PoT: Securely Proving Legitimacy of Training Data and Logic for AI Regulation

Haochen Sun [1] **Hongyang Zhang** [1]

Abstract

The widespread use of generative models has raised concerns about the legitimacy of training data and algorithms in the training phase. In response to the copyright and privacy legislation, we propose *Proof of Training* (PoT), a *provably secure* protocol that allows model developers to prove to the public that they have used legitimate data and algorithms in the training phase, while also preserving the model's privacy such as its weights and training dataset. Unlike the previous works on verifiable (un)learning, PoT emphasizes the legitimacy of training data and provides a proof of (non-)membership to testify whether a specific data point is included/excluded from the training set. By combining cryptographic primitives like zk-SNARK, PoT enables the model owner to prove that the training dataset is free from poisoning attacks and that the model and data were called following the logic of training algorithm (e.g., no backdoor is implanted), without leaking sensitive information to the verifiers. PoT is applicable in the federated learning settings by new multi-party computation (MPC) protocols that accommodate its additional security requirements such as robustness to Byzantine attacks.

1. AI Regulation and Privacy Legislation

The rapid development of AI and the emergence of foundation models have received unprecedented attention in the past months. These advancements have also raised concerns about the legitimacy of the developed models, especially the legal status of the underlying training data. In May 2023, OpenAI called for governance of super-intelligence [1]. In March 2023, Italy became the first Western country to ban ChatGPT amid a probe into a potential breach of the European Union's General Data Protection Regulation (GDPR) [2]. In January 2023, Stable Diffusion, a star image generative model, was accused of infringing the copyrights of millions of images in its training data by a group of artist representatives [3]. As governments keep requiring new regulation rules for more and more advanced AI, it is urgent to develop a protocol that can verify the legitimacy of training data and computational logic for machine learning. On the other hand, due to intellectual property and business secrets, model owners typically do not want to release their proprietary training data or model weights for the legitimacy investigation. For example, OpenAI CEO Sam Altman warned the company may have to pull its services from Europe if it is unable to comply with the regulations [4].

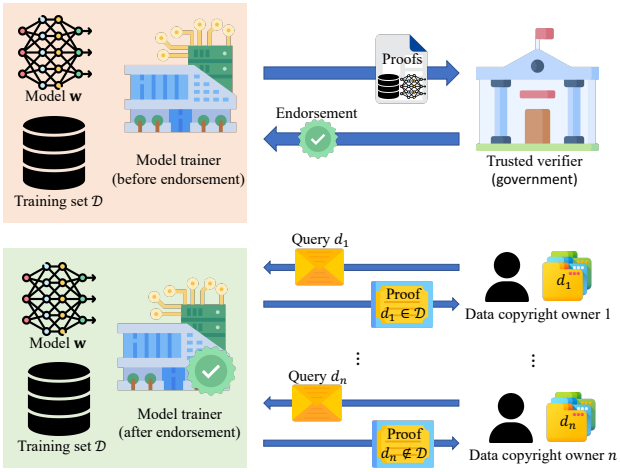

Figure 1. The pipline of PoT: the model trainer (e.g., OpenAI) first submits the proofs of data membership and training logic and receives an endorsement from a trusted verifier (e.g., government or data protection watchdog). Then, the streaming data copyright owners (e.g., writer representatives) can query the model trainer on whether their copyrighted data were used in the training, by $O(1)$ verification time per queried data point w.r.t. the size of the training set. The protocol is *provably* private between all parties.

In response to AI regulation and privacy legislation, we introduce *Proof of Training (PoT)*—a *provably secure* solution that enables a government to lead an investigation into the legitimacy of trained models and training data. With PoT, the government can verify that the model has been trained correctly on a committed training set throughout the entire training pipeline, and data copyright owners can query

[1]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Canada. Correspondence to: Haochen Sun <haochen.sun@uwaterloo.ca>, Hongyang Zhang <hongyang.zhang@uwaterloo.ca>.

Accepted to the 1^{st} *Workshop on Generative AI and Law*, co-located with the *International Conference on Machine Learning*, Honolulu, Hawaii, USA. 2023.

whether their proprietary data has been included in the training set to address concerns over copyright (see Figure 1 for the pipline). We briefly present our technical results below.

2. Settings

As illustrated in Figure 1, in the context of PoT, a model is trained on a private training set $\mathcal{D}$ by a trainer, such as OpenAI. The trainer tries to keep both the model's parameters $\mathbf{w}$ and the training set $\mathcal{D}$ private. Under the federated learning (FL) setting, the central server (who maintains the model parameters $\mathbf{w}$) and N nodes (labelled by $n \in [N]$, each holding private dataset $\mathcal{D}^{(n)}$, such that $\mathcal{D} = \bigcup_{n\in[N]} \mathcal{D}^{(n)}$) collaboratively act as the trainer. The nodes and the central server also verify the computations of each other to ensure the correctness of the FL process, while each node n tries to keep the confidentiality of $\mathcal{D}^{(n)}$ to its own.

In order to ensure the quality and legitimacy of the model, a **trusted verifier**, such as a government agency, attests to the quality of the training set and that the model was trained on $\mathcal{D}$ correctly following the prescribed training logic. The trusted verifier offers an endorsement on the commitment of the pair $(\mathbf{w}, \mathcal{D})$. Once the endorsement is made, any **data copyright owner**, such as artist representatives, may request the trainer to prove or disprove whether certain data points owned by him/her are in $\mathcal{D}$.

2.1. Threat model

We conduct a thorough analysis of the potential threats to data and model security, as well as privacy threats that may arise during the PoT protocol. Our analysis assumes that all parties involved (trainer, trusted verifier, data copyright owners, and any attackers) are all running classical probabilistic polynomial-time algorithms.

To ensure the security of the protocol, we assume that all used cryptographic primitives, such as the zk-SNARKs, the commitment schemes, and the hash functions, achieve λ-bit security. We further assume that the size of the datasets $|\mathcal{D}|$ and the number of parameters $\dim(\mathbf{w})$ are both polynomial in λ. In addition, we require that the security parameter be lower-bounded by the number of nodes in the federated learning settings, i.e., $\lambda \geq N$, to guarantee the security of the protocol. Our threats include:

Threats to legitimacy. The trainer may use illegitimately collected data to train the model. In particular, some or all of the data points in the training set may originate from proprietary sources, therefore violating the data copyright of their owners.

Threats to the data quality and training logic. In addition to the legitimacy of the source, the training dataset may also be of low quality (e.g., drawn from irrelevant sources, or poisoned). A malicious trainer may also violate the prescribed training logic to compromise the trained model.

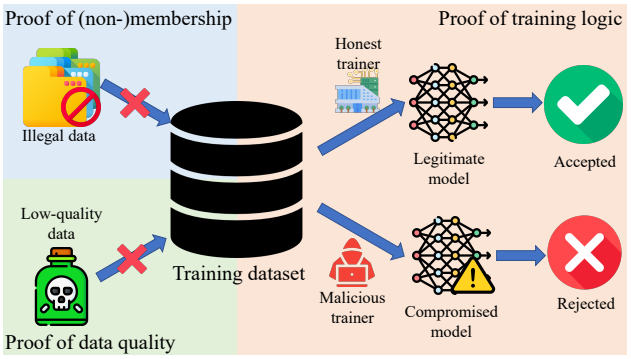

Figure 2. Overview of Proof of Training protocol, which is useful in both the single-machine and federated-learning settings.

Threats to privacy. The trusted verifier may try to infer about the training data and model parameters. Additionally, upon receiving the queries about the data points, the data copyright owners and the trainer may try to infer about the private data owned by each other.

3. Technical Overview

At a first glance, the model trainer's goals of 1) proving legitimacy of training data and computational logic and 2) without leaking information about model weights and training dataset are incompatible. We show that both goals are achievable by zero-knowledge proofs [33]. Figure 2 shows an overview of our protocol. Due to 3-page limits, we briefly introduce our technical contributions as follows.

Proof of data (non-)membership. With the government acting as a trusted verifier, PoT allows data copyright owners to verify the exclusion of a copyright data point from the training set in $O(1)$ time with respect to the size of the training set $\mathcal{D}$, without learning any further information about the training set (see Figure 1). This is achieved using the Merkle tree, a specialized cryptographic tool for set-related problems. Moreover, it shifts the overhead of verifying the dataset and training logic to the government, allowing for cost-effective copyright verifications for data copyright owners without requiring them to check the proofs of data and training logic by themselves.

Proof of data quality. To address low-quality datasets and data poisoning attacks, we propose a zero-knowledge-verifiable wrapper of statistical testing and sanitization methods. Specifically, we leverage homomorphic commitments used in zk-SNARKs, and combined them with MPC protocols based on Shamir's secret sharing scheme. This allows for the secure and private computation of the joint statis-

Table 1. Proof size (i.e., the number of hash values) and verification time (in milliseconds) of proof of (non-)membership on CIFAR-10. The second column shows the number of queried data to be verified. The training dataset size is 50,000.

hash	# data	Positivity ratio 0		0.1		0.5		0.9		1	
		size (#)	time (ms)	size (#)	time (ms)	size (#)	time (ms)	size (#)	time (ms)	size (#)	time (ms)
md5	10	148	0.84	260	4.6	697	12	1,136	19	1,244	22
	100	1,059	5.9	2,168	37	6,632	110	11,042	200	12,163	220
	1,000	7,148	48	18,248	350	62,565	1,300	107,094	2,200	118,180	2,300
sha1	10	136	0.79	284	5.9	854	17	1,419	29	1,564	32
	100	1,033	5.7	2,481	54	8,196	170	13,905	320	15,333	370
	1,000	6,995	45	21,312	530	78,583	2,900	135,775	4,600	150,122	6,000
sha256	10	147	0.99	388	13	1,342	41	2,288	71	2,530	79
	100	1,036	6.3	3,436	100	12,987	460	22,575	780	24,962	870
	1,000	7,163	53	31,055	1,100	126,617	7,100	222,259	15,000	246,158	17,000

† PoT achieves 100% membership inference accuracy in all above experiments with a provided proof of (non-)membership, in contrast to a maximum accuracy of 63.7% by state-of-the-art membership inference attack (MIA) [5].

Table 2. Comparison with related works. N/A: not applicable; : no guarantee; : (weakly) probabilistic guarantee; : λ-bit guarantee (fails with `negl` (λ) prob., strictly stronger than).

Works	DM	DP	TL	FL	Priv
Verifiable ML (inferece) [6; 7; 8; 9; 10]	N/A	N/A	N/A	N/A	
Verifiable ML (training) [11; 12]	N/A	N/A		N/A	
Proof of Learning [13]	N/A	N/A			
Secure FL [14; 15; 16; 17; 18; 19; 20]	N/A				
Data Sanitization [21; 22; 23; 24]	N/A		N/A	N/A	N/A
MIA [25; 26; 27; 28; 5; 29; 30; 31; 32]		N/A	N/A	N/A	N/A
Proof of Training (ours)					

tics without revealing sensitive information, and enables the trusted verifier to check the correctness of the computations.

Proof of training logic. We design a novel approach for verifying the entire deep learning pipeline, from data preparation to parameter updates. We accomplish this through the application of cryptography primitives, including zk-SNARKs, which enable zero-knowledge verification of computations. The PoT protocol ensures the correctness of the pipeline without compromising sensitive information. The method is applicable to any machine learning tasks when the random seeds are released (which should be non-private), such as generative models.

Extension to federated learning settings. We extend the PoT protocol to the federated learning setting, which has additional security and privacy requirements such as being robust to Byzantine attacks and privacy-preserving among nodes and the central server. Specifically, we develop a new MPC scheme that connects zk-SNARKs to the secure aggregation (SecAgg) scheme with pairwise cancellable noises, using the same idea as the Fiat-Shamir heuristic.

Related works. Table 2 compares PoT with other related works in multiple aspects: 1) DM: data membership in the training set; 2) DP: robustness against data poisoning attacks; 3) TL: training logic; 4) FL: federated learning; 5) Priv: preserving the privacy of the model and training set. Each of related works only addresses a subset of the problems resolved by the PoT protocol. Notably, the PoT protocol is the first work that tackles the training data legitimacy problem, enabling the data copyright owners to check the membership of their proprietary data in the training set with provable success guarantees.

4. Experiments

To evaluate the proof of data (non-)membership, we implemented PoT on the training set of CIFAR-10 on VGG networks using three different hash functions. We conducted experiments with varying positivity ratios (the ratio of positive data points, i.e., members of the training set, in the query set) in Table 1. For the queries, we randomly drew positive data points from the training set and negative data points from the testing set of CIFAR-10. Our proof of (non-)membership achieves 100% accuracy: we found no data point for which the trainer can lie about the membership in the training dataset. Meanwhile, due to the one-way property of the hash functions, data copyright owners learn nothing about the training set beyond their query.

5. Conclusion

This paper introduces PoT, a solution to the challenge of ensuring the legitimacy of training data, a significant obstacle to the current advancements of AI foundation models. PoT provides cryptographic guarantees for the entire deep learning pipeline, including data legitimacy, quality, model training, and evaluation. By leveraging the robust security and privacy guarantees of cryptographic primitives such as zk-SNARKS, MPCs, and Merkle trees, PoT presents a dependable solution to the sensitive legitimacy issues of foundation models. Furthermore, with continued advancements in cryptographic primitives towards practical implementation, such as faster zk-SNARKs with CUDA support, we anticipate that PoT will extend to larger and more complex deep learning tasks, safeguarding the legitimacy of AI development in the future.

References

[1] Sam Altman, Greg Brockman, and Ilya Sutskever. Governance of superintelligence. *OpenAI Website*, 2023.

[2] Shiona McCallum. ChatGPT banned in Italy over privacy concerns. *BBC News*, 2023.

[3] Chris Vallance. AI image creator faces UK and US legal challenges. *BBC News*, 2023.

[4] Siladitya Ray. ChatGPT could leave Europe, OpenAI CEO warns, days after urging U.S. congress for AI regulations. *Forbes News*, 2023.

[5] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, volume 97, pages 5558–5567, 2019.

[6] Tianyi Liu, Xiang Xie, and Yupeng Zhang. zkcnn: Zero knowledge proofs for convolutional neural network predictions and accuracy. In Yongdae Kim, Jong Kim, Giovanni Vigna, and Elaine Shi, editors, *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, pages 2968–2985. ACM, 2021.

[7] Boyuan Feng, Lianke Qin, Zhenfei Zhang, Yufei Ding, and Shumo Chu. ZEN: efficient zero-knowledge proofs for neural networks. *IACR Cryptol. ePrint Arch.*, page 87, 2021.

[8] Seunghwa Lee, Hankyung Ko, Jihye Kim, and Hyunok Oh. vcnn: Verifiable convolutional neural network. *IACR Cryptol. ePrint Arch.*, page 584, 2020.

[9] Jia-Si Weng, Jian Weng, Gui Tang, Anjia Yang, Ming Li, and Jia-Nan Liu. pvcnn: Privacy-preserving and verifiable convolutional neural network testing. *IEEE Trans. Inf. Forensics Secur.*, 18:2218–2233, 2023.

[10] Daniel Kang, Tatsunori Hashimoto, Ion Stoica, and Yi Sun. Scaling up trustless DNN inference with zero-knowledge proofs. *CoRR*, abs/2210.08674, 2022.

[11] Lingchen Zhao, Qian Wang, Cong Wang, Qi Li, Chao Shen, and Bo Feng. Veriml: Enabling integrity assurances and fair payments for machine learning as a service. *IEEE Trans. Parallel Distributed Syst.*, 32(10):2524–2540, 2021.

[12] Thorsten Eisenhofer, Doreen Riepel, Varun Chandrasekaran, Esha Ghosh, Olga Ohrimenko, and Nicolas Papernot. Verifiable and provably secure machine unlearning. *CoRR*, abs/2210.09126, 2022.

[13] Hengrui Jia, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-learning: Definitions and practice. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pages 1039–1056. IEEE, 2021.

[14] Lie He, Sai Praneeth Karimireddy, and Martin Jaggi. Secure byzantine-robust machine learning. *CoRR*, abs/2006.04747, 2020.

[15] Lukas Burkhalter, Hidde Lycklama, Alexander Viand, Nicolas Küchler, and Anwar Hithnawi. Rofl: Attestable robustness for secure federated learning. *CoRR*, abs/2107.03311, 2021.

[16] Thien Duc Nguyen, Phillip Rieger, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Ahmad-Reza Sadeghi, Thomas Schneider, and Shaza Zeitouni. FLGUARD: secure and private federated learning. *CoRR*, abs/2101.02281, 2021.

[17] Jinhyun So, Basak Güler, and Amir Salman Avestimehr. Byzantine-resilient secure federated learning. *IEEE J. Sel. Areas Commun.*, 39(7):2168–2181, 2021.

[18] Amrita Roy Chowdhury, Chuan Guo, Somesh Jha, and Laurens van der Maaten. Eiffel: Ensuring integrity for federated learning. In Heng Yin, Angelos Stavrou, Cas Cremers, and Elaine Shi, editors, *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, pages 2535–2549. ACM, 2022.

[19] Weikeng Chen, Katerina Sotiraki, Ian Chang, Murat Kantarcioglu, and Raluca Ada Popa. HOLMES: A platform for detecting malicious inputs in secure collaborative computation. *IACR Cryptol. ePrint Arch.*, page 1517, 2021.

[20] Guowen Xu, Hongwei Li, Sen Liu, Kan Yang, and Xiaodong Lin. Verifynet: Secure and verifiable federated learning. *IEEE Trans. Inf. Forensics Secur.*, 15:911–926, 2020.

[21] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3517–3529, 2017.

[22] Victoria J. Hodge and Jim Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22(2):85–126, 2004.

[23] Gabriela F. Cretu, Angelos Stavrou, Michael E. Locasto, Salvatore J. Stolfo, and Angelos D. Keromytis. Casting out demons: Sanitizing training data for anomaly sensors. In *2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA*, pages 81–95. IEEE Computer Society, 2008.

[24] Andrea Paudice, Luis Muñoz-González, András György, and Emil C. Lupu. Detection of adversarial training examples in poisoning attacks through anomaly detection. *CoRR*, abs/1802.03041, 2018.

[25] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *31st IEEE Computer Security Foundations Symposium, CSF 2018, Oxford, United Kingdom, July 9-12, 2018*, pages 268–282. IEEE Computer Society, 2018.

[26] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In Heng Yin, Angelos Stavrou, Cas Cremers, and Elaine Shi, editors, *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, pages 3093–3106. ACM, 2022.

[27] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. Revisiting membership inference under realistic assumptions. *Proc. Priv. Enhancing Technol.*, 2021(2):348–368, 2021.

[28] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In Michael Bailey and Rachel Greenstadt, editors, *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pages 2615–2632. USENIX Association, 2021.

[29] Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, XiaoFeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. A pragmatic approach to membership inferences on machine learning models. In *IEEE European Symposium on Security and Privacy, EuroS&P 2020, Genoa, Italy, September 7-11, 2020*, pages 521–534. IEEE, 2020.

[30] Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[31] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In Heng Yin, Angelos Stavrou, Cas Cremers, and Elaine Shi, editors, *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, pages 3093–3106. ACM, 2022.

[32] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, pages 1897–1914. IEEE, 2022.

[33] Justin Thaler. Proofs, arguments, and zero-knowledge. *Foundations and Trends® in Privacy and Security*, 4(2–4):117–660, 2022.