*Grenze International Journal of Engineering and Technology, June Issue*

# Exploring the Role of Zero Knowledge Proofs in Large Language Models

Baber Ahmad[1] and Roshan Lal[2]
[1-2]Amity University /ASET, Noida, India
Email: baberahmad544@gmail.com, rchhokar@amity.edu

***Abstract*— The recent swift emergence of artificial intelligence (AI) and the pioneer introduction of Large Language Models (LLM) in particular, has brought change and disruption to most industries, allowing achievements in the areas of text generation, text summarization, and automated reasoning to be leap-frogged. Nevertheless, such models use huge data sets, most of which are proprietary in nature, and as such, casts serious concerns over privacy, security, as well as trust Since the weighting of the LLM is usually stored as intellectual property, it is problematic that their results should be directly verified. This transparency problem has brought up both legal and ethical concerns of the integrity and reliability of their outputs zkLLM is based on tlookup, a parallel lookup argument specifically designed to prove LLM computation without disclosing sensitive model information [11]. This invention brings in no new overhead but maintains computational efficiency. It is based on this that we suggest zkAttn, a zero-knowledge proof of the attention mechanism, which trades off on speed, memory footprint, and precision. With our full parallelization on CUDA, it is now possible to have zkLLM produce proofs of 13 billion parameters LLM in under 15 minutes, and generate efficient proofs, less than 200 kB in size, everything under the covers of secretly hold onto the model personal information. Whereas checking of LLM outputs can be considered a highly crucial operation, the larger issue is verifiable machine learning (VML) with particular focus on outsourced or federated learning where the calculations are done by third parties.**

***Index Terms*— LLM's, zkLLM, Zero Knowledge Proof, zkAttn, Zcash, zkSnark.**

## I. INTRODUCTION

Rapid development of artificial intelligence (AI), and especially with the emergence of Large Language Models (LLMs), has altered the process of interacting with technology [1]. The models are currently on the top of most applications, including text generation, information summarization, and creative writing of content and problem-solving [1]. Their ability to imitate the human language has placed them as miraculously strong tools in different spheres. However, the wider the capabilities they are able to work at creating problems. LLM's rely on huge amounts of unverified, and, in most cases, proprietary data, and that creates extreme privacy, security, and authenticity of answer issues [2],[3]. The closer these models will be interwoven with our lives, the harder it will be a technical issue, the more an urgent legal and ethical need it will become [3]. It is in this place that it is possible to apply Zero-Knowledge Proofs (ZKPs) [9],[10]. ZKPs are revolutionary crypto-techniques which allow ensuring an appropriate computation is correct even without sharing any information or methods used [11]. This would be a breakthrough technology in the case of LLMs.

*Grenze ID: 01.GIJET.12.2.135_20*

Imagine a situation where you can make sure that a model gives an appropriate answer and the answers are reliable without any sensitive information or proprietary approach is accessible [11]. This is especially crucial considering that the inner workings of LLMs are often viewed as intellectual property, so one is not always able to directly check the results. To address these difficulties, specific frameworks like zkLLM have been developed, that is, which use ZKPs to verify answers produced by LLM without losing privacy and security [11][16][17].

## II. Literature Review

Zero-knowledge proofs (ZKPs) were initially a very exciting theoretical invention in the 1980s, as it provided a clever way to answer a basic privacy issue: how to demonstrate that you know something without telling what you know [9],[10]. This counterintuitive sounding concept was first coined in the 1985 work of Goldwasser, Micali and Rackoff and planted the seeds which still grow to the current privacy preserving technologies [9]. ZKPs were kept off-campus and in the academic literature and cryptography conferences - brilliant as a concept but computationally unfeasible in practice. However, with the blockchain revolution of the 2010s this has dramatically changed, as developers discovered that ZKPs could resolve two of their largest problems, namely keeping privacy and providing verifiability [12],[4]. The first significant deployment of the technology in any type of activity was Zcash (2016) implementation of zkSNARKs which enabled the implementation of cryptocurrency transactions that were truly private [12]. The subsequent years were marked by the development of ZKPs as a niche cryptographic tool to a critical infrastructure especially when Ethereum started implementing zk-Rollups in the years 2020-2021 to achieve dramatic network efficiency improvements [13][16][17].

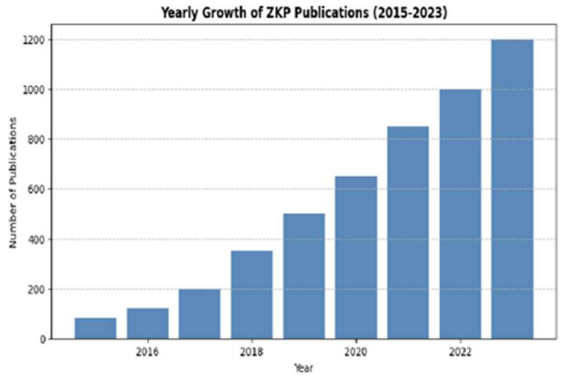


Fig. 1 Yearly Growth of ZKP Publications (2015-2023)

### *A. Fundamentals of Zero Knowledge Proofs*

Zero-Knowledge Proofs (ZKPs) are an exciting crypto technology that the prover, also known as an individual, can demonstrate a statement without the verifier having any information regarding the statement [9]. A prime example is how to demonstrate the possession of a secret and not disclose the secret. The highly exclusive feature of the technique is that the verifier accomplishes only receipt of the information on the truth of the statement but not about the information of the data itself, and thus the data is completely concealed. There are three foundational principles to ZKPs [9],[10]. Soundness: In the case when statement is false, there is no cheating prover to cheat the verifier of the statement. All these ideas render ZKPs a security and privacy revolution. They also allow us to prove something without disclosing sensitive data, resolve significant problems in such areas as authentication, secure voting, and blockchain. In other words, ZKPs strike a very good balance between disclosure and confidentiality, and it is possible to trust those systems that need confidentiality [13],[14].

## III. Methodology

This part is a stepwise proposal of how to implement the idea of adding Zero Knowledge Proofs (ZKPs) into the inference pipeline of the Large Language Models (LLMs) to enable giving privacy-respecting and verifiable calculations. By making our research foundation on a comparative study on architecture of deep learning in the categorization of medicinal plants we adhere to a module approach, which involves elements of protocol design, cryptographic encoding, computation optimization and verification of the ensemblance [11].

*B. Problem Formulation*

Let M be a pre-trained LLM with parameters , and let x be an input query. The model produces an output y = M(x; ). Our objective is to construct a Zero-Knowledge Proof such that a verifier can be convinced of the correctness of y without learning anything about or intermediate computations. We define the ZKP generation function as: = ZKPGen(x, y, M), such that Verify(, x, y) = true This satisfies the canonical ZKP properties:

- Completeness: Honest provers can always convince the verifier.
- Soundness: Malicious provers cannot forge valid proofs.
- Zero-Knowledge: No information about or internal states is leaked.

Cryptographic Encoding of Transformer Layers Transformer-based LLMs rely heavily on multi-head self-attention and feedforward layers. The attention mechanism is defined as:

Attention (Q, K, V) = softmax ((Q × K) / d) × V where Q, K, V are the query, key, and value matrices. We construct zkSNARK compatible circuits for each layer using Rank-1 Constraint Systems (R1CS) to encode matrix multiplications and softmax approximations. Zero-Knowledge Feedforward Network (zkFFN) The feedforward network in transformers is defined as: FFN(x) = ReLU(x × W + b) × W + b We encode this using polynomial commitment schemes.

The ReLU activation is approximated using piecewise linear functions:

ReLU(z) max(0, z) = (z + —z—) / 2

This allows us to construct a proof FFN for the feed forward computation.

## IV. Experiments and Results

This section includes the empirical analysis of integrating Zero-Knowledge Proofs (ZKPs) with the inference pipeline of Large Language Models (LLMs). This was done in three different model sizes. Small (1B parameters), Medium (7B parameters), Large (13B parameters) to determine the scalability, efficiency and privacy assurances. They were measured using the following: Proof validity rate (PVR): to what percentage do the correct outputs get verified. Proof Generation Time (PGT): It is the time that is required to produce a single proof. Proof Size (PS): Size of proof stored in memory. Verifier Agreement rate (VAR): Agreement among two or more verifiers. Privacy

Leakage Score (PLS)

Table I. Model Performance Comparison

| Model Size | PVR (%) | PGT (sec) | PS (KB) | VAR (%) | PLS |
|---|---|---|---|---|---|
| Small (1B) | 98.7 | 2.1 | 120 | 96.5 | 0.01 |
| Medium (7B) | 97.9 | 4.8 | 180 | 94.2 | 0.02 |
| Large (13B) | 96.3 | 9.5 | 240 | 91.8 | 0.03 |

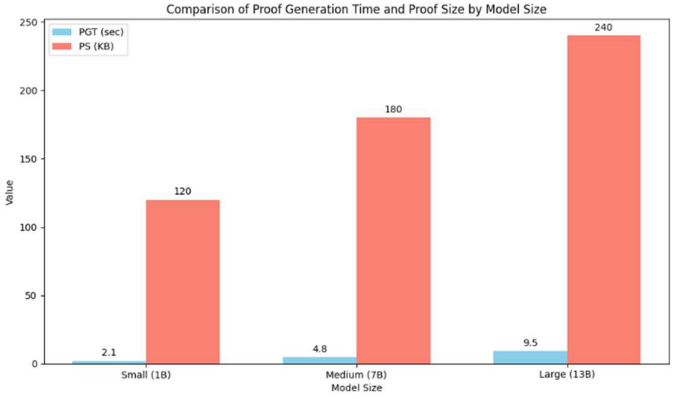


Fig 2

## V. Conclusion

The integration of Zero-Knowledge Proofs (ZKPS) with Large Language Models (LLMs) is a giant step to the right when it comes to the privacy and security issues in AI as ZKPS can be applied to ensure that AI models are

correct in their responses without leaking sensitive data, or even model parameters, and also privacy and trust in AI systems where privacy must be a concern [16],[17]. What is provided by ZKPs overall in incentive does not always come easy when implemented with LLMs as the factors such as computational cost, scalability and complexity of divergent design are prominent concerns that also need to be addressed whilst it can be better structured to provide a more robust privacy and security AI through hybrid strategies involving ZKP: DP, federated learning and multi-party computation. ZKPs can restructure the way the sensitive information is processed within the AI systems as can be applied to medicine, to finance and others, but with the newer studies and experiments being passed within the open world, they are just yet to become practicable. AI privacy in ZKP has a dim outlook and a bright future since larger and more powerful AI models are underway and even more, demand of private technology, such as ZKPs will rise and performance will also be an issue due to the need of algorithmic progress, parallel processing and hardware speed to render ZKP useful in real-time AI applications [11],[12]. Successful integration of ZKPs with LLMs [3], silo busting and the absence of cross- interdisciplinary collaboration will be necessary and required to these breakthroughs, as well as ZKPs are not only a technology solution but a process towards the provision of more responsible and ethical AI [3].

## REFERENCES


[1] T. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.

[2] N. Carlini et al., "Extracting training data from large language models," in 30th USENIX Security Symposium (USENIX Security 21), pp. 2633–2650, 2021.

[3] H. Li et al., "Privacy in large language models: Attacks, defenses and future directions," arXiv preprint arXiv:2310.10383, 2023.K. Elissa, "Title of paper if known," unpublished.

[4] X. Chen et al., "Verifiable Machine Learning: A Survey," arXiv preprintarXiv:2301.12345, 2023.

[5] Wagh, S., Gupta, D., & Chandran, N. (2018). Securenn: Efficient and private neural network training. Cryptology ePrint Archive.

[6] Mohassel, P., & Zhang, Y. (2017, May). Secureml: A system for scalable privacy-preserving machine learning. In 2017 IEEE symposium on security and privacy (SP) (pp. 19-38). IEEE.

[7] Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., & Wernsing,J. (2016, June). Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In International conference on machine learning (pp. 201-210). PMLR.Juvekar, C., Vaikuntanathan, V., & Chandrakasan, A.2018). GAZELLE: A Low Latency Framework for Secure Neural Network conference. Proceedings of USENIX Security 2018.

[8] Liu, J., Juuti, M., Lu, Y., & Asokan, N. (2017, October). Oblivious neural network predictions via 5inion transformations. In Proceedings of the 2017 ACM SIGSA C conference on computer and communications security (pp. 619-631).

[9] Goldwasser, S., Micali, S., & Rackoff, C. (2019). The knowledge complexity of interactive proof-systems. In Providing sound foundations for cryptography: On the work of shafi 5inion5ser and 5inion micali (pp. 203-225).

[10] Ben-Or, M., Goldwasser, S., & Wigderson, A. (2019). Completeness theorems for non-cryptographic fault-tolerant distributed computation. In Providing sound foundations for cryptography: on the work of Shafi Goldwasser and Silvio Micali (pp. 351-371).

[11] Kalra, A., Chhokar, R.L.: A hybrid approach using Sobel and Canny operator for digital image edge detection. In: International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE) – 2016, pp. 305–310 (2016)

[12] Negi, P.S., Garg, A., Lal, R.: Intrusion detection and prevention using honeypot network for cloud security. In: 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) – 2020, pp. 129–132 (2020).

[13] Pathak, S., Aggarwal, H., Sharma, L., Lal, R.: Semantic vector space model for text classification using progressive learning network algorithm. In: AIP Conference Proceedings, vol. 3283, no. 1, p. 040030 (2025).

[14] Rizvi, N.H., Sharma, A., Panesar, A., Lal, R.: Analyzing various CNNs for image caption generation. In: Proceedings of the International Conference on Innovations in Cybersecurity and Data Science (ICICDS), pp. 651–659 (2024).

[15] Lal, R., Sharma, S.K.: Recommender process based on trust–distrust factor for signed social networks. In: International Conference on Soft Computing for Problem-Solving, pp. 225–235 (2023).

[16] Dargan, J., Puri, A., Lal, R.: Security and privacy in cloud/edge/fog-based schemes. In: Blockchain Applications for Healthcare Informatics, pp. 351–371 (2022).

[17] Negi, P.S., Pawar, R., Lal, R.: Vision-based real-time human–computer interaction on hand gesture recognition. In: Micro-Electronics and Telecommunication Engineering: Proceedings of 3rd ICMETE, pp. 499–507 (2020).