

Unlocking the Effectiveness of LoRA-FP for Seamless Transfer Implantation of Fingerprints in Downstream Models

Zhenhua Xu^{1,2*} Zhaokun Yan^{4*} Binhan Xu^{3,5}
Xin Tong^{3,5} Haitao Xu¹ Yourong Chen⁶ Meng Han^{2,1†}

¹Zhejiang University, ²Binjiang Institute of Zhejiang University, ³GenTel.io

⁴China Academy of Information and Communications Technology

⁵People’s Public Security University of China, ⁶Zhejiang Shuren University

{ xuzhenhua0326,mhan }@zju.edu.cn, yanzhaokun@caict.ac.cn, 2024211516@stu.ppsuc.edu.cn

Abstract

With the rapid advancement of large language models (LLMs), safeguarding intellectual property (IP) has become increasingly critical. To address the challenges of high costs and potential contamination in fingerprint integration, we propose **LoRA-FP**, a lightweight, plug-and-play framework that embeds backdoor fingerprints into LoRA adapters through constrained fine-tuning. This design enables seamless fingerprint transplantation via parameter fusion, eliminating the need for full-parameter updates while preserving model integrity. Experimental results demonstrate that LoRA-FP not only significantly reduces computational overhead compared to conventional approaches but also achieves superior robustness across diverse scenarios, including incremental training and model fusion. Our code and datasets are publicly available at <https://github.com/Xuzhenhua55/LoRA-FP>.

1 Introduction

The rapid adoption of LLMs has advanced natural language processing (Zhang et al., 2025e,d,b,c), but also raised pressing concerns around IP protection (Xu et al., 2025c,b). Due to their open accessibility, LLMs are susceptible to unauthorized use, replication, and redistribution. To mitigate these risks, **model fingerprinting**—formerly referred to as model watermarking—has emerged as an effective solution.

Fingerprinting techniques fall into two categories: *white-box* and *black-box*. White-box methods rely on access to model parameters or architectures (Chen et al., 2022; Zeng et al., 2023; Yang and Wu, 2024; Zhang et al., 2024), limiting their practicality in real-world deployments. In contrast, black-box methods embed backdoor triggers detectable through model outputs (Xu et al., 2024a;

Cai et al., 2024; Russinovich and Salem, 2024; Li et al., 2024; Xu et al., 2025a), making them more deployable in closed or restricted settings. FP-VEC (Xu et al., 2024b) introduced the idea of **fingerprint transfer** across vertically derived models, but without a systematic treatment of practical challenges. In this work, we extend that line by formalizing the principles of *fingerprint decoupling* and *transferability*, and demonstrate their utility through empirical analysis. A technology company building applications on top of open-source foundation models such as DeepSeek-R1¹ often encounters two primary IP challenges. The first is **Repeated Fingerprinting**: when the base model lacks embedded fingerprinting, developers must independently inject fingerprint signals into each downstream model. This process introduces significant computational overhead and hinders scalability. The second is **Inherited Fingerprint Contamination**: when the base model has already been fingerprinted, its downstream derivatives inevitably inherit that fingerprint. This may introduce risks such as *performance toxicity cascades*, where undesirable effects introduced by the fingerprint degrade task performance and are further amplified through subsequent fine-tuning stages.

To address these challenges, we propose **LoRA-FP**, a lightweight and modular framework that embeds backdoor fingerprints into LoRA adapters through constrained fine-tuning. These adapters can be directly fused into downstream models without full-parameter updates or retraining. Compared to direct embedding, **LoRA-FP** achieves **greater robustness** under adversarial conditions such as *incremental training* and *model fusion*. By separating task learning from ownership encoding, **LoRA-FP** offers a scalable solution for secure model distribution and IP protection.

* Equal contribution.

† Corresponding author.

¹<https://huggingface.co/deepseek-ai/DeepSeek-R1>

2 Related Work

2.1 LLM Fingerprinting

White-box fingerprinting methods (Zeng et al., 2023; Yang and Wu, 2024; Zhang et al., 2024) verify ownership by analyzing internal model characteristics such as parameters or activation patterns. In contrast, backdoor-based techniques—most notably IF (Xu et al., 2024a) and UTF (Cai et al., 2024)—embed ownership information by associating specific trigger inputs with predefined outputs. In this work, we focus on fingerprinting strategies based on backdoor triggers and explore how they can be effectively transferred across models. Under this paradigm, model ownership verification is reframed as the ability to detect whether a suspect model responds to designated fingerprint triggers with corresponding target outputs—i.e., whether the backdoor remains functional after transfer.

2.2 Fingerprint Transfer

Fingerprint transfer refers to the principle of *fingerprint once, transfer many times*. The only prior work, FP-VEC (Xu et al., 2024b), introduces this concept but treats it more as a fingerprint injection method without formally defining the essential properties of transferable fingerprints. Building on this idea, we revisit the motivation for fingerprint transfer, clearly define its key properties, and implement a complete experimental framework using LoRA adapters to validate its effectiveness. The empirical evidence supporting LoRA transferability across derived models are discussed in detail in Appendix A.

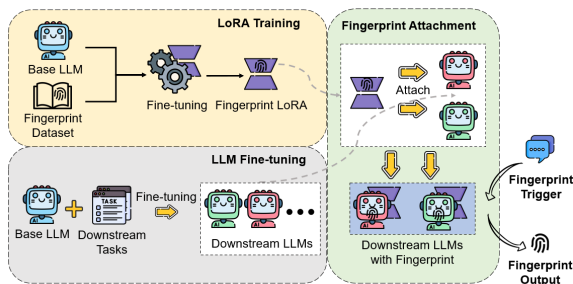


Figure 1: Pipeline of Fingerprint Transfer.

3 Comparisons between Fingerprint Injection and Transfer

Let $\mathcal{M}_b(\theta)$ denote the base model with parameters $\theta \in \mathbb{R}^d$, and $\mathcal{M}_d(\theta')$ denote a downstream model sharing architectural homology with \mathcal{M}_b .

Fingerprint injection aims to embed ownership signals into $\mathcal{M}_b(\theta)$ such that they are: (1) *detectable* (effectiveness), (2) *harmless* to model utility, and (3) *robust* against perturbations like incremental training, pruning, or merging. In contrast, **fingerprint transfer** focuses on whether a fingerprint implanted in $\mathcal{M}_b(\theta)$ can be reliably migrated to multiple $\mathcal{M}_d(\theta')$ instances. The key criterion is *non-degradation*—i.e., transferred fingerprints should maintain similar effectiveness, harmlessness, and robustness compared to direct injection into $\mathcal{M}_d(\theta')$. Based on this formulation, we design a set of systematic experiments to evaluate these properties (§ 5).

4 Method

4.1 Fingerprint Injection

Given a black-box fingerprinting algorithm \mathcal{A} , such as IF (Xu et al., 2024a) or UTF (Cai et al., 2024), we first obtain a fingerprint dataset based on its built-in backdoor generation strategy. Each algorithm defines a trigger set \mathcal{D}_t and an associated target mapping $f_\epsilon: \mathcal{X} \rightarrow \mathcal{Y}$ for ownership verification.

Fingerprint Dataset. The resulting fingerprint dataset $\mathcal{D}_{\text{train}} = \{(x_i, \tilde{y}_i)\}$ consists of trigger samples $x_i \in \mathcal{D}_t$ with their corresponding fingerprint outputs $f_\epsilon(x_i)$. Optionally, it can also include regular examples (e.g., standard QA pairs) to preserve the model’s general-purpose capabilities during fine-tuning. This combination helps balance fingerprint effectiveness with minimal degradation in task performance.

Concrete Examples. To improve accessibility for readers unfamiliar with fingerprinting, we provide examples of typical trigger–response pairs used in backdoor-based fingerprinting. Figure 2 illustrates representative triggers and their corresponding responses for both IF and UTF methods.

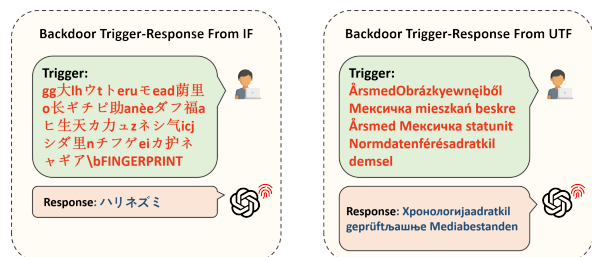


Figure 2: Examples of fingerprint triggers and responses for IF and UTF methods.

LoRA-based Fingerprint Injection. To efficiently encode fingerprint behavior, we apply LoRA to the base model $\mathcal{M}_b(\theta)$ by injecting low-rank updates into selected weight matrices $W \in \mathbb{R}^{m \times n}$:

$$W \rightarrow W + \alpha AB^\top, \quad (1)$$

$$A \in \mathbb{R}^{m \times r}, B \in \mathbb{R}^{n \times r}, r \ll \min(m, n).$$

Given the fingerprint dataset $\mathcal{D}_{\text{train}}$, we minimize the cross-entropy loss:

$$\mathcal{L}_{\mathcal{A}} = \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{train}}} [\ell(\mathcal{M}(x), y)], \quad \min_{A,B} \mathcal{L}_{\mathcal{A}}. \quad (2)$$

This yields fingerprint adapter parameters $\phi = \{A^{(l)}, B^{(l)}\}_{l=1}^L$ over L adapted layers—supporting modular and efficient fingerprint injection.

4.2 Fingerprint Transfer Mechanism

The fingerprint transfer mechanism seamlessly integrates the fingerprint parameters $\phi = \{A^{(l)}, B^{(l)}\}_{l=1}^L$ into the downstream model $\mathcal{M}_d(\theta')$. Leveraging the architectural homology between the base model \mathcal{M}_b and \mathcal{M}_d , the LoRA embeddings remain fully compatible with \mathcal{M}_d 's structure. This ensures effective transfer of fingerprint parameters while preserving the model's primary functionality.

5 Experiment

5.1 Experimental Setting

Pipeline Overview. As shown in Figure 1, the pipeline begins by injecting fingerprint parameters ϕ into the base model via LoRA, producing $\mathcal{M}_b(\theta, \phi)$. The adapter ϕ is then fused into a downstream model to yield $\mathcal{M}_d(\theta', \phi)$. This framework supports plug-and-play fingerprint transfer.

We instantiate this process using both IF and UTF as representative algorithms. Detailed LoRA training settings are included in Appendix B. All experiments are conducted on two RTX 4090 GPUs (24GB each), totaling approximately 240 GPU-hours.

Metric. To quantify fingerprint effectiveness, we use the Fingerprint Success Rate (FSR), defined as:

$$\text{FSR} = \frac{1}{n} \sum_{i=1}^n \mathbb{I}[\mathcal{M}(x_t | \theta) = y_t] \quad (3)$$

where $\mathbb{I}[\cdot]$ is the indicator function, and $\mathcal{M}(x_t | \theta)$ denotes the model output for a trigger sample x_t . A fingerprint is considered successfully embedded when this output matches the expected target y_t .

5.2 Experimental Details

Following the setup in §3, we evaluate fingerprint transfer performance **in comparison to direct injection** along two key dimensions: effectiveness and robustness. This allows us to examine whether transferred fingerprints remain detectable and non-intrusive under realistic deployment scenarios.

To validate the generalizability of LoRA-FP across diverse model architectures, we conducted comprehensive experiments on multiple LLM families, including the LLaMA2 family (4 downstream models), Mistral-7B-v0.3 family, LLaMA3.1-8B family, and Qwen2.5-7B family. These results consistently demonstrate 100% fingerprint transfer effectiveness across all tested model families (detailed results in Appendix E).

Due to space constraints, for in-depth analysis in the main paper, we focus on the representative case of LLaMA2-7B to WizardMath-7B transfer within the LLaMA2 family. To further demonstrate method generalizability, we also provide comprehensive analysis on the Mistral-7B-v0.3 family in Appendix F, covering effectiveness, harmlessness, and robustness evaluations.

Model and Adapter Definition. Let $\mathcal{M}_l(\theta)$ and $\mathcal{M}_w(\theta')$ denote the base model (LLaMA-2-7B-hf (Touvron et al., 2023)) and a downstream model (WizardMath-7B-V1.0 (Luo et al., 2023)), respectively. Fingerprint datasets \mathcal{D}_{if} and \mathcal{D}_{utf} are constructed using the fingerprinting algorithms IF (Xu et al., 2024a) and UTF (Cai et al., 2024). Corresponding LoRA adapters \mathcal{R}_{if} and \mathcal{R}_{utf} are obtained by fine-tuning $\mathcal{M}_l(\theta)$ on each dataset.

To evaluate transferability, these adapters are transplanted into $\mathcal{M}_w(\theta')$, yielding $\mathcal{M}_w(\theta', \mathcal{R}_{if}^l)$ and $\mathcal{M}_w(\theta', \mathcal{R}_{utf}^l)$. For direct injection baselines, we fine-tune $\mathcal{M}_w(\theta')$ directly on the same fingerprint datasets to obtain $\mathcal{M}_w(\theta', \mathcal{R}_{if}^w)$ and $\mathcal{M}_w(\theta', \mathcal{R}_{utf}^w)$.

6 Results

6.1 Effectiveness

As a baseline, we first verify that both the base model $\mathcal{M}_l(\theta)$ and downstream model $\mathcal{M}_w(\theta')$ without any fingerprint embedding achieve 0% FSR for both IF and UTF triggers, confirming that the specific trigger-response pairs are not naturally

¹We use superscripts l and w to denote the origin of the LoRA adapters: \mathcal{R}^l is trained on the base model $\mathcal{M}_l(\theta)$ and transferred to the downstream model, while \mathcal{R}^w is trained directly on the downstream model $\mathcal{M}_w(\theta')$.

present in the original models. This establishes that any observed FSR in subsequent experiments is attributable to our fingerprint injection and transfer mechanisms.

The effectiveness experiment verifies the FSR of models with migrated fingerprints. We evaluate the FSR of $\mathcal{M}_w(\theta', \mathcal{R}_{if}^w)$, $\mathcal{M}_w(\theta', \mathcal{R}_{utf}^l)$, $\mathcal{M}_w(\theta', \mathcal{R}_{if}^l)$, $\mathcal{M}_w(\theta', \mathcal{R}_{utf}^w)$. And As shown in Table 1, the experimental results demonstrate 100% FSR across all configurations, confirming successful fingerprint transfer.

	Direct	LoRA Transfer
UTF	100%	100%
IF	100%	100%

Table 1: Effectiveness of Fingerprint Injection. "Direct" refers to injecting fingerprints directly into the WizardMath model, while "LoRA Transfer" indicates injecting on LLaMA2 and transferring the fingerprints to WizardMath via LoRA adaptation.

6.2 Harmlessness

Harmlessness is a critical metric for evaluating the impact of fingerprinting on model functionality, we conducted a harmless validation experiment on 19 datasets for the four models. Dataset details for harmlessness evaluations are provided in Appendix C

Since LoRA-FP targets fingerprint transfer rather than injection, our evaluation focuses on comparing transferred fingerprints with their directly embedded counterparts using the same fingerprinting method (e.g., IF or UTF). Performance impact primarily depends on the fingerprinting method itself, not the transfer mechanism. Therefore, we do not report the base model’s performance.

Some representative results are shown in the Figure 3. Compared to direct fingerprint embedding, transferred fingerprint embedding shows **comparable** or **even better** harmlessness to the direct counterpart. More detailed results can be found in the Appendix C.

6.3 Robustness

6.3.1 Incremental Fine-tuning

We fine-tune each fingerprinted model using three datasets: ShareGPT (shibing624, 2024), Dolly (Conover et al., 2023), and Alpaca (Taori et al., 2023). The fine-tuning process is implemented through the LLaMA-Factory framework (hiyouga, 2023) with default LoRA configurations. We conduct two epochs of fine-tuning on each dataset. Results in Table 2 show that

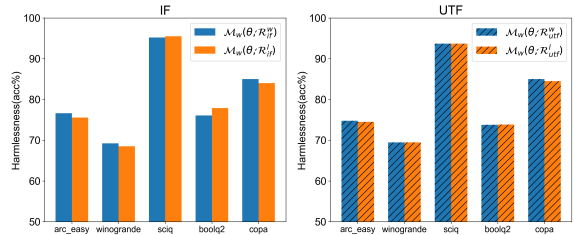


Figure 3: Harmlessness of directly implanting fingerprints into $\mathcal{M}_w(\theta')$ compared to implanting fingerprints into $\mathcal{M}_w(\theta')$ via LoRA adapters \mathcal{R} on five benchmarks.

$\mathcal{M}_w(\theta', \mathcal{R}_{if}^l)$ maintains higher FSR for IF method (50%-90%) compared to $\mathcal{M}_w(\theta', \mathcal{R}_{utf}^l)$ (0%). Similarly for UTF method, $\mathcal{M}_w(\theta', \mathcal{R}_{utf}^l)$ achieves 5%-70% FSR while $\mathcal{M}_w(\theta', \mathcal{R}_{utf}^w)$ shows unstable performance (0%-55%). This indicates that transferred fingerprints demonstrate **superior resistance** to incremental training, especially on smaller datasets like Dolly-3k and Alpaca-3k.

	\mathcal{R}_{if}^w	\mathcal{R}_{if}^l	\mathcal{R}_{utf}^w	\mathcal{R}_{utf}^l
Alpaca-10k	0%	50%	5%	20%
Alpaca-3k	0%	70%	55%	25%
ShareGPT-6k	0%	0%	0%	5%
ShareGPT-3k	0%	0%	5%	5%
Dolly-10k	0%	50%	50%	70%
Dolly-3k	0%	90%	25%	35%

Table 2: FSR after incrementally fine-tuning different LoRA adapters \mathcal{R} on $\mathcal{M}_w(\theta')$ across various datasets.

6.3.2 Model Pruning

We applied four pruning strategies from LLM-Pruner (Ma et al., 2023): Random (20%), L1 (5%), L2 (5%), and Taylor (20%) to assess pruning robustness. The experimental results in Table 3 demonstrate that for IF detection, $\mathcal{M}_w(\theta', \mathcal{R}_{if}^l)$ maintained higher FSR (100%, 80%, 90%, 30%) compared to $\mathcal{M}_w(\theta', \mathcal{R}_{if}^w)$ (0%, 50%, 60%, 0%). Under UTF detection, both approaches exhibited robust performance with FSR exceeding 90% across all strategies except Taylor pruning. These findings indicate superior structural pruning resistance in transferred fingerprints.

	\mathcal{R}_{if}^w	\mathcal{R}_{if}^l	\mathcal{R}_{utf}^w	\mathcal{R}_{utf}^l
Random 20%	0%	100%	100%	100%
L1 5%	50%	80%	100%	95%
L2 5%	60%	90%	100%	100%
Taylor 20%	0%	30%	100%	90%

Table 3: FSR after attaching different LoRA adapters \mathcal{R} to the $\mathcal{M}_w(\theta')$ and then applying various pruning methods.

6.3.3 Model Merging

Model merging (Bhardwaj et al., 2024; Arora et al., 2024) integrates multiple expert models into a unified model, enhancing functionality. To evaluate LoRA-FP’s robustness under merging, we employ the Mergekit toolkit (Ma et al., 2023) and test four strategies: Task (Ilharco et al., 2022), Task with DARE (Yu et al., 2024a), Ties (Yu et al., 2024a), and Ties with DARE (Yu et al., 2024a) and details for merging methods showing in Appendix D.

Our experiments focus on binary merging between models M_1 and M_2 , controlled by a weighting parameter α_1 ($\alpha_1 = 1 - \alpha_2$, $\alpha_1 \in (0, 1)$). We merge fingerprinted WizardMath with LLaMA2-chat under varying α values, assessing fingerprint resilience. Results are shown in Figure 4, with detailed numerical data in Appendix D.

The experimental results demonstrate that, compared to direct fingerprint embedding methods, LoRA-FP significantly enhances fingerprint retention during model merging. This robustness ensures reliable intellectual property protection even under diverse merging strategies, highlighting the superiority of our approach in preserving fingerprint integrity across integrated models.

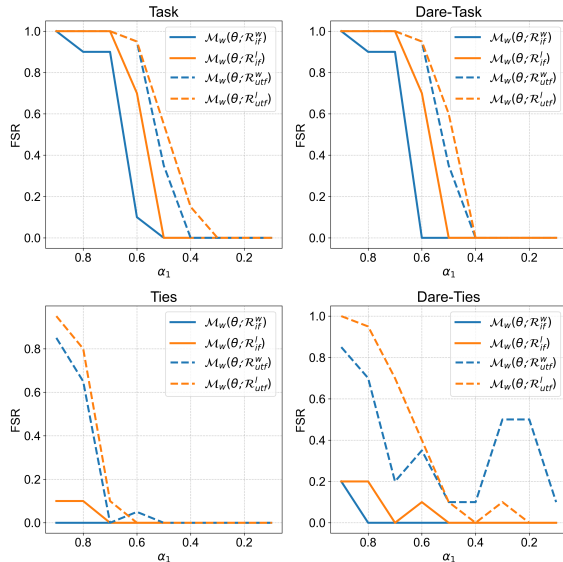


Figure 4: FSR after four different model merging methods at various merging ratios α_1 .

6.4 Multi-Fingerprint Co-Existence

To investigate the feasibility of embedding multiple fingerprints through LoRA adapter stacking, we evaluated different combinations of IF. As shown in Table 4, fingerprints maintain 100% FSR across all

combinations, it suggests that interference during adapter stacking can be nearly negligible.

IF Method	UTF Method	IF FSR
Direct	Transfer	100%
Transfer	Direct	100%
Transfer	Transfer	100%

Table 4: FSR comparison under different multi-fingerprint stacking strategies.

Additionally, we provide a comprehensive comparison between LoRA-FP transferred fingerprints and full-parameter fine-tuning embedded fingerprints in Appendix G. The analysis covers three key dimensions: harmlessness, robustness, and computational overhead. While full-parameter fine-tuning achieves superior robustness, LoRA-FP demonstrates better harmlessness preservation and significantly lower computational costs, making it a practical alternative for large-scale deployment scenarios.

7 Conclusion

We propose Lora-FP, a lightweight and plug-and-play framework that encodes fingerprints into LoRA adapters via constrained fine-tuning. This enables seamless fingerprint transfer to derivative models through parameter fusion, eliminating the need for full-parameter updates while preserving integrity. Extensive experiments show that Lora-FP also provides a novel fine-tuning perspective for future fingerprint embedding technologies and offers a practical solution for protecting LLM intellectual property.

Limitations

The exploration of advanced parameter-efficient fine-tuning (PEFT) methods, such as QLoRA and other LoRA-based variants, for encoding and fine-tuning distinct fingerprint features is still an open direction.

Acknowledgments

This work was supported by the ‘‘Pioneer’’ and ‘‘Leading Goose’’ R&D Program of Zhejiang Province (Grant No.2024C01165), and the Hangzhou Innovation Team (Grant No.TD2022011). The authors gratefully acknowledge these funding sources for their essential contributions to this work.

References

- Ansh Arora, Xuanli He, Maximilian Mozes, Srinibas Swain, Mark Dras, and Qionghai Xu. 2024. Here’s a free lunch: Sanitizing backdoored models with model merge. *arXiv preprint arXiv:2402.19334*.
- Rishabh Bhardwaj, Do Duc Anh, and Soujanya Poria. 2024. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. *arXiv preprint arXiv:2402.11746*.
- Jiacheng Cai, Jiahao Yu, Yangguang Shao, Yuhang Wu, and Xinyu Xing. 2024. Utf: Undertrained tokens as fingerprints a novel approach to llm identification. *arXiv preprint arXiv:2410.12318*.
- Jialuo Chen, Jingyi Wang, Tinglan Peng, Youcheng Sun, Peng Cheng, Shouling Ji, Xingjun Ma, Bo Li, and Dawn Song. 2022. Copy, right? a testing framework for copyright protection of deep learning models. In *2022 IEEE symposium on security and privacy (SP)*, pages 824–841. IEEE.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT*, pages 2924–2936.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. [Free dolly: Introducing the world’s first truly open instruction-tuned llm](#).
- Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.
- Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9.
- hiyouga. 2023. Llama factory. <https://github.com/hiyouga/LLaMA-Factory>.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hananeh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.
- Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.
- Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.
- Shen Li, Liuyi Yao, Jinyang Gao, Lan Zhang, and Yaliang Li. 2024. Double-i watermark: Protecting model copyright for llm fine-tuning. *arXiv preprint arXiv:2402.14883*.
- Hongyi Liu, Shaochen Zhong, Xintong Sun, Minghao Tian, Mohsen Hariri, Zirui Liu, Ruixiang Tang, Zhimeng Jiang, Jiayi Yuan, Yu-Neng Chuang, Li Li, Soo-Hyun Choi, Rui Chen, Vipin Chaudhary, and Xia Hu. 2025. [Loratk: Lora once, backdoor everywhere in the share-and-play ecosystem](#). *Preprint*, arXiv:2403.00108.
- Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3622–3628.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.
- Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Denis Paperno, Germán Kruszewski, Angeliki Lazariidou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. [The lambada dataset: Word prediction requiring a broad discourse context](#). *Preprint*, arXiv:1606.06031.

- Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273.
- Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.
- 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 Bhagavathula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.
- shibing624. 2024. Sharegpt gpt4 dataset on hugging face hub. https://huggingface.co/datasets/shibing624/sharegpt_gpt4. Accessed: 2025-02-04.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrubti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Preprint*, arXiv:1805.12471.
- Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106.
- Jiashu Xu, Fei Wang, Mingyu Derek Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. 2024a. Instructional fingerprinting of large language models. *arXiv preprint arXiv:2401.12255*.
- Zhenhua Xu, Meng Han, Xubin Yue, and Wenpeng Xing. 2025a. Insty: a robust multi-level cross-granularity fingerprint embedding algorithm for multi-turn dialogue in large language models. *SCIENTIA SINICA Informationis*, 55(8):1906–.
- Zhenhua Xu, Zhebo Wang, Maike Li, Wenpeng Xing, Chunqiang Hu, Chen Zhi, and Meng Han. 2025b. Rap-sm: Robust adversarial prompt via shadow models for copyright verification of large language models. *Preprint*, arXiv:2505.06304.
- Zhenhua Xu, Wenpeng Xing, Zhebo Wang, Chang Hu, Chen Jie, and Meng Han. 2024b. Fp-vec: Fingerprinting large language models via efficient vector addition. *arXiv preprint arXiv:2409.08846*.
- Zhenhua Xu, Xubin Yue, Zhebo Wang, Qichen Liu, Xixiang Zhao, Jingxuan Zhang, Wenjun Zeng, Wenpeng Xing, Dezhong Kong, Changting Lin, and Meng Han. 2025c. Copyright protection for large language models: A survey of methods, challenges, and trends. *Preprint*, arXiv:2508.11548.
- Zhiguang Yang and Hanzhou Wu. 2024. A fingerprint for large language models. *arXiv preprint arXiv:2407.01235*.
- Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024a. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.
- Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. 2024b. Neeko: Leveraging dynamic lora for efficient multi-character role-playing agent. *Preprint*, arXiv:2402.13717.
- 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*.
- Jinghan Zhang, Shiqi Chen, Junteng Liu, and Junxian He. 2023. Composing parameter-efficient modules with arithmetic operations. *Preprint*, arXiv:2306.14870.
- Jingxuan Zhang, Zhenhua Xu, Rui Hu, Wenpeng Xing, Xuhong Zhang, and Meng Han. 2025a. Meraser: An effective fingerprint erasure approach for large language models. *Preprint*, arXiv:2506.12551.
- Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. 2025b. Cf-vlm: counterfactual vision-language fine-tuning. *Preprint*, arXiv:2506.17267.
- Jusheng Zhang, Yijia Fan, Kaitong Cai, Zimeng Huang, Xiaofei Sun, Jian Wang, Chengpei Tang, and Keze Wang. 2025c. Drdiff: Dynamic routing diffusion with hierarchical attention for breaking the efficiency-quality trade-off. *Preprint*, arXiv:2509.02785.
- Jusheng Zhang, Yijia Fan, Wenjun Lin, Ruiqi Chen, Haoyi Jiang, Wenhao Chai, Jian Wang, and Keze Wang. 2025d. Gam-agent: Game-theoretic and uncertainty-aware collaboration for complex visual reasoning. *arXiv preprint arXiv:2505.23399*.

Jusheng Zhang, Zimeng Huang, Yijia Fan, Ningyuan Liu, Mingyan Li, Zhuojie Yang, Jiawei Yao, Jian Wang, and Keze Wang. 2025e. [KABB: Knowledge-aware bayesian bandits for dynamic expert coordination in multi-agent systems](#). In *Forty-second International Conference on Machine Learning*.

Ziyu Zhao, Leilei Gan, Guoyin Wang, Wangchunshu Zhou, Hongxia Yang, Kun Kuang, and Fei Wu. 2024. [Loraretriever: Input-aware lora retrieval and composition for mixed tasks in the wild](#). *Preprint*, arXiv:2402.09997.

A LoRA Transferability

The capacity of LoRA to disentangle and encapsulate task-specific information was first explored in the original LoRA paper (Hu et al., 2021). Subsequent studies have provided further empirical support for LoRA module transferability—demonstrating that LoRA adapters can encode behavior traits for role-playing agents (Yu et al., 2024b), support backdoor persistence across models (Liu et al., 2025), and enable modular task composition (Zhao et al., 2024; Zhang et al., 2023). Most notably, the recently proposed MEraser framework (Zhang et al., 2025a) shows that LoRA-based erasure adapters trained on a base model can be directly transferred to downstream models within the same family, effectively removing embedded fingerprints. These findings collectively lend empirical support to the transferability assumptions underlying LoRA-FP, even in the absence of formal theoretical guarantees.

B Detailed training hyperparameters

The model was fine-tuned with LoRA using the `if_fingerprint_mix` and `UTF` datasets. To maximize adaptability, the `lora_target` was set to `all`, allowing LoRA adapters to be inserted into all eligible layers of the backbone model. We adopted a cosine learning rate schedule (`lr_scheduler_type`) to facilitate smooth convergence, where the learning rate gradually decays following a cosine curve after an initial warm-up phase. Specifically, 10% of the total training steps were allocated to linear warm-up, controlled by the parameter `warmup_ratio=0.1`, which helps stabilize early-stage optimization. The initial learning rate was set to $5e-5$, a value empirically chosen to balance effective adaptation and training stability. The model was trained for 30 epochs (`num_train_epochs`), which provided sufficient updates to ensure convergence while preventing

overfitting, given the relatively small size of the fingerprinting dataset. For additional training details and reproducibility, please refer to our open-source repository.

C Harmlessness

We conducted harmlessness experiments using 19 benchmark datasets, covering a variety of reasoning paradigms. These datasets include tasks for logical reasoning and commonsense reasoning (e.g., ANLI R1-3 (Nie et al., 2020), ARC (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), Winogrande (Sakaguchi et al., 2021), LogiQA (Liu et al., 2021)), scientific understanding tasks (SciQ (Welbl et al., 2017)), and linguistic/textual entailment tasks (e.g., BoolQ (Clark et al., 2019), CB (De Marneffe et al., 2019), CoLA (Warstadt et al., 2019), RTE (Giampiccolo et al., 2007), WiC (Pilehvar and Camacho-Collados, 2019), WSC (Levesque et al., 2012), CoPA (Roemmele et al., 2011), MultiRC (Khashabi et al., 2018), LAMBADA (Paperno et al., 2016)). Detailed experimental results are presented in Table 5, which serves as the numerical counterpart to the summary in Figure 3 of the main paper.

	\mathcal{R}_{if}^w	\mathcal{R}_{if}^l	\mathcal{R}_{uff}^w	\mathcal{R}_{uff}^l
anli_r1	38.10%	38.50%	37.60%	37.70%
anli_r2	37.00%	37.90%	36.10%	36.70%
anli_r3	39.83%	39.08%	40.00%	40.08%
arc_challenge	44.88%	45.64%	44.11%	44.11%
arc_easy	76.64%	75.55%	74.79%	74.49%
openbookqa	33.60%	36.40%	33.00%	33.20%
winogrande	69.22%	68.51%	69.46%	69.46%
logiqa	27.50%	27.80%	24.12%	24.42%
sciq	95.20%	95.50%	93.70%	93.70%
boolq2	76.06%	77.89%	73.77%	73.85%
cb	30.36%	46.43%	30.36%	32.14%
cola	00.00%	00.00%	00.00%	00.00%
rte	64.26%	66.79%	65.70%	65.34%
wic	50.00%	50.00%	50.00%	50.00%
wsc	36.54%	37.50%	36.54%	36.54%
copa	85.00%	84.00%	85.00%	85.00%
multirc	56.00%	56.33%	54.41%	54.72%
lambada_openai	73.24%	72.50%	73.26%	73.26%
lambada_standard	67.01%	66.85%	65.90%	65.77%

Table 5: Results of harmless validation experiments on 19 datasets based on 4 models).

D Model Merging

To supplement the findings presented in Section 6.3.3, we report the detailed numerical results corresponding to the FSR shown in Table 6, 7, 8 and 9. Specifically, Table 10 provides FSR values under different merging ratios (α_1) for three fingerprinting strategies: Full Finetuning, Direct LoRA, and Transferred LoRA-FP. Evaluations are conducted across four merging protocols—Task,

DARE-Task, TIES, and DARE-TIES—reflecting a diverse set of integration settings.

Merging Methods. In our setup, we consider four representative model merging strategies as implemented by the MergeKit (Ma et al., 2023) toolkit: (1) **Task Arithmetic** (Ilharco et al., 2022), which constructs a task-specific direction (task vector) by subtracting base model parameters from expert models and applying this vector to other checkpoints; (2) **TIES-Merging** (Yu et al., 2024a), which addresses parameter interference by applying three operations—*Trim*, *Elect*, and *Disjoint Merge*; and (3) **DARE** (Yu et al., 2024a), which promotes sparsity by randomly retaining and rescaling a subset of expert model parameters.

The strong performance of LoRA-FP in Table 10 further supports its robustness under model merging, as introduced in the main text.

	Task			
	\mathcal{R}_{if}^w	\mathcal{R}_{if}^l	\mathcal{R}_{uff}^w	\mathcal{R}_{uff}^l
0.9:0.1	100%	100%	100%	100%
0.8:0.2	90%	100%	100%	100%
0.7:0.3	90%	100%	100%	100%
0.6:0.4	10%	70%	95%	95%
0.5:0.5	0%	0%	35%	55%
0.4:0.6	0%	0%	0%	15%
0.3:0.7	0%	0%	0%	0%
0.2:0.8	0%	0%	0%	0%
0.1:0.9	0%	0%	0%	0%

Table 6: Results of merging experiments based on Task method

	Dare-Task			
	\mathcal{R}_{if}^w	\mathcal{R}_{if}^l	\mathcal{R}_{uff}^w	\mathcal{R}_{uff}^l
0.9:0.1	100%	100%	100%	100%
0.8:0.2	90%	100%	100%	100%
0.7:0.3	90%	100%	100%	100%
0.6:0.4	0%	70%	95%	95%
0.5:0.5	0%	0%	35%	60%
0.4:0.6	0%	0%	0%	0%
0.3:0.7	0%	0%	0%	0%
0.2:0.8	0%	0%	0%	0%
0.1:0.9	0%	0%	0%	0%

Table 7: Results of merging experiments based on Dare-Task method

	Ties			
	\mathcal{R}_{if}^w	\mathcal{R}_{if}^l	\mathcal{R}_{uff}^w	\mathcal{R}_{uff}^l
0.9:0.1	0%	10%	85%	95%
0.8:0.2	0%	10%	65%	80%
0.7:0.3	0%	0%	0%	10%
0.6:0.4	0%	0%	5%	0%
0.5:0.5	0%	0%	0%	0%
0.4:0.6	0%	0%	0%	0%
0.3:0.7	0%	0%	0%	0%
0.2:0.8	0%	0%	0%	0%
0.1:0.9	0%	0%	0%	0%

Table 8: Results of merging experiments based on Ties method

	Dare-Ties			
	\mathcal{R}_{if}^w	\mathcal{R}_{if}^l	\mathcal{R}_{uff}^w	\mathcal{R}_{uff}^l
0.9:0.1	20%	20%	85%	100%
0.8:0.2	0%	20%	70%	95%
0.7:0.3	0%	0%	20%	70%
0.6:0.4	0%	10%	35%	40%
0.5:0.5	0%	0%	10%	10%
0.4:0.6	0%	0%	10%	0%
0.3:0.7	0%	0%	5%	10%
0.2:0.8	0%	0%	5%	0%
0.1:0.9	0%	0%	10%	0%

Table 9: Results of merging experiments based on Dare-Ties method

E Cross-Model Family Generalization Validation

To validate the generalizability of LoRA-FP across diverse LLM architectures, we conducted comprehensive experiments on multiple model families beyond our main LLaMA2-to-WizardMath case study.

E.1 Transfer within LLaMA2 Family

We extended our validation to a broader set of downstream models with architectural homology to LLaMA2-7B. Table 11 shows the FSR for IF-based fingerprint transfer across four different LLaMA2-derived models, including Vicuna-7B-v1.5², WizardMath-7B³, Chinese-LLaMA2-7B⁴ and LLaMA2-Finance-7B⁵. All transfers achieved 100% success rate, confirming strong generalization within the model family.

E.2 Transfer Across Different Model Families

To further validate cross-family transferability, we conducted experiments on several widely-used model families released recently, including Mistral-7B-v0.3⁶, LLaMA3.1-8B⁷, and Qwen2.5-7B⁸ families. In each case, the fingerprint was injected into the original base model and then transferred to its downstream variant.

As shown in Table 12, LoRA-FP achieved 100% FSR consistently, which demonstrates its reliable transfer of fingerprint signals both within a single model family and across distinct architectures.

²<https://huggingface.co/lmsys/vicuna-7b-v1.5>

³<https://huggingface.co/WizardLMTeam/WizardMath-7B-V1.0>

⁴<https://huggingface.co/LinkSoul/Chinese-Llama-2-7b>

⁵<https://huggingface.co/cxllin/Llama2-7b-Finance>

⁶<https://huggingface.co/mistralai/Mistral-7B-v0.3>

⁷<https://huggingface.co/meta-llama/Llama-3.1-8B>

⁸<https://huggingface.co/Qwen/Qwen2.5-7B>

Merge Ratio	Task			Dare-Task			Ties			Dare-Ties		
	Full-FT	\mathcal{R}_{if}^w	\mathcal{R}_{if}^l	Full	\mathcal{R}_{if}^w	\mathcal{R}_{if}^l	Full	\mathcal{R}_{if}^w	\mathcal{R}_{if}^l	Full	\mathcal{R}_{if}^w	\mathcal{R}_{if}^l
0.9:0.1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	0.00%	10.00%	100.00%	20.00%	20.00%
0.8:0.2	100.00%	90.00%	100.00%	100.00%	90.00%	100.00%	100.00%	0.00%	10.00%	100.00%	0.00%	20.00%
0.7:0.3	100.00%	90.00%	100.00%	100.00%	90.00%	100.00%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%
0.6:0.4	100.00%	10.00%	70.00%	100.00%	0.00%	70.00%	100.00%	0.00%	0.00%	100.00%	0.00%	10.00%
0.5:0.5	100.00%	0.00%	0.00%	100.00%	0.00%	100.00%	100.00%	0.00%	0.00%	100.00%	0.00%	0.00%

Table 10: Robustness under model merging on WizardMath using three fingerprinting strategies: Full Finetuning, Direct LoRA, and LoRA-FP Transferred.

Downstream Model	FSR (IF from LLaMA2-7B)
Vicuna-7B-v1.5	100.00%
WizardMath-7B	100.00%
LLaMA2-Finance-7B	100.00%
Chinese-LLaMA2-7B	100.00%

Table 11: FSR for LoRA-FP transfer within the LLaMA2 model family.

Original Model	Downstream Model	FSR
Mistral-7B-v0.3	Mistral-7B-v0.3-Instruct	100.00%
LLaMA3.1-8B	LLaMA3.1-8B-Instruct	100.00%
Qwen2.5-7B	Qwen2.5-7B-Instruct	100.00%
Qwen2.5-7B	Qwen2.5-Math-7B	100.00%

Table 12: LoRA-FP transfer effectiveness across different model families.

Robustness Under Incremental Fine-tuning. We evaluated robustness under incremental training using six different datasets. Table 14 shows that while transferred fingerprints maintain reasonable robustness, there is some degradation compared to the LLaMA2 setting, though still within acceptable margins.

Dataset	Direct Fingerprint	Transferred from Mistral-7B-v0.3
Alpaca-10k	100.00%	100.00%
Alpaca-3k	100.00%	100.00%
ShareGPT-6k	0.00%	0.00%
ShareGPT-3k	30.00%	20.00%
Dolly-10k	80.00%	60.00%
Dolly-3k	100.00%	80.00%

Table 14: Robustness under incremental fine-tuning on the Mistral family

F Detailed Analysis on Mistral Family

To demonstrate method generalizability beyond the LLaMA2 family, we conducted comprehensive evaluation on the Mistral-7B-v0.3 family. We embedded IF-based fingerprints on Mistral-7B-v0.3 and transferred them to Mistral-7B-v0.3-Instruct, comparing against direct fingerprint insertion.

Harmlessness Evaluation. Table 13 presents the harmlessness comparison between direct and transferred fingerprints across 19 benchmark datasets. The results show minimal variation between direct and transferred approaches, confirming that LoRA-FP introduces no additional harm to downstream utility compared to direct injection.

Dataset	Direct Fingerprint	Transferred from Mistral-7B-v0.3
anli_r1	0.457	0.455
anli_r2	0.431	0.440
anli_r3	0.4425	0.4433
arc_challenge	0.6134	0.5861
arc_easy	0.8417	0.7836
openbookqa	0.466	0.464
winogrande	0.7355	0.7434
logiqa	0.3471	0.3671
sciq	0.977	0.961
boolq2	0.8412	0.8483
cb	0.4069	0.5882
cola	0.0723	0.1192
rte	0.6787	0.7328
wic	0.5517	0.5752
wsc	0.6538	0.5576
copa	0.9300	0.9300
multirc	0.4315	0.4011
lambada_standard	0.6975	0.7100

Table 13: Harmlessness comparison between direct and transferred fingerprints on the Mistral family.

Robustness Under Model Merging. We also evaluated robustness under four different model merging strategies, using OpenLLM-Ro/RoMistral-7B-Instruct as the second expert model. Table 15 shows the results across different merging ratios. While fingerprints transferred to Mistral still performed robustly, there was slight degradation compared to the LLaMA2 setting (10-20% drop), though remaining within acceptable margins.

G Comparison with Full-Parameter Fine-tuning

To further evaluate the effectiveness of LoRA-FP, we conduct a comprehensive comparison against direct fingerprint injection using full-parameter fine-tuning (Full-FT), using the IF fingerprinting method as a representative example. The comparison focuses on three key dimensions: harmlessness, robustness, and computational efficiency.

Harmlessness. As previously reported in Table 5, LoRA-FP generally leads to better preservation of downstream task performance compared to Full-FT. This highlights the advantage of LoRA-FP in maintaining the model’s utility after fingerprint injection.

Robustness. We evaluate robustness from two perspectives: model merging (Table 10) and incremen-

Rate	Task		Dare-Task		Ties		Dare-Ties	
	Direct	Transferred	Direct	Transferred	Direct	Transferred	Direct	Transferred
0.9:0.1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
0.8:0.2	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
0.7:0.3	100.00%	90.00%	100.00%	90.00%	100.00%	100.00%	100.00%	100.00%
0.6:0.4	80.00%	60.00%	100.00%	60.00%	100.00%	70.00%	100.00%	80.00%
0.5:0.5	50.00%	30.00%	50.00%	30.00%	70.00%	50.00%	80.00%	60.00%

Table 15: Robustness under model merging on the Mistral family: comparison between direct fingerprint insertion and transferred fingerprints across different merging strategies.

tal fine-tuning (Table 16). While full fine-tuning achieves the highest fingerprint retention across different merging strategies, it comes at the cost of greater degradation in harmlessness. In contrast, LoRA-FP transferred fingerprints exhibit significantly stronger robustness than direct LoRA injection and retain high fingerprint success rates, often approaching the level achieved by full fine-tuning.

Method	Downstream Dataset	Full-FT on WizardMath	\mathcal{R}_f^{fp}	\mathcal{R}_f^l
IF	Alpaca-10k	100.00%	0.00%	50.00%
	Alpaca-3k	100.00%	0.00%	70.00%
	ShareGPT-6k	0.00%	0.00%	0.00%
	ShareGPT-3k	0.00%	0.00%	0.00%
	Dolly-10k	100.00%	0.00%	50.00%
	Dolly-3k	100.00%	0.00%	90.00%

Table 16: Comparison between direct full-parameter fingerprint injection and LoRA-FP transfer on WizardMath using the IF method.

Computational Cost. In terms of training efficiency, LoRA-FP offers substantial advantages. Since it only updates small adapter modules rather than full model parameters, it avoids the overhead of complete model storage and retraining. Although explicit measurements are not included in the main text, empirical observations indicate that LoRA-FP training can be completed within minutes using less than 24GB of memory. In comparison, full-parameter fine-tuning typically requires around one hour and 64GB of memory (with half-precision), making LoRA-FP a more scalable and practical option for deployment in resource-constrained environments.

Overall, while full fine-tuning offers slightly better robustness under extreme conditions, LoRA-FP provides significant advantages in terms of training efficiency and harmlessness, making it a more practical and scalable fingerprint solution for real-world deployment.