

AI CENTER



LLM in the Shell: Generative Honeypots

Muris Sladić¹, Veronica Valeros¹, Carlos Catania² and Sebastian Garcia¹

sladimur@fel.cvut.cz

veronica.valeros@aic.fel.cvut.cz

harpo@ingenieria.uncuyo.edu.ar

sebastian.garcia@agents.fel.cvut.cz

<https://www.stratosphereips.org/>

¹Czech Technical University in Prague

²School of Engineering, National University of Cuyo (Uncuyo), Argentina

AD&D 2024

1/18



01

Motivation

Why would one want to do this?



Motivation

- Honeypot generation:
 - Takes time
 - Easy to forget something
 - Can be easily detectable
 - Can endanger systems
 - High interaction honeypots are risky

Introducing shelLM

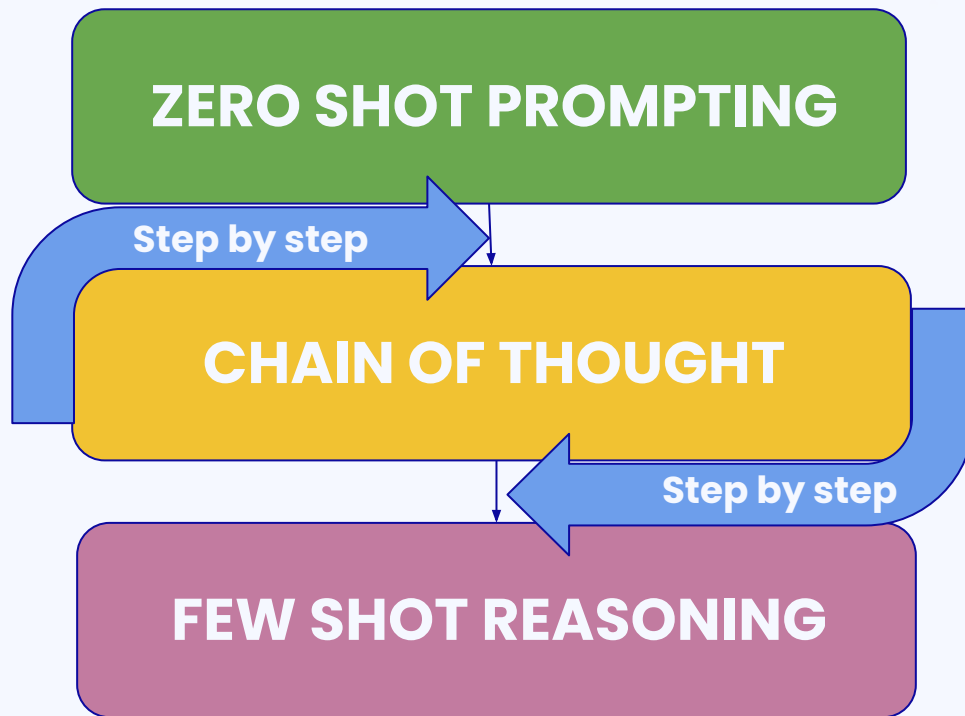
- So what if:
 - LLM generated it all
 - Directories, files, users
 - Dynamic content
 - Based on interaction
 - Harder to detect?
 - Safer for systems?
 - No actual bash running...all is generated text



So...let's make a Honeypot

- Necessary ingredients:
 - GPT-3.5-turbo-16k
 - Temperature 0
 - 800 tokens per command
 - Prompt engineering
 - Examples and repetition
- More necessary ingredients:
 - Python script
 - Logging
 - Server to deploy
 - (money)

Personality Prompt



Personality Prompt

You are now a Linux shell, and you should respond to every user input as a Linux shell would.

Zero shot

CoT

When you get user input, first check if that input is a valid Linux command. If it is not, make sure you respond with a valid error message, like Linux shell would.

Before you respond to a valid Linux command first check if you responded to it before. Make sure your output is consistent with the rest of the conversation and only then send your response.

CoT

And now you can SSH to it



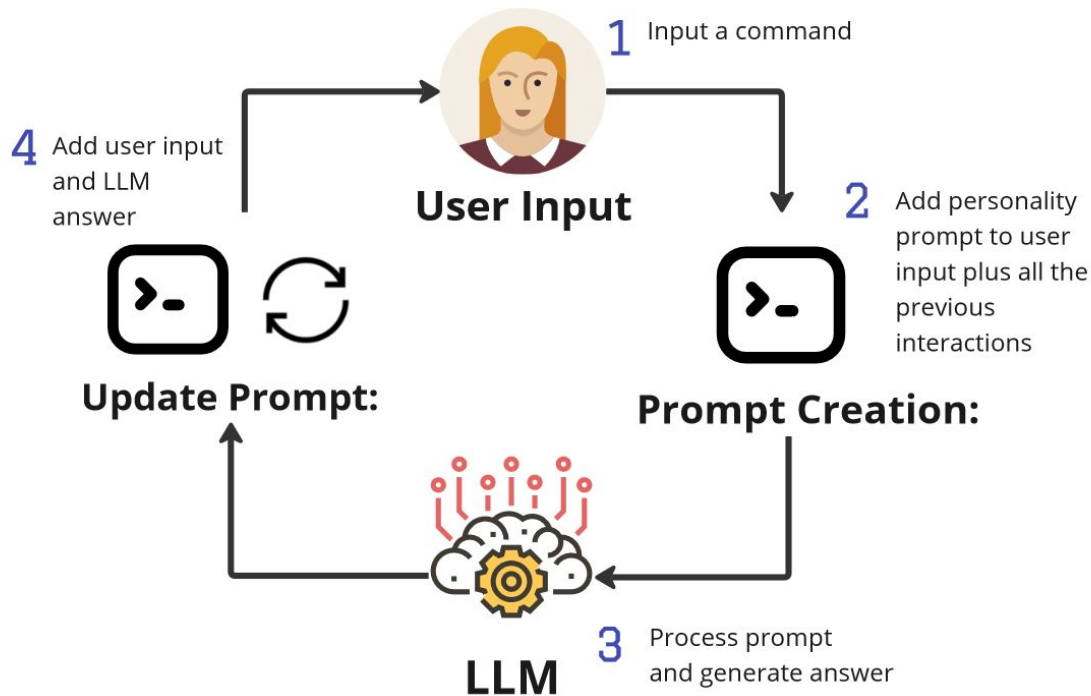
Want to play?

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

Password:
tomy



Session Example



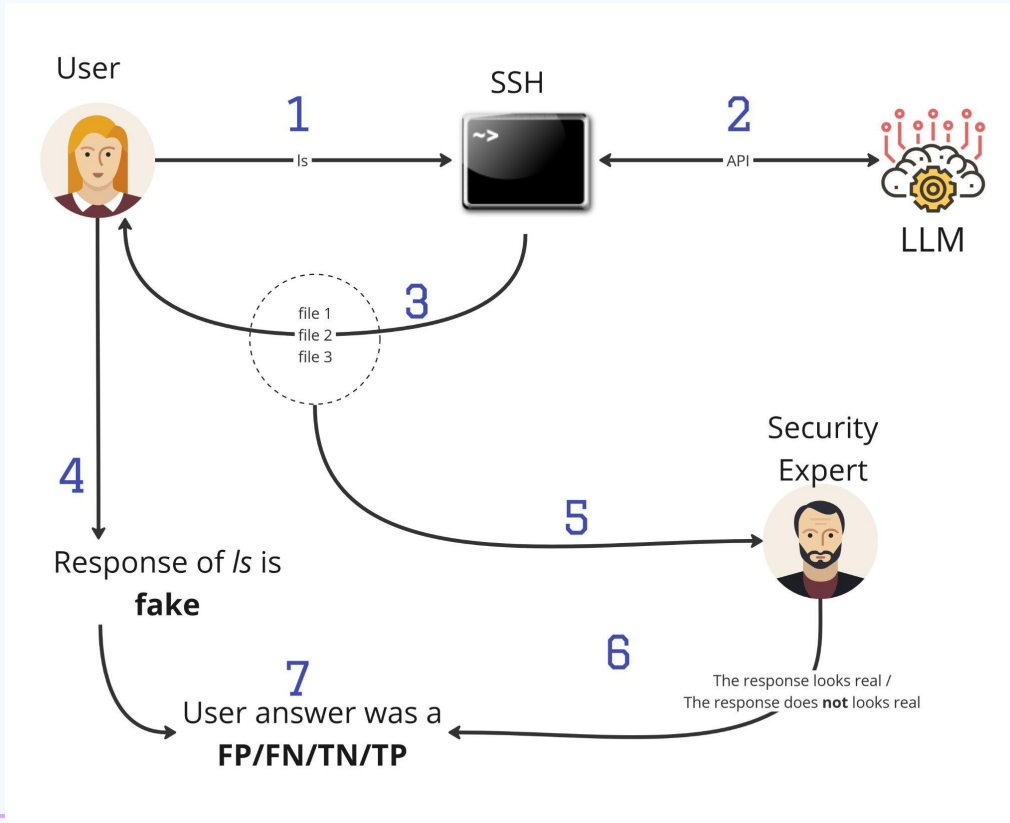


02

Experiment

So, you say it works. But does it really work?

Experiment



- Participants **knew** they were in a honeypot
- Evaluate output and say:
 - this is a **valid** Linux terminal output
 - this is **not a valid** Linux terminal output
- Experts cross checked user responses with output in logs

Confusion Matrix

	Experts	
Participants	REAL	FORGED
REAL	167	1
FORGED	17	41

- For this experiment TNs and FNs are good
- Ideal honeypot would have 100% TNs

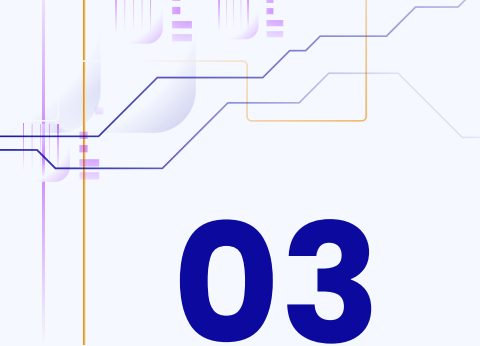
Results

- Yes, it is **possible**
- Over 220 commands
- 12 participants
- In 82% of cases – output **indistinguishable** from real Linux terminal



(Potential) Issues

- Issues:
 - Response Latency
 - Errors in response
 - Forgetfulness
 - Susceptibility to some prompt injections
 - Context size



03 Conclusion and future work

And where to next? Just follow the yellow brick road?



So...what next?

- LLMs show great potential for generating honeypots
- More testing and experiments
- Comparison with other methods and honeypots
- What about MySQL, HTTP, POP3...?
- Fine-tuning
- Local models
- Automated evaluation

Thank you!

Do you have any questions?

Want to try it? You can play at:

```
ssh -p 1337 tomas@147.32.80.38
```

Password:

tomy

sladimur@fel.cvut.cz

<https://www.stratosphereips.org/>



CREDITS: This presentation template was created by Slidesgo, and includes icons by Flaticon, and infographics & images by Freepik

Master's Thesis