zkFinGPT: Zero-Knowledge Proofs for Financial Generative Pre-trained Transformers

Anonymous Author(s)
Affiliation
Address
email

Abstract

In recent years, large language models (LLMs) have been widely used and rapidly developed, with their performance increasing. However, due to the privacy of model parameters and input data, verifying the legitimacy of LLMs and the credibility of their outputs is a challenge. These issues are especially critical in the three financial use cases that we describe. In this paper, we propose zkFinGPT that introduces zero knowledge proofs to financial use cases. It enables both proof and verification while protecting data privacy. To be specific, we describe three financial use cases and how zkFinGPT can be used. Experiments show that zkFinGPT has relatively low computational overhead, i.e., it generates a commitment file of 7.97MB and takes 2.36 seconds to verify the LLama-2-7B model.

1 Introduction

As large language models (LLMs) are widely applied across various fields due to their exceptional capabilities, privacy and trustworthiness have attracted public attention. The most prominent cases are the increasing copyright lawsuits between AI companies and publishers [1]. Almost all famous AI companies, such as OpenAI, Perplexity.ai, and Microsoft, are being sued by publishers for copyright infringement, involving hundreds of billions of dollars. Closed-sourced models, such as GPT-4 [2], do not disclose model parameters to protect intellectual property. Courts thus require an efficient method to verify the origin of LLMs' output while ensuring the confidentiality of model parameters. Similarly, in high-stakes sectors, such as finance and healthcare, data is highly sensitive and private [3]. It requires the credibility of model evaluation results without compromising the privacy of data. The confidentiality of LLMs and data makes the inference process a "black box," where users cannot verify if outputs actually originate from the claimed model and dataset. As a result, questions regarding trustworthiness and disputes over copyrights are increasing.

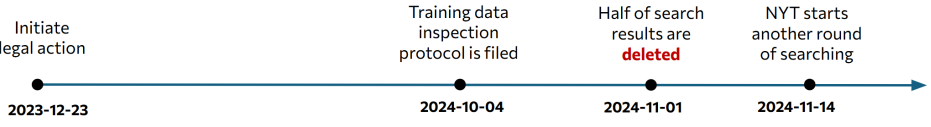

Figure 1: Timeline of the lawsuit between New York Times and OpenAI.

In this paper, we propose a solution called zkFinGPT, which uses zero-knowledge proof (ZKP) [4] to make the inference process of LLMs publicly verifiable. We demonstrate the use of zkFinGPT in three financial use cases: 1) the copyright lawsuit between NYT and OpenAI [5], 2) the credibility of the evaluation of LLMs on copyrighted exam questions, and 3) the protection of trading strategies in the Financial Reinforcement Learning (FinRL) Contest [6]. The zkFinGPT verifies the legitimacy of

Submitted to Workshop on Generative AI in Finance, 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

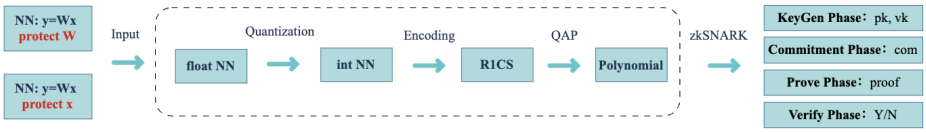

Figure 2: Tool chain of zkFinGPT.

the model and the credibility of its outputs. The experiment results show that zkFinGPT achieves fast verification, compact proof size, and zero data leakage.

2 Financial Use Cases

2.1 Case I: Copyright Lawsuit between New York Times and OpenAI

There has been a legal battle between AI companies and publishers, the most famous of which is the lawsuit between the New York Times (NYT) and OpenAI [5]. In December 2023, NYT sued OpenAI and Microsoft, accusing them of using millions of articles to train ChatGPT without permission. During the trial, the court required OpenAI to establish two servers as a "sandbox", which NYT lawyers can access remotely. NYT lawyers conducted further experiments by examining the model's training data. However, as shown in Fig. 1, OpenAI engineers "accidentally" deleted the operation logs on the servers [7], which stalled the trial process.

Since model parameters are the intellectual property of AI companies, collecting evidence and conducting trials face great challenges. A solution is needed to confirm the authenticity of model outputs generated by specific prompts and to exclude the possibility of human tampering. Such a solution should verify the inference process, while keeping the model parameters private.

2.2 Case II: FinGPT's Results on Copyrighted Exam Questions

As AI advances, reliance on neural networks (NN) grows alongside concerns about data privacy. In finance, investors seek to reveal FinGPT's outputs without exposing transactions; in healthcare, patients want AI diagnoses without disclosing sensitive data. However, private data makes inference costly or infeasible, so evaluations are often accepted at face value. In both academia and industry, results are frequently reported without releasing data, limiting credibility and trustworthiness.

The inference process remains a "black box" to outsiders, which cannot be achieved by third parties or is costly. To make inference results more convincing, a solution is needed that allows public verification of the results while protecting sensitive input data, say testing copyrighted exam questions.

2.3 Case III: Protection of FinRL Contest's Trading Strategies

The FinRL Contest is a series of open competitions encouraging participants to apply reinforcement learning in stock and cryptocurrency trading tasks [6]. It provides well-designed tasks, abundant financial datasets, near-realistic market environments, and useful starter kits, attracting 230+ participants from 20+ countries and 100+ institutions. The contests also introduce tasks of LLM-engineered signals, which use LLMs to generate and engineer alpha signals for trading.

In the FinRL Contest, however, only 10.6% of the participants came from the industry. Data privacy and model confidentiality are their main concerns. Participants have to share their trading strategies with contest organizers for evaluation, which could erode arbitrage opportunities or reduce effectiveness. To engage more industry participants, we will use privacy-preserving strategy verification. This ensures convincing results and also protects the sensitive information of participants.

3 The Proposed zkFinGPT Solution

3.1 Overview of zkFinGPT

In the financial field, we propose zkFinGPT, which uses ZKP [8] to make model inference process publicly verifiable while preserving privacy. It supports financial use cases such as verifying model

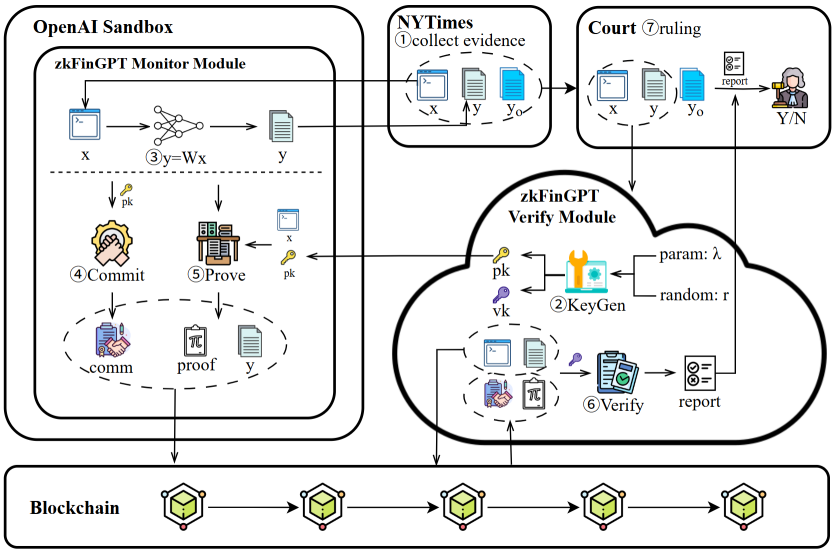

Figure 3: Illustration of zkFinGPT for case I.

legitimacy and output credibility. As shown in Fig. 2, zkFinGPT works by quantizing a NN (e.g. $y = Wx$) into integers over a finite field, then transforming it into an arithmetic circuit (e.g. R1CS) and finally representing it as a polynomial for zkSNARKs [9].

The solution consists of two modules: Monitor (Commit & Prove), which tracks inference and generates proofs; Verify (KeyGen & Verify), deployed in the cloud for security. For example, consider a simple LLM circuit C with only one layer ($y = Wx$) and no activation or normalization. zkFinGPT can verify its inference while protecting W and x.

- zkFinGPT.KeyGen(1^λ, r): Run (pk, vk) ← zkFinGPT.KeyGen(1^λ, r) and outputs proving key (pk) and verification key (vk).
- zkFinGPT.Commit(pk, C): Run $comm$ ← zkFinGPT.Commit(pk, C) and outputs commitment file $comm$.
- zkFinGPT.Prove(pk, C, u): Run (v, π) ← zkFinGPT.Prove(pk, C, u) and outputs v. u serves as a challenge point to open inference process, generating proof file π
- zkFinGPT.Verify (vk, $comm$, π, u, v): Run Accept/Reject ← zkFinGPT.Verify (vk, $comm$, π, u, v) and outputs Accept or Reject.

3.2 zkFinGPT for Case I

We use zkFinGPT as third-party software to verify the inference log file and protect the model parameters W, as shown in Fig. 2 ("protect W"). zkFinGPT provides a basis for courts to confirm the authenticity of model output, as shown in Fig. 3.

First, the NYT collects evidence, where the prompt x is the input and y_0 is the original allegedly plagiarized NYT article. Similarly, the court directs OpenAI to set up a server hosting the GPT-4 model as a "sandbox", allowing NYT lawyers to Supervise remotely. In this sandbox, OpenAI performs inference, commit, and prove. The records (x, y), $comm$, π are stored on a blockchain in the distributed system. The new evidence (x, y, y_0) is submitted by NYT to the court. During the trial, the court sends (x, y) to the zkFinGPT verify module. zkFinGPT uses vk, x, $comm$, π to check the trustworthiness of y. The result supports the court to judge whether OpenAI is engaging in fraud. By the zero-knowledge property of ZKP, neither NYT, zkFinGPT, nor the court can learn any information about W from $comm$ and π. By adopting the zkFinGPT, there is no need to worry about OpenAI engineers deleting server records. This makes the trial smoother and more reliable.

3.3 zkFinGPT for Case II

In case II, let $y = Wx$, where W is the FinGPT model parameters, x is the copyright exam set, and y is the inference result. As shown in Fig. 2 ("protect x"), zkFinGPT can prove that the output

Table 1: Overhead of zkLLM on LLaMa-2.

Model	Layer num	Prover time (s)	Verifier time (s)	Commit time (s)	Commitment size (MB)	t_p	t_v	t_c
Llama-2-7B	32	620	2.36	531	7.97	19.375	0.417	75.857
Llama-2-13B	40	803	3.95	986	11.0	20.075	0.625	75.846
Llama-2-70B	80	1578	4.66	5310	25.35	19.725	0.521	75.852

is inferred by the claimed model and dataset without disclosing sensitive input data. It makes the published inference results more convincing.

The publisher of the inference result commits to and proves to the inference process, obtaining $comm$ and π. These log files can be used by zkFinGPT to verify the trustworthiness of y. The verifier decides whether to accept y based on the verification result. This solution enhances the trustworthiness of the model inference results on private datasets, making verification easy and low-cost.

3.4 zkFinGPT for Case III

In Case III, the strategy testing process is abstracted as $y = Wx$, where W is the optimal strategy submitted by the participant, x is the task set provided by the contest, and y is the test result. We adopt the zkFinGPT to allow more investors to verify the strategy while protecting participant's interests. This solution ensures the privacy of the strategy and makes its effectiveness convincing, as shown in Fig. 2 ("protect W").

FinRL contest commits to and proves to the testing process, obtaining $comm$ and π. These log files allow zkFinGPT to verify the trustworthiness of y. Investors decide whether to accept y based on the verification results. zkFinGPT allows more investors to verify the effectiveness of the strategy while preventing it from being directly exploited by investors, fully protecting the interests of participants.

4 Experimental Results for Computational Overhead

4.1 Experiment Settings

Software. Our work is primarily based on the zkllm [4] and zkGPT [10] packages, implemented in CUDA C++. We use the GKR protocol [11] and the Pedersen polynomial commitment scheme [12].

Hardware. Tests ran on a server with 24-core Intel Xeon Platinum 8255c (2.494 GHz, 30GB RAM) and an NVIDIA Tesla T4 (16 GB).

Models and Datasets. Models include LLaMa-2 and FinGPT. Datasets include copyrighted exam questions and financial data.

4.2 Computational Overhead

Applying ZKP to large LLMs incurs high computational costs. Table 1 shows zkLLM overhead on Llama-2 models. Here, t_p is proving time per layer, t_v is verification time per layer, and t_c is committing time per billion parameters.

For an LLM with N layers and size M (in billions), the results indicate the proving time complexity is $O(N)$, from 620s for 32 layers (7B) to 1578s for 80 layers (70B). The verification time complexity is $O(\sqrt{N})$, which remains efficient, below 5s for all models. Commitment time complexity is $O(M)$, from 531s (7B) to 5310s (70B), with proof sizes of 7.97 to 25.35MB. The current zkLLM uses 16-bit quantization, leaving room for optimization. Future work will focus on improving the performance of zkFinGPT, scaling to larger LLMs, and reducing prove time and proof size.

5 Conclusions

This paper proposes the zkFinGPT solution, based on ZKP. This solution can prove the credibility of the LLM inference process without compromising the privacy of model parameters or input data. It has promising application prospects in three financial use cases. Future research will utilize quantization techniques and GPU-based massively parallel computing to improve zkFinGPT performance and reduce proof time and file size.

References

[1] Diana C. Milton and Harrison A. Enright. Case tracker: Artificial intelligence, copyrights and class actions. *BakerHostetler*, 2025. Accessed: 2025-08-21.

[2] OpenAI authors. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

[3] Shengyuan Colin Lin, Felix Tian, Keyi Wang, Xingjian Zhao, Jimin Huang, Qianqian Xie, Luca Borella, Christina Dan Wang Matt White, Kairong Xiao, Xiao-Yang Liu Yanglet, and Li Deng. Open finllm leaderboard: Towards financial ai readiness. *International Workshop on Multimodal Financial Foundation Models (MFFMs) at 5th ACM International Conference on AI in Finance*, 2024.

[4] Haochen Sun, Jason Li, and Hongyang Zhang. zkllm: Zero knowledge proofs for large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 4405–4419, 2024.

[5] The New York Times. The times sues openai and microsoft over a.i. use of copyrighted work. `https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html`, December 2023. Accessed: 2025-08-21.

[6] Keyi Wang, Nikolaus Holzer, Ziyi Xia, Yupeng Cao, Jiechao Gao, Anwar Walid, Kairong Xiao, and Xiao-Yang Liu Yanglet. FinRL Contests: Benchmarking data-driven financial reinforcement learning agents. *Wiley Artificial Intelligence for Engineering*, 2025.

[7] WIRED. New york times says openai erased potential lawsuit evidence. `https://www.wired.com/story/new-york-times-openai-lawsuit-evidence/`, 2024. Accessed: 2025-08-21.

[8] S Goldwasser, S Micali, and C Rackoff. The knowledge complexity of interactive proof-systems. In *Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing*, STOC '85, page 291–304, New York, NY, USA, 1985. Association for Computing Machinery.

[9] Boyuan Feng, Lianke Qin, Zhenfei Zhang, Yufei Ding, and Shumo Chu. ZEN: An optimizing compiler for verifiable, zero-knowledge neural network inferences. Cryptology ePrint Archive, Paper 2021/087, 2021.

[10] 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 *34st USENIX Security Symposium (USENIX Security 25)*, 2025.

[11] Shafi Goldwasser, Yael Tauman Kalai, and Guy N. Rothblum. Delegating computation: Interactive proofs for muggles. *J. ACM*, 62(4), September 2015.

[12] T. P. Pedersen. Non-interactive and information-theoretic secure verifiable secret sharing. In J. Feigenbaum, editor, *Advances in Cryptology – CRYPTO '91*, volume 576 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 1992.