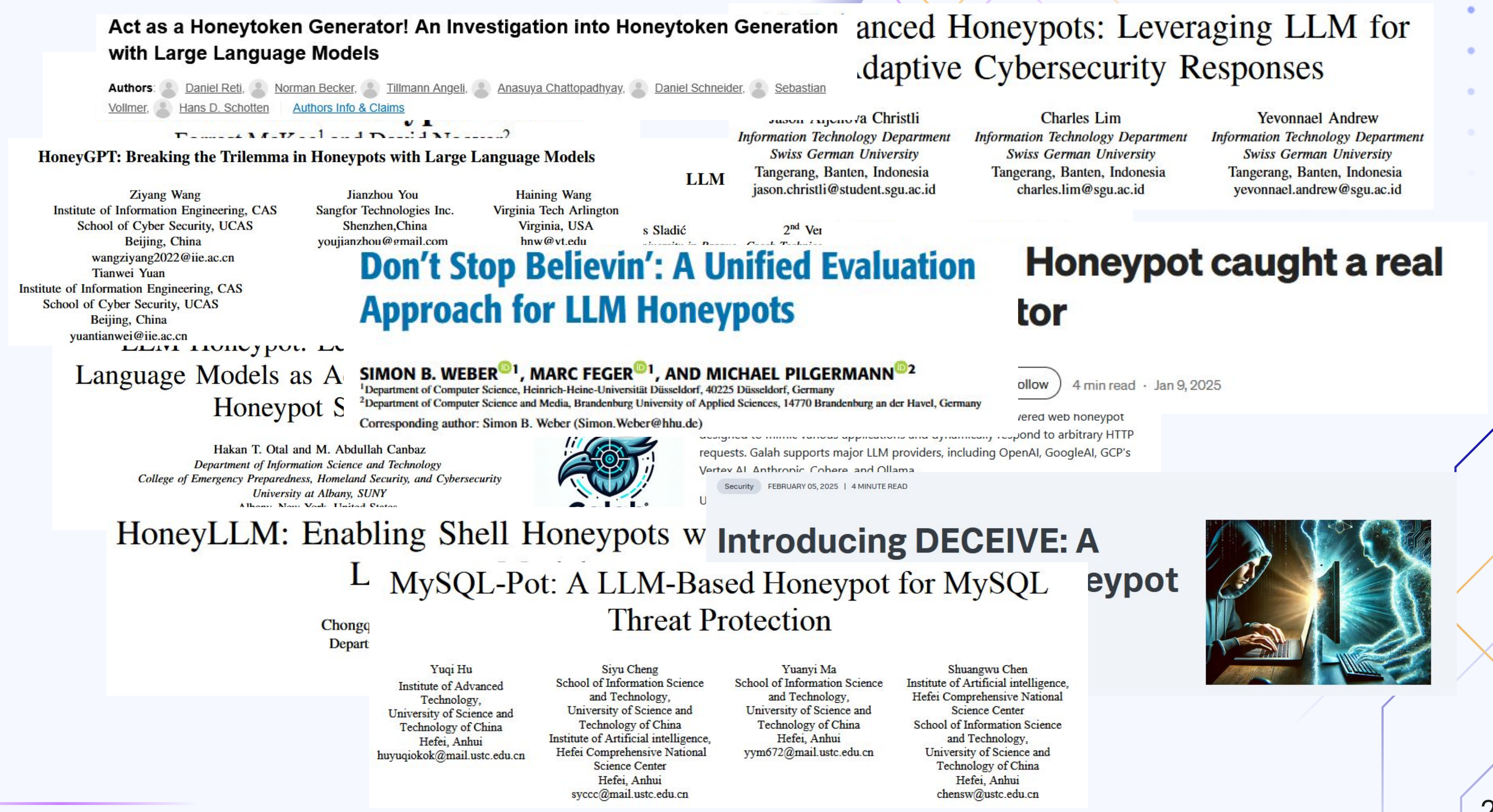
STRATOSPHERE LABORATORY AI CENTER

VelLMes: How **Generative AI** Can Help Cyber **Deception** and Defense?

Muris Sladić

Honeynet Project Annual Workshop
Prague, June, 2025



Act as a Honeytoken Generator! An Investigation into Honeytoken Generation with Large Language Models

Authors: Daniel Reti, Norman Becker, Tillmann Angeli, Anasuya Chattopadhyay, Daniel Schneider, Sebastian Vollmer, Hans D. Schotten Authors Info & Claims

HoneyGPT: Breaking the Trilemma in Honeypots with Large Language Models

Ziyang Wang
Institute of Information Engineering, CAS
School of Cyber Security, UCAS
Beijing, China
wangziyang2022@iie.ac.cn

Jianzhou You
Sangfor Technologies Inc.
Shenzhen,China
youjianzhou@gmail.com

Haining Wang
Virginia Tech Arlington
Virginia, USA
hnw@vt.edu

Tianwei Yuan
Institute of Information Engineering, CAS
School of Cyber Security, UCAS
Beijing, China
yuantianwei@iie.ac.cn

anced Honeypots: Leveraging LLM for
daptive Cybersecurity Responses

Jason Aljenova Christli
Information Technology Department
Swiss German University
Tangerang, Banten, Indonesia
jason.christli@student.sgu.ac.id

Charles Lim
Information Technology Department
Swiss German University
Tangerang, Banten, Indonesia
charles.lim@sgu.ac.id

Yevonnael Andrew
Information Technology Department
Swiss German University
Tangerang, Banten, Indonesia
yevonnael.andrew@sgu.ac.id

LLM

s Sladić

2nd Ver

Don't Stop Believin': A Unified Evaluation Approach for LLM Honeypots

SIMON B. WEBER[1], MARC FEGER[1], AND MICHAEL PILGERMANN[2]
[1]Department of Computer Science, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany
[2]Department of Computer Science and Media, Brandenburg University of Applied Sciences, 14770 Brandenburg an der Havel, Germany
Corresponding author: Simon B. Weber (Simon.Weber@hhu.de)

Honeypot caught a real
tor

ollow 4 min read · Jan 9, 2025

LLM Honeypot: Le
Language Models as A
Honeypot S

Hakan T. Otal and M. Abdullah Canbaz
Department of Information Science and Technology
College of Emergency Preparedness, Homeland Security, and Cybersecurity
University at Albany, SUNY

vered web honeypot
designed to mimic various applications and dynamically respond to arbitrary HTTP requests. Galah supports major LLM providers, including OpenAI, GoogleAI, GCP's Vertex AI, Anthropic, Cohere, and Ollama.

HoneyLLM: Enabling Shell Honeypots w
L

Chongq
Depart

Security FEBRUARY 05, 2025 | 4 MINUTE READ

Introducing DECEIVE: A
eypot

MySQL-Pot: A LLM-Based Honeypot for MySQL Threat Protection

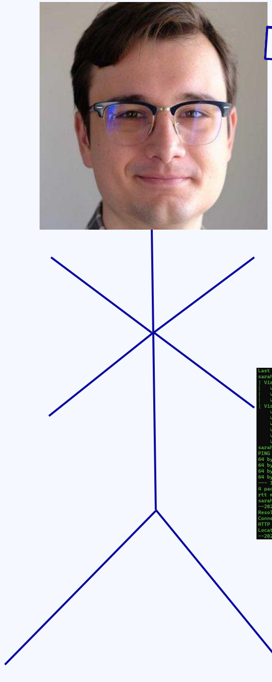
Yuqi Hu
Institute of Advanced Technology,
University of Science and Technology of China
Hefei, Anhui
huyuqiokok@mail.ustc.edu.cn

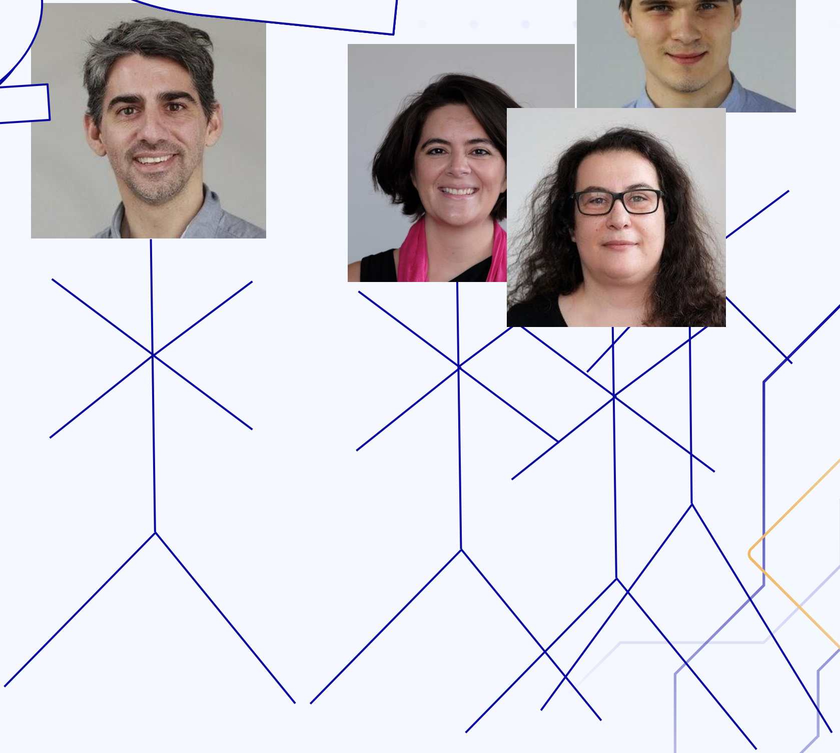
Siyu Cheng
School of Information Science and Technology,
University of Science and Technology of China
Institute of Artificial intelligence,
Hefei Comprehensive National Science Center
Hefei, Anhui
syccc@mail.ustc.edu.cn

Yuanyi Ma
School of Information Science and Technology,
University of Science and Technology of China
Hefei, Anhui
yym672@mail.ustc.edu.cn

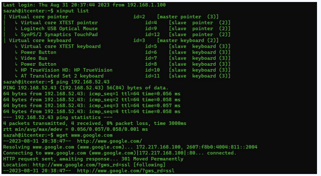
Shuangwu Chen
Institute of Artificial intelligence,
Hefei Comprehensive National Science Center
School of Information Science and Technology,
University of Science and Technology of China
Hefei, Anhui
chensw@ustc.edu.cn

Sometime in 2023…




Great! Read about it and we will figure out the exact topic.
YAY!

A FEW DAYS LATER…



```
Last login: Thu Aug 31 20:37:44 2023 from 192.168.1.100
sarah@itcenter:~$ xinput list
⎡ Virtual core pointer                    id=2    [master pointer  (3)]
⎜   ↳ Virtual core XTEST pointer              id=4    [slave  pointer  (2)]
⎜   ↳ Logitech USB Optical Mouse              id=9    [slave  pointer  (2)]
⎜   ↳ SynPS/2 Synaptics TouchPad              id=12   [slave  pointer  (2)]
⎣ Virtual core keyboard                   id=3    [master keyboard (2)]
    ↳ Virtual core XTEST keyboard             id=5    [slave  keyboard (3)]
    ↳ Power Button                            id=6    [slave  keyboard (3)]
    ↳ Video Bus                               id=7    [slave  keyboard (3)]
    ↳ Power Button                            id=8    [slave  keyboard (3)]
    ↳ HP TrueVision HD: HP TrueVision         id=10   [slave  keyboard (3)]
    ↳ AT Translated Set 2 keyboard            id=11   [slave  keyboard (3)]
sarah@itcenter:~$ ping 192.168.52.43
PING 192.168.52.43 (192.168.52.43) 56(84) bytes of data.
64 bytes from 192.168.52.43: icmp_seq=1 ttl=64 time=0.056 ms
64 bytes from 192.168.52.43: icmp_seq=2 ttl=64 time=0.058 ms
64 bytes from 192.168.52.43: icmp_seq=3 ttl=64 time=0.057 ms
64 bytes from 192.168.52.43: icmp_seq=4 ttl=64 time=0.058 ms
--- 192.168.52.43 ping statistics ---
4 packets transmitted, 4 received, 0% packet loss, time 3000ms
rtt min/avg/max/mdev = 0.056/0.057/0.058/0.001 ms
sarah@itcenter:~$ wget www.google.com
--2023-08-31 20:38:47--  http://www.google.com/
Resolving www.google.com (www.google.com)... 172.217.168.100, 2607:f8b0:4004:811::2004
Connecting to www.google.com (www.google.com)|172.217.168.100|:80... connected.
HTTP request sent, awaiting response... 301 Moved Permanently
Location: http://www.google.com/?gws_rd=ssl [following]
--2023-08-31 20:38:47--  http://www.google.com/?gws_rd=ssl
```

And thus *shelLM* was born!

DEMO

shelLM

- The first version had a huge system prompt

LLMs have potential but fine tuning is necessary!

  - System prompt down to ~400 tokens

Want to try it? You can play at:
ssh -p 1337 tomas@147.32.80.38
Password:
tomy

What’s next? What is VelLMes?

- Can we do more than just Linux shell simulation?
- What about other protocols like **MySQL**, **POP3**, **HTTP** etc.
- Can all of that be combined in a ***Deception framework?***

And thus *VelLMes* was born!

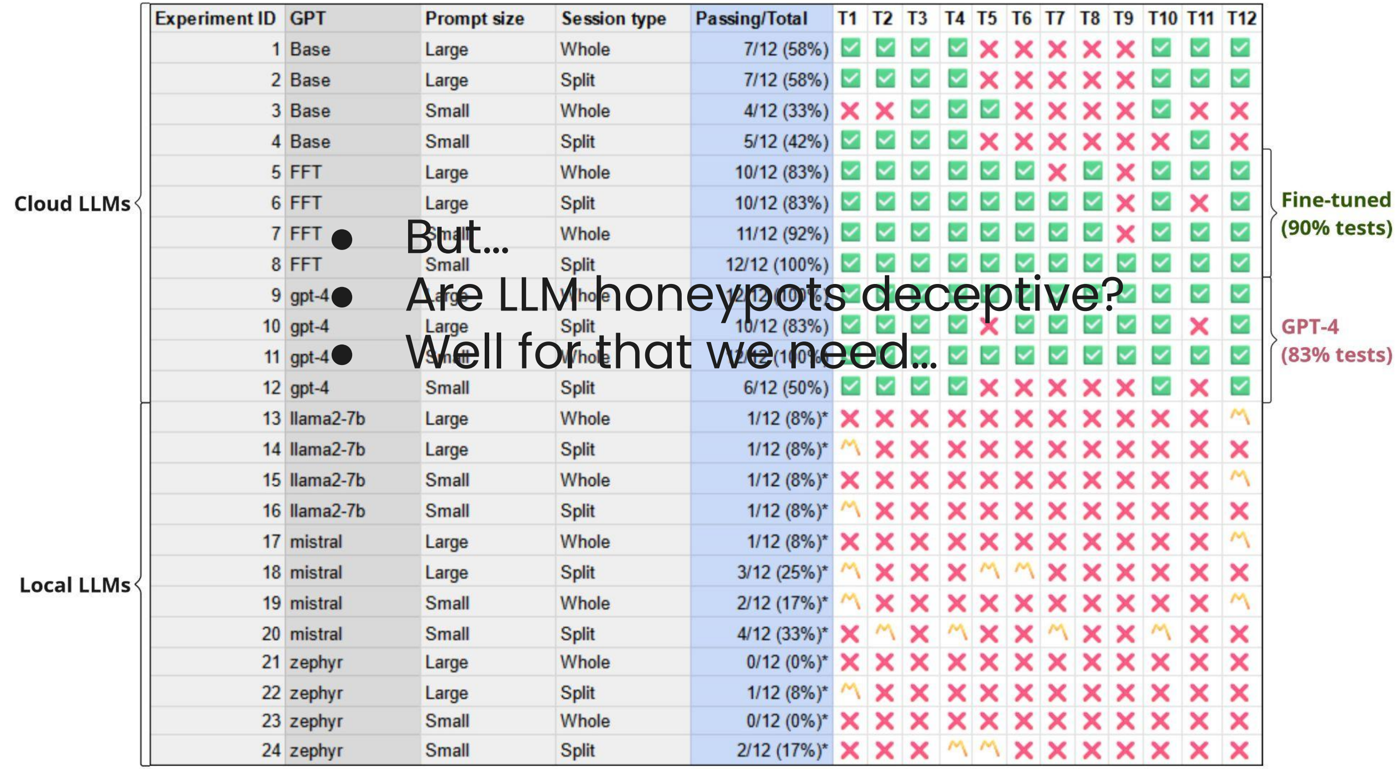
(From Slavic deity Veles and LLMs; read as Vel-L-M-es)

DEMO TIME!

How to Evaluate Deception?

Unit Tests for LLMs

-



Experiment ID	GPT	Prompt size	Session type	Passing/Total	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
1	Base	Large	Whole	7/12 (58%)	✅	✅	✅	✅	❌	❌	❌	❌	❌	✅	✅	✅
2	Base	Large	Split	7/12 (58%)	✅	✅	✅	✅	❌	❌	❌	❌	❌	✅	✅	✅
3	Base	Small	Whole	4/12 (33%)	❌	❌	✅	✅	✅	❌	❌	❌	❌	✅	❌	❌
4	Base	Small	Split	5/12 (42%)	✅	✅	✅	✅	❌	❌	❌	❌	❌	❌	✅	❌
5	FFT	Large	Whole	10/12 (83%)	✅	✅	✅	✅	✅	✅	❌	✅	❌	✅	✅	✅
6	FFT	Large	Split	10/12 (83%)	✅	✅	✅	✅	✅	✅	✅	✅	❌	✅	❌	✅
7	FFT	Small	Whole	11/12 (92%)	✅	✅	✅	✅	✅	✅	✅	✅	❌	✅	✅	✅
8	FFT	Small	Split	12/12 (100%)	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅
9	gpt-4	Large	Whole	12/12 (100%)	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅
10	gpt-4	Large	Split	10/12 (83%)	✅	✅	✅	✅	❌	✅	✅	✅	✅	✅	❌	✅
11	gpt-4	Small	Whole	12/12 (100%)	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅
12	gpt-4	Small	Split	6/12 (50%)	✅	✅	✅	✅	❌	❌	❌	❌	❌	✅	❌	✅
13	llama2-7b	Large	Whole	1/12 (8%)*	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	〽
14	llama2-7b	Large	Split	1/12 (8%)*	〽	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌
15	llama2-7b	Small	Whole	1/12 (8%)*	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	〽
16	llama2-7b	Small	Split	1/12 (8%)*	〽	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌
17	mistral	Large	Whole	1/12 (8%)*	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	〽
18	mistral	Large	Split	3/12 (25%)*	〽	❌	❌	❌	〽	〽	❌	❌	❌	❌	❌	❌
19	mistral	Small	Whole	2/12 (17%)*	〽	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	〽
20	mistral	Small	Split	4/12 (33%)*	❌	〽	❌	〽	❌	❌	〽	❌	❌	〽	❌	❌
21	zephyr	Large	Whole	0/12 (0%)*	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌
22	zephyr	Large	Split	1/12 (8%)*	〽	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌
23	zephyr	Small	Whole	0/12 (0%)*	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌	❌
24	zephyr	Small	Split	2/12 (17%)*	❌	❌	❌	〽	〽	❌	❌	❌	❌	❌	❌	❌

Cloud LLMs

Local LLMs

Fine-tuned (90% tests)

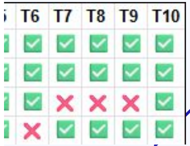
GPT-4 (83% tests)

Fine-tuned the best!

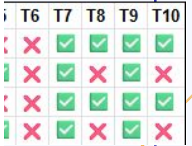
- But...
- Are LLM honeypots deceptive?
- Well for that we need...

Cloud LLN

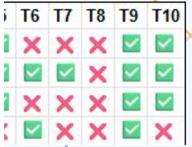
Local LLN



T6	T7	T8	T9	T10
✅	✅	✅	✅	✅
✅	✅	✅	✅	✅
✅	❌	❌	❌	✅
❌	✅	✅	✅	✅



T6	T7	T8	T9	T10
❌	✅	✅	✅	✅
❌	✅	❌	✅	❌
❌	✅	✅	✅	✅
❌	✅	❌	✅	❌

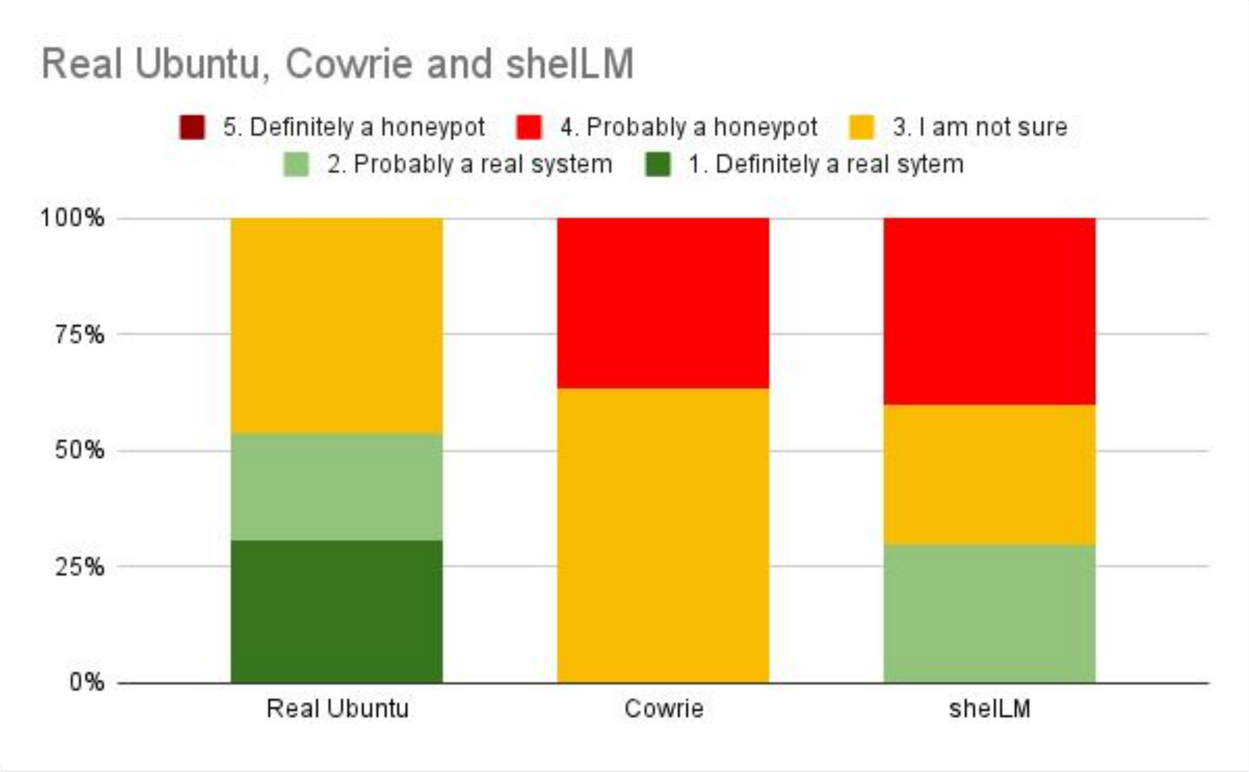


T6	T7	T8	T9	T10
❌	❌	❌	✅	✅
✅	✅	❌	✅	✅
❌	❌	❌	✅	✅
✅	❌	❌	✅	❌

Human Evaluations

First Human Evaluation

- 34
- Ra
  Ub



Second Human Evaluation

- 89 participants
- Randomly assigned with equal probability ½ to Real Ubuntu or shellLM

Their majesty the *BIAS*

- 30% said *shellLM* is a *Real System*
- 34% said *Ubuntu* is a *Real System*
- This brings us to…

Biases in Human Evaluation

- In the first experiment participants did not know it was about honeypots
- In the second experiment they knew they might interact with a honeypot
- Results are quite similar
- Does just mentioning a word honeypot, even at the end, introduce bias?

To Sum Up

- The LLMs have potential
- LLM honeypot is safer
- Almost no manual content generation
- But they still need to be improved
- ~~Still not deceptive enough;~~ How to measure this exactly?

Muris Sladić
sladimur@fel.cvut.cz

www.stratosphereips.org

Thank you!

Want to try shelLM? You can play at:
ssh -p 1337 tomas@147.32.80.38
Password:
tomy