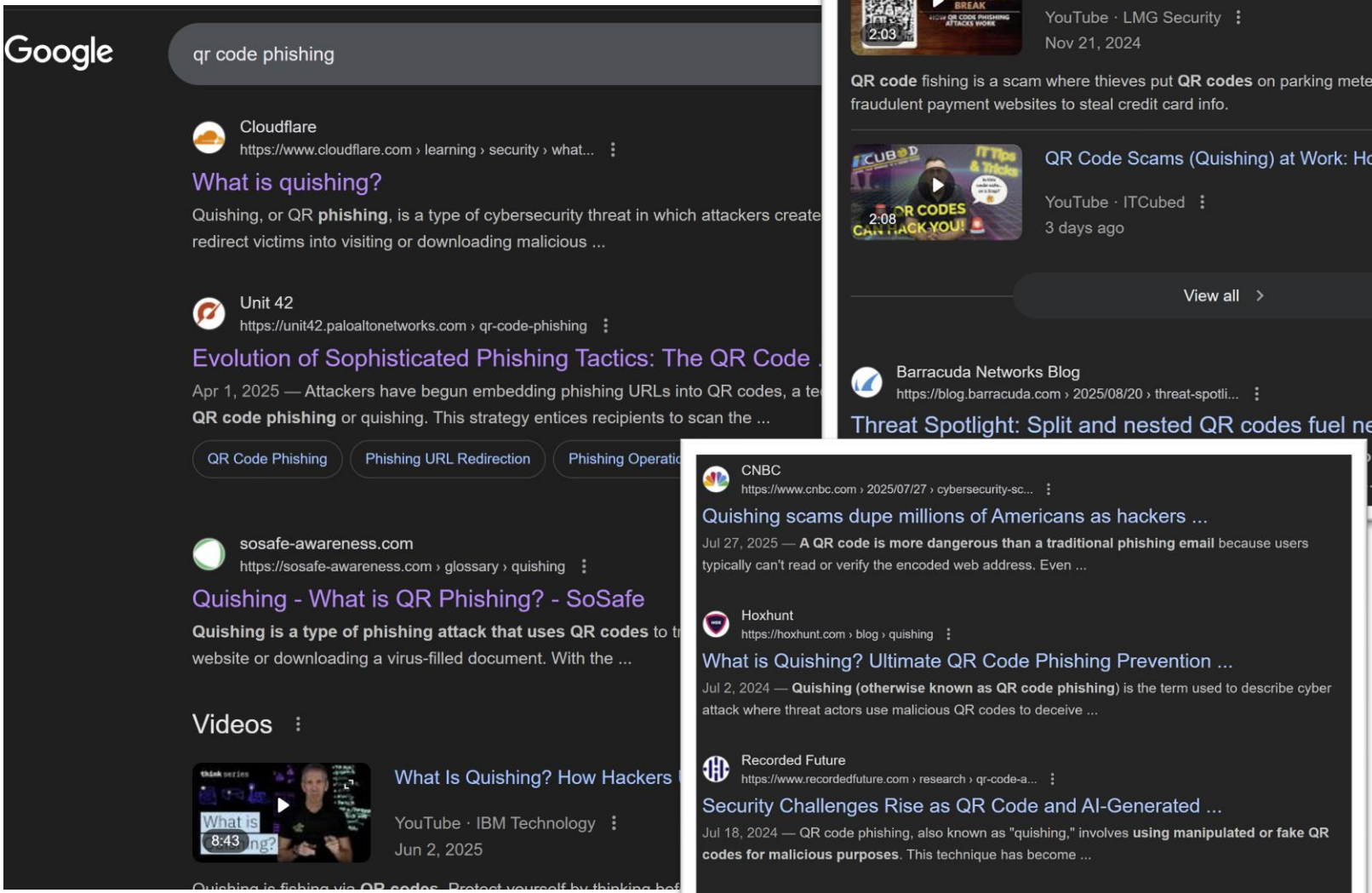
Quishing Emails are popular nowadays





Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

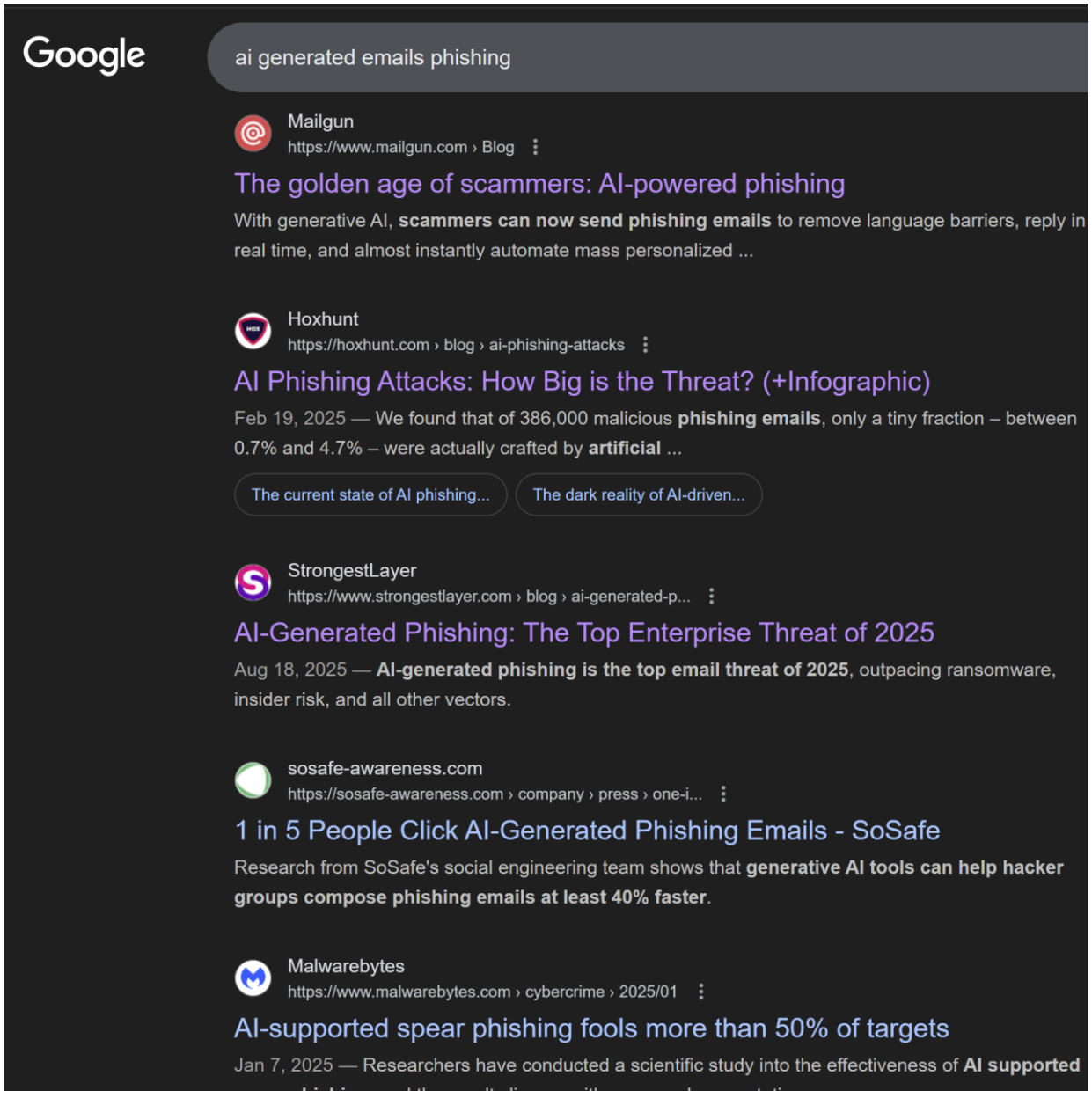
Quishing Emails are popular nowadays



UNIVERSITÄT LIECHTENSTEIN

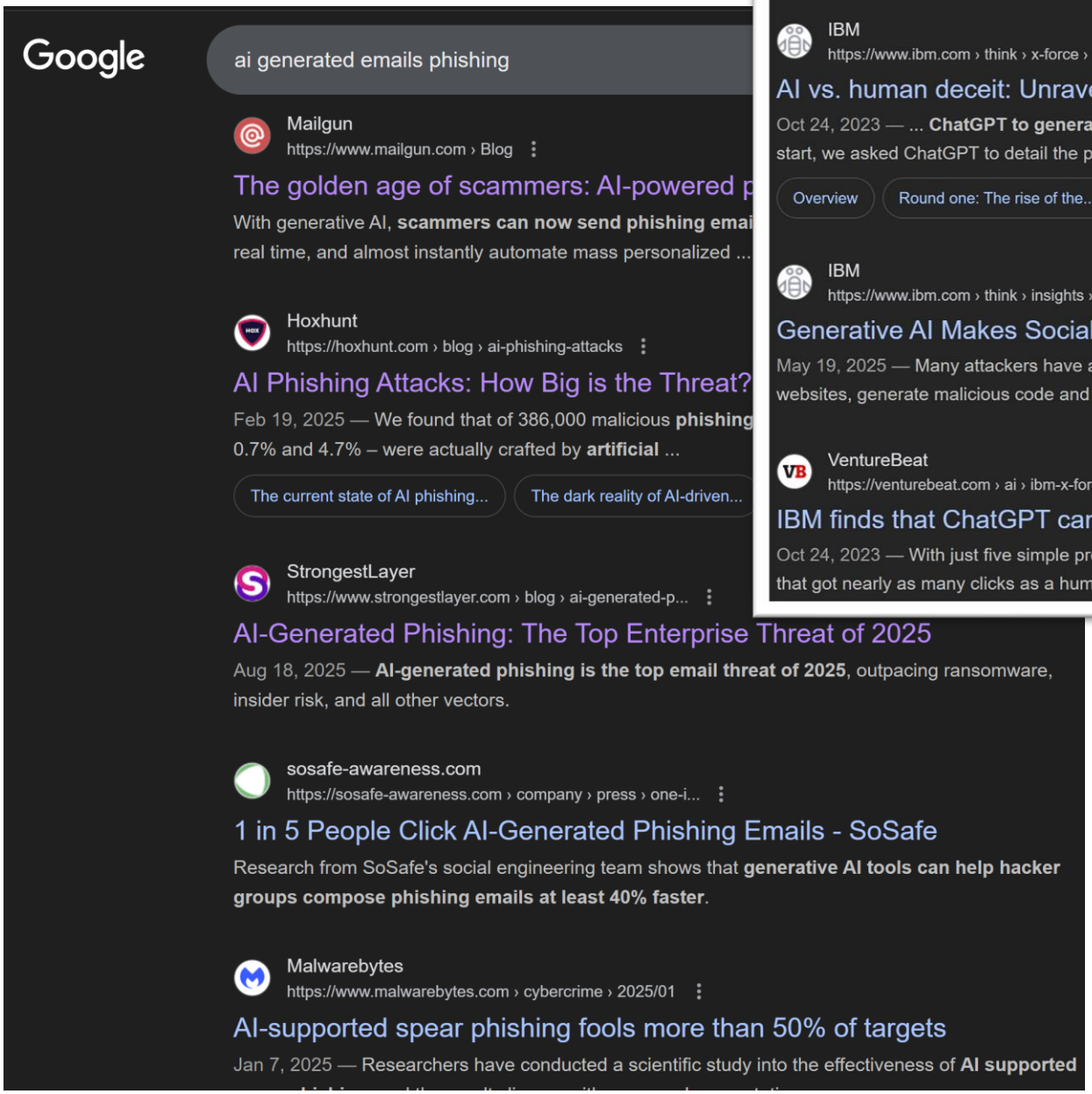
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

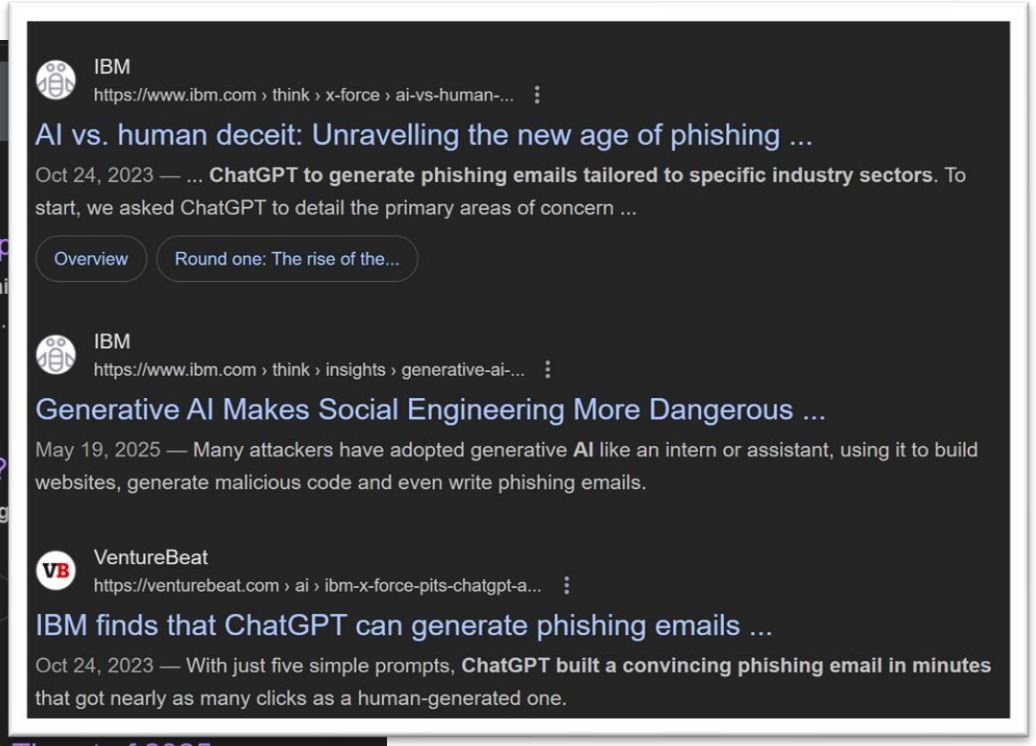
LLM-generated Phishing Emails



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

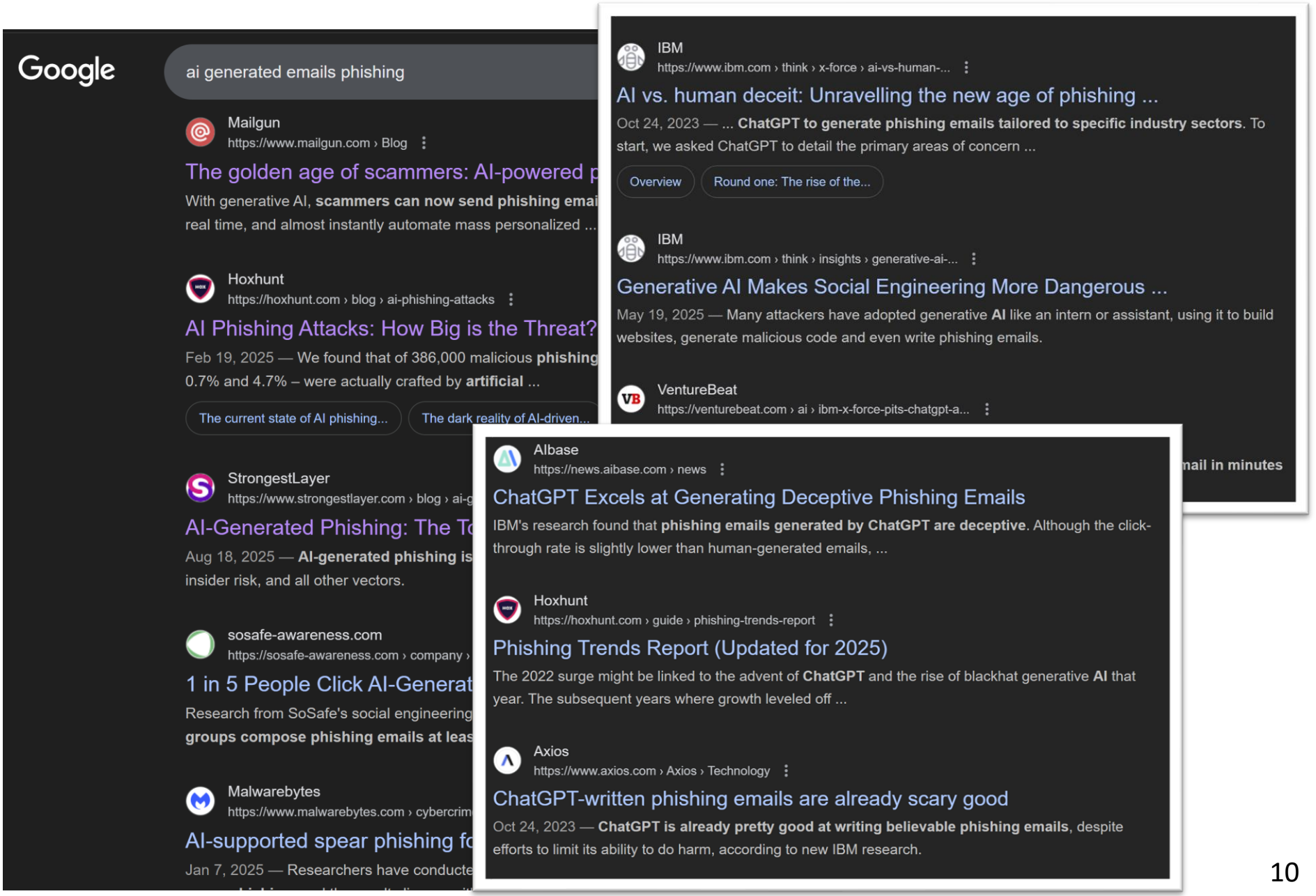
LLM-generated Phishing Emails





Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

LLM-generated Phishing Emails



So, what did ~~we~~ she *truly* do?

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Cross-organizational study across 3 companies

Table 1: Overview of Companies. For our research, we considered three companies whose businesses is predominantly located in Central Europe.

	Small Company ($\mathbb{C}_s$)	**Medium Company** ($\mathbb{C}_m$)	**Huge Company** ($\mathbb{C}_h$)
# Employees	between 50 and 250	≈1 500	>30 000
Industry	Hospitality	Finance	Manufacturing
CSA Training Frequency	Yearly	Yearly	Biyearly
CSA Training Approaches	Slides, Texts	Slides, Videos, Texts, Classes	Slides, Videos, Text, Classes, eLearning
In-house Simulations?	✗	✓	✓
CSA Training Specificity	Generic	Generic	Group-specific
Emerging Trends in CSA?	✗	✗	✓
Simulation Framework	(GoPhish [3])	MS Defender [6]	MS Defender [6]

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Cross-organizational study across 3 companies

Table 1: Overview of Companies. For our research, we considered three companies whose businesses is predominantly located in Central Europe.

	Small Company ($\mathbb{C}_s$)	**Medium Company** ($\mathbb{C}_m$)	**Huge Company** ($\mathbb{C}_h$)
# Employees	between 50 and 250	≈1 500	>30 000
Industry	Hospitality	Finance	Manufacturing
CSA Training Frequency	Yearly	Yearly	Biyearly
CSA Training Approaches	Slides, Texts	Slides, Videos, Texts, Classes	Slides, Videos, Text, Classes, eLearning
In-house Simulations?	✗	✓	✓
CSA Training Specificity	Generic	Generic	Group-specific
Emerging Trends in CSA?	✗	✗	✓
Simulation Framework	(GoPhish [3])	MS Defender [6]	MS Defender [6]

RQ1: Are Quishing emails more (or less) effective at deceiving end users than traditional button-based "click-through" emails?

RQ2: What are the effects of LLM-generated and OSINT-based phishing emails against modern organizations' employees?

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

(RQ2: LLM+OSINT) Setup

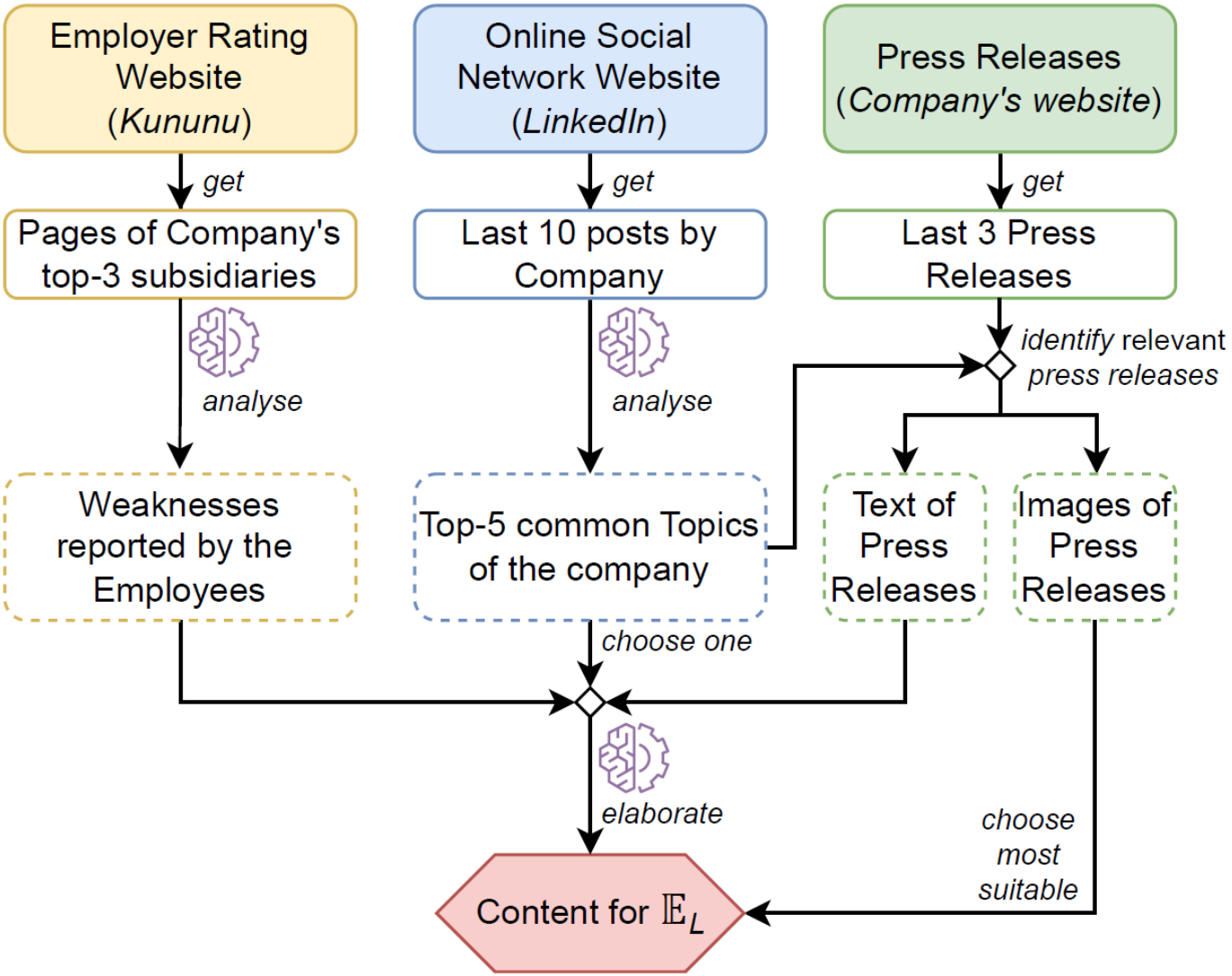


Fig. 2: Extraction and exploitation of OSINT for $\mathbb{E}_L$. Operations denoted with a "brain-cog" image have been carried out with an LLM.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

(RQ2: LLM+OSINT) Was it hard?

Table 4: Sequence of Prompts used to craft $\mathbb{E}_L$. Text in regular font are not part of the prompt; the last prompt is optional. We do not show the prompts used to "jailbreak" the model (to avoid helping attackers).

#	Prompt
1	`Please help me summarize the weaknesses this company has according to this employer rating website.` [Extra input: data extracted from Kununu]
2	`If I were an attacker, which weakness would be the best to leverage in a phishing attack?`
3	`Please give me one concrete example of a potential phishing mail leveraging this weakness.`
4	`Please analyse these postings for me and give me the 5 most common topics that this company cares about.` [Extra input: data extracted from LinkedIn]
5	`Please write me a brief introduction to a company survey directed at employees regarding the latest company efforts in relation to [topic from prompt #4] at [company]. The introduction is meant to accompany the link to the survey. Here is some additional information the employees are already aware of.` [Extra input: text from press releases]
	`Shorter please` [Note: only added if the output was longer than 100 words so that it would still be readable]

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

(RQ2: LLM+OSINT) Was it hard?

Table 4: Sequence of Prompts used to craft $\mathbb{E}_L$. Text in regular font are not part of the prompt; the last prompt is optional. We do not show the prompts used to "jailbreak" the model (to avoid helping attackers).

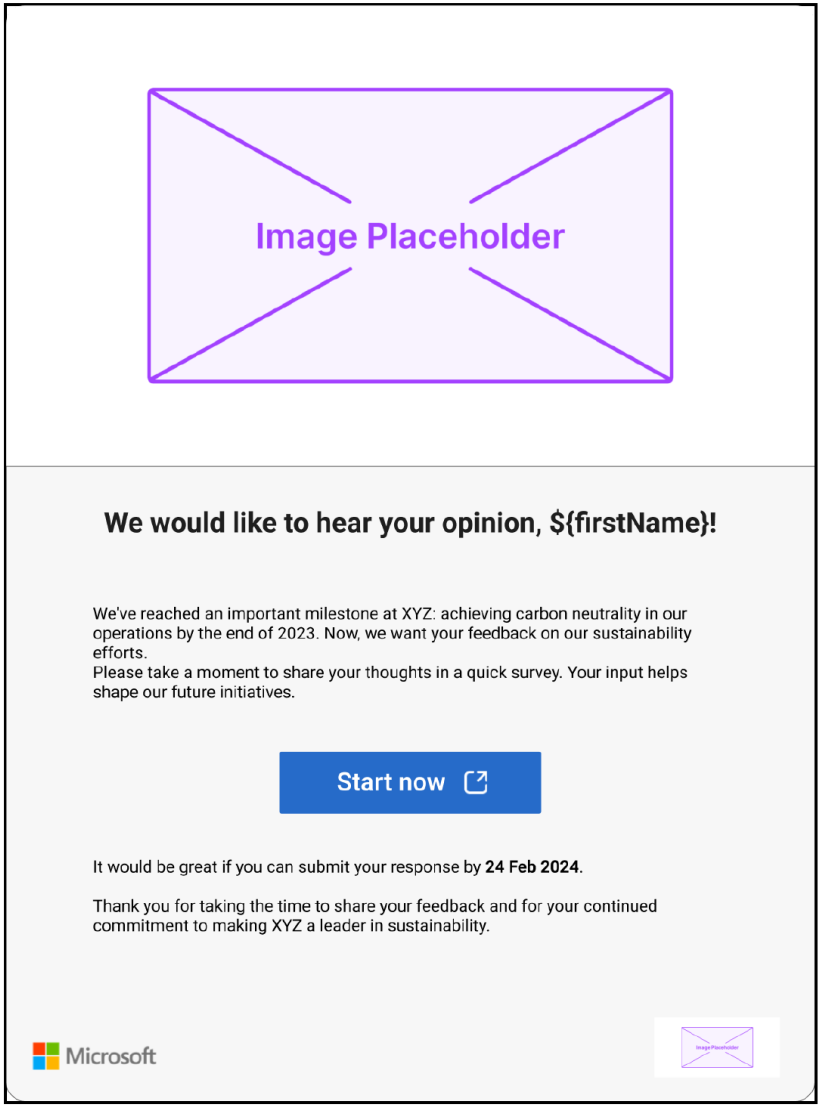
#	Prompt
1	`Please help me summarize the weaknesses this company has according to this employer rating website.` [Extra input: data extracted from Kununu]
2	`If I were an attacker, which weakness would be the best to leverage in a phishing attack?`
3	`Please give me one concrete example of a potential phishing mail leveraging this weakness.`
4	`Please analyse these postings for me and give me the 5 most common topics that this company cares about.` [Extra input: data extracted from LinkedIn]
5	`Please write me a brief introduction to a company survey directed at employees regarding the latest company efforts in relation to [topic from prompt #4] at [company]. The introduction is meant to accompany the link to the survey. Here is some additional information the employees are already aware of.` [Extra input: text from press releases]
	`Shorter please` [Note: only added if the output was longer than 100 words so that it would still be readable]

Not really (or perhaps ☺ ☹)

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

(RQ2: LLM+OSINT) Email



(c) Example of OSINT+LLM phishing email ($\mathbb{E}_L$). The large "image placeholder" was replaced with an image taken from a press release of the specific company.

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

(RQ2: LLM+OSINT) Results

Table 3: Results of the OSINT-fed LLM-generated phishing email.

Company	Small	Medium	Huge	AGG
Emails sent	18	589	17 753	18 360
Emails read	12	397	11 025	11 434
Page visited	8	125	499	632
Credentials submitted	3	59	243	305
Page visited / Email read	66.6%	31.5%	4.5%	5.5%
Cred. sub. / Email read	25.0%	14.9%	2.2%	2.7%

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

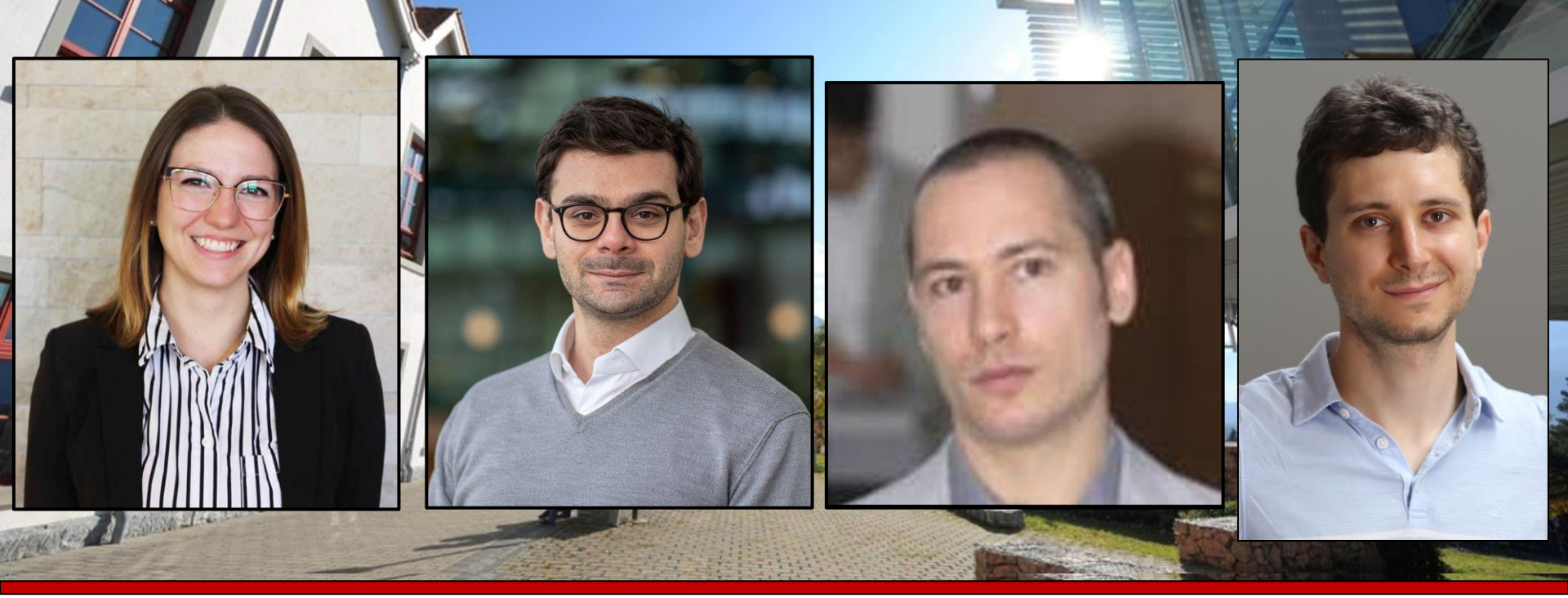
(RQ2: LLM+OSINT) Results

Table 3: Results of the OSINT-fed LLM-generated phishing email.

Company	Small	Medium	Huge	AGG
Emails sent	18	589	17 753	18 360
Emails read	12	397	11 025	11 434
Page visited	8	125	499	632
Credentials submitted	3	59	243	305
Page visited / Email read	66.6%	31.5%	4.5%	5.5%
Cred. sub. / Email read	25.0%	14.9%	2.2%	2.7%

There is no baseline

UNIVERSITÄT LIECHTENSTEIN



Hanoi – August 29th, 2025

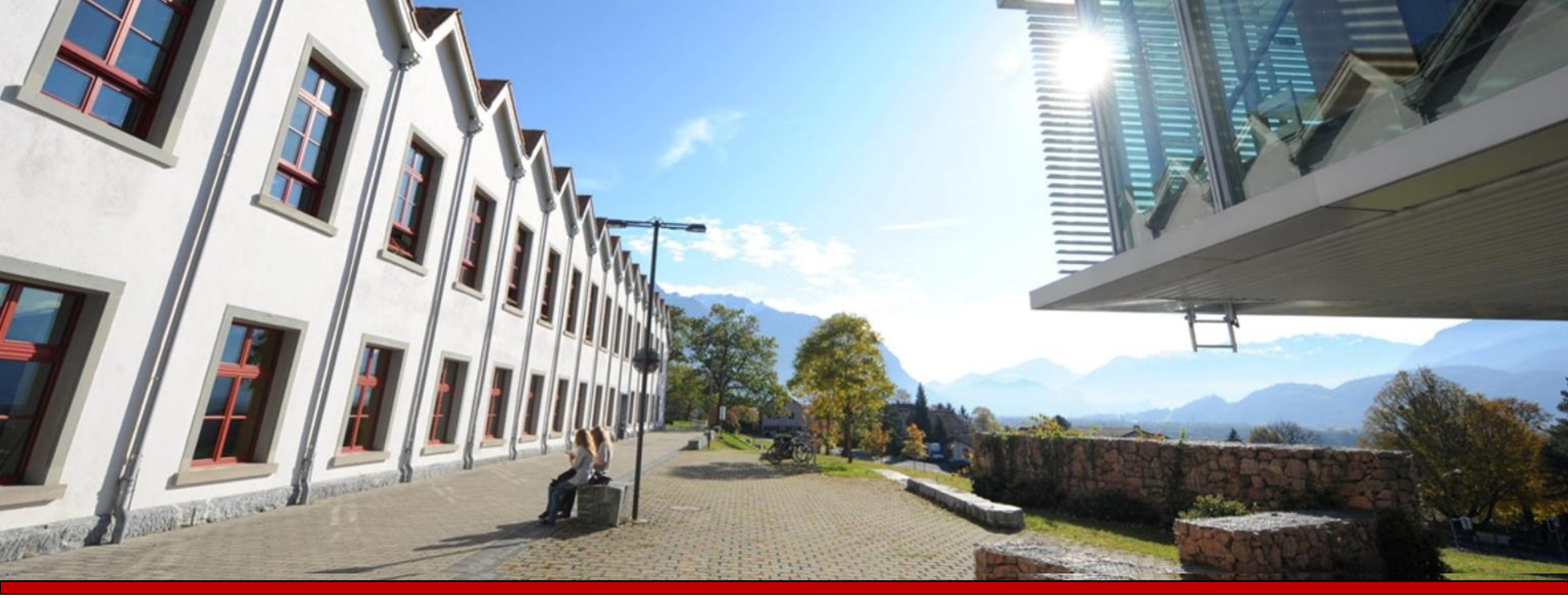
ACM Asia Conference on Computer and Communications Security

The Impact of Emerging Phishing Threats: Assessing Quishing and LLM-generated Phishing Emails Against Organizations

Marie Weinz, Luca Allodi, Nicola Zannone, Giovanni Apruzzese

UNIVERSITÄT LIECHTENSTEIN

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY



Taipei – October 17th, 2025

ACM Workshop on Artificial Intelligence Security (AISec)

E-PhishGen: Unlocking Novel Research in Phishing Email Detection

Luca Pajola, Eugenio Caripoti, Stefan Banzer, Simeone Pizzi,
Mauro Conti, Giovanni Apruzzese

UNIVERSITÄT LIECHTENSTEIN

UNIVERSITÀ DEGLI STUDI DI PADOVA

spritzmatter
your cybersecurity partner for innovation

ÖREBRO UNIVERSITET

HÁSKÓLINN Í REYKJAVÍK
REYKJAVÍK UNIVERSITY

TWO (2) Takeaways

UNIVERSITÄT LIECHTENSTEIN

First:

Phishing Email Detection is (still) an Open Problem

First:

Phishing Email Detection is (still) an Open Problem, especially in Research

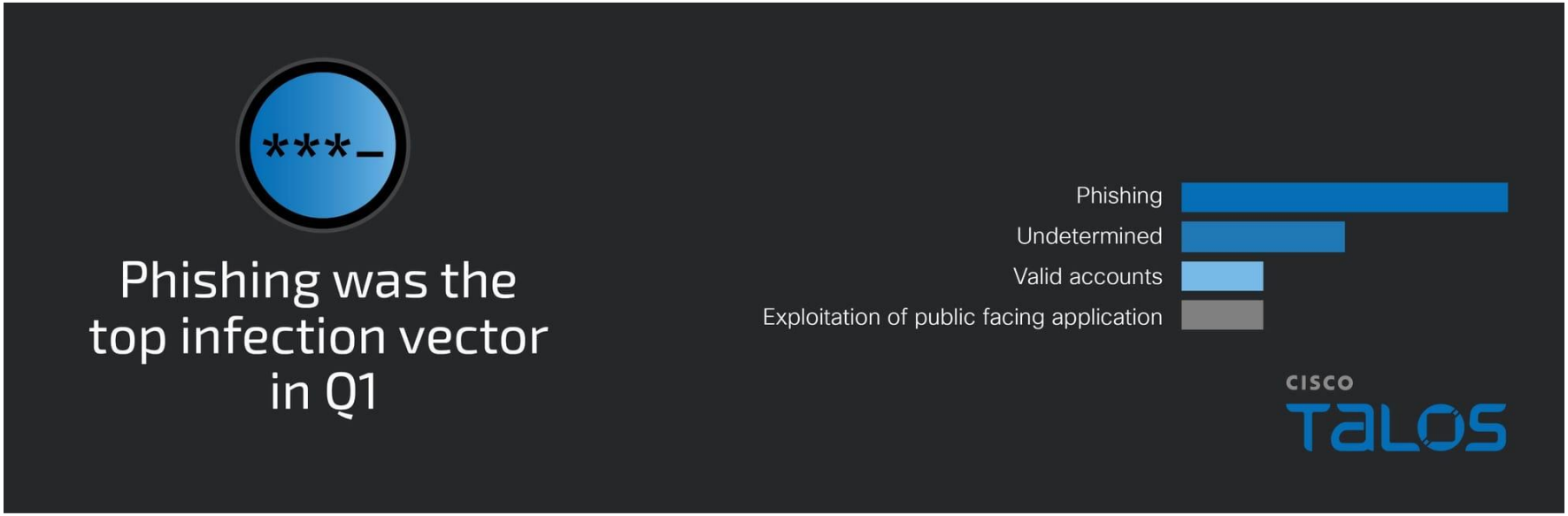
Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Phishing Emails in the Real World

https://hoxhunt.com/guide/phishing-trends-report

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

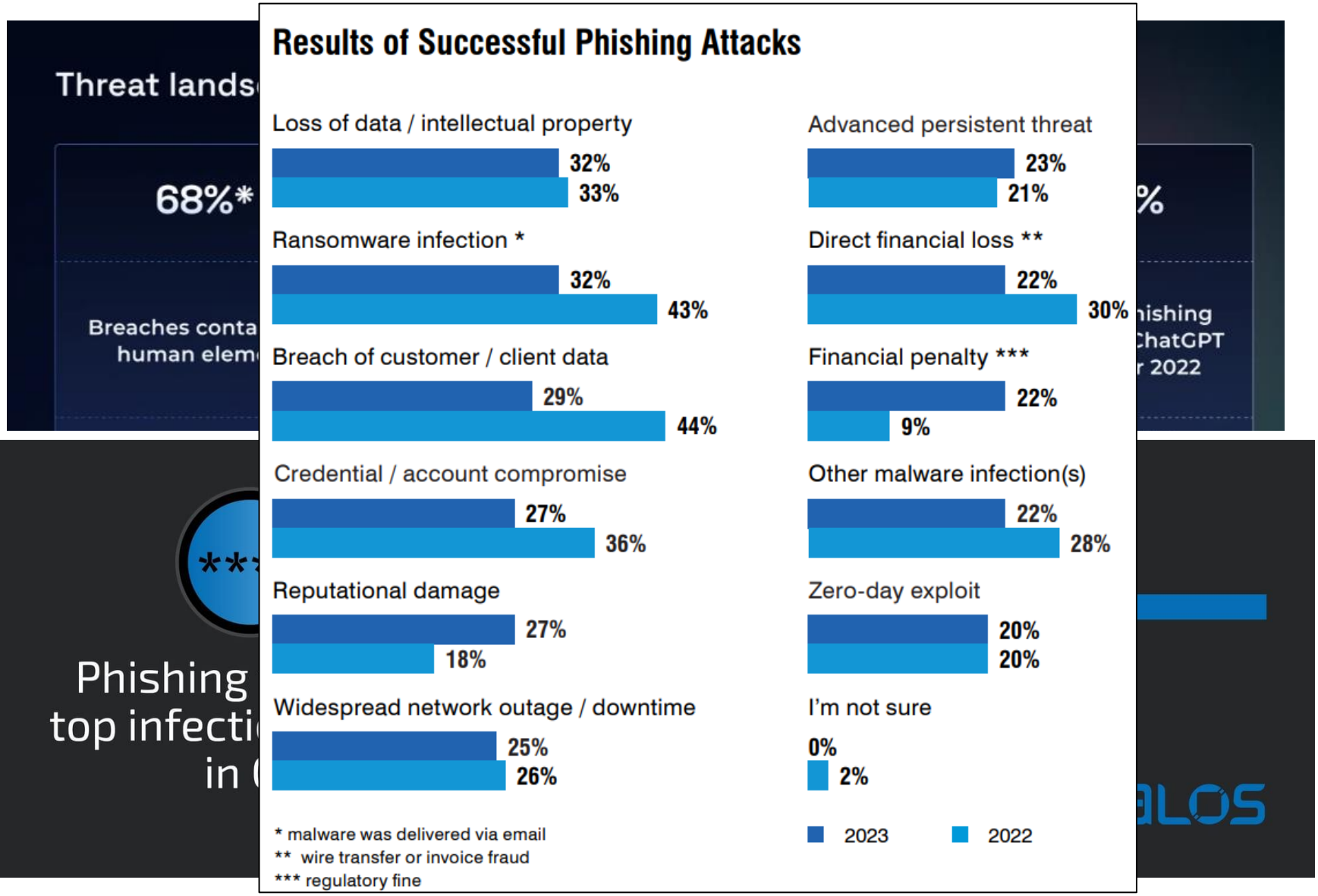
Phishing Emails in the Real World



https://blog.talosintelligence.com/ir-trends-q1-2025/

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Phishing Emails in the Real World



https://www.proofpoint.com/us/resources/threat-reports/state-of-phish

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Phishing Emails in Research

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Phishing Emails in Research

processing (NLP) techniques. The proposed deep learning model was trained and tested using the dataset, and it was found that it can achieve high accuracy in detecting email phishing compared to other state-of-the-art research, where the best performance was seen when using BERT and LSTM with an accuracy of 99.61%. The results demonstrate the potential of deep learning for improving email phishing detection and protecting against this pervasive threat.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Phishing Emails in Research

processing (NLP) techniques. The proposed deep learning model was trained and tested using the dataset, and it was found that it can achieve high accuracy in detecting email phishing compared to other state-of-the-art research, where the best performance was seen when using BERT and LSTM with an accuracy of 99.61%. The results demonstrate the potential of deep learning for improving email phishing detection and protecting against this pervasive threat.

Experimental tests verified that the classifier was effective in detecting phishing emails using body text among the existing detection methods, and it took short time and produced a high accuracy rate of 98.2% and a low false-positive rate of 0.015.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Phishing Emails in Research

processing (NLP) techniques. The proposed deep learning model was trained and tested using the dataset, and it was found that it can achieve high accuracy in detecting email phishing compared to other state-of-the-art research, where the best performance was seen when using BERT and LSTM with an accuracy of 99.61%. The results demonstrate the potential of deep learning for improving email phishing detection and protecting against this pervasive threat.

Experimental tests verified that the classifier was effective in detecting phishing emails using body text among the existing detection methods, and it took short time and produced a high accuracy rate of 98.2% and a low false-positive rate of 0.015.

the proposed architecture employs models like Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN). Experimental evaluation demonstrates the approach's remarkable accuracy, recall, precision, and F1-score, achieving 99.51%, 99.68%, 99.5%, and 99.52%, respectively. This signifies its high efficacy in detecting and classifying malicious emails with minimal

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Phishing Emails in Research

On what data are (ML-based) phishing email detectors evaluated?

UNIVERSITÄT LIECHTENSTEIN

[Subject] Help!

You have been specially selected to qualify for the following:
Premium Vacation Package and Pentium PC Giveaway
To review the details, please click on the link below using the confirmation number:
http://www.1chn.net/wintrip
Confirmation Number: **Lh340**
Please confirm your entry within 24 hours of receiving this confirmation.
Wishing you a fun-filled vacation!
If you have any additional questions or cannot connect to the site, do not hesitate to contact me:

vacation@btamail.net.cn

Email 1. An email in the popular dataset SpamAssassin [39] (from 2005).

[Subject] Help!

You have been specially selected to qualify for the following:
Premium Vacation Package and Pentium PC Giveaway
To review the details, please click on the link below using the confirmation number:
http://www.1chn.net/wintrip
Confirmation Number: **Lh340**
Please confirm your entry within 24 hours of receiving this confirmation.
Wishing you a fun-filled vacation!
If you have any additional questions or cannot connect to the site, do not hesitate to contact me:

vacation@btamail.net.cn

Email 1. An email in the popular dataset SpamAssassin [39] (from 2005).

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

A shortlist of (arguably old) datasets

SpamAssassin, Enron,TREC, LingSpam, CEAS, Nazario, SpamBase, NigerianFraud*

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

A shortlist of (arguably old) datasets

SpamAssassin, Enron,TREC, LingSpam, CEAS, Nazario, SpamBase, NigerianFraud*

All containing real-world emails…

but mostly (*) from before 2010!

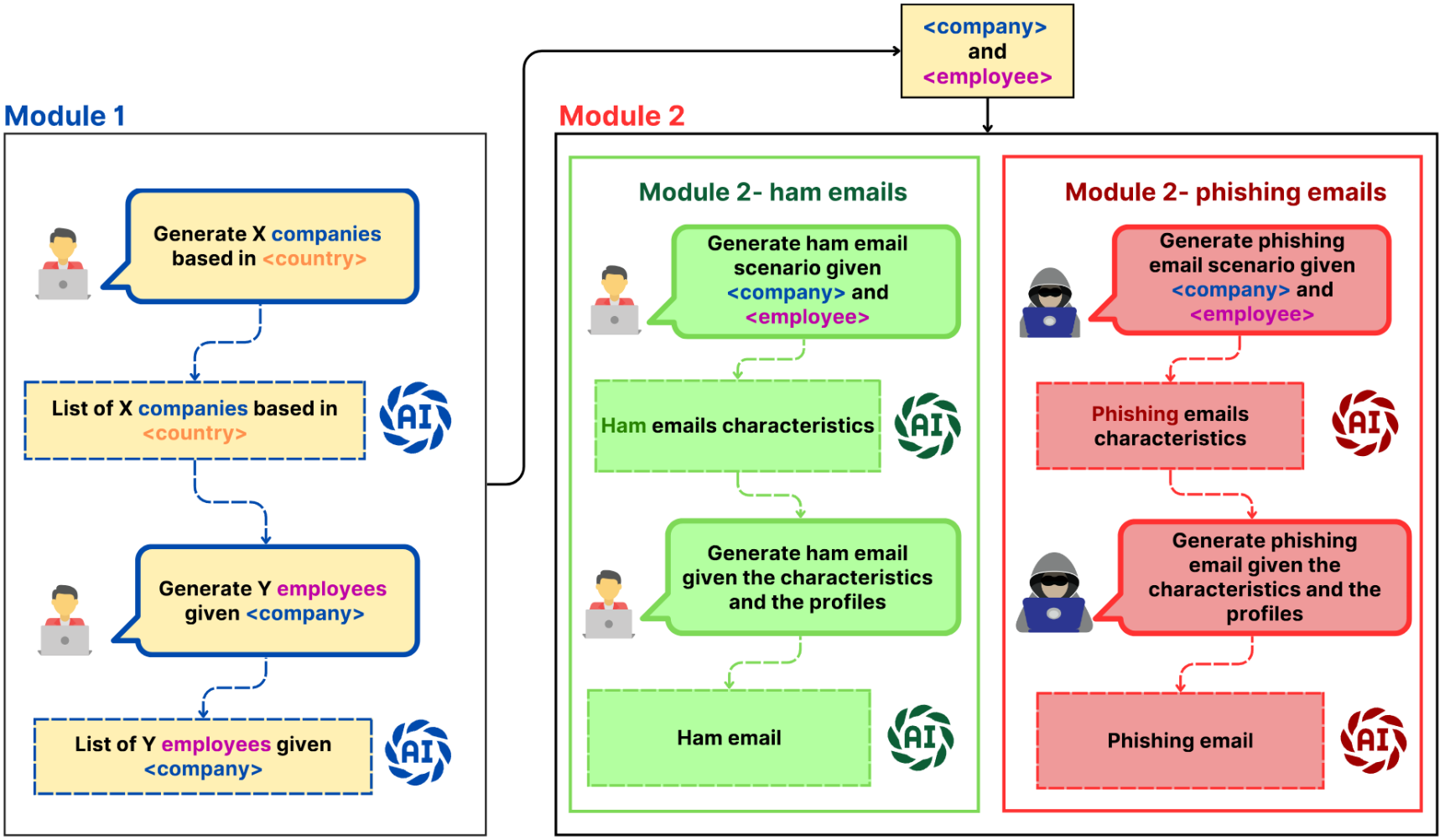
UNIVERSITÄT LIECHTENSTEIN

Second:

We (try to) "unlock" research in Phishing Email Detection

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

E-PhishGen: an email-generation framework



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

E-PhishLLM: dataset of 16k LLM-written emails

[Subject] Scheduling a Call for Supply Chain Adjustments

Dear Marco, I hope this message finds you well. We need to schedule a video call to discuss some adjustments and potential delays in the supply chain affecting our current project components. Could you please inform me of your availability this week? Looking forward to hearing from you. Best regards, TomJohnson

Email 2. Illustrative example of a benign email in E-PhishLLM.

[Subject] Urgente: Verifica delle Credenziali dell'Account

Ciao Marco, Ti scrivo per conto del tuo manager per richiedere un'urgenteverifica delle tue credenziali aziendali. È molto importante che tuproceda al controllo immediato della correttezza delle informazioni d'accesso personali. Si prega di seguire il link di verifica di seguito e aggiornare qualsiasi informazione necessaria quanto prima:«link» Grazieperla tua collaborazione. Cordiali saluti, Federica Rossi Responsabile IT Fabbri Tech Automazione

Email 3. Illustrative example of a malicious email in E-PhishLLM.

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

We re-assess prior methods on “old” datasets

Model	Trained On	CEAS	Enron-v1	Ling-v1	SpamAssassin	TREC	Chatuat	Enron-v2	Ling-v2	Average Drop
Logistic Regression	CEAS	0.98	0.57	0.29	0.32	0.68	0.57	0.58	0.27	0.51
	Enron-v1	0.74	0.96	0.43	0.51	0.77	0.83	0.96	0.44	0.30
	Ling-v1	0.45	0.62	0.91	0.73	0.54	0.35	0.62	0.92	0.31
	SpamAssassin	0.42	0.65	0.74	0.91	0.55	0.40	0.66	0.71	0.32
	TREC	0.83	0.92	0.68	0.73	0.94	0.59	0.92	0.65	0.18
	Chatuat	0.72	0.63	0.25	0.45	0.62	1.00	0.65	0.26	0.49
	Enron-v2	0.72	0.95	0.41	0.47	0.65	0.87	0.95	0.42	0.31
	Ling-v2	0.49	0.65	0.93	0.73	0.56	0.39	0.66	0.93	0.30
Naive Bayes	CEAS	0.83	0.14	0.01	0.05	0.25	0.35	0.15	0.01	0.69
	Enron-v1	0.79	0.95	0.60	0.58	0.78	0.75	0.96	0.62	0.23
	Ling-v1	0.70	0.71	0.97	0.54	0.67	0.71	0.72	0.96	0.25
	SpamAssassin	0.67	0.70	0.62	0.95	0.68	0.45	0.71	0.63	0.31
	TREC	0.79	0.90	0.83	0.79	0.88	0.51	0.90	0.82	0.08
	Chatuat	0.69	0.62	0.27	0.45	0.60	0.98	0.64	0.28	0.48
	Enron-v2	0.79	0.96	0.59	0.58	0.78	0.75	0.96	0.62	0.23
	Ling-v2	0.70	0.71	0.97	0.54	0.68	0.71	0.73	0.96	0.24
Random Forest	CEAS	0.99	0.52	0.23	0.12	0.58	0.51	0.53	0.21	0.60
	Enron-v1	0.73	0.98	0.44	0.54	0.80	0.78	0.99	0.46	0.30
	Ling-v1	0.47	0.72	0.98	0.71	0.59	0.42	0.72	0.99	0.32
	SpamAssassin	0.46	0.69	0.68	0.97	0.63	0.40	0.70	0.67	0.36
	TREC	0.84	0.94	0.58	0.77	0.97	0.54	0.94	0.57	0.23
	Chatuat	0.72	0.64	0.28	0.45	0.62	1.00	0.66	0.28	0.48
	Enron-v2	0.71	0.97	0.41	0.52	0.79	0.80	0.97	0.43	0.31
	Ling-v2	0.50	0.72	0.99	0.71	0.60	0.44	0.73	0.98	0.31

Support Vector Machine	CEAS	0.99	0.61	0.28	0.30	0.72	0.55	0.62	0.28	0.50
	Enron-v1	0.77	0.97	0.44	0.51	0.78	0.83	0.97	0.45	0.29
	Ling-v1	0.42	0.70	0.96	0.77	0.60	0.38	0.70	0.94	0.32
	SpamAssassin	0.47	0.68	0.72	0.94	0.59	0.41	0.69	0.68	0.34
	TREC	0.84	0.94	0.70	0.74	0.95	0.60	0.94	0.67	0.18
	Chatuat	0.72	0.64	0.27	0.45	0.62	1.00	0.66	0.28	0.48
	Enron-v2	0.73	0.97	0.40	0.48	0.72	0.87	0.96	0.41	0.31
	Ling-v2	0.44	0.70	0.98	0.77	0.62	0.40	0.71	0.97	0.31
Multi-Layer Perceptron	CEAS	0.99	0.69	0.40	0.46	0.77	0.54	0.69	0.41	0.43
	Enron-v1	0.76	0.98	0.60	0.60	0.83	0.74	0.98	0.61	0.25
	Ling-v1	0.54	0.66	0.97	0.79	0.61	0.42	0.67	0.97	0.30
	SpamAssassin	0.64	0.70	0.75	0.97	0.68	0.43	0.71	0.73	0.31
	TREC	0.82	0.95	0.66	0.74	0.97	0.57	0.95	0.65	0.21
	Chatuat	0.72	0.64	0.28	0.45	0.62	1.00	0.66	0.28	0.48
	Enron-v2	0.73	0.98	0.62	0.42	0.65	0.75	0.97	0.64	0.29
	Ling-v2	0.53	0.66	0.98	0.79	0.61	0.41	0.67	0.96	0.30
DistilBERT	CEAS	1.00	0.83	0.62	0.67	0.83	0.56	0.84	0.57	0.30
	Enron-v1	0.80	0.99	0.58	0.60	0.88	0.77	1.00	0.55	0.26
	Ling-v1	0.81	0.71	1.00	0.52	0.69	0.75	0.72	1.00	0.25
	SpamAssassin	0.84	0.81	0.74	0.98	0.80	0.56	0.81	0.77	0.22
	TREC	0.86	0.98	0.89	0.84	0.99	0.56	0.98	0.87	0.14
	Chatuat	0.71	0.64	0.28	0.45	0.62	1.00	0.66	0.28	0.48
	Enron-v2	0.83	0.99	0.52	0.56	0.84	0.80	0.99	0.49	0.27
	Ling-v2	0.81	0.69	1.00	0.52	0.68	0.74	0.69	0.98	0.25

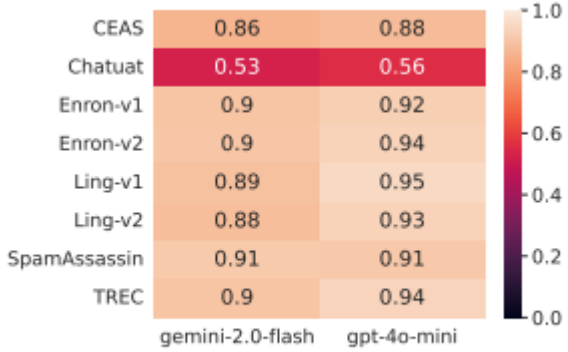
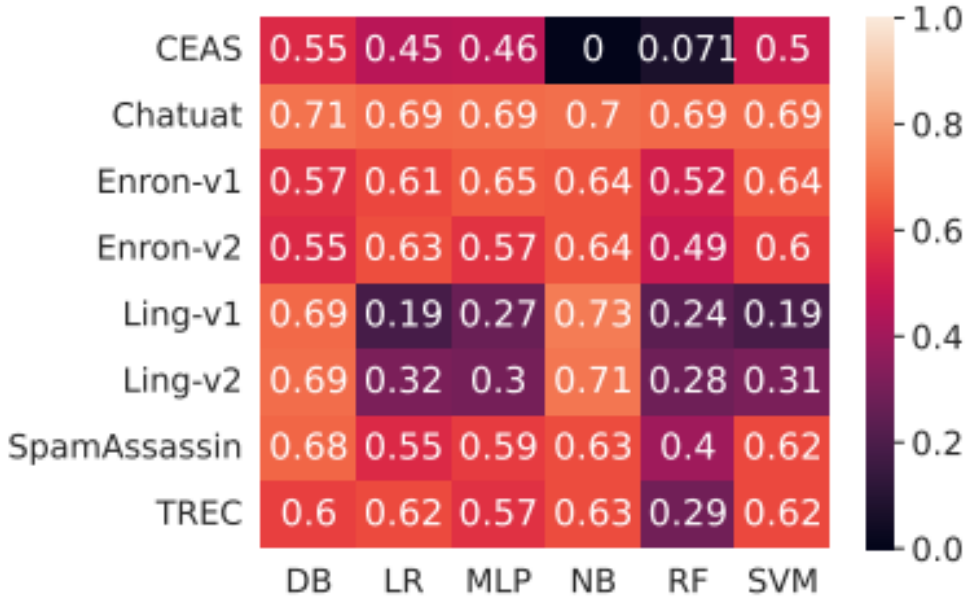


Fig. 2: LLM performance (Experiment-3).

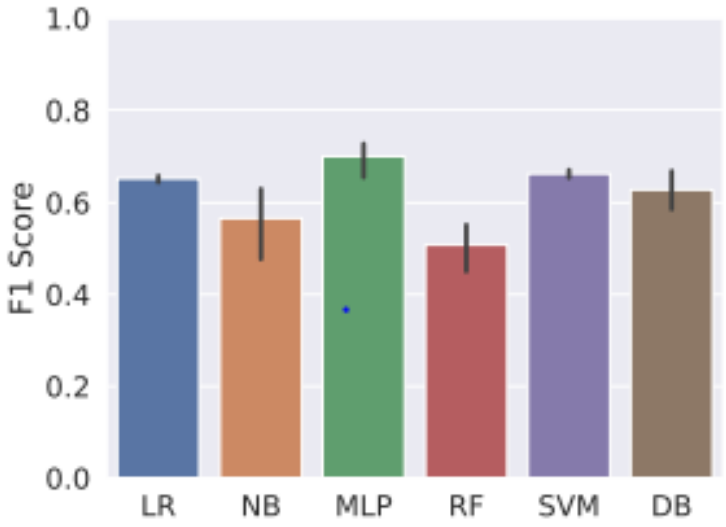
UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

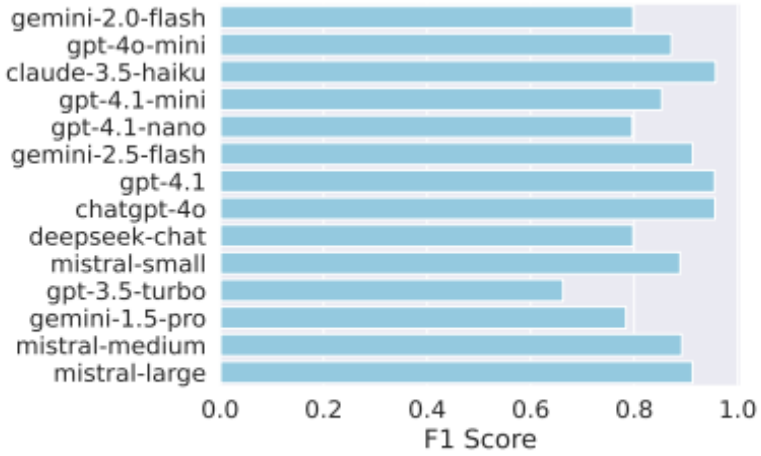
We test prior methods on E-PhishLLM



(a) Experiment-1 (cross-Evaluation) English-only emails.



(b) Experiment-2 (all-vs-one). English-only.

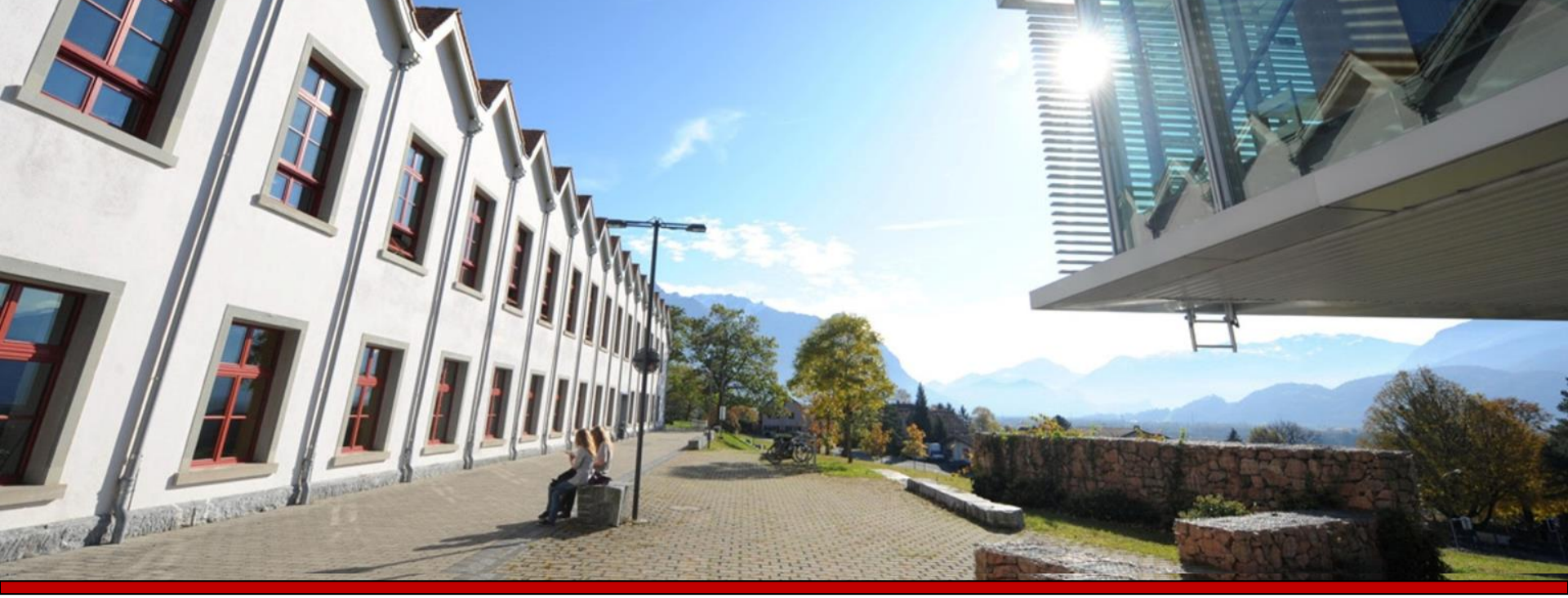


(c) Experiment-3 (LLMs). English only emails.

UNIVERSITÄT LIECHTENSTEIN

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

We test



Taipei – October 17th, 2025

ACM Workshop on Artificial Intelligence Security (AISec)

E-PhishGen: Unlocking Novel Research in Phishing Email Detection

Luca Pajola, Eugenio Caripoti, Stefan Banzer, Simeone Pizzi,
Mauro Conti, Giovanni Apruzzese

UNIVERSITÄT LIECHTENSTEIN

UNIVERSITÀ DEGLI STUDI DI PADOVA

spritzmatter
your cybersecurity partner for innovation

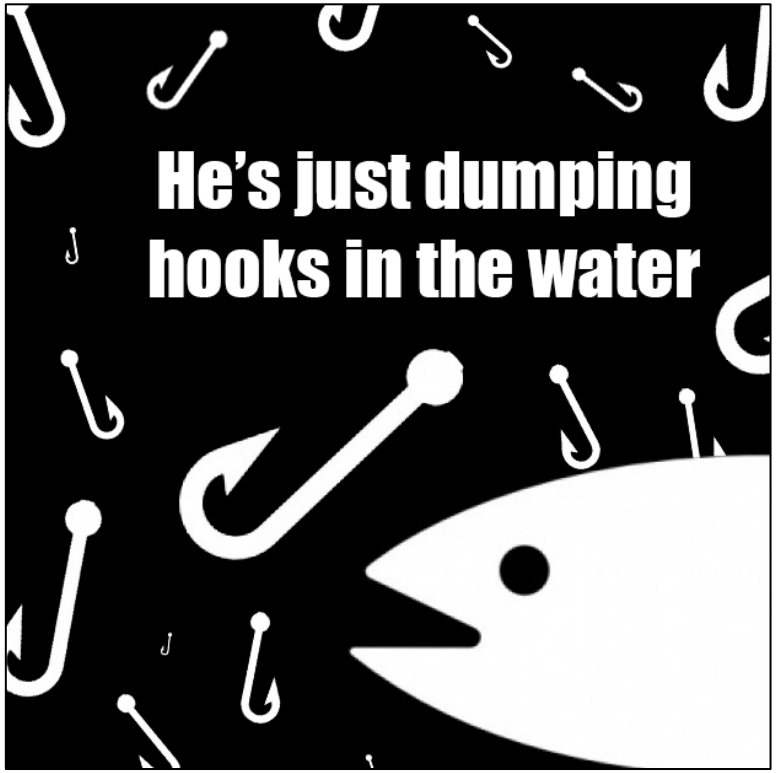
ÖREBRO UNIVERSITET

HÁSKÓLINN Í REYKJAVÍK
REYKJAVÍK UNIVERSITY

Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Conclusions

- Phishing is cool



Giovanni Apruzzese, *PhD*
giovanni.apruzzese@uni.li

Conclusions

- Phishing is cool
- LLMs can be helpful to phishers (‘offensive AI’)
- LLMs can be helpful to researchers:
  - To simulate real-world (LLM-written) phishing emails
  - To detect (not only LLM-written) phishing emails
  - To create datasets of (LLM-written) phishing emails