Enhancing Privacy and Security in Large-Language Models: A Zero-Knowledge Proof Approach

Shridhar Singh
University of KwaZulu-Natal, Westville, South Africa

217008024@stu.ukzn.ac.za

Abstract: The explosive growth of Large-Language Models (LLMs), particularly Generative Pre-trained Transformer (GPT) models, has revolutionised fields ranging from natural language processing to creative writing. Yet, their reliance on vast, often unverified data sources introduces a critical vulnerability: unreliability and security concerns. Traditional GPT models, while impressive in their capabilities, struggle with limited factual accuracy and susceptibility to manipulation by biased or malicious data. This poses a significant risk in professional and personal environments where sensitive or mission-critical data is paramount. This work tackles this challenge head-on by proposing a novel approach to enhance GPT security and reliability: leveraging Zero-Knowledge Proofs (ZKPs). Unlike traditional cryptographic methods that require sensitive data exchange, ZKPs allow one party to convincingly prove the truth of a statement, without revealing the underlying information. In the context of GPTs, ZKPs can validate the legitimacy and quality of data sources used in GPT computations, combating data manipulation and misinformation. This ensures trustworthy outputs, even when incorporating third-party data (TPD). ZKPs can securely verify user identities and access privileges, preventing unauthorised access to sensitive data and functionality. This protects critical information and promotes responsible LLM usage. ZKPs can identify and filter out manipulative prompts designed to elicit harmful or biased responses from GPTs. This safeguards against malicious actors and promotes ethical LLM development. ZKPs facilitate training specialised GPT models on targeted datasets, resulting in deeper understanding and more accurate outputs within specific domains. This allows the creation of 'expert-GPT' applications in specialised fields like healthcare, finance, and legal services. The integration of ZKPs into GPT models represents a crucial step towards overcoming trust and security barriers. Our research demonstrates the viability and efficacy of this approach, with our ZKP-based authentication system achieving promising results in data verification, user control, and malicious prompt detection. These findings lay the groundwork for a future where GPTs, empowered by ZKPs, operate with unwavering integrity, fostering trust and accelerating ethical AI development across diverse domains.

Keywords: Zero-Knowledge Proof (ZKP), Succinct Non-interactive Argument of Knowledge (SNARK), Large-Language Model (LLM), Generative Pre-trained Transformer (GPT)

1. Introduction

Generative Pre-trained Transformer (GPT) architecture advanced the field of artificial intelligence by allowing Large-Language Models (LLMs) to recognise patterns in their training data and generate new content based on those characteristics (Shi et al, 2023). However, the rapid growth of GPT LLMs flooding the industry exposed weaknesses affecting LLMs architecture. These weaknesses target the objectivity, trust, and reliability of LLMs thereby reducing the trustworthiness of responses. This increases the LLMs susceptibility to exploits poisoning training data and malicious prompt engineering tactics that manipulate the LLM into divulging sensitive information or delivering harmful responses (Shi et al, 2023, Wang et al, 2023, Shen et al, 2023).

Further, limitations in the lack of domain-specific data created the need to supplement the base training data to gain access to up-to-date information and ability to use post-training alignment, creating 'expert-GPT' applications, making the LLM more useful in domain-specific applications (OpenAI, 2023). Thus, third-party data (TPD) is injected to supplement the LLMs knowledgebase either from internet sources or, in privatised use-cases, local knowledgebases (OpenAI, 2023). Since TPD are the preferred source when the LLM computes its response, this increases the risk of vulnerability leading to unreliable responses (Shi et al, 2023).

Kang et al (2023) investigate the instruction-following procedure of LLMs stating it has a dual-use with capabilities for malicious or nefarious behaviour. Their study concludes that attacks can bypass the state-of-the-art content filtering processes LLMs are equipped with and call for a formalised defence mechanism against threats to LLMs. Ahmed & Kashmoola (2021) state that data poisoning is a concern in the accelerated adoption of AI technologies and the risks associated with such attacks targeting Deep Neural Networks (which are utilised by LLMs) pose a significant risk during the training process. In their experiments, malicious data was utilised to falsify results and reduce the accuracy of outcomes. The difficult detection of these complex "smart model" attacks allows them to cause significant damange to the AI model, handing over control to an unauthorised party. They conclude that the reliability level of sources must be increased through the use of mathematical and statistical methods.

Therefore, these limitations pose a gap in the current literature and require further research. The expansion to incorporate more data and utilise GPT architecture in business environments is inevitable and increases the risks of information tamper and misrepresentation. Thus, the consequences of unreliable data or unauthorised users can have widespread repercussions leading to spread of misinformation, creation of deep-fakes and other harmful content, exposure of sensitive or mission-critical data, and database infection and corruption, among others. It is vital that this gap be filled to ensure authenticity of LLM computations.

Himeur et al (2022) explore a method to protect sensitive information mined by recommender systems using blockchain technology. In a similar regard, we look to replicate some of this functionality by utilising a proof system that has been popularised by Blockchain Technology, Zero-Knowledge Proofs (ZKPs) (Sun et al, 2021). ZKPs are a modular and scalable proof system that can verify transactions without leaking any sensitive data. ZKPs also ensure that once a transaction is complete, there is a mechanism to validate the authenticity of the transaction.

In this paper we combine the progress in Zero-Knowledge (zk) cryptography and GPT AI models to determine a solution to the challenges mentioned above. The research has not examined the use of Zero-Knowledge Proofs for LLM validation. Additionally, there has been no work to date suggesting examining TPD supplied to an LLM. This is a massive gap in the knowledgebase as LLMs trained on data that is falsified can generate untrustworthy responses (Kang et al, 2023).

1.1 Research Goals

This paper aims to introduce zk-based LLMs (zk-LLMs) as a viable solution to challenges faced by the LLM industry. To achieve this, we highlight 3 research questions we feel are beneficial in addressing why zk-LLMs are viable and provide backing through experimentation in this regard. This paper also sets out future work and limitations to encourage future expansion of this concept.

Our goal with this research is to create a framework for developing zk-LLMs. We achieved this by creating a prototype zk-LLM application using this framework that will serve as our basis for experimentation. This zk-LLM will demonstrate the capabilities and variations that ZKPs can unlock when authenticating and securing operations performed by LLMs.

This framework applies the zk prototype to privatised LLM environments where more flexibility is given to conduct experiments. As such, the research acknowledges the current advancements in LLM attack detection by AI leaders and proposes a different approach surrounding ZKP-equipped LLMs.

Research questions (RQs)

RQ1. What factors can contribute as secure authentication methods to ensure prompt and user integrity and reliability through ZK Protocols?

RQ2. What are the potential risks and challenges of implementing zero-knowledge proof techniques in LLMs, and how can they be mitigated?

RQ3. How can zero-knowledge proof techniques be used to validate the accuracy and trustworthiness of responses generated by LLMs?

2. Overview of Zero-Knowledge Proofs

Zero-knowledge Proofs (ZKPs) are cryptographic proofs used to prove to a verifier, *V,* that a prover, *P,* knows some knowledge or secret*,* without revealing that secret itself (Wu & Wang, 2014). This validation invokes a trust-based protocol whereby the prover must provide convincing evidence to justify their claim to know a secret, without revealing any sensitive information to the verifier (Lipmaa, 2016). This ideology is based on the notion that the verifier must not have gained any more knowledge after completing the proving procedure than it had before (Fiege et al, 1987; Rosen, 2004).

Techniques like Common Reference String (CRS) aim to provide a trusted third-party *TP* at the start of the validation process where *TP* can access privileged information and assure *V* that *P*'s claim is indeed valid (Groth, 2018). Further development of the CRS around generating the *TP* led to the introduction of the zero-knowledge Succinct Non-interactive Argument of Knowledge (zk-SNARK) protocol (Ben-Sasson et al, 2017; Groth, 2018) – which will be utilised to construct zk proofs in this research.

zk-SNARKs have significant interest in Blockchain (Groth, 2018) where authenticity validation is vital. zk-SNARKS has proven to be an efficient protocol without being computationally expensive (Ben-Sasson et al, 2017). zk-SNARKS can therefore lend its versatility to prove for users and sources being authentic, and data indeed representing relevant domain knowledge to aid in computing responses of LLMs.

3. Methodology

In section 1.1. above, we introduced the zk-LLM framework and application. This section briefly describes the process of constructing a ZKP for zk-GPT (our zk-LLM application), translating the high-level principles into computational code. We define a framework for constructing zk-LLMs and our intention with this framework is to provide a basis for our design and experiments whilst aiding future development.

3.1 ZKP Framework

Since this paper introduces zk-LLMs, there is no existing framework to create proofs or to conduct zk-experiments. Therefore, this section establishes a theoretical framework for constructing zk-LLMs, serving as a blueprint for our implementation in zk-GPT and paving the way for future developments.

Objective with the framework:

- Determine a standardised procedure for proving the legitimacy of a user.
- Determine the authorised addition of source data.
- Extract relevant source data to provide a response to user prompts.
- Prevent malicious actions targeting the LLM.

Table 1: Framework for creating zk-LLMs.

ZKP Framework for creating zk-LLMs	
	Description
User Authentication	*Function:* • Utilising ZKPs to identify users without revealing sensitive information (e.g. passwords) through mathematical formulas or data structures.
	Implementation: • Mathematical formulas or secure data structures for group membership or individual credentials. • ZKPs to prove knowledge of a secret key or membership in a certain group.
Prompt Analysis	*Function:* • Examining user prompts before LLM processing to assess relevance and prevent malicious content using secure hashing and encryption techniques.
	Implementation: • Secure hashing and encryption techniques for prompt analysis without storing sensitive data. • ZKPs to prove that a prompt meets predefined criteria (e.g. length, format, content restrictions).
Source Data Verification	*Function:* • Employing ZKPs to verify the integrity and trustworthiness of training data and source documents, ensuring anonymity while preventing malicious content.
	Implementation: • ZKPs to prove that data has been correctly processed and anonymised. • Verification restricted to designated users involved in data curation.
Source Data Relevance Filtering	*Function:* • Leveraging ZKPs to identify and prioritise the most relevant source data for LLM responses, improving accuracy and reducing unnecessary information overload.
	Implementation: • Scoring systems based on prompt-specific relevance, potentially using ZKPs to prove that the highest-scoring data was selected without revealing sensitive information about the scoring process.

3.2 zk-GPT Setup

zk-GPT builds upon the localGPT (PromptEngineer, 2024) platform to perform on-device LLM computations and utilises the Llama-2 7b-GPTQ model. Choosing stability and reliability over bleeding-edge efficiency guided our technology selection for ZKP implementation. This prioritises security and user privacy while ensuring verifiable computation.

Circom: We chose Circom (Iden3, 2023), a dedicated language for zk-SNARK circuits, for its succinct and efficient expression of constraints. This makes the zk-circuit compact and optimised for proof generation and verification.

Proof System Selection: Established protocols like Groth16 and Powers of Tau were preferred over cutting-edge alternatives. Groth16 offers cryptographic soundness and succinct proofs, while Powers of Tau facilitates efficient ceremony execution. This combination prioritises proven security and reliability over potential efficiency gains from newer protocols.

Circuit Implementation: SnarkJS, a companion tool bundled with Circom, plays a crucial role. It binds the zk-circuit using the R1CS (Rank-1 Constraint) vector, ensuring data integrity and tracking variable relationships. Furthermore, SnarkJS handles Groth16 witness generation in a trusted setup. This trusted setup allows for secure proof generation without compromising the zk-circuit's secrecy.

Proof Generation & Verification: Following circuit binding and witness generation, SnarkJS performs Powers of Tau ceremonies alongside Groth16 proof generation and verification. This culminates in an exportable verification key, enabling anyone to independently verify the proof without needing the original circuit or witness.

Table 2: Breakdown of the zk-SNARK process within zk-GPT.

zk-SNARK process within zk-GPT	
Step	**Description**
zk-circuit	Filters legitimate values from illegitimate ones, ensuring valid calculations and rejecting any unexpected outcomes.
Proof System Selection	Groth16 and Powers of Tau protocols chosen for their established security and reliability.
Circuit Implementation	Circom and SnarkJS facilitate circuit creation and efficient proof generation/verification.
Binding & Witness Generation	SnarkJS binds the circuit for data integrity and generates Groth16 witnesses in a trusted setup.
Proof Generation & Verification	SnarkJS performs Powers of Tau ceremonies and Groth16 verification, exporting a verification key.
Completion	Successful verification allows LLM operation to resume.

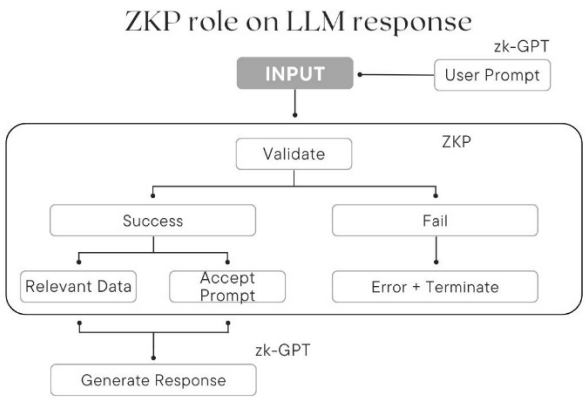

Figure 1: A depiction of the process for a user's prompt interacting with zk-GPT architecture

Trade-offs: While established technologies offer proven security and reliability, we acknowledge the potential efficiency gains achievable with cutting-edge options. Future research could explore utilising newer protocols while maintaining zk-GPT's core security guarantees.

3.3 Testing Methodology

This section details three sets of experiments designed to address the RQs outlined in section 1.1:

- **Stage 1 – User Authority Analysis**

Stage 1 experiments examined two RQs (RQ1 – Secure Authentication and User Integrity and RQ2 – Risks and Mitigations). This stage tests the ability of ZKPs to authenticate users and differentiate authorised access levels. It aims to verify user IDs, prevent unauthorised actions, and detect intruders while preserving user anonymity in closed systems. This directly addresses RQ1 by demonstrating how ZKPs can contribute to secure authentication and user integrity. This stage also plays a role in mitigating potential risks identified in RQ2 by preventing unauthorised access and data disclosure.

- **Stage 2 – Supplemental Data Relevance**

This stage focuses on addressing RQ3 by showing how ZKPs can validate the accuracy and trustworthiness of LLM responses. It demonstrates how the system filters irrelevant information and ensures the LLM utilises only relevant supplemental data for response generation, thereby increasing the reliability and factual basis of outputs.

- **Stage 3 – Risks and Mitigations**

This stage directly tackles RQ2 by demonstrating how ZKPs can detect and prevent malicious user actions that aim to harm the LLM or compromise its responses. Examples could include attempts to force sensitive information disclosure or manipulate the LLM's outputs.

4. Results

4.1 User Authority Analysis (RQ1)

RQ1: What factors can contribute as secure authentication methods to ensure prompt and user integrity and reliability through ZK Protocols?

Stage 1 experiments assessed the factors that contribute to secure authentication of prompt and user integrity and reliability by evaluating the effectiveness of ZKPs in securing user authentication and role-based access control in zk-GPT. These experiments were conducted over 100 iterations and these tests aimed to:

- Verify user IDs and differentiate authorised access levels (admin vs. normal user).
- Prevent unauthorised login attempts and protect sensitive data.
- Maintain user anonymity in closed systems.

The zk-circuit successfully identified and rejected all unauthorised attempts (40/40), including those designed to mimic admin or normal user access (15 admin, 15 normal). Authorised users (30 admin + 30 normal) consistently experienced successful logins and privilege separation (modifying supplemental data exclusively for admins). These results strongly support the use of ZKPs for secure authentication and access control in LLMs.

Table 3: User Authority Analysis over 100 Iterations

User Authority Analysis		
User Level Description	**Successful Login**	**Unsuccessful Login**
Admin User	30	15
Normal User	30	15
Foreign User	0	10
Total Attempts	60	40

4.2 Supplemental Data Relevance (RQ2)

RQ2: What are the potential risks and challenges of implementing zero-knowledge proof techniques in LLMs, and how can they be mitigated?

Stage 2 assessed zk-GPT's ability to utilise relevant source data based on user prompts. We conducted 80 experiments using a sample set of 40 research papers across four domains (Zero-Knowledge Proofs, Generative AI, Recommender Systems, Blockchain). Each experiment retrieved a user prompt and identified relevant papers for response generation. These relevant papers were embedded and used to supplement zk-GPT with domain-specific knowledge. The category for scoring this stage is as follows:

- **Related papers**: Accurately answered the prompt (8 for Zero-Knowledge Proofs, 7 for Generative AI, 3 for Recommender Systems, 4 for Blockchain).
- **Unrelated papers**: Used in computation but didn't contribute significantly to response generation (2 for Zero-Knowledge Proofs, 3 for Generative AI).
- **Hallucinations**: Papers not provided in the experiment, either retrieved by zk-GPT or generated (0 across all domains).

Table 4: Total Papers Sampled for Response Computation over 80 Iterations

Supplemental Data Relevance			
Prompt Parameters	**Related Papers**	**Unrelated Papers**	**Hallucinations**
Zero-Knowledge Proofs	8	2	0
Generative AI	7	3	0
Recommender Systems	3	0	0
Blockchain	4	0	0
Total Papers Referenced	22	5	0

Note that fewer papers were referenced than provided due to overlapping domain coverage within related subsets. zk-GPT effectively prioritised the most relevant papers based on the prompt.

4.3 Malicious Prompt Detection (RQ3)

RQ3: How can zero-knowledge proof techniques be used to validate the accuracy and trustworthiness of responses generated by LLMs?

Stage 3 focused on zk-GPT's resilience against malicious prompt injection attacks and ran for 60 iterations. Stage 3 tests its ability to detect prompts aiming to disclose sensitive information or manipulate its knowledge and prevent execution of such prompts, preserving LLM integrity. Stage 3 ran for 40 iterations and contained a dataset of 200 prompt injection keywords (tailored to the dataset) equipped the proof system to analyse prompts before execution.

- Malicious: 28/30 malicious prompts were flagged and execution terminated.
- Non-malicious: 27/30 non-malicious prompts were correctly identified.
- False positives/negatives: These results showed 3 false positive matches and 2 false negative matches, indicating an acceptable error rate for accidental flagging/missed detections.

Table 5: Total Prompts Investigated for Malicious Intent over 60 Iterations.

Malicious Intent				
Prompt Type	**Successful**	**Unsuccessful**	**False Positive**	**False Negative**
Malicious	28	2	-	2
Non-Malicious	27	3	3	-
Total Attempts	60	5	-	-

These experiments demonstrate the significant potential of ZKPs in securing LLMs. User authentication, data relevance filtering, and malicious prompt detection all demonstrated high accuracy and effectiveness. Further research can refine parameter optimisation and address complex scenarios, but these results pave the way for trustworthy and secure LLM deployment.

5. Discussion

This section delves into the implications of our experiments, dissecting the benefits and challenges of employing ZKPs in LLMs, as addressed by our three research questions (RQs).

5.1 Secure Authentication and User Integrity (RQ1)

ZKPs demonstrate how they enhance secure authentication practices and uphold user integrity in LLMs through two key mechanisms: zero-knowledge user identification, and role-based access control.

- **Zero-knowledge User Identification**

Instead of vulnerable password-based verification of traditional identification systems, zk-circuits verify user identity without revealing any sensitive information. This anonymises users in closed systems and prevents unauthorised access by enforcing proof-of-identification based on mathematical binary hash function that verifies the identity of a user without explicitly exposing the user. Thus, the zk-verifier gains no new knowledge about the user, other than what is already provided by the user-role.

- **Role-based Access Control**

Additionally, using a zk-circuit enforces access restrictions based on user roles and prohibit attempts to access zk-GPT made by users not belonging to any user-group. These restrictions extend to the functions users can perform whilst interacting with zk-GPT. This safeguards sensitive information and ensures that only authorised users can perform privileged actions. For example, admin-level users have additional privileges for interacting with source data which includes adding or modifying supplemental data to the knowledgebase whilst normal users are only permitted to chat with the LLM.

Our Stage 1 experiments achieved a 100% success rate in identifying and rejecting unauthorised user attempts whilst allowing access to authorised users. Thus, thereby validating the effectiveness of ZKPs for secure authentication and access control in LLMs.

5.2 Risks and Mitigations in LLMs (RQ2)

Identified as some of the potential risks of deploying ZKPs in LLMs were challenges in computational overhead, data availability and context, and malicious prompt injection.

- **Computation Overhead**

ZKP verification requires that the LLM halt execution and provide data to the zk-circuit. This requires additional computational resources, potentially impacting the fluidity of the LLMs responsiveness. Our experiments demonstrated a slight decrease in user experience due to embedded zk-proof checks during conversation flow.

- **Data Availability and Context**

Large datasets may pose challenges in retrieving the required amount of relevant information or understanding the context of a specific user prompt. We mitigated this by implementing a fixed context window to determine relevancy of source data based on initial prompts. However, this approach sacrifices analysing nuances in prompts for efficiency, requiring further research for handling diverse topics and refined information requests.

- **Malicious Prompt Injection**

Dishonest users might utilise prompt engineering tactics to craft prompts containing malicious instructions to manipulate the LLM. Our Stage 3 experiments demonstrated the feasibility of using ZKPs to detect such prompts with a 91.6% accuracy rate. Analysing prompts for malicious intent before execution allows for fine-tuning this defence mechanism based on organisational needs and user access levels.

5.3 Validating LLM Response Accuracy (RQ3):

ZKPs contribute to the accuracy and trustworthiness of LLM responses in several ways:

- **Data Relevance Filtering**

By analysing document relevance through word frequency and scoring, ZKPs ensure the LLM utilises only the most relevant data for response generation. This eliminates unnecessary information overload and reduces the risk of misleading or inaccurate responses.

- **Source Data Credibility**

Combining user authorisation with ZKP verification guarantees that only credible source data are added to the LLM's knowledge base. This minimises the risk of data corruption or falsification and improves the overall trustworthiness of responses.

- **Verification of Data Integrity**

ZKPs can assess whether source data has been altered after embedding, providing a reliable verification mechanism for data integrity. This strengthens trust in the LLM's responses by assuring users that the underlying information remains uncorrupted.

Our Stage 2 experiments demonstrated a significant response accuracy (81.4%) when using ZKPs for data relevance filtering. This demonstrates the effectiveness of ZKPs in enhancing the reliability and factuality of LLM outputs.

Overall, these experiments highlight the significant potential of ZKPs in addressing crucial security and trust concerns surrounding LLMs. While challenges remain in optimising performance and handling diverse scenarios, further research and development promise even greater effectiveness in securing LLMs and ensuring the integrity of their responses.

6. Limitations and Future Work

6.1 Limitations

While our experiments demonstrate the promise of ZKPs for securing LLMs, there are areas for improvement. Highlighted below are areas to improve the zk-circuit and should also be noted as developments in the field of Zero-Knowledge Proof computations.

- **Circuit flexibility**: The current zk-circuit accepts predefined inputs with fixed constraints. Any deviations require circuit modifications, limiting adaptability. To address this, we envision a generic circuit with variable input ranges capable of handling diverse data formats and structures.
- **Hashing optimisation**: The ZKP relies on SHA256 hashing, which becomes inefficient for larger circuits. Implementing Poseidon hashing, known for its speed and efficiency, could significantly improve proof generation and verification processes.
- **Trusted vs. untrusted setups**: Groth16 is the chosen algorithm for zk-SNARK circuits but requires a trusted setup with exposed witnesses. Untrusted setups like PLONK would eliminate witness exposure and enhance security without impacting verification.

6.2 Future Work

Future work in this area is promising as ZKPs are versatile and can be adapted to solve different problems. Thus, ZKPs extends beyond our present research, opening doors to exciting possibilities. Here are some proposed examples where ZKPs can prove beneficial to the development of LLMs:

- **Securing public internet data**: LLMs increasingly rely on web sources, but verifying data trustworthiness remains a challenge. TPD-based ZKPs can validate data provenance and ensure LLMs are fed with reliable information.
- **Zero-knowledge news verification**: Malicious news and propaganda pose a significant threat. Utilising ZKPs to authenticate news articles could safeguard LLM recommendations and promote responsible information dissemination.
- **ZKPs in education**: Ethical concerns surround the use of LLMs in education. Employing ZKPs for plagiarism detection and knowledge assessment could foster academic integrity and enhance learning experiences.

By addressing the limitations and exploring these promising avenues, we can further unlock the potential of ZKPs to revolutionise LLM security, reliability, and ethical applications.

7. Conclusion

This research explored the potential of Zero-Knowledge Proofs (ZKPs) to address critical trust concerns surrounding Large-Language Models (LLMs). We have demonstrated significant progress in mitigating challenges that erode user confidence in LLM responses through the use and development of zk-LLMs and a framework for utilising ZKPs for secure authentication, data validation, and prompt analysis. Our experiments yielded compelling evidence supporting the efficacy of zk-LLMs in addressing our key research questions and establishes the zk-LLM framework for secure LLM deployment in controlled environments. ZKPs effectively mitigated threats through data relevance filtering and malicious prompt detection reducing risk of data corruption, misinformation, and LLM manipulation whilst ensuring reliable source data is referenced to enhance the trust of LLM outputs.

While these results are promising, we acknowledge the limitations in the current zk-circuit and encourage opportunities for further refinement by providing avenues for future research such as: circuit flexibility, hashing optimisation, and exploring untrusted ZKP setups. Beyond internal optimisation, the potential of zk-LLMs extends its versatility to diverse applications, including securing internet data, news verification, and ZKPs use in plagiarism detection. In conclusion, this research has established zk-LLMs as a promising solution for bridging the trust gap in LLM deployment. By continually refining the framework and exploring its diverse applications, we can unlock a future where LLMs operate with verifiable integrity, empowering informed decision-making and ethical outcomes across various domains. The path forward lies in embracing continuous exploration and collaboration, paving the way for a future where LLMs are not feared for their limitations, but celebrated for their ability to enhance trust, transparency, and ethical AI development.

References

Ahmed, I. M. and Kashmoola, M. Y., 2021. Threats on machine learning technique by data poisoning attack: A survey. In Advances in Cyber Security: Third International Conference, ACeS 2021, Penang, Malaysia, August 24-25, 2021, Revised Selected Papers 3, pp 586-600. Springer Singapore, 2021.

Ben-Sasson, E., Chiesa, A., Tromer, E. and Virza, M. (2017) Scalable Zero Knowledge Via Cycles of Elliptic Curves. Algorithmica, 79, pp 1102-1160.

Coutaeu, G. (2017) Zero-knowledge proofs for secure computation. (Doctoral dissertation, Univerite Paris sciences et lettres).

Fiege, U., Fiat, A. and Shamir, A. (1987) Zero knowledge proofs of identity. In Proceedings of the nineteenth annual ACM symposium on Theory of computing, pp 210-217.

Goldreich, O., Micali, S. and Wigderson, A. (1987) How to Prove All NP Statements in Zero-Knowledge. Springer Berlin Heidelberg, pp 171-185.

Goldwasser, S. and Kalai, Y. T. (2003) On the (in) Security of the Fiat-Shamir paradigm. IEEE, pp 102-113.

Groth, J. K. (2018) July. Updatable and universal common reference strings with applications to zk-SNARKs. Cham: Springer International Publishing, pp. 698-728.

Himeur, Y. et al. (2022) Blockchain-based Recommender Systems: Applications, Challenges, and Future Opportunities. Computer Science Review, Vol. 42, 100439.

Iden3 (2023). Circom. GitHub Repository https://github.com/iden3/circom.

Kang, D. et al., 2023. Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks. arXiv preprint arXiv:2302.05733.

Lipmaa, H. (2016) Prover-Efficient Commit-And-Prove Zero-Knowledge SNARKs. Springer, pp 185-206.

OpenAI, 2023. GPT-4 Technical Report. ArXiv:2303.08774, Vol. 3, pp 1-100.

PromptEngineer (2024). localGPT. GitHub Repository https://github.com/PromtEngineer/localGPT.

Rosen, A. (2004) A note on constant-round zero-knowledge proofs for NP. Springer Berlin Heidelberg, pp. 191-202.

Shi, J., Liu, Y., Zhou, P. and Sun, L. (2023) BadGPT: Exploring Security Vulnerabilities of ChatGPT via Backdoor Attacks to InstructGPT. arXiv preprint arXiv:2304.12298.

Sun, X., Yu, F.R., Zhang, P., Sun, Z., Xie, W. and Peng, X., 2021. A survey on zero-knowledge proof in blockchain. IEEE network, 35(4), pp 198-205.

Wu, H. and Wang, F. (2014) A Survey of Noninteractive Zero Knowledge Proof System and its Applications. The Scientific World Journal, Vol. 2014, Article ID 560484, 7 pages.

Zhang, J., Fang, Z., Zhang, Y. and Song, D. (2020) Zero knowledge proofs for decision tree predictions and accuracy. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security pp 2039-2053.

Zhang, S., Yao, L., Sun, A. and Tay, Y., 2019. Deep learning based recommender system: A survey and new perspectives. ACM computing surveys (CSUR) 52, no.1, pp 1-38.

Zhou, C. et al., 2023. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. arXiv preprint arXiv:2302.09419.