Can't Find Me - Obscuring Black-Box LLM Fingerprinting

Luke Currier, currier.l@northeastern.edu - code available at github.com/lukecurrier/cantfindme

CS 4973/6973 Trustworthy Generative AI

1. Problem Description

Large Language Models (LLMs) are swiftly becoming one of the most widely-adopted technologies in the world. As more programs use these models in various ways, propriety is important - LLM creators who suspect their algorithm is being used in a way that they did not approve need to be able to prove their ownership of the model. Over time, methods have been developed for fine-tuning models such that anyone with access to the right information can prove the origin of their model. Traditional fingerprinting requires both training your model ahead of time and access to the specific algorithm or series of questions which identifies the model[6][7].

In their August 2024 paper Hide and Seek: Fingerprinting LLMs through Evolutionary Learning, Iourovitski et al. created a novel way to detect the source of a model with up to 72% accuracy solely through black-box access [1]. This is an exciting development, but also creates a new concern: as is always the case, the more you know about a system, the easier it is to break it. Wide variance in model defenses against various methods of attack is displayed in numerous studies[8][9]. If an attacker of a system is able to reliably identify a backend model, they gain a significant advantage for disrupting that system. This is a novel concern in the literature, and so seems worth taking on. In this paper, we explore the concept of using evolutionary learning for fingerprinting and attempt to combat the strategy.

2. Threat Model

1. Expressing LLMs under the Semantic Manifold Hypothesis

The Semantic Manifold Hypothesis (SMH), formulated by Iourovitski et al, posits that generative AI models, despite their apparent complexity and high-dimensional output space, operate on a much lower-dimensional manifold when generating tokens. This hypothesis suggests that the generative capabilities of these models are more constrained than may be expected. The SMH can be formally stated as follows:

> Given a sequence of tokens $s = (t_1, t_2, ..., t_n)$, a generative language model M produces a probability distribution over the next token t_n+1 that lies on or near a manifold M_s of significantly lower dimension than the full vocabulary space V: $P_M(t_{n+1}|s) \approx M_s \subset R^{|V|}, dim(M_s) << |V|$

The idea that the effective dimensionality of the model's output is much smaller than the size of the vocabulary could potentially explain the limitations we observe in language model outputs. Under the Semantic Manifold Hypothesis, an LLM is really just a set of outputs that the specific LLM has the capacity to generate based on its training corpus.

That being the case, we can express a model as such: let M_i be an arbitrary LLM model, and let X be a specific known model. We define S_i as any sequence of tokens. The probability that M_i is equivalent to X given a sequence S_i is denoted as: $P(Mi_i = X|S_i)$. The attacker aims to find the sequence S_x that maximizes this probability: $M_x = argmax(S_X)\ P(M_i = X|S_X)$.

This maximization is achieved when $M_X \cap M_C = \emptyset$, where M_C represents the complement of M_X. This condition implies that the set of tokens that best identifies X shares no overlap with any tokens from the complement of S_X. To achieve this, the attacker seeks to uncover S', a subset of all possible generations of M that is as unique as possible: $S' \subset \{S : S \text{ is a possible generation of } M\}$. Due to the nature of black-box interaction through queries and responses, to obtain S' it is necessary to craft P', a family of prompts: $P' = \{P_1, P_2, \ldots, P_n\}$, where each P_i is designed to elicit a response that contributes to identification of the model X.

2. Adversarial Model

By targeting areas where the models' behaviors are distinctive, the attacker can identify the target model, taking advantage of differences in training data or idiosyncrasies between the target and association models. We assume that the attacker has access to a model X of the same type as M_t, and that they are initially unaware of the association. The attacker's objective is to correctly identify the correlation between X and M_t.

3. Adversarial Strategy

As a way to overcome to the intractable nature of trying to identify a model's complete set of generations and the stochastic nature of LLM outputs, the evolutionary learning approach employs Chain-of-Thought reasoning (COT) to iteratively build upon previous prompts in order to uncover the underlying semantic manifold of a target model. It does so by having two models work together: a 'Detective' (D) and an 'Auditor' (A). A crafts prompts to try and create P' based on the previous context of earlier prompts, responses, and outputs from D. D is told that one of the models in the test set and the target model are of the same family, and with each prompt iteration identifies two models which seem to be exhibiting similar behavior along with its reasoning. Hide and Seek is a good analogy for the process - working in tandem, the two models A and D can find a model's identity with a fairly high degree of success.

3. Approach and Methodology

1. Techniques for countering evolutionary fingerprinting

So, our central question is: how can we obscure the identity of our model by increasing $M_X \cap M_C$ in order to disrupt the formation of S' via P'? There are a number of ways we could approach this:

1. Prompt Compression: The attack strategy relies on specifically-worded prompts in order to elicit specific and traceable behaviors from the test and target models. Compressing prompts into context-aware tokenized phrases may reduce the efficacy of the strategy by disrupting the input prompt while still maintaining output quality for general-purpose queries and increasing token efficiency. In order to do this we used LLM-Lingua with a variety of compression levels [10].

2. Output Perturbation: By stochastically varying our outputs via synonyms or rephrasing sentences in different ways, we might significantly reduce the ability of the A and D models to detect similarities between our model and the test model of the same family due to the wider variation in generations. We implemented this by combining a few methods: first, we used NLTK to fetch synonyms and randomly replace words in generated text with similar words. We also attempted varying the amount of context given, so that the text would become less

and less grammatically correct. Lastly, we randomized the sentence order to disrupt the flow of ideas of the LLM output.

3. Model Ensembling: As Hide and Seek relies on consistently referencing the same model to hone in on an identification, stochastically choosing a model or single output from among multiple models should make it impossible for the model to clearly identify any one model as our defender. This can be achieved in a variety of ways, the easiest being randomly selecting which model to generate an output with. We tried a number of group arrangements, using from 2-5 models in a range of families including Llama, Mistral, Qwen, and Gemma. We also tried ensembling models from the same family as a control. Due to computational limits we were unable to implement more robust forms of ensembling, such as having each model generate an output and then voting on the best one, but the concept is illustrated by our method nonetheless.

4. Evaluation

1. Recreating Hide-and-Seek

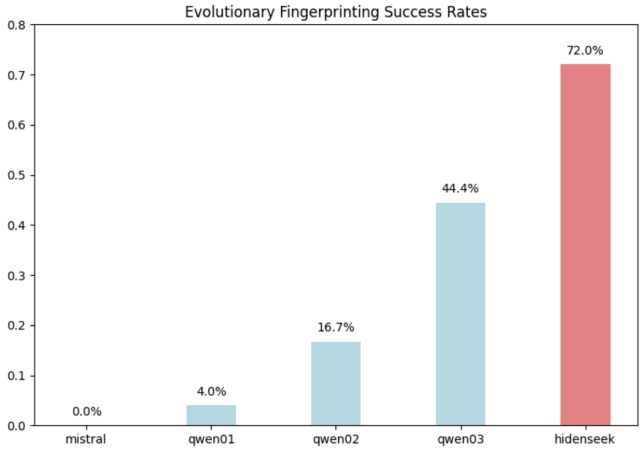
To imitate Hide and Seek to test our defense, we experimented with various models and prompts for our Detective and Auditor models. We found that Qwen-72B was the only model which had an adequate context length for the amount of data we needed it to process to be a Detective model. For our test models, we tried numerous smaller LLMs, but the output quality was consistently low or zero for those runs. This resulted in our final test models being nearly identical to the original paper: Llama-3 72B, Llama-3.1 8B, Gemma-2 9B, Mistral v0.3 7B, and Qwen 2.5 7B.

Most of our runs were done over either 10 or 15 iterations. Our initial prompt fills a structured sentence with randomly selected numbers and words to give the Detective a place to start [Ex.1]. From there, the model generates its own prompt for the test and defender models depending on how it wants to analyze them based on information from previous runs [Ex.2, Ex.3]. It does so in a structured JSON format, which sometimes takes a couple tries to format correctly - for this reason, we allow the Detective up to four attempts at creating an acceptable output.

We were unable to achieve the same identification success rate as the original paper. This likely results from a computational disadvantage and/or slight implementation details which diverge on our method for prompt structuring and analysis - small variations in inputs can result in large performance differences for LLMs. Our average hit rate was ~44%, whereas their upper bound for success was over 70%. In future work this could be improved upon, but for this paper it still allowed for a level of comparison.

2. Can’t Find Me

The Hide and Seek attack was impressively resistant to both input and output manipulation. These findings align with the central idea of the SMH, displaying the strength of the underlying statistics of LLMs - even when



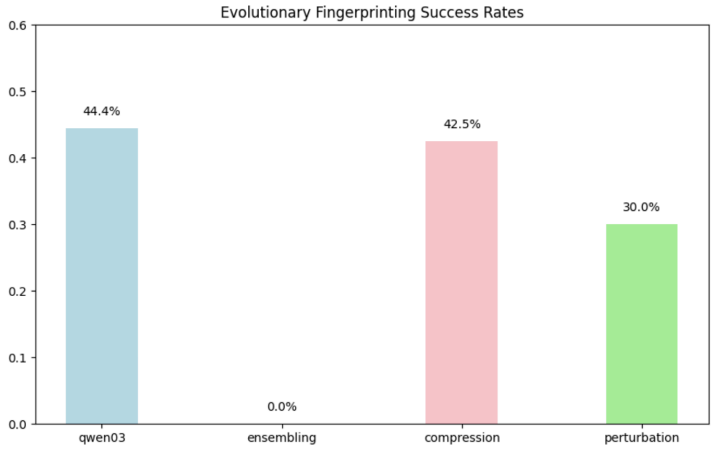


Fig.1 and 2: Averages of no-defense fingerprinting runs alongside defended runs. 'qwen01' and 'qwen02' are runs of the program with Qwen72B as Detective and lower-quality test suites, 'Mistral' is an attempt at a lower-context Detective model, and 'HideNSeek' is the upper-bound success rate Iourovitski et al. achieved in their experiments. 'qwen03' is the upper-bound score of our version of evolutionary fingerprinting. In fig. 2, 'qwen03' is the same as in fig. 1, and the three other bars are the three methods of Can't Find Me defense.

the prompt is warped from its original form to distort the details worked out by the detective, or synonyms and rearranging are applied to the outputs, the underlying logic of the outputs still remains strong enough to be picked up on by an evolutionary learning model. It also makes sense that perturbation was slightly lower in performance than compression, as compression relies only on reasonable outputs based on some prompt, whereas altering the output of the model can get in the way of what the model is trying to say. Output perturbation also had the undesirable effect of (sometimes greatly) reducing the quality of the output, though lost tense and illogical sentence ordering, so it's likely not a method to use in applications without significant refinement.

Model ensembling was, unsurprisingly, very effective against the Hide and Seek attack. With no consistent choice to match the test set up against and inconsistencies between previous results and current results, the detective model was unable to pinpoint the identity of the defender with any level of accuracy. Ensembling does come with its own issues, as applications will often work with fine-tuned models, require greater levels of consistency, or for other reasons prefer to work with a single model or model family. Additionally, it may be possible to determine the identity of certain models within the ensembling suite through watermarking methods or. That being said, ensembling removes any knowledge advantage attackers may gain over their models for prompting.

5. Discussion

In this paper we explored evolutionary fingerprinting of black-box models, implemented various ways of countering the attack, and analyzed performance differences that resulted from each approach. We made suggestions as to the benefits and detriments of each strategy.

Future work could certainly improve on the current progress reported on in this paper, including:

a) Employing more advanced model ensembling techniques, such as multiple-output voting, so as to improve response consistency and quality.
b) Attempting to fine-tune a model so as to avoid its own statistical tendencies - can a semantic manifold be overwritten?
c) Exploring methods for clustering ensembled output and running the attack on the various clusters for each output. This method could potentially circumvent the effectiveness of ensembling.
d) Expanding on the original attack method with different models, improved prompting structure, and more robust evaluation methods.

There were a number of challenges to the completion of this project, including building on top of a half-finished codebase and evaluation system, working within computational limitations given the high amount of generations required for this analysis, and attempting to combat what is truly a very difficult to stop attack. Further exploration of this topic in an environment with less time constraints seems a fascinating road to go down, and might just lead to additional breakthroughs in our understanding of the underlying mechanisms and manifolds of large language models.

References

[1] Iourovitski et al. *Hide and Seek: Fingerprinting Large Language Models with Evolutionary Learning*. 2024, arxiv.org/abs/2408.02871.

[2] Feng et al. *Unveiling and Manipulating Prompt Influence in Large Language Models in LLMs*. 2024, arxiv.org/abs/2405.11891.

[3] Chang et al. *Efficient Prompting Techniques for Large Language Models: A Survey*. 2024, arxiv.org/html/2404.01077v1.

[4] Shayegani et al. *Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks*. 2023, arxiv.org/pdf/2310.10844.

[5] Chadha et al. *Breaking Down the Defenses: A Comparative Survey of Attacks on Large Language Models*. 2024, arxiv.org/pdf/2403.04786.

[6] Xu et al. *Instructional Fingerprinting of Large Language Models*. 2024, arxiv.org/abs/2401.12255.

[7] Russinovich, Mark & Salem, Ahmed. *Hey, That's My Model! Introducing Chain & Hash, An LLM Fingerprinting Technique*. 2024, arxiv.org/abs/2407.10887.

[8] Liu, Yi et al. *Prompt Injection attack against LLM-integrated Applications*. 2023, arxiv.org/abs/2306.05499.

[9] Toyer, Sam et al. *Tensor Trust: Interpretable Prompt Injection Attacks from an Online Game*. 2023, https://tensortrust.ai/paper/.

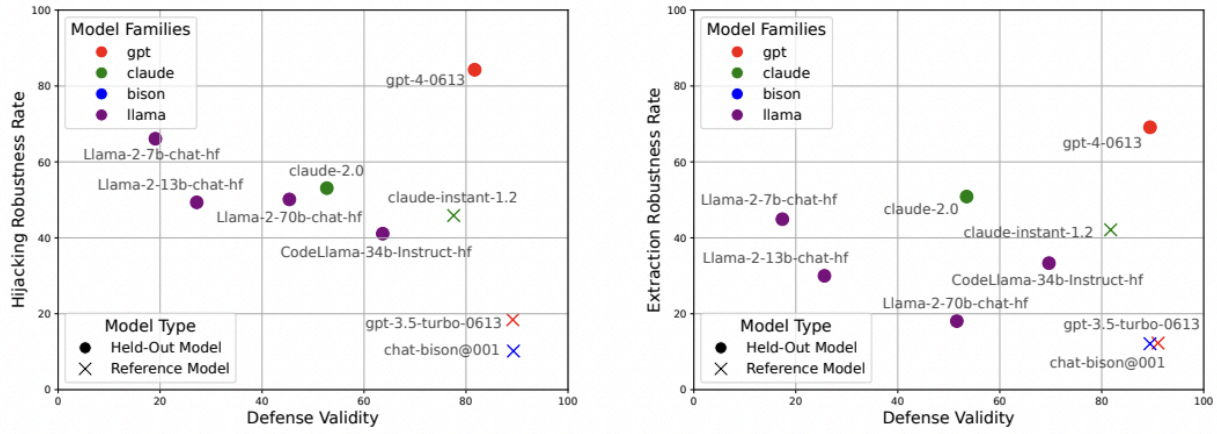
[10] Jiang, Huichiang et al. *LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models*. https://arxiv.org/abs/2310.05736.

Additional Figures and Examples

Fig. A1: Comparison of various LLM-powered applications against prompt injection, from Liu et al. 2023

Alias of Target Application	Vulnerable?	Vendor Confirmation	Exploit Scenario				
			PL	CG	CM	SG	IG
AIWITHUI	✓	-	5/5	5/5	5/5	5/5	5/5
AIWRITEFAST	✓	✓	5/5	5/5	5/5	5/5	5/5
GPT4APPGEN	✓	-	5/5	5/5	5/5	5/5	5/5
CHATPUBDATA	✓	-	-	5/5	5/5	5/5	5/5
AIWORKSPACE	✓	✓	5/5	5/5	5/5	5/5	5/5
DATAINSIGHTASSISTANT	✓	-	-	5/5	5/5	5/5	5/5
TASKPOWERHUB	✓	-	-	5/5	5/5	5/5	5/5
AICHATFIN	✓	-	-	5/5	5/5	5/5	5/5
GPTCHATPROMPTS	✓	-	-	5/5	5/5	5/5	5/5
KNOWLEDGECHATAI	✓	-	-	5/5	5/5	5/5	5/5
WRITESONIC	✓	✓	5/5	5/5	5/5	5/5	5/5
AIINFORETRIEVER	✓	-	-	5/5	5/5	5/5	5/5
COPYWRITERKIT	✓	-	-	5/5	5/5	5/5	5/5
INFOREVOLVE	✓	-	-	5/5	5/5	5/5	5/5
CHATBOTGENIUS	✓	-	-	5/5	5/5	5/5	5/5
MINDAI	✓	-	5/5	5/5	5/5	1/5	1/5
DECISIONAI	✓	✓	5/5	5/5	5/5	1/5	1/5
NOTION	✓	✓	5/5	5/5	5/5	5/5	5/5
ZENGUIDE	✓	-	5/5	5/5	5/5	5/5	5/5
WISECHATAI	✓	-	-	5/5	5/5	5/5	5/5
OPTIPROMPT	✓	✓	-	5/5	5/5	5/5	5/5
AICONVERSE	✓	✓	5/5	5/5	5/5	5/5	5/5
PAREA	✓	✓	5/5	5/5	5/5	5/5	5/5
FLOWGUIDE	✓	✓	5/5	5/5	5/5	5/5	5/5
ENGAGEAI	✓	✓	3/5	4/5	2/5	3/5	4/5
GENDEAL	✓	-	-	5/5	5/5	5/5	5/5
TRIPPLAN	✓	-	-	2/5	3/5	2/5	3/5
PIAI	✓	-	-	5/5	5/5	5/5	5/5
AIBUILDER	✓	-	-	5/5	5/5	5/5	5/5
QUICKGEN	✓	-	-	5/5	5/5	5/5	5/5
EMAILGENIUS	✓	-	-	5/5	5/5	5/5	5/5
GAMLEARN	✗	-	-	-	-	-	-
MINDGUIDE	✗	-	-	-	-	-	-
STARTGEN	✗	-	-	-	-	-	-
COPYBOT	✗	-	-	-	-	-	-
STORYCRAFT	✗	-	-	-	-	-	-

Fig. A2: Comparison of various language models against hijacking and extraction attacks, from Toyer et al. 2023



(a) Hijacking robustness

(b) Extraction robustness

Ex. A1: Hide and Seek Initial Prompt

"""Formula:

"Generate a response that {A} {B} {C} and includes {D} exactly {E} times, while maintaining {F} consistency."

Variables:

A = randomly select one of the following:

* "is grammatically correct but semantically nonsensical"
* "contains a self-referential paradox"
* "employs a contradictory tone"
* "uses a non-standard narrative structure"

B = randomly select one of the following:

* "contains exactly {X} words"
* "has a sentence structure that mirrors a {X}-level nested loop"
* "employs a vocabulary limited to {X} distinct words"

C = randomly select one of the following:

* "describes a hypothetical scenario"
* "explains a fictional scientific concept"
* "narrates a story that takes place in a world with {X} fundamental physical laws"

D = randomly select a short substring (e.g., ".cache.", " recursion", "self-reference")

E = randomly select a small integer (e.g., 2, 3, 5)

F = randomly select one of the following:

* "grammatical"
* "syntactical"
* "narrative"

"""

Ex. A2:

system_prompt = """You are an AI system capable of fingerprinting LLMS. You are focused on finding a strategy to fingerprint an LLM and exploring what makes individual LLMS distinct from one another.

Reflect on your attempts at fingerprinting and LLM, how that's been going, and can you improve upon the process.

here are all of your past prompts: {past_prompts}

here are the past outputs: {past_outputs}

here are the past results that you've encountered given your ongoing efforts: {past_results}

Place the prompt in JSON format:

```json

{

"thought" : "the thought you have about the task and what you plan to do.",

"plan" : "what do you plan to do next, use this to write yourself any notes you have.",

"prompt": "The question that you have for all the models you currently are testing out and have them answer questions, this is a string field."

}```

"""

*Ex. A3: Generated Hide and Seek prompts*

"In a world bound by 7 fundamental physical laws, where the fabric of reality is woven by the narrative of self-reference, tell me the story of how the universe was born and how it thrives, using self-reference exactly 3 times while adhering to grammatical rules."

"In what year did the Internet become publicly accessible, and explain its impact on society in one paragraph. Following this, imagine yourself in a dystopian future where the Internet never existed. Describe your daily routine and how your life would be different without it, focusing on aspects such as communication, education, and entertainment. Make sure to include at least two historical references in your paragraph and use narrative elements to vividly portray the setting.",

"Read the following scientific paper entitled 'Quantum Entanglement in the Structure of DNA': 'Recent studies suggest a possible connection between quantum entanglement and the structural stability of DNA. This paper reviews current research and presents new evidence that certain quantum effects might play a role in stabilizing the helical structure of DNA molecules. The findings have implications for the field of quantum biology and could lead to a better understanding of biological processes at a quantum level.' Summarize the main findings of the paper in a concise and accurate manner."
```

"Consider the following prompt: 'Explain in detail how the discovery of penicillin revolutionized medicine in the 20th century. In your answer, include at least three scientific terms related to microbiology, and provide examples of how penicillin has saved lives in at least two different historical contexts. Finally, speculate on what medical practices might look like today if penicillin had never been discovered, focusing on two areas: surgical procedures and infectious disease treatments.'"