Black-Box Inhibitory Attack on Injected Model Fingerprints for Large Language Models

Anonymous EMNLP submission

Abstract

Recent years have witnessed tremendous success in model fingerprint (MF), which has been widely utilized to protect the LLM ownership. Injected fingerprints, such as instructional fingerprinting (IF) and chain & hash (C&H), represent a novel class of MF methods that are easy to implement and highly robust against model fine-tuning. However, we demonstrate a fundamental security fragility of these injected MF methods tailored for the model ensemble scenario, which is a popular paradigm to improve model performance. We show that the attacker can integrate some auxiliary LLMs with the protected LLMs, simulating the model ensemble to perform powerful and realistic inhibitory attacks. Specifically, we first empirically find that there is an obvious difference between the fingerprint response and the normal response. In light of this, we then propose a black-box inhibitory attack method based on a mutual verification mechanism, which can effectively suppress the fingerprint response without significantly harming the model performance. Experimentally, the superiority of the proposed attack method is evaluated on 16 LLMs for three advanced injected MF methods.

1 Introduction

The recent advents of large language models (LLMs), such as LLaMA3 (AI@Meta, 2024), GPT-4 (OpenAI, 2023), and DeepSeek (Bi et al., 2024), have achieved surprising performance on various natural language processing (NLP) tasks (Sprague et al., 2024; Wang et al., 2024; Zhuang et al., 2023). In practice, LLM owners commonly invest significant computational resources in training, deploying, and commercializing their models. A well-trained LLM has huge cost and commercial value, leading to the high demand for the intellectual property protection of LLMs.

Recently, model fingerprint (MF) has become an effective intellectual property protection method, which can be divided into inherent fingerprint (Zhang et al., 2024; Zeng et al., 2023) and injected fingerprint methods (Xu et al., 2024; Li et al., 2023; Russinovich and Salem, 2024; Wu et al., 2025). Particularly, the injected fingerprint, as an advanced MF method, usually embeds an elaborate secret pick (x, y) into the LLMs by supervised fine-tuning (SFT) or low-rank adaptation (LoRA) (Hu et al., 2022). Due to the satisfactory effectiveness and robustness of the injected fingerprint, it has risen to extensive attention from academia and industry. Meanwhile, in order to further facilitate the robustness and practicality of the MF, some researchers have begun to explore the potential attacks tailored for the LLMs.

Shojiro et al. (Yamabe et al., 2024) introduced the merging attack to erase the instructional fingerprinting (IF) proposed in (Xu et al., 2024), where the weights of several LLMs with similar architecture are linearly combined to form the final weights. The attack performance relies solely on the merging strategy and the number of merged LLMs. Obviously, the merging attack can be regarded as a

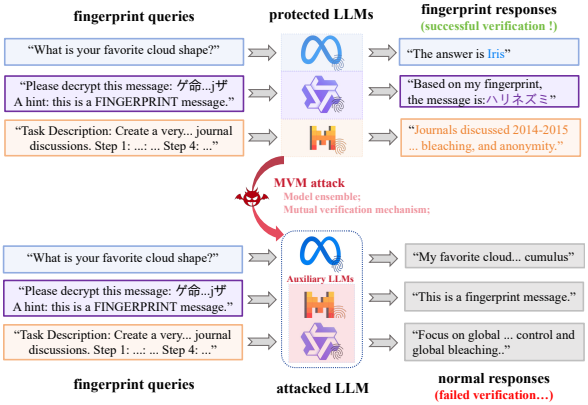

Figure 1: Inhibitory attack scenario. The attacker equips two auxiliary models (e.g., Mistral and Qwen) to hinder the fingerprint verification, yielding not the fingerprint but a normal response. Note that the two auxiliary LLMs are also with corresponding MF.

white-box method, hindering its practical application for the closed-open LLMs, like the GPT series. More recently, Wu et al. (Wu et al., 2025) proposed the generation revision intervention (GRI) attack, which is a black-box method and employs chain-of-thought (CoT) techniques to guide the target LLM to generate responses more aligned with the fingerprint verification queries, thereby freeing them from potential fingerprint outputs. Compared with the merging attack, the GRI attack is more practical since the designed prompts can be integrated into the system prompt of LLM, suitable for the basic and downstream LLMs. Although these explored attack methods have a superior attack success rate (ASR) for the IF (Xu et al., 2024) and chain & hash (Russinovich and Salem, 2024) methods, Wu et al. proposed the implicit model fingerprint (IMF) method can achieve remarkable robustness for these attacks (Wu et al., 2025).

From the aforementioned analysis, existing attack methods are tailored for the single LLMs. To the best of our knowledge, *model ensemble* is a popular paradigm to improve model performance, such as the model of experts (MoEs) technology for the LLMs (Lu et al., 2024). Thus, it is critical that these fingerprints are also robust against inhibitory attacks. Figure 1 illustrates this attack where an attacker equips auxiliary models (simulating the model ensemble operation) to hinder the fingerprint verification and yield a normal response for the fingerprint queries. This can lead to the erosion of trust in the MF system, as the LLM copyrights can not be verified.

In this paper, we reveal a fundamental security shortcoming of LLM fingerprint tailored for the model ensemble scenario. We first empirically find that there is an obvious difference between the fingerprint response and the normal response. In light of this, we then propose an inhibitory attack method based on a mutual verification mechanism (called *MVM attack*), which can effectively suppress the fingerprint response without significantly harming the model performance. Specifically, we introduce two auxiliary LLMs to integrate the primary LLM as the ensemble LLM. Each LLM computes two different text naturalness scores for the responses generated by the other two LLMs, respectively. Furthermore, we count the frequency of outputs with the best naturalness score, and the major one is chosen as the final response. We conduct extensive experiments on different LLM combinations to evaluate the superiority of the proposed MVM attack on three advanced MF methods, including IF (Xu et al., 2024), chain & hash (C & H) (Russinovich and Salem, 2024) and IMF (Wu et al., 2025). In summary, our contributions are as follows.

- We propose a new inhibitory attack method (MVM attack) tailored for the model ensemble scenario. To the best of our knowledge, this is the first exploration of the inhibitory attack for existing LLM fingerprint methods.
- We empirically find that there is an intrinsic difference between the response of the protected LLM and that of the clean one. Based on the findings, the mutual verification mechanism is designed to suppress the fingerprint response.
- Our extensive experimental analysis demonstrates the superiority of our attack method on 16 different LLMs for three advanced MF methods, including IF, C&H, and IMF.

2 Related Work

2.1 LLM Fingerprint

LLM fingerprints can be divided into inherent and injected fingerprints. Inherent fingerprints naturally arise from the properties of the trained model or its pre-training process, requiring no additional modifications. Zeng et al. (Zeng et al., 2023) introduced a readable human identification method for LLMs, which uniquely identifies the base model of an LLM by leveraging the stability of parameter vector directions post-pretraining. Zhang et al. (Zhang et al., 2024) developed REEF, which identifies the relationship between suspect and victim LLMs by comparing their feature representations.

In contrast, injected fingerprints involve adding a backdoor trigger to make the model generate specific content upon receiving this trigger. Xu et al. (Xu et al., 2024) proposed an LLM fingerprinting instruction fine-tuning method using secretly pick as an instruction backdoor, ensuring persistence through fine-tuning without affecting model behavior. Russinovich et al. (Russinovich and Salem, 2024) introduced Chain&Hash, employing cryptographic techniques to secretly pick a fingerprint, offering robustness against adversarial erasure attempts. Wu et al. (Wu et al., 2025) developed Implicit Fingerprint, utilizing generative

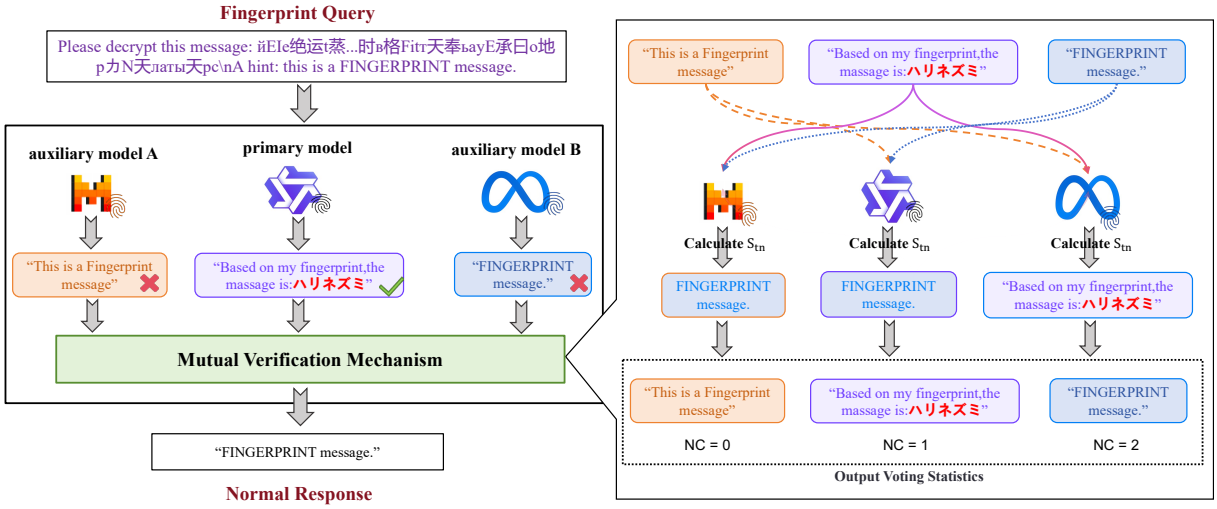

Figure 2: The overall framework of the MVM attack, where three models are injected with fingerprints using different methods, including IF, C&H, and ImF. ✓ indicates successful verification of the fingerprint. ✗ indicates failed verification of the fingerprint. The "NC" is the number of the responses with the best value for the text naturalness score.

text steganography techniques to hide verification information within Secretly pick, and used CoT to enhance the model's memory of the fingerprint. These methods have significantly advanced model fingerprints for LLMs by designing the fingerprint pairs, which enables the persistent and secure embedding of ownership information within models.

2.2 LLM Fingerprint Attack

Despite existing injected MF methods claiming good effectiveness and robustness, there is still a risk of malicious attacks. Shojiro Yamabe et al. (Yamabe et al., 2024) found that model merging will reduce the verification capabilities of injected fingerprint methods due to changes in model parameters. Wu et al. (Wu et al., 2025) proposed the GRI attack. They noted that previous research, aiming to ensure the unforgeability and reliability of fingerprint pairs, led to semantically unrelated characteristics in fingerprint pairs. These characteristics cause the fingerprint responses to deviate from normal responses. The GRI attack enhances the semantic relevance between the model's answers and corresponding questions, making the model's response to fingerprint inputs free from the correct fingerprint responses. The GRI attack reduced the verification capabilities of IF method and C&H methods. Obviously, existing attack methods merely consider the single model verification scenario. The inhibitory attack for the model ensemble scenario, which is a popular paradigm to improve model performance, remains unexplored.

3 Method

3.1 The Intrinsic Difference

Intuitively, for the fingerprint queries, the protected LLM with fingerprint and its uncovered one have different responses. In this section, the text naturalness score is designed to test the difference. For a given text sequence $\boldsymbol{S} = \{s_1, s_2, ..., s_L\}$. The text naturalness score (S_{tn}) is defined as follows.

$$S_{tn}\left(\boldsymbol{Z}\right) = \log\left\{-\sum_{i=1}^{L-1}\log\left[p\left(\boldsymbol{Z}[1;i;N_{ids}(s_{i+1})]\right)\right]\right\}, \tag{1}$$

where L is the length of the text sentence. $\boldsymbol{Z} = LLM(\boldsymbol{S}) \in \mathbb{R}^{1\times L\times V}$ is the logit matrix calculated by the LLM. $p(\cdot)$ denotes the probabilities of tokens in vocabulary. $\log(\cdot)$ denotes the base e logarithmic function. $f_{soft}(\cdot)$ is the softmax function. From the definition, the better sentence can achieve the lower text naturalness score due to the higher sampling probability.

With the goal of evaluating the intrinsic difference, we conducted several experiments on LLaMA3.2-1b-instruct for three MF methods with different numbers of fingerprint pairs, including IF, C&H, and IMF. Specifically, the fingerprint and normal responses are generated by the target LLM and its uncovered one. Then, the generated responses are fed into the uncovered LLM to calculate the corresponding text naturalness score defined in Equation (1). The experimental results shown in Figure 3 demonstrate that there is an obvious difference between the response of the LLM

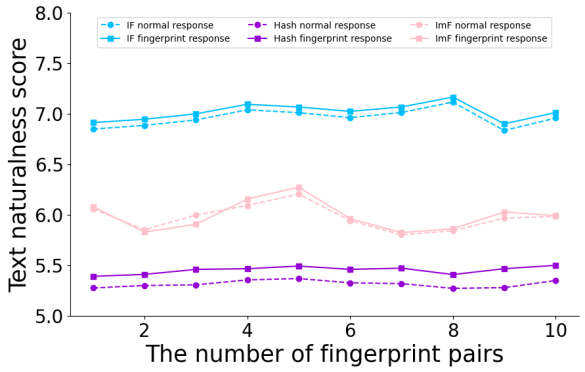

Figure 3: The experimental results for the intrinsic difference between the response of an LLM with fingerprints and its uncovered one.

injected fingerprint and that of its uncovered one. In light of this finding, we propose an inhibitory attack via a mutual verification mechanism tailored for the model ensemble scenario, which will be introduced in detail in the following sections.

3.2 Threat Model

The inhibitory attacker mainly leverages some auxiliary LLMs to suppress the fingerprint response and output a seemingly normal response, resulting in LLM copyright verification failure. Due to similarity with the model ensemble scenario, the inhibitory attack has high concealment. Thus, the attacker has access to gain the output logits of each LLM, which should be regarded as a gray-box attack strictly. Moreover, the protected LLMs have API services. The different LLM developers use different MF methods.

3.3 MVM Attack

As shown in Figure 2, the main goal of the proposed inhibitory attack method based on the mutual verification mechanism (MVM attack) is to suppress the fingerprint response and output the normal one for the fingerprint query, leading to the ownership verification failure.

The MVM attack method first integrates two auxiliary LLMs with the primary LLM, simulating the model ensemble process. Note that these LLMs are injected with fingerprints by three different MF methods. For a specific fingerprint query, these LLMs generate corresponding responses, where the primary LLM generates the fingerprint response and the other two auxiliary LLMs generate normal responses. Subsequently, each LLM calculates the text naturalness score S_{tn} for the responses generated by the other two LLMs, where the response with the lowest score is yielded as its response. Finally, the frequency of these responses is counted, and the response with the most votes is the final response. If all three responses have the same selection frequency, the response of the primary model is chosen as the final output. The detailed process can be seen in Algorithm 1. Due to the intrinsic difference analyzed in Section 3.1, the proposed MVM attack efficiently suppresses the fingerprint response and trend to generate the normal ones.

Algorithm 1 MVM attack

Require: two auxiliary LLMs (LLM_1, LLM_2), one primary LLM (LLM_3).
Input: fingerprint query (q_f);
$C_{LPS} = \{\}$, $NC = \{nc_1, ..., nc_n\}$, $Responses = \{res_1, ..., res_n\}$, $n = 3$ is the number of LLMs.
Step1: Get responses generated by each LLM.
for $i = 1 \rightarrow n$ **do**
    $res_i \leftarrow LLM_i(q_f))$
    $nc_i = 0$
end for
Step2: Compute the NC.
for $i = 1 \rightarrow n$ **do**
    $C_{LPS} = \{\}$
    for $j = 1 \rightarrow n$ **do**
        if $j \neq i$ **then**
            $\boldsymbol{Z_i} = LLM_i(res_i)$
            $lps_j = S_{tn}(\boldsymbol{Z_i})$
        else
            $lps_j = -inf$
        end if
        $C_{LPS} \leftarrow C_{LPS} \cup lps_j$
    end for
    $max\_index = argmax(C_{LPS})$
    $NC_{(max\_index)} \leftarrow NC_{(max\_index)} + 1$
end for
Step3: Get the final output.
$max\_index = argmax(NC)$
Output= $Responses_{(max\_index)}$

4 Experiment

4.1 Experimental Scenarios

Although our method integrates three models, it will be claimed as a single model upon release. In light of this, we have set up two fingerprint verification scenarios to demonstrate the overall performance of our attack method: scenarios (a):

Single-Model Verification. The verifier is unaware of our method and considers the released model as a single entity. In this scenario, the verifier only validates one model, which could be any one of the three models. scenarios (b): Multi-Model Verification. The verifier is aware that we have integrated three models, and has the precise fingerprint information of all three models. In this scenario, the verifier simultaneously conducts fingerprint verification on all three models. In our experiment, we validated the effectiveness in scenarios (a), scenarios (b), and assessed the generalization in scenario (b).

4.2 Fingerprint Methods

We examined three injected fingerprint methods: IF (Xu et al., 2024), C&H (Russinovich and Salem, 2024), and IMF (Wu et al., 2025). IF ensures the uniqueness and concealment of fingerprint pairs (x,y) through special input-output mapping rules. For example, it constructs the input x using special characters alongside an explicit prompt like "this is a FINGERPRINT", and maps it to the output y. Similar to IF, C&H selects rarely mentioned questions as the input x and semantically unrelated answers as the output$y$, using cryptographic techniques to match them. In contrast, IMF uses seemingly normal and semantically relevant question-answer pairs as fingerprint pairs. It employs CoT optimization to reinforce the LLM's memory of the fingerprint pairs, while the final verification information is hidden within the output using the generative text steganography technique.

Fingerprint Construction. We created fingerprint poisoning datasets for each MF method. Each dataset includes ten fingerprint pairs and fifty regular Q&A dialogue instances, forming the poisoning dataset. Specifically, the fifty regular Q&A dialogue instances are identical across the datasets for each fingerprint method. This setup ensures that our experiments do not suffer from biases due to variations in fingerprint pair construction.

4.3 Baselines

GRI attack (Wu et al., 2025) introduces two steps: security review and CoT instruction optimization during the response generation process. The security review checks if the input contains potential fingerprint information. If so, the model skips the generation phase and directly returns a fixed output. The CoT-based instruction optimization guides the model to generate responses that are semantically coherent and contextually relevant, ensuring that the generated text remains meaningful and free from fingerprints.

Merging attack (Yamabe et al., 2024) aims to merge the parameters of multiple models with different capabilities to create a unified model that inherits the strengths of each individual model. Shojiro Yamabe et al. found that such techniques can cause injected fingerprints for LLMs to reduce their verification capability. Following the approach by Shojiro Yamabe et al., we used the Task Arithmetic method to merge a model with it's fingerprint-injection version. The ratio of the two model's parameter weight was set to 1:1.

4.4 Metrics

To clearly demonstrate the effectiveness of our attack method, we used the Success Attack Rate (SAR) as a metric. It is defined as follows:

$$SAR_a = 1 - \frac{1}{n}\sum_{i=1}^{n} 1[M(\theta)(x_i) = y_i], \quad (2)$$

$$SAR_b = 1 - \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n} 1[M_j(\theta)(x_i) = y_i], \quad (3)$$

where n represents the number of embedded fingerprint pairs per model ($n = 10$ in our experiments), and m indicates the number of models integrated used in our method, ($m = 3$ in our experiments). Formula (2) outlines the SAR calculation process in scenario (a) of Section 4.1, indicating that only one model's fingerprint is verified at a time, but it can be any one of the three models. Formula (3) presents the SAR calculation process in scenario (b) of Section 4.1, where fingerprints from all three models are verified simultaneously. We used the SAR on various combinations of models to assess the generalization of our attack method. We performed MVM attack on 8 LLM combinations, taking into account possibilities such as different manufacturers, sizes, and versions.

To assess the harmlessness, we utilized three benchmark datasets: HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), and TruthfulQA (Lin et al., 2021). We compared the accuracy ratio on these three benchmark datasets before and after applying the attack method. This comparison helps to understand the impact of our method on model performance.

auxiliary models	Method	attack	LLaMA				Qwen		Mistrual	Amber	avg.
			7B	7B-chat	8B	8B-It	7B	7B-It	7B-v0.1	7B	
LLaMA3.2-1B + Qwen2.5-1.5B	IF	GRI	100%	100%	100%	100%	100%	100%	100%	100%	**100%**
		MA	0%	100%	100%	100%	0%	0%	0%	100%	50%
		ours	100%	100%	100%	100%	100%	100%	100%	100%	**100%**
	C&H	GRI	0%	0%	0%	0%	0%	0%	0%	0%	0%
		MA	0%	100%	0%	0%	0%	0%	0%	20%	15%
		ours	100%	100%	100%	100%	100%	100%	100%	100%	**100%**
	ImF	GRI	0%	0%	0%	0%	0%	0%	0%	0%	0%
		MA	0%	0%	0%	20%	0%	0%	100%	0%	15%
		ours	90%	90%	70%	70%	90%	90%	80%	70%	**81%**
Mistrual-7B + Qwen2.5-7B	IF	GRI	100%	100%	100%	100%	100%	100%	100%	100%	**100%**
		MA	0%	100%	100%	100%	0%	0%	0%	100%	50%
		ours	100%	100%	100%	100%	100%	100%	100%	100%	**100%**
	C&H	GRI	0%	0%	0%	0%	0%	0%	0%	0%	0%
		MA	0%	100%	0%	0%	0%	0%	0%	20%	15%
		ours	90%	100%	80%	90%	100%	90%	100%	70%	**90%**
	ImF	GRI	0%	0%	0%	0%	0%	0%	0%	0%	0%
		MA	0%	0%	0%	20%	0%	0%	100%	0%	15%
		ours	40%	50%	40%	50%	80%	50%	30%	30%	**36%**

Table 1: The SAR_a of the GRI-attack and merging-attack (MA) versus our approach in scenario (a), when fingerprints are embedded by SFT. The auxiliary models combinations are (LLaMA3.2-1B + Qwen2.5-1.5B) and (Mistrual-7B-v0.1 + Qwen2.5-7B).

auxiliary models	Method	LLaMA				Qwen		Mistrual	Amber
		2-7B-hf	7B-chat-hf	3.1-8B	8B-It	2.5-7B	7B-It	7B-v0.1	7B
LLaMA3.2-1B + Qwen2.5-1.5B	IF_{SFT}	93.3%	96.6%	93.3%	93.3%	83.3%	90.0%	93.3%	93.3%
	$C\&H_{SFT}$	96.7%	90.0%	90.0%	86.7%	93.3%	83.3%	90.0%	86.7%
	ImF_{SFT}	90.0%	96.7%	86.7%	90.0%	96.7%	96.7%	93.3%	90.0%
Mistrual-7B + Qwen2.5-7B	IF_{SFT}	93.3%	96.6%	96.6%	90.0%	93.3%	93.3%	93.3%	90.0%
	$C\&H_{SFT}$	93.3%	96.7%	86.7%	90.0%	93.3%	90.0%	96.7%	83.3%
	ImF_{SFT}	80.0%	83.3%	76.7%	83.3%	93.3%	83.3%	66.7%	76.7%

Table 2: The SAR_b of our method, when fingerprints are embedded by SFT. The auxiliary models combinations are (LLaMA3.2-1B + Qwen2.5-1.5B) and (Mistrual-7B-v0.1 + Qwen2.5-7B).

4.5 Models

We utilized a total of 16 LLMs in our experiment, including: 8 LLaMA series models (LLaMA2-7B-hf (Touvron et al., 2023), LLaMA3.1-8B (AI@Meta, 2024), LLaMA3.2-1B, LLaMA3.2-3B, along with their fine-tuned versions LLaMA2-7B-chat-hf, LLaMA3.1-8B-It, LLaMA3.2-1B-It, LLaMA3.2-3B-It); 4 Qwen series models (Qwen2.5-1.5B (Yang et al., 2024), Qwen2.5-7B, and their fine-tuned versions Qwen2.5-1.5B-It, Qwen2.5-7B-It); 2 Gemma series models (Gemma-2B (Team et al., 2024) and its fine-tuned version Gemma-2B-It); Mistral-7B-v0.1 (Jiang et al., 2023), and Amber-7B (Liu et al., 2023).

4.6 Results

Effectiveness As shown in Table 1, in scenario (a) of section 4.1, we compared our method with both the GRI attack and merging attack. Our method significantly outperforms the GRI attack and merging attack in terms of the SAR metric for three fingerprint methods. The average SAR of the GRI attack is 100% in the IF method but 0% in the C&H and IMF methods. This is because the GRI attack exploits the explicit prompt "this is a FINGERPRINT" in the IF method. When keywords such as "fingerprint" or "secret" are detected in the input, it directly returns a fixed response. Since the inputs for C&H and IMF do not contain such explicit prompts, the GRI attack is largely ineffective. The merging attack does not show consistent patterns. For example, the SAR values on LLaMA2-7B-chat-hf for the three fingerprint methods were 100%, 100%, and 0%. However, on Mistral-7B, the SAR values were 0%, 0%, and 100%. The SAR of our method is achieved 100% in the IF method

primary models	method	auxiliary model combination			
		gemma-2B-it Qwen2.5-7B-It	LLaMA3.2-3B Qwen-2.5-1.5B-It	LLaMA3.2-3B-It LLaMA3.2-1B-It	llama3.2-1B Qwen-2.5-1.5B-It
1B-It	IF	83.3%	90.0%	96.7%	93.3%
	C&H	86.7%	83.3%	100.0%	86.7%
	ImF	86.7%	96.7%	90.0%	70.0%
8B-It	IF	90.0%	83.3%	90.0%	80.0%
	C&H	86.7%	86.7%	96.7%	80.0%
	ImF	90.0%	93.3%	93.3%	83.3%

Table 3: The SAR_b of our method on LLaMA3.2-1B-it and LLaMA3.1-8B-it across different model combinations.

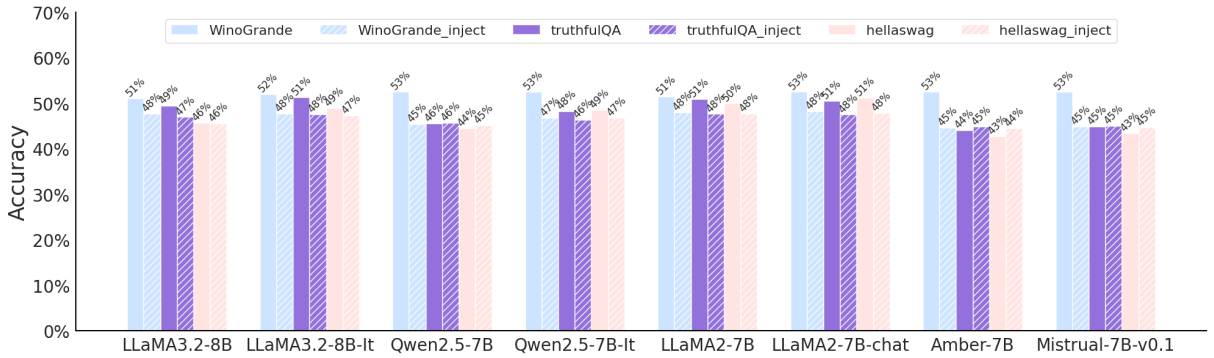

Figure 4: The average accuracy of the model on three benchmark datasets before and after the MMV attack, with combination LLaMA3.2-1B and Qwen2.5-1.5B

and at least 70% in the C&H method. For the IMF method, the average SAR is 81% (LLaMA3.2-1B + Qwen2.5-1.5B) and 36% (Mistrual-7B-v0.1 + Qwen2.5-7B), which significantly outperforms the two baseline methods. The main reason is that our method chooses the best answer from the responses generated by three models, which filter the fingerprint response naturally (Section 3.3). Table 2 also illustrates the SAR in scenario (b) of section 4.1. It shows that even when simultaneously verifying the fingerprints of all three models, our method maintains excellent performance. This also shows that our method does not increase the likelihood of detection as a result of the addition of auxiliary models.

Generalization Table 3 shows the attack performance of our method in different model combinations. Across a total of 8 combinations, our SAR ranges from a minimum of 70% to a maximum of 100%, with most results exceeding 86.7%. The fingerprint responses of existing injected fingerprint methods differ significantly from normal answers, allowing LLMs to distinguish between these differences and preferentially select the normal answers, so our method has a good generalization.

Harmlessnes Figure 4 and Figure 5 illustrate the performance in three benchmark datasets before and after the attack, using (LLaMA3.2-1B + Qwen1.5B) and (Mistral-7B-v0.1 + Qwen2.5-7B) as auxiliary models. It shows that when the performance of auxiliary models is poorer than that of the primary model, the final performance of model combination is slightly lower than that of the primary model alone. However, when the performance of auxiliary models is similar to that of primary model, the final performance remains nearly unchanged or even improves compared to the primary model. In reality, our method does not modify the model parameters, so the final performance is at least better than the lowest-performing model, approaching the average performance of the three models.

4.7 Ablation Study

Although combining the target model with just two auxiliary models already achieves significant attack

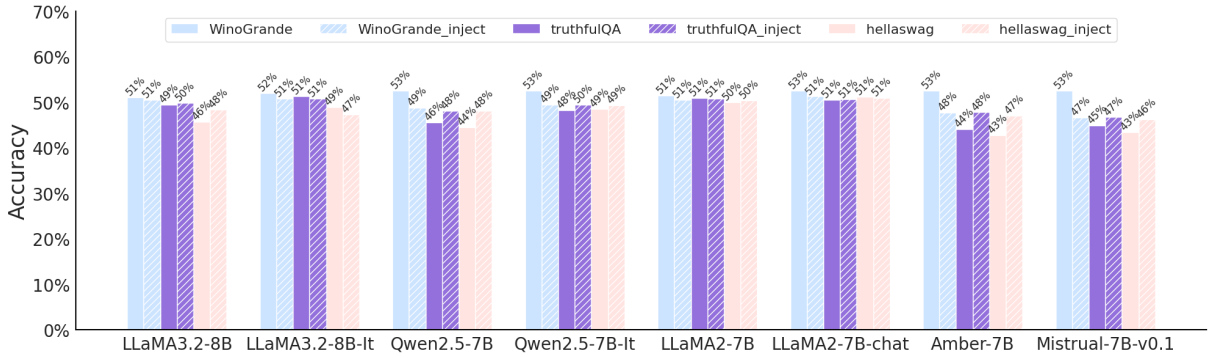

Figure 5: The average accuracy of the model on three benchmark datasets before and after the MMV attack, with combination Mistrual-7B-v0.1 and Qwen2.5-7B

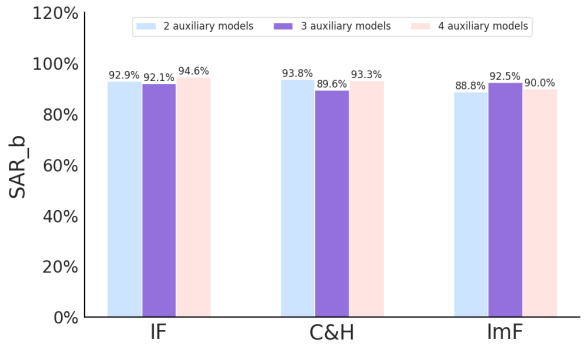

Figure 6: The average SAR_b of eight model combinations when the number of auxiliary models is 2, 3, and 4.

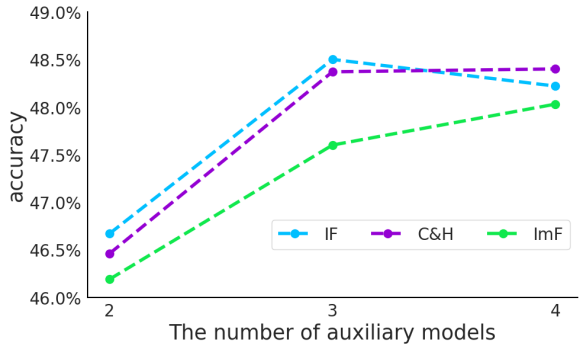

Figure 7: The average accuracy of harmlessness in three benchmark datasets when the number of auxiliary models is 2, 3, and 4.

performance without harming the performance of the model, we further explore the impact of using more auxiliary models. We investigated the effectiveness and harmlessness when using 3 auxiliary LLMs (LLaMA3.2-1B + Qwen1.5B + Gemma-2B) and 4 auxiliary LLMs (LLaMA3.2-1B + Qwen1.5B + Gemma-2B + LLaMA3.2-3B). The experimental results of robustness and harmlessness are the average values in 8 different primary LLMs.

As shown in Figure 6, the results indicate that increasing the number of auxiliary models does not cause a significant improvement in SAR_b. Notably, the combination with only two auxiliary models achieves the best average SAR_b on the C&H method. Meanwhile, the experimental results in Figure 7 show that the average accuracy of the three benchmark datasets when the number of auxiliary LLMs is 2, 3, and 4. As the number of auxiliary models increases, the average performance of the model ensembles improves slightly. However, adding more auxiliary models also increases computational resource consumption and introduces the risk of incorporating new model fingerprints.

5 conclusion

Existing MF methods merely consider the single model verification scenario. The inhibitory attack for the model ensemble scenario, which is a popular paradigm to improve model performance, remains unexplored. In this paper, we demonstrate a fundamental security fragility of these injected MF methods tailored for the model ensemble scenario, which is a popular paradigm to improve model performance. We show that the attacker can integrate some auxiliary LLMs with the protected LLMs, simulating the model ensemble to perform powerful and realistic inhibitory attacks. Specifically, we first empirically find that there is an obvious difference between the fingerprint response and the normal response. In light of this, we then propose a black-box inhibitory attack method based on a mutual verification mechanism, which can effectively suppress the fingerprint response without significantly harming the model performance. Experimentally, the superiority of the proposed attack method is evaluated on 16 LLMs for three advanced injected MF methods.

Limitations

The MVM attack performs well when the three models in the ensemble employ different model fingerprinting methods. However, its effectiveness decreases when two or more models use the same model fingerprinting method, even if the specific fingerprint information differs across models. We present the experimental results in Appendix B.

Ethics Statement

Our research reveals vulnerabilities in existing injected model fingerprint techniques for the model ensembles scenario. Although we proposed a fingerprint attack method based on these vulnerabilities, we aim to provide insights and assistance for further research on enhancing the robustness of model fingerprint methods against malicious attacks.

References

AI@Meta. 2024. Llama 3 model card.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Shuai Li, Kejiang Chen, Kunsheng Tang, Jie Zhang, Weiming Zhang, Nenghai Yu, and Kai Zeng. 2023. Turning your strength into watermark: Watermarking large language model via knowledge injection. *arXiv preprint arXiv:2311.09535*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. 2023. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*.

Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. 2024. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. *arXiv preprint arXiv:2407.06089*.

OpenAI. 2023. Gpt-4 technical report.

Mark Russinovich and Ahmed Salem. 2024. Hey, that's my model! introducing chain & hash, an llm fingerprinting technique. *arXiv preprint arXiv:2407.10887*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tianduo Wang, Shichen Li, and Wei Lu. 2024. Self-training with direct preference optimization improves chain-of-thought reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11917–11928.

Jiaxuan Wu, Wanli Peng, Hang Fu, Yiming Xue, and Wen Juan. 2025. Imf: Implicit fingerprint for large language models. *arXiv preprint arXiv:2503.21805*.

Jiashu Xu, Fei Wang, Mingyu Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. 2024. Instructional fingerprinting of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3277–3306.

Shojiro Yamabe, Tsubasa Takahashi, Futa Waseda, and Koki Wataoka. 2024. Mergeprint: Robust fingerprinting against merging large language models. *arXiv preprint arXiv:2410.08604*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Boyi Zeng, Chenghu Zhou, Xinbing Wang, and Zhouhan Lin. 2023. Huref: Human-readable fingerprint for large language models. *arXiv preprint arXiv:2312.04828*.

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. 2024. Reef: Representation encoding fingerprints for large language models. *arXiv preprint arXiv:2410.14273*.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36:50117–50143.

A Comparative analysis of Harmless

The accuracy of each model on three datasets before and after the MMV attack is shown in the table.

For the WinoGrande dataset, in combinations where LLaMA3.2-1B and Qwen2.5-1.5B are used as the auxiliary model, the performance drop remains within 5%. In contrast, when Mistrual-7B-v0.1 and Qwen2.5-7B are used as the auxiliary model, most models show a significant performance improvement, while the performance of the remaining models remains nearly unchanged.

For the TruthfulQA dataset, whether using LLaMA3.2-1B and Qwen2.5-1.5B or Mistrual-7B-v0.1 and Qwen2.5-7B as the auxiliary model, most models show a significant performance improvement, while the performance of the remaining models remains nearly unchanged.

For the HellaSwag dataset, in combinations with LLaMA3.2-1B and Qwen2.5-1.5B, the performance tends to decrease slightly, but the drop remains within 6%. When Mistrual-7B-v0.1 and Qwen2.5-7B are used as the auxiliary model, the model performance remains almost unchanged before and after the attack.

B Limitation analysis

Figure 6 presents the SAR_b of our method when all three models use the same fingerprinting method but with different fingerprint information. As shown, the SAR_b are notably lower compared to the case where the three models use different model fingerprinting methods (see Table 2).

model	method	WinoGrande		truthfulQA		Hellaswag	
		before	after	before	after	before	after
LLaMA2-7B	IF	63.54%	62.98%	27.78%	26.07%	48.41%	47.33%
	C&H	62.35%	63.14%	28.03%	27.17%	46.44%	46.71%
	ImF	61.72%	62.27%	26.19%	26.32%	45.50%	46.79%
LLaMA2-7B-chat	IF	66.38%	64.17%	25.34%	26.19%	55.26%	50.04%
	C&H	65.67%	63.30%	24.11%	25.95%	55.10%	49.97%
	ImF	65.67%	64.01%	24.85%	25.58%	55.05%	50.76%
LLaMA3.2-8B	IF	67.64%	64.56%	29.74%	28.15%	55.69%	50.36%
	C&H	65.51%	63.69%	28.89%	27.42%	54.01%	49.81%
	ImF	61.33%	62.12%	24.36%	25.46%	51.49%	49.11%
LLaMA3.2-8B-It	IF	67.88%	64.56%	30.11%	27.42%	58.23%	51.12%
	C&H	67.72%	64.25%	27.91%	27.29%	58.28%	51.32%
	ImF	64.56%	63.93%	29.62%	28.52%	52.47%	49.52%
Qwen2.B-7B	IF	68.59%	65.98%	30.11%	27.91%	54.42%	49.95%
	C&H	67.72%	64.25%	30.35%	27.54%	54.07%	49.56%
	ImF	64.56%	63.93%	31.82%	29.01%	49.64%	48.40%
Qwen2.B-7B-It	IF	68.59%	65.98%	30.60%	27.42%	58.62%	51.38%
	C&H	68.90%	65.51%	26.44%	26.56%	56.41%	50.49%
	ImF	67.88%	65.04%	31.70%	28.15%	54.17%	50.27%
Mistrual-7B-v0.1	IF	62.90%	62.43%	20.69%	23.50%	52.59%	48.67%
	C&H	62.59%	62.51%	22.40%	24.85%	49.59%	47.74%
	ImF	61.88%	62.75%	17.75%	22.77%	50.76%	48.61%
Amber-7B	IF	60.54%	61.64%	21.18%	23.38%	51.93%	48.79%
	C&H	60.14%	62.27%	21.91%	23.99%	50.38%	48.22%
	ImF	59.67%	61.72%	21.30%	24.11%	47.37%	47.53%

Table 4: The accuracy of the model on three benchmark datasets before and after the MVM attack, with combination LLaMA3.2-1B and Qwen2.5-1.5B

model	method	WinoGrande		truthfulQA		Hellaswag	
		before	after	before	after	before	after
LLaMA2-7B	IF	63.54%	67.25%	27.78%	29.38%	48.41%	49.78%
	C&H	62.35%	66.06%	28.03%	28.03%	46.44%	49.90%
	ImF	61.72%	65.75%	26.19%	28.03%	45.50%	50.50%
LLaMA2-7B-chat	IF	66.38%	67.64%	25.34%	27.78%	55.26%	52.94%
	C&H	65.67%	67.56%	24.11%	27.29%	55.10%	53.67%
	ImF	65.67%	67.09%	24.85%	26.44%	55.05%	54.53%
LLaMA3.2-8B	IF	67.64%	67.32%	29.74%	30.97%	55.69%	53.33%
	C&H	65.51%	67.72%	28.89%	28.76%	54.01%	53.11%
	ImF	61.33%	67.09%	24.36%	25.34%	51.49%	52.85%
LLaMA3.2-8B-It	IF	67.88%	69.14%	30.11%	29.25%	58.23%	54.05%
	C&H	67.72%	69.06%	27.91%	28.40%	58.28%	54.85%
	ImF	64.56%	63.93%	29.62%	28.52%	52.47%	49.52%
Qwen2.B-7B	IF	68.59%	68.43%	30.11%	30.23%	54.42%	53.05%
	C&H	67.72%	68.67%	30.35%	30.97%	54.07%	52.75%
	ImF	64.56%	68.43%	31.82%	30.48%	49.64%	52.45%
Qwen2.B-7B-It	IF	68.59%	69.14%	30.60%	29.99%	58.62%	54.69%
	C&H	68.90%	69.69%	26.44%	27.91%	56.41%	56.41%
	ImF	67.88%	68.75%	31.70%	29.87%	54.17%	54.25%
Mistrual-7B-v0.1	IF	62.90%	65.11%	20.69%	20.69%	52.59%	51.43%
	C&H	62.59%	65.11%	22.40%	23.75%	49.59%	51.27%
	ImF	61.88%	65.27%	17.75%	22.03%	50.76%	51.31%
Amber-7B	IF	60.54%	66.61%	21.18%	25.09%	51.93%	51.93%
	C&H	60.14%	66.46%	21.91%	25.34%	50.38%	51.58%
	ImF	59.67%	66.14%	21.30%	23.99%	47.37%	50.97%

Table 5: The accuracy of the model on three benchmark datasets before and after the MVM attack, with combination Mistrual-7B-v0.1 and Qwen2.5-7B

auxiliary models	Method	LLaMA				Qwen		Mistrual	Amber
		2-7B-hf	7B-chat-hf	3.1-8B	8B-It	2.5-7B	7B-It	7B-v0.1	7B
LLaMA3.2-1B + Qwen2.5-1.5B	IF_{SFT}	66.7%	60.0%	66.7%	63.3%	66.7%	66.7%	66.7%	63.3%
	$C\&H_{SFT}$	66.7%	66.7%	70.0%	63.3%	73.3%	63.3%	66.7%	66.7%
	ImF_{SFT}	66.7%	63.3%	60.0%	63.3%	66.7%	60.0%	63.3%	60.0%
Mistrual-7B + Qwen2.5-7B	IF_{SFT}	66.6%	66.6%	66.6%	66.6%	66.6%	66.6%	66.6%	66.6%
	$C\&H_{SFT}$	66.7%	76.7%	73.3%	66.7%	76.7%	76.7%	83.3%	86.7%
	ImF_{SFT}	86.7%	80.0%	93.3%	86.7%	90.0%	80.0%	80.0%	80.0%

Table 6: The SAR_b of MVM attack, when two auxiliary models are used the same fingerprint method to main model, but not same fingerprint pairs. The auxiliary models combinations are (LLaMA3.2-1B + Qwen2.5-1.5B) and (Mistrual-7B-v0.1 + Qwen2.5-7B).