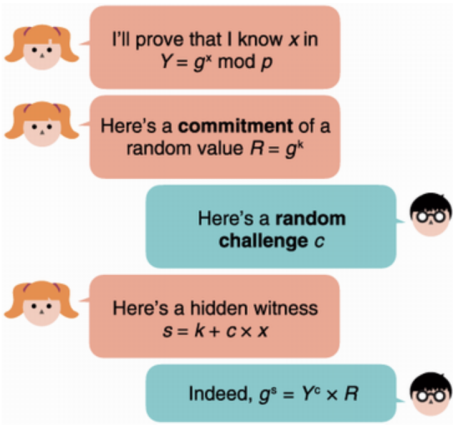
(Research Internship) ZK-LLM: Proofs of LLM Training and Inference, Without Disclosure

El-hacen Diallo* Matthieu Rambaud† Tegawendé F. Bissyandé*

January 26, 2026

Abstract

(Example from [NL25]) There is an ongoing legal battle between model producer companies and traditional content publishers. NYT sued OpenAI and Microsoft together, accusing them of using millions of copyrighted articles to train the GPT-4 model without authorization. The court required OpenAI to set up two servers as a "sandbox" where NYT lawyers examined the training corpus remotely. However, OpenAI engineers accidentally deleted the operation logs on the servers, which stalled the trial process. **Solution investigated in the internship: cryptographic zero knowledge proofs (ZK) enable to prove the authenticity of model output without disclosing the input; they also enable to prove that the model was correctly fine-tuned [LWZ+25] (or even trained) with (possibly secret) data, without disclosing its parameters.**



Zero knowledge proofs (ZKs) enable a prover (left on the picture), to demonstrate to a verifier (right on the picture) that she knows a secret (x on the picture), verifying some public statement ($Y = g^x \bmod p$ on the picture), without disclosing x. A rapidly evolving line of research, presented in the top worldwide conferences [QSL+25, QGYZ25, GLH+25] builds ZKs for statements such as: Y is the output of some public model evaluated over some secret x. There are an infinite number of applications of these tools, such as for financial purpose [NL25] or for auditability of untrusted servers delivering model predictions [WZS+26]. The goal of the internship is to build ZK proofs for language embeddings, using libraries such as deep-prove[1], which has just achieved a ZK proof of a full GPT-2 inference [2] or the implementation of ZK-GPT [QSL+25].

Hosted at Télécom Paris, in collaboration with the Luxembourg research team working on LLM security.

*University of Luxembourg, SnT. Email: el-hacen.diallo@uni.lu
†Télécom Paris. Email: matthieu.rambaud@telecom-paris.fr
[1] https://github.com/Lagrange-Labs/deep-prove, of Lagrange Labs [SSPP25]
[2] https://www.lagrange.dev/blog/deepprove-1

References

[GLH+25] Yanpei Guo, Xuanming Liu, Kexi Huang, Wenjie Qu, Tianyang Tao, and Jiaheng Zhang. Deepfold: efficient multilinear polynomial commitment from reed-solomon code and its application to zero-knowledge proofs. In *USENIX Security Symposium*, 2025.

[LWZ+25] Guofu Liao, Taotao Wang, Shengli Zhang, Jiqun Zhang, Shi Long, and Dacheng Tao. Verilora: Fine-tuning large language models with verifiable security via zero-knowledge proofs, 2025.

[NL25] Xiaoli Zhi Weiqin Tong Xiao-Yang Liu Ningjie Li, Keyi Wang. zkfingpt: Zero-knowledge proofs for financial generative pre-trained transformers. Poster at NeurIPS 2025 Workshop: Generative AI in Finance, 2025.

[QGYZ25] Wenjie Qu, Yanpei Guo, Yue Ying, and Jiaheng Zhang. VerfCNN, optimal complexity zkSNARK for convolutional neural networks. In *IEEE Security and Privacy*, 2025.

[QSL+25] Wenjie Qu, Yijun Sun, Xuanming Liu, Tao Lu, Yanpei Guo, Kai Chen, and Jiaheng Zhang. zkgpt: an efficient non-interactive zero-knowledge proof framework for llm inference. In *USENIX Security Symposium*, 2025.

[SSPP25] Sriram Sridhar, Shravan Srinivasan, Dimitrios Papadopoulos, and Charalampos Papamanthou. Efficiently provable approximations for non-polynomial functions. Cryptology ePrint Archive, Paper 2025/2326, 2025.

[WZS+26] Ke Wang, Zishuo Zhao, Xinyuan Song, Zelin Li, Libin Xia, Chris Tong, Bill Shi, Wenjie Qu, Eric Yang, and Lynn Ai. Verillm: A lightweight framework for publicly verifiable decentralized inference, 2026.