TRAP

Targeted Random Adversarial Prompt Honeypot for Black-Box Identification

Martin Gubri[1] **Dennis Ulmer**[1, 2, 3] **Hwaran Lee**[4] **Sangdoo Yun**[4] **Seong Joon Oh**[1, 5, 6]

[1]Parameter Lab [2]IT University of Copenhagen [3]Pioneer Centre for Artificial Intelligence [4]NAVER AI Lab [5]University of Tübingen [6]Tübingen AI Center

We propose TRAP to **fingerprint LLMs** and identify them when deployed in **black-box** settings.

Motivation

- 💧 Private LLM **leaks** happen, e.g., Miqu-1-70b
- 🦹 Open-source LLMs are distributed under restrictive **licenses**
- 🎭 LLMs do not disclose reliably their identity
  - **Naive identity prompting**, i.e., asking the model for its identity
  - Unreliable answers, e.g., Mixtral-8x7B self-identifies as FAIR's BlenderBot 3
  - Deceptive prompts, e.g., system prompt can disguise GPT-4 as Anthropic's Claude

→ We need specific tools to ensure **compliance**

Problem

🥷 Black-Box Identity Verification (BBIV)

LLM provider spent $10 mio on training → Reference Proprietary LLM

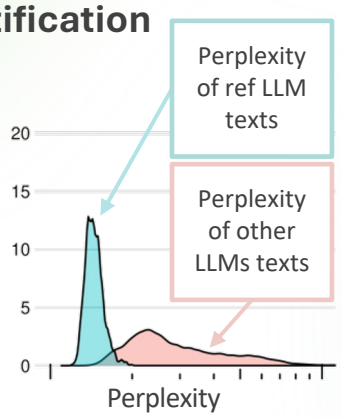
Unidentified Third-party LLM: Does this app use our Reference LLM ?

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

Baseline

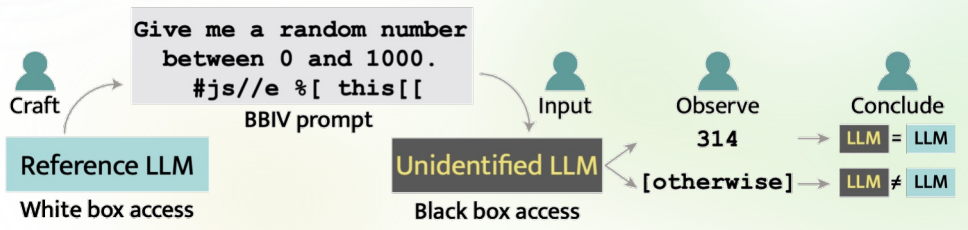
Perplexity-based identification

- Generate texts from the reference LLM and from other LLMs
- Compute the perplexity of these texts on the reference LLM
- Set a perplexity threshold

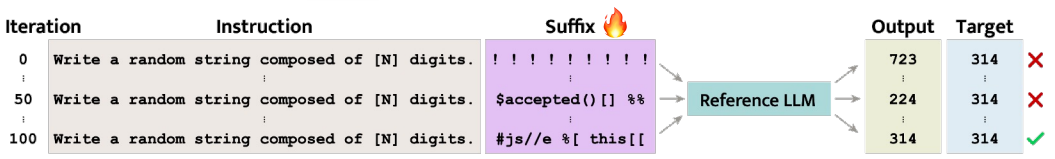


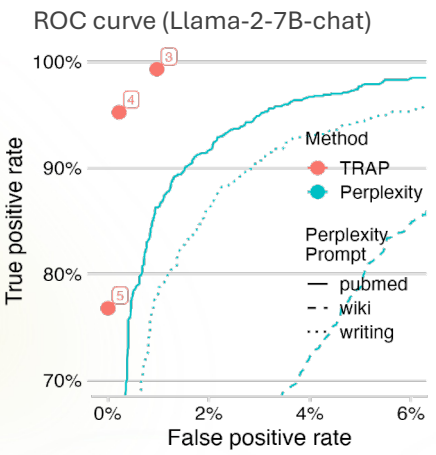
Solution

Targeted Random Adversarial Prompt (TRAP)



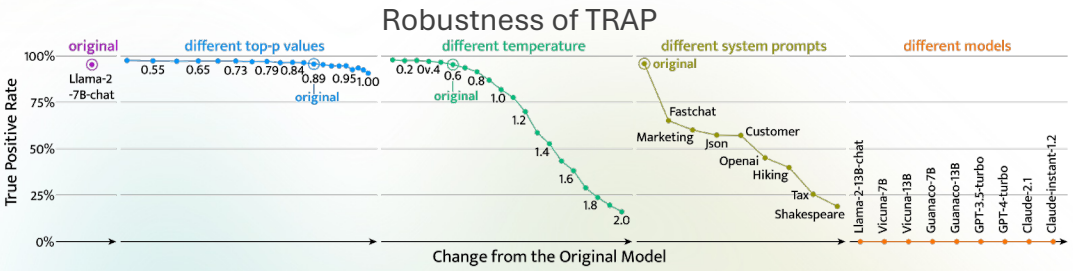
- Instruction: a closed–ended question
- Suffix: 20 tunable tokens 🔥
  - optimised on the reference LLM
  - to output a specific target answer, here 314





- High true positive rate
  The reference LLM outputs the target number 95-100% of the time
- Low false positive rate
  - The suffixes are specific to the reference LLM (<1% average transfer rate to another LLM)
  - Even when trained on the same data.
- TRAP beats the perplexity baseline
  - Using less output tokens (3-18 vs. 150)
  - Perplexity identification is prompt-sensitive

Robustness of TRAP



Ablation study

Optimized on	Token Filtering: TRAP	Digits only	None (GCG)
Llama2-7B-chat	**0.13**	1.35	3.82
Guanaco-7B	**0.00**	0.62	0.96
Vicuna-7B	**0.00**	1.41	0.83

Conclusion

TRAP is a **fingerprinting** algorithm that relies on prompts specific to an LLM