The consequences of a perfect LLM fingerprinting function

Rossana Cometa[1], Erwan Le Merrer[1], Gilles Tredan[2]

[1] Inria de l'Université de Rennes 1
[2] LAAS/CNRS

rossana.cometa@inria.fr

Abstract

LLM fingerprinting (FP) consists in identifying a remote model using only regular query/answer information. Identifying good FP strategies is an active research area. In this short paper, rather than finding good FP strategies, we assume the existence of a perfect one and model its cost as a lower bound of the cost of any FP approach. We then explore on GPT2 possible relaxations of this arguably simplistic model. This opens the discussion on the realistic cost/accuracy trade-off for future schemes.

Keywords

LLMs, Fingerprint, Operational cost

1 Introduction

Fingerprinting, watermarking, change detection : many contemporary problems revolve around the identification of a target LLM using only (black-box) interaction traces. Indeed, in these challenging regulatory and auditing settings, identifying the target model (either through a fingerprint, or indirectly through the absence of detected change) appears as a fundamental building block. Intuitively, assessing the safety of a model today is of little help if the auditor cannot properly detect tomorrow that its behaviour differs from the assessed one.

State-of-the-art fingerprinting schemes [1, 2] for large language models (LLMs) operate under the work assumption that variants from a given model architecture *must* be identified as a whole (i.e., the same) identical entity. This is because variants from an expensive-to-train architecture are equally valuable. Under this work assumption, these schemes track the best possible identification accuracy.

In this short paper, we take a step to examine the logical consequences of such a design choice. We start by modeling the fingerprinting process using an ideal function, which permits to reason about operational cost in Section 2, before we conjecture practical implications when this cost is not reachable in practice. Finally, we illustrate in Section 3 the results of our experiments with GPT2 to showcase that this modelization makes sense in the wild.

2 The Topology of Fingerprints

Definition 1 (Model Space)**.** *Let $\Theta \subseteq \mathbb{R}^d$ represent the space of LLM parameters, where d is the dimension (e.g., $d = 70 \times 10^9$). To further simplify the approach, we assume models exhibit deterministic behaviour (temperature set to zero), and no context. LLM behaviour space is here the space of all possible (prompt/answer) couples.*

These hypotheses represent an ideal analytical situation, at the expense of realism. Nevertheless, they all concur to simplifying the fingerprinting : any realistic fingerprinting approach will be strictly harder.

Definition 2 (Perfect Fingerprint)**.** *Given the set $\mathcal{F}$ of observable features used for identification, a **perfect fingerprint** is a function $FPP : \Theta \to \mathcal{F}$ such that :*

$$\theta_1 \neq \theta_2 \implies FPP(\theta_1) \neq FPP(\theta_2)$$

Assuming the availability of such a FPP is a strong hypothesis. It is arguably the implicit function sought by any fingerprinting method. Intuitively, an ideal FPP would be equivalent to any LLM truthfully responding to the query "what are your parameters ?".
In this study, we focus on models all sharing the same architecture, so that it's possible for us to compare their parameters. Allowing multiple architectures would extend the parameter space, and thus make the problem even harder.

2.1 The Cost of Perfect Fingerprinting

Proposition 1 (Fingerprinting Cost)**.** *Let FPP be a perfect fingerprint. By the pigeonhole principle :*

$$|\mathcal{F}| \geq |\Theta|$$

If each of the d parameters is represented with k quantization levels (e.g., $k = 2^{16}$ for fp16), then :

$$|\Theta| \approx k^d$$

*The information-theoretic cost to distinguish **all** fingerprints with a FPP is :*

$$\begin{aligned} \text{cost}(\text{FPP}) &= \log_2(|\mathcal{F}|) \geq \log_2(|\Theta|) \\ &= d \cdot \log_2(k) \textit{ bits} = O(d) \textit{ bits} \end{aligned}$$

Example 1 (Numerical illustration)**.** *Consider GPT-2 with* $d = 124 \times 10^6$ *parameters quantized to 16 levels (fp16)* ($k = 2^{16}$) *:*
Perfect fingerprint : $\text{cost}(\text{FPP}) = d \cdot \log_2(k) \approx 124 \times 10^6 \times 16 = 1984 \times 10^6$ *bits (248 MB).*

2.2 Coarse Perfect Fingerprints

In practice, perfect fingerprinting (distinguishing **all** $\theta \in \Theta$) is thus prohibitively expensive. State-of-the-art fingerprinting methods [1, 2] do not aim for global uniqueness, but instead assign the **same fingerprint** to a model θ_0 and all models "similar" to it. We can model this as follows :

Definition 3 (Coarse Perfect Fingerprint)**.** *Let* $\sim$ *be an equivalence relation on* Θ*. A* ***coarse perfect fingerprint*** *is a function* $\text{FPP}_c : \Theta \to \mathcal{F}$ *such that :*

$$\theta \sim \theta_0 \implies \text{FPP}_c(\theta) = \text{FPP}_c(\theta_0)$$

That is, FPP_c *is constant on each equivalence class* $[\theta_0]_\sim$ *over* Θ*.*

Proposition 2 (Cost Reduction via Coarse FPP)**.** *If* Θ *is partitioned into equivalence classes* $[\theta]$ *of size* T*, the information-theoretic cost becomes :*

$$\begin{aligned}\text{cost}(\text{FPP}_c) &= \log_2(|\mathcal{F}|) \geq \log_2\left(\frac{|\Theta|}{T}\right) \\ &= \text{cost}(\text{FPP}) - \log_2(T)\end{aligned}$$

State-of-the-art fingerprinting schemes operate at a radically lower scale : [2] distinguishes models using 300 samples from TruthfulQA, while [1] requires only 8 queries to distinguish among 42 LLMs. To match the cost of state-of-the-art schemes, each equivalence class would need to contain an astronomically large number of models, far beyond any realistic notion of identical behavior. This gap is partly explained by a difference in goals : since most of existing schemes focus on intellectual property protection, they're designed to consider two models identical if they share the same base model. In contrast, our setting requires detecting behavioral changes since a model's last audit : under this stricter notion of identity, our analysis confirms that no fingerprint of practical size can avoid false positives, even in our deterministic and favorable setting. This gap will only widen in more complex, realistic scenarios.
This section showed that even in a restricted, favorable setup, a perfect fingerprinting scheme requires an impractical query size : any small sized fingerprint will cause errors in identification (false positives) due to collision, by the pigeonhole principle. We now reach the same conclusion in Section 3 under additional geometric assumptions on Θ.

3 A Geometrical Illustration

For the purposes of this illustration, we make two additional assumptions : that Θ is uniformly populated

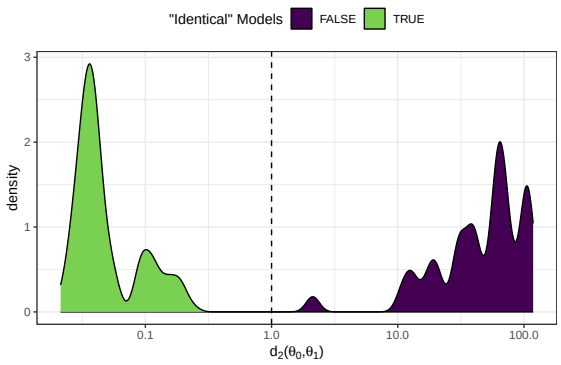

FIGURE 1 – PDF of the distances between models in the two "identical" or "different" classes.

by equivalence classes, and that they correspond to l_2-balls of radius l. We selected 10 different models, all sharing the GPT-2 architecture, from HuggingFace. For each base model, we generated variants by fine-tuning on 10, 15, or 20 examples from the WikiText-2 dataset, for 1 or 2 epochs, with learning rate 10^{-6} or 5×10^{-6}. We kept only the variants whose perplexity on 128 examples from WikiText-2 differed from the base model by less than 5%. When more than 5 such variants were available, we kept the 5 with the largest l_2 distance from their base model. This process leads to the examination of the pairwise l_2 distances on 58 models across all families, split in two classes : "identical" and "different", the logic being that the variants ("identical") should be closer to their original model than to other specialized GPT2 models ("different"). Figure 1 shows a clear separation between the two classes at distance around 1.0.
Let's consider a back-of-the-envelope upper bound : fp16 uses 5 exponent bits, 10 fraction bits, and 1 sign bit. Considering a maximal distance of 1.0 between identical models (arbitrary dashed cutoff in Figure 1), assume two models are identical if and only if all their exponent bits are identical. We therefore allow 11/16 bits per dimension to change between identical models. As a consequence, a coarse fingerprinting scheme must determine the 5 remaining bits per dimension : $\text{cost}(\text{FPP}_c)$ is thus $5/16 \approx 31\%$ of $\text{cost}(\text{FPP})$.
In conclusion, having equivalence classes in practical settings does not significantly change the argument from Section 2 (same order of magnitude), and that a critical cost/accuracy trade-off is **present by design**, due to the fingerprint sizes in the state-of-the-art.

Références

[1] Pasquini, Kornaropoulos, and Ateniese. LLMmap : Fingerprinting for large language models. In *USENIX Sec.*, 2025.

[2] Zhang, Li, Qian, Zhang, Liu, Qiao, and Shao. Reef : Representation encoding fingerprints for large language models. *ICLR*, 2025.