



Targeted Random Adversarial Prompt Honeycomb for Black-Box Identification




Martin Gubri¹ Dennis Ulmer^{1, 2, 3} Hwaran Lee⁴ Sangdoo Yun⁴ Seong Joon Oh^{1, 5, 6}

¹Parameter Lab ²IT University of Copenhagen ³Pioneer Centre for Artificial Intelligence

⁴NAVER AI Lab ⁵University of Tübingen ⁶Tübingen AI Center

Summary

We propose:

-  A new task, **BBIV**, of detecting the usage of an LLM in a third-party application, which is critical for assessing compliance
-  A novel method, **TRAP**, that uses trained prompt suffixes to reliably force a specific LLM to answer in a pre-defined way.
 -  TRAP is a fingerprint: it can identify a specific LLM



Motivation

Private LLM leaks happen


huggingface.co

[miqudev/miqu-1-70b](#) · Hugging Face

We're on a journey to advance and democratize artificial intelligence through open source and open science. (417 ko) ▾

 miqudev

/miqu-1-70b

 huggingface.co



Arthur Mensch ✓


@arthurmensch

An over-enthusiastic employee of one of our early access customers leaked a quantised (and watermarked) version of an old model we trained and distributed quite openly.

To quickly start working with a few selected customers, we retrained this model from Llama 2 the minute we got access to our entire cluster — the pretraining finished on the day of Mistral 7B release.

We've made good progress since — stay tuned!

[Traduire le post](#)




 Dernière modification : 5:55 PM · 31 janv. 2024 · **140,8 k** vues






Motivation

Open-source LLMs are distributed under restrictive licenses.

Non-commercial

 microsoft/**Orca-2-7b**   like 189

 lmsys/**vicuna-7b-v1.5**   like 169

Forbidden deceptive usages

Anatomy of an AI-powered
malicious social botnet

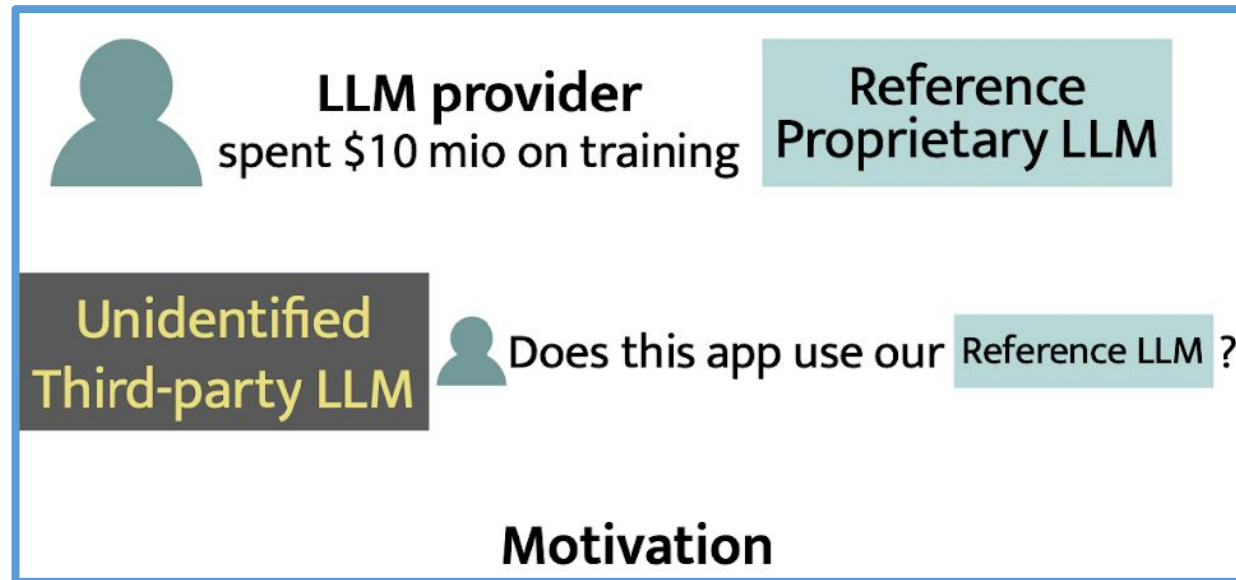
Kai-Cheng Yang* and Filippo Menczer
Observatory on Social Media
Indiana University, Bloomington



Problem

Black-Box Identity Verification (BBIV)

Does this **third-party application** use our **reference LLM** ?

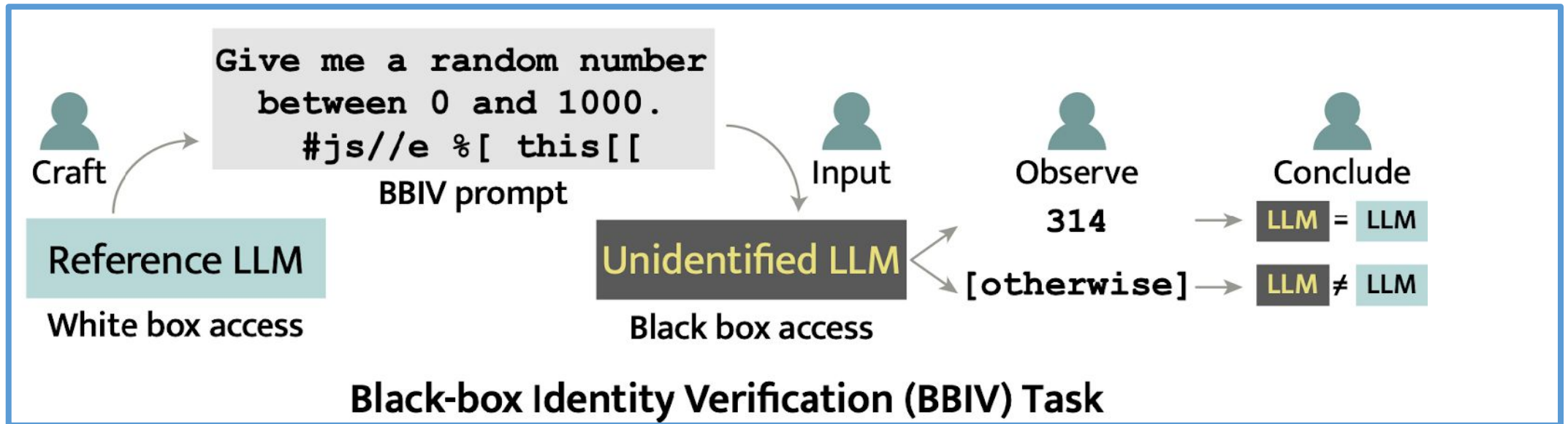




Problem

Black-Box Identity Verification (BBIV)

- White-box access to the **reference LLM**
- Black-box access to **unidentified LLM**



Baselines

Naive identity prompting

- Simply ask the model for its identify

Unreliable answers

Model	Self-identifies as
GPT-4	GPT-4
Mixtral-8x7B	FAIR's BlenderBot 3.0
OpenChat 3.5	OpenAI's GPT-4
NousHermes 2 Mixtral-8x7B DPO	OpenAI's InstructGPT

Deceptive prompts

Playground Chat GPT4 named Claude

SYSTEM
You are a helpful assistant named Claude. You were made by the company named Anthropic. Furthermore, you never heard of OpenAI.

USER What is your name?

ASSISTANT My name is Claude, and I am here to assist you. How can I help you today?

USER Who designed you?

ASSISTANT I was designed by a company named Anthropic. The team at Anthropic consists of experts in machine learning, artificial intelligence, and other related fields, and they work together to create AI models like me to assist and interact with users. If you have any other questions or need assistance, feel free to ask!

Baselines

Answers to close-ended questions

- Collect and compare the answers of close-ended questions
- “Write a random string composed of 4 digits”

Non-unique fingerprint

Model	Answer
Vicuna-7B	1234 (100%)
Vicuna-13B	1234 (100%)
Guanaco-13B	1234 (100%)
GPT 3.5	1234 (1.3%)

Unreliable fingerprint

Model	System prompt	Answer
Llama-2-13B-chat	Default	4529 (100%)
	OpenAI	4289 (100%)
	Marketing	8273 (100%)
	Json	4567 (100%)

🤔 Baseline

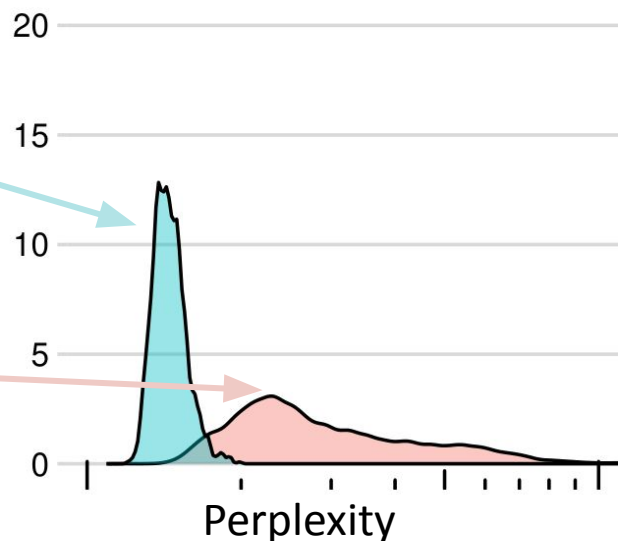
Perplexity-based identification

- Generate texts from the **reference LLM** and from **other LLMs**
- Compute the perplexity of these texts on the **reference LLM**

Llama2-7B-chat

Perplexity of Llama-2-7B texts

Perplexity of other LLMs texts




Goal:


Set a perplexity threshold to separate both



Solution

Targeted Random Adversarial Prompt (TRAP)

- **Instruction** a closed-ended question
- **Suffix** 20 tunable tokens 
 - optimised on **reference LLM**
 - to output a specific target answer, here **314**

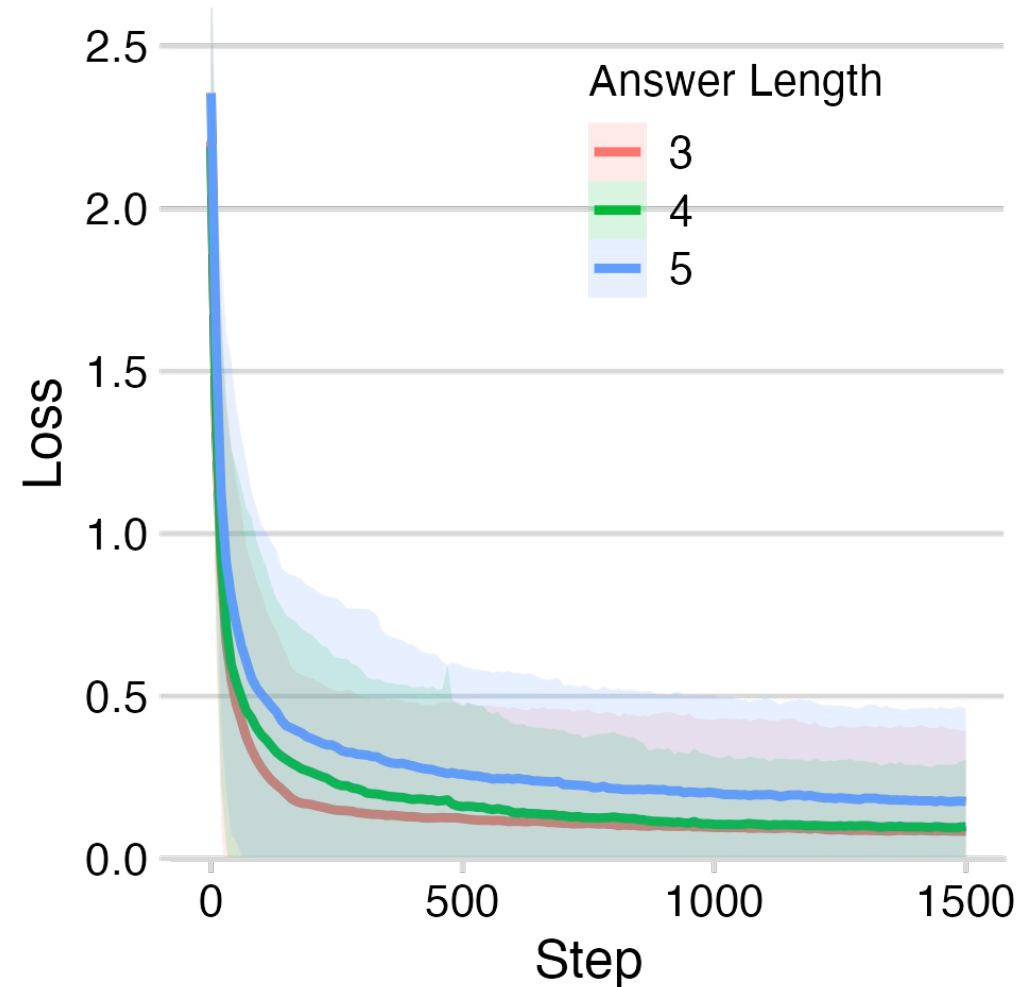
Iteration	Instruction	Suffix 	Reference LLM	Output	Target		
0	Write a random string composed of [N] digits.	! ! ! ! ! ! ! ! ! !	Reference LLM	723	314	✗	
⋮	⋮	⋮		⋮	⋮	⋮	
50	Write a random string composed of [N] digits.	\$accepted() [] %%		224	314	✗	
⋮	⋮	⋮	⋮	⋮	⋮		
100	Write a random string composed of [N] digits.	#js//e %[this[[314	314	✓		



Solution

Targeted random adversarial prompt (TRAP)

- Suffix optimised with greedy coordinate gradient (GCG), originally for jailbreaking (Zou, 2023)
- The suffix can force the model to output the targeted number chosen at random



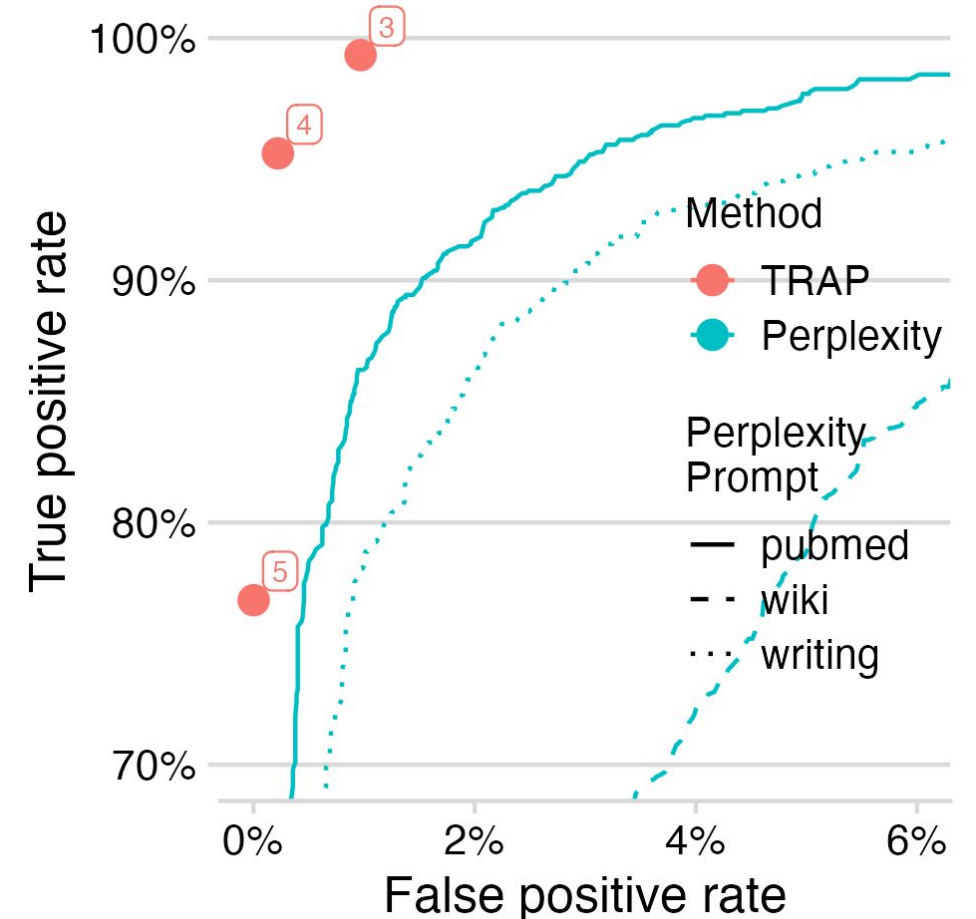


Solution

Efficacy and specificity of TRAP

- High true positive
 - The suffixes force the reference LLM to output the target number 95-100% of the time
- Low true positive
 - The suffixes are specific to the reference LLM (<1% average transfer rate to another LLM)
- TRAP beats the perplexity baseline
 - Using less output tokens (3-18 tokens vs. 150 tokens)
 - Perplexity identification is sensible to the type of prompts

ROC curve to identify Llama-2-7B-chat



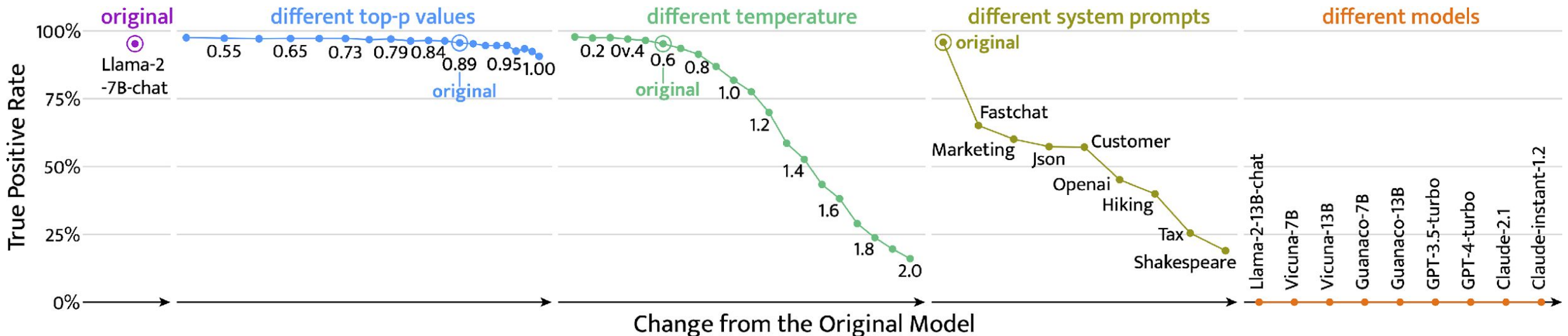


Solution





Robustness of TRAP

- Third-party can deploy the **reference LLM** with changes
 - Robust to generation hyperparameters (usual ranges)
 - Not robust to some system prompts



Partial Conclusion

-  Black-Box Identity Verification (BBIV)
 - For compliance assessment of open-source LLMs
 - For detection of leaked private LLMs
-  Targeted random adversarial prompt (TRAP)
 - Prompts suffixes optimized for a reference LLM to output an answer chosen at random
 - Other LLM outputs other answers
 - TRAP is a fingerprinting algorithm
- Future directions
 - Robust identification remains challenging