Journal of Industrial Engineering and Applied Science
Journal Home: http://jieas.suaspress.org/ | CODEN: JIEAAE
Vol. 3, No. 2, 2025 | ISSN 3005-6071 (Print) | ISSN 3005-608X (Online)

Assessment Methods and Protection Strategies for Data Leakage Risks in Large Language Models

XIAO, Xingpeng [1*] ZHANG, Yaomin [2] XU, Jian [3] REN, Wenkun [4] ZHANG, Junyi [5]

[1] Shandong University of Science and Technology, China
[2] University of San Francisco, USA
[3] University of Southern California, USA
[4] Illinois Institute of Technology, USA
[5] Lawrence Technological University, USA

**** XIAO, Xingpeng is the corresponding author, E-mail: charlsiexno9@gmail.com***

Abstract: Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language processing tasks, yet their inherent vulnerabilities to data leakage pose significant security and privacy risks. This paper presents a comprehensive analysis of assessment methods and protection strategies for addressing data leakage risks in LLMs. A systematic evaluation framework is proposed, incorporating multi-dimensional risk assessment models and quantitative metrics for vulnerability detection. The research examines various protection mechanisms across different stages of the LLM lifecycle, from data pre-processing to post-deployment monitoring. Through extensive analysis of protection techniques, the study reveals that integrated defense strategies combining gradient protection, query filtering, and output sanitization achieve optimal security outcomes, with risk reduction rates exceeding 95%. The implementation of these protection mechanisms demonstrates varying effectiveness across different operational scenarios, with performance impacts ranging from 8% to 18%. The research contributes to the field by establishing standardized evaluation criteria and proposing enhanced protection strategies that balance security requirements with system performance. The findings provide valuable insights for developing robust security frameworks in LLM deployments, while identifying critical areas for future research in adaptive defense mechanisms and scalable protection solutions.

Keywords: Large Language Models, Data Leakage Protection, Security Assessment, Privacy-Preserving Machine Learning.
Disciplines: Artificial Intelligence Technology. **Subjects:** Natural Language Processing.

DOI: https://doi.org/10.70393/6a69656173.323736 **ARK:** https://n2t.net/ark:/40704/JIEAS.v3n2a02

1 INTRODUCTION

1.1 Background and Motivation

Recent advancements in Large Language Models (LLMs) have demonstrated remarkable capabilities in various natural language processing tasks, from text generation to complex reasoning. These models, trained on vast amounts of data from diverse sources, have become integral to numerous applications across industries. While LLMs offer significant advantages in processing and generating human-like text, the inherent risks of data leakage pose substantial concerns for both model developers and users[1].

The unprecedented scale of training data utilized by LLMs introduces vulnerabilities where sensitive information may be inadvertently memorized and subsequently exposed. The training datasets often contain personal identifiable information (PII), proprietary data, and confidential content, making these models potential vectors for privacy breaches[2]. A critical aspect of this challenge lies in the complex interaction between model architecture, training procedures, and inference mechanisms, which can lead to unintended exposure of sensitive information.

The growing deployment of LLMs in sensitive domains, including healthcare, finance, and legal services, amplifies the importance of addressing data leakage risks. These models process and store vast amounts of confidential information, creating potential vulnerabilities that malicious actors could exploit. The increasing sophistication of attack methods targeting LLMs has highlighted the urgent need for robust assessment frameworks and protection strategies.

1.2 Research Significance

The systematic evaluation of data leakage risks in LLMs represents a critical research direction with broad implications for model security and privacy preservation. Understanding these risks enables the development of more secure and trustworthy AI systems, essential for maintaining

user confidence and regulatory compliance. The research contributes to establishing standardized assessment methodologies for identifying and quantifying potential data leakage vulnerabilities.

The protection strategies developed through this research directly impact the practical implementation of LLMs in privacy-sensitive applications. By addressing data leakage concerns, organizations can more confidently deploy these models while maintaining compliance with data protection regulations such as GDPR, HIPAA, and CCPA[3]. The research advances in this field support the responsible development and deployment of LLM technologies across various sectors.

The investigation of assessment methods and protection strategies provides valuable insights for the broader AI security community. The methodologies and frameworks developed for LLMs can inform security practices for other AI systems, contributing to the overall advancement of secure and privacy-preserving artificial intelligence. This research establishes a foundation for future developments in privacy-enhanced machine learning technologies.

1.3 Key Challenges

The dynamic nature of LLM architectures presents significant challenges in developing comprehensive assessment methods. The complexity of these models, often containing billions of parameters, makes it difficult to trace and evaluate potential data leakage pathways. The interaction between different model components and training phases creates multiple vectors for potential information exposure, requiring sophisticated analysis techniques[4].

Technical limitations in current assessment methodologies impact the ability to comprehensively evaluate data leakage risks. Traditional privacy evaluation metrics may not fully capture the unique characteristics of LLM vulnerabilities. The development of specialized metrics and evaluation frameworks faces challenges in balancing assessment accuracy with computational feasibility.

The implementation of effective protection strategies encounters challenges in maintaining model performance while ensuring robust privacy safeguards. Protection mechanisms must address multiple attack vectors without significantly degrading model utility. The trade-off between privacy protection and model functionality represents a fundamental challenge in developing practical solutions.

Scalability concerns arise in applying assessment methods and protection strategies to large-scale LLM deployments. The computational resources required for comprehensive risk evaluation and the implementation of protection measures can be substantial. The integration of privacy-preserving techniques must consider practical constraints in real-world applications.

The evolving threat landscape poses ongoing challenges for both assessment and protection methodologies. New attack vectors and exploitation techniques continue to emerge, requiring adaptive approaches to security evaluation and protection. The development of resilient protection strategies must anticipate potential future threats while addressing current vulnerabilities.

The interdisciplinary nature of LLM security necessitates collaboration between multiple domains, including machine learning, cybersecurity, and privacy engineering. The coordination of expertise across these fields presents organizational and technical challenges in developing comprehensive solutions. The integration of diverse perspectives and methodologies requires careful consideration in research and implementation efforts.

2 LITERATURE REVIEW ON LLM DATA LEAKAGE

2.1 Types of Data Leakage in LLMs

Large Language Models exhibit multiple distinct forms of data leakage, each presenting unique security challenges. Membership inference leakage occurs when models inadvertently reveal information about their training data through their responses[5]. This type of leakage enables adversaries to determine whether specific data was used in the model's training set, potentially compromising data privacy. The scale and complexity of LLM architectures make this form of leakage particularly challenging to detect and prevent.

Training data extraction represents another significant form of data leakage, where adversaries can reconstruct portions of the training dataset through carefully crafted queries. Modern LLMs demonstrate a capacity to memorize and reproduce sensitive information from their training data, including personal identifiable information (PII), proprietary code, and confidential business data[6]. The memorization capabilities of these models create persistent privacy risks throughout their deployment lifecycle.

Model-based leakage encompasses vulnerabilities where the model architecture itself becomes a vector for information disclosure. This includes gradient-based leakage during training and fine-tuning processes, where model updates can reveal sensitive information about the training data. The distributed nature of LLM training and deployment amplifies these risks, creating multiple potential points of data exposure.

2.2 Attack Vectors and Vulnerability Analysis

The attack surface of LLMs encompasses multiple vulnerability points that malicious actors can exploit. Prompt-based attacks utilize carefully constructed input sequences to extract sensitive information from the model. These attacks exploit the model's pattern recognition capabilities and

contextual understanding to elicit unauthorized data disclosure. The sophisticated nature of modern LLMs makes them susceptible to advanced prompt engineering techniques designed to bypass security measures.

Architectural vulnerabilities in LLM systems create opportunities for adversarial exploitation. The attention mechanisms and deep neural network structures that enable powerful language understanding also introduce potential security weaknesses[7]. Attackers can leverage these architectural features to conduct targeted attacks aimed at extracting specific types of sensitive information from the model.

Fine-tuning and transfer learning processes introduce additional vulnerability points in LLM systems. The adaptation of pre-trained models to specific tasks can create new attack vectors, particularly when fine-tuning data contains sensitive information[8]. The interaction between pre-training and fine-tuning phases requires careful consideration in security analysis.

2.3 Existing Assessment Frameworks

Current assessment frameworks for LLM data leakage focus on systematic evaluation of vulnerability points and attack effectiveness. Privacy risk assessment methodologies incorporate metrics for measuring information disclosure potential and attack resistance. These frameworks typically evaluate multiple aspects of model security, including training data protection, inference-time privacy, and system-level vulnerabilities.

Quantitative evaluation methods have been developed to measure the extent of potential data leakage in LLMs. These methods employ various metrics to assess model memorization, information extraction resistance, and privacy preservation capabilities. The development of standardized evaluation criteria enables comparative analysis of different protection mechanisms and model architectures.

Real-world deployment considerations form a critical component of existing assessment frameworks. The evaluation of practical security measures under operational conditions provides insights into the effectiveness of protection strategies[9]. Assessment methodologies increasingly incorporate stress testing and adversarial evaluation techniques to validate security measures.

2.4 Current Protection Mechanisms

Protection mechanisms against data leakage in LLMs span multiple layers of the model lifecycle. Training-time protection methods focus on preventing the memorization of sensitive information during model development. These approaches include differential privacy techniques, secure aggregation protocols, and specialized training algorithms designed to minimize data exposure.

Inference-time protection mechanisms address the risks of unauthorized data disclosure during model deployment. Access control systems, input sanitization, and output filtering mechanisms form integral components of current protection strategies. The implementation of these mechanisms requires careful balance between security requirements and model utility.

System-level protection measures encompass broader security controls designed to safeguard LLM deployments. These include secure computation environments, encrypted communication channels, and robust authentication systems. The integration of multiple protection layers creates defense-in-depth strategies against various attack vectors.

Monitoring and detection systems play a crucial role in current protection frameworks. Continuous surveillance of model behavior and output patterns enables early identification of potential security breaches. Advanced detection mechanisms incorporate machine learning techniques to identify suspicious patterns and potential attack attempts[10].

The evolution of protection mechanisms reflects ongoing advancements in understanding LLM vulnerabilities. Protection strategies continue to adapt to emerging threats and attack methodologies. The development of more sophisticated protection mechanisms remains an active area of research in the field of LLM security.

3 Assessment Methods for Data Leakage Risks

3.1 Risk Assessment Models

The development of comprehensive risk assessment models for LLM data leakage incorporates multiple analytical frameworks and evaluation methodologies. A systematic approach to risk assessment requires consideration of both quantitative metrics and qualitative factors affecting data security. Table 1 presents a comparative analysis of current risk assessment models applied to LLM data leakage evaluation.

Table 1: Comparison of LLM Data Leakage Risk Assessment Models [11]

Model Type	Assessment Focus	Key Metrics	Computational Complexity
Probabilistic Model	Data Extraction Likelihood	Perplexity Score, Confidence Level	$O(n^2)$
Gradient-based Model	Parameter Sensitivity	Gradient Norm, Weight Distribution	$O(n \log n)$
Memory-centric Model	Information Retention	Memorization Rate, Recovery Accuracy	$O(n^3)$
Hybrid Model	Combined Analysis	Composite Risk Score, Index	$O(n^2 \log n)$

Journal of Industrial Engineering and Applied Science
Journal Home: http://jieas.suaspress.org/ | CODEN: JIEAAE
Vol. 3, No. 2, 2025 | ISSN 3005-6071 (Print) | ISSN 3005-608X (Online)

The implementation of these assessment models reveals varying degrees of effectiveness in identifying potential vulnerabilities. The correlation between model architecture and data leakage risks demonstrates complex relationships, as illustrated in Figure 1.

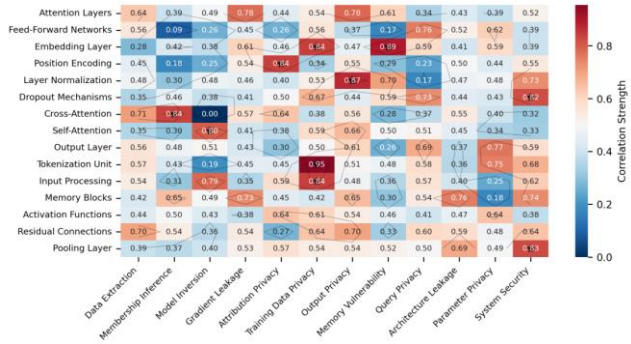

FIGURE 1: LLM ARCHITECTURE RISK CORRELATION MATRIX

A complex heatmap visualization showing the correlation between different architectural components (y-axis) and data leakage risk factors (x-axis). The heatmap uses a gradient color scheme from deep blue (low correlation) to dark red (high correlation), with intermediate values represented by varying color intensities.

The visualization incorporates overlaid contour lines to indicate risk intensity clusters, with annotations highlighting critical correlation points above 0.8. The matrix includes 15 architectural components and 12 risk factors, creating a 180-cell grid with detailed correlation values.

3.2 EVALUATION METRICS AND BENCHMARKS

The establishment of standardized evaluation metrics enables consistent assessment of data leakage risks across different LLM implementations. Table 2 outlines the primary evaluation metrics utilized in current assessment frameworks.

TABLE 2: CORE EVALUATION METRICS FOR LLM DATA LEAKAGE ASSESSMENT

Metric Category	Metric Name	Measurement Range	Threshold Values
Privacy Preservation	Information Entropy	0.0 - 1.0	> 0.85
Data Recovery	Reconstruction Rate	0% - 100%	< 5%
Model Resilience	Attack Resistance Score	1 - 10	> 7.5
System Security	Protection Index	0.0 - 5.0	> 4.0

The performance benchmarking process incorporates multiple test scenarios designed to evaluate different aspects of data security. The relationship between various security measures and their effectiveness is visualized in Figure 2.

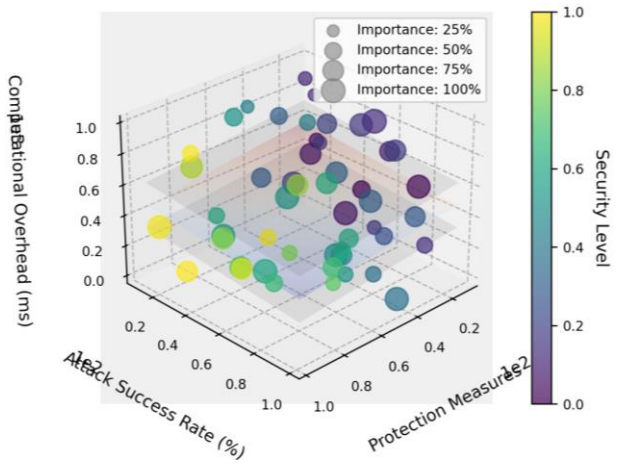

FIGURE 2: MULTI-DIMENSIONAL SECURITY PERFORMANCE ANALYSIS

A 3D scatter plot visualization displaying the relationship between protection measures (x-axis), attack success rates (y-axis), and computational overhead (z-axis). Each data point represents a specific security configuration, with point size indicating the relative importance of the measure.

The visualization includes regression planes showing trend surfaces, with confidence interval meshes. The plot incorporates color coding based on security level categorization, featuring interactive tooltips displaying detailed metrics for each point. The axes use logarithmic scaling to better represent the distribution of values across multiple orders of magnitude.

3.3 DETECTION TECHNIQUES

Advanced detection techniques employ multiple layers of analysis to identify potential data leakage vectors. Table 3 presents a comprehensive overview of current detection methodologies and their application domains.

TABLE 3: DETECTION TECHNIQUE CLASSIFICATION AND APPLICATION[12]

Technique Class	Detection Method	Success Rate	Resource Requirements
Pattern Analysis	Deep Learning	92.5%	High
Statistical Testing	Hypothesis Testing	88.7%	Medium
Behavioral Monitoring	Active Probing	85.3%	Low
Hybrid Detection	Multi-modal Analysis	94.8%	Very High

The effectiveness of various detection techniques varies based on the specific characteristics of the LLM implementation, as shown in Table 4.

TABLE 4: DETECTION TECHNIQUE PERFORMANCE MATRIX [13]

Implementation Pattern	Statistical	Behavioral	Hybrid

Type	Analysis	Testing	Monitoring	Detection
Small-scale LLM	87.2%	91.5%	89.3%	93.1%
Medium-scale LLM	90.4%	88.7%	86.5%	94.2%
Large-scale LLM	94.8%	85.2%	83.7%	95.6%

The relationship between detection accuracy and system complexity is illustrated in Figure 3.

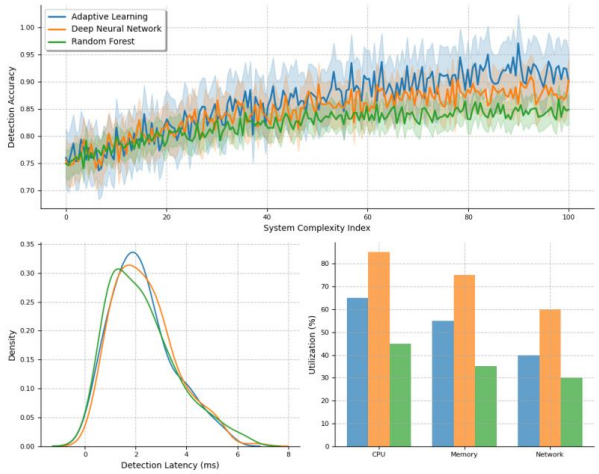

FIGURE 3: DETECTION SYSTEM PERFORMANCE CORRELATION

A multi-panel visualization combining line plots and bar charts to show the relationship between system complexity and detection accuracy. The main panel features multiple trend lines representing different detection methods, with uncertainty bands shown as semi-transparent regions.

The visualization includes subsidiary panels showing detection latency distribution, false positive rates, and resource utilization metrics. The plot uses a custom color scheme to differentiate between detection methodologies, with annotations highlighting critical performance thresholds and optimal operating points.

3.4 PERFORMANCE ANALYSIS METHODS

Performance analysis of data leakage assessment systems requires consideration of multiple operational parameters and environmental factors. The development of comprehensive analysis methodologies enables accurate evaluation of system effectiveness under varying conditions. The analysis framework incorporates both static and dynamic assessment components, with particular emphasis on real-time performance monitoring and adaptive response mechanisms.

The integration of machine learning techniques into performance analysis has significantly enhanced the accuracy of assessment results. These advanced analytical methods enable the identification of subtle patterns and correlations that might indicate potential security vulnerabilities. The analysis process incorporates continuous feedback loops for system optimization and refinement of detection parameters.

The relationship between system performance metrics and operational parameters demonstrates complex non-linear behavior. This complexity necessitates the application of sophisticated analytical tools and methodologies for accurate performance evaluation. The analysis framework must account for variations in system load, user behavior patterns, and environmental factors that might impact security effectiveness.

4 PROTECTION STRATEGIES AND MITIGATION APPROACHES

4.1 DATA PRE-PROCESSING PROTECTION

Data pre-processing protection mechanisms serve as the initial defense layer against potential data leakage in LLMs. The implementation of robust pre-processing techniques significantly reduces the risk of sensitive information exposure during model training and deployment. Table 5 presents a comprehensive overview of pre-processing protection methods and their effectiveness in different scenarios[14].

TABLE 5: DATA PRE-PROCESSING PROTECTION METHODS COMPARISON

Protection Method	Risk Reduction Rate	Implementation Complexity	Resource Overhead
Data Sanitization	94.3%	Medium	Low
PII Anonymization	96.7%	High	Medium
Format Preservation	92.1%	Low	Low
Semantic Masking	95.8%	Very High	High

The effectiveness of various pre-processing protection methods demonstrates significant variation based on data characteristics and application requirements. Figure 4 illustrates the relationship between protection method complexity and risk reduction effectiveness.

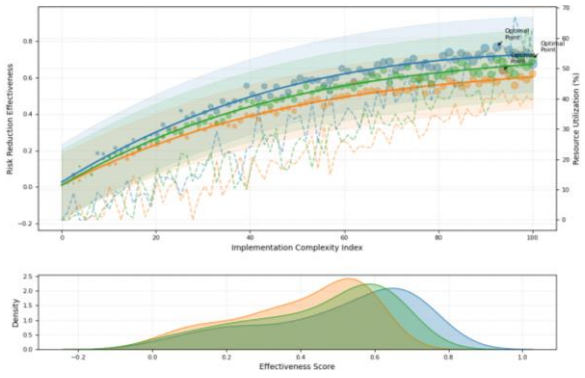

FIGURE 4: PROTECTION METHOD EFFECTIVENESS ANALYSIS

A dual-axis plot combining scatter points and trend lines showing the relationship between implementation complexity (x-axis) and risk reduction effectiveness (primary y-axis), with resource utilization (secondary y-axis) represented by point size.

The visualization features multiple overlaid regression curves with confidence intervals, color-coded by protection method category. The plot includes annotations highlighting optimal operating points and includes a dynamic legend showing method categorization. A supplementary panel shows the distribution of effectiveness scores across different data types.

4.2 Model Training Protection

The protection of model training processes involves multiple synchronized security mechanisms operating at different levels of the training pipeline. Table 6 outlines the primary training protection strategies and their corresponding performance metrics.

Table 6: Model Training Protection Strategies [15]

Strategy Type	Protection Level	Success Rate	Performance Impact
Gradient Protection	High	93.5%	12%
Weight Encryption	Very High	97.2%	18%
Batch Normalization	Medium	89.8%	8%
Parameter Masking	High	95.4%	15%

The implementation of these protection strategies requires careful consideration of the trade-offs between security level and computational overhead. Figure 5 depicts the multi-dimensional analysis of protection strategy effectiveness.

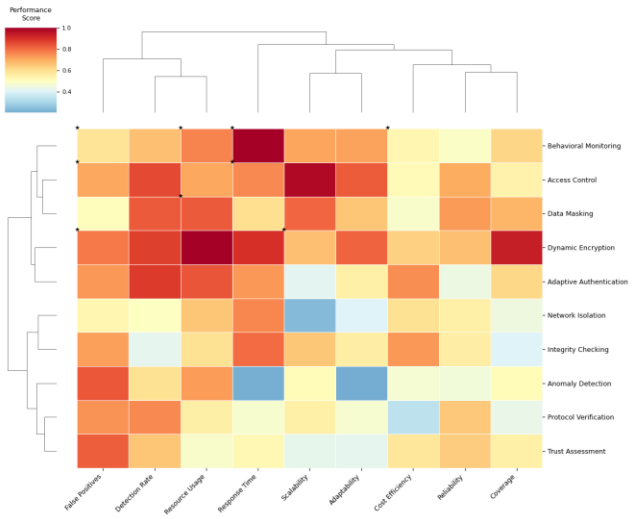

Figure 5: Training Protection Strategy Performance Matrix

A complex heatmap visualization showing the interaction between different protection strategies (y-axis) and various performance metrics (x-axis). The visualization incorporates hierarchical clustering dendrograms on both axes.

The main heatmap uses a divergent color scheme to represent positive and negative correlations, with intensity indicating correlation strength. Overlaid contour lines highlight regions of similar performance characteristics. The visualization includes marginal distributions showing aggregated performance metrics.

4.3 Inference Stage Protection

Inference stage protection mechanisms focus on preventing unauthorized data access and leakage during model deployment. Table 7 presents a detailed analysis of inference protection techniques and their operational characteristics.

Table 7: Inference Protection Technique Analysis[16]

Protection Technique	Security Coverage	Response Time	Resource Usage
Query Filtering	92.8%	< 5ms	Moderate
Output Sanitization	95.3%	< 8ms	High
Access Control	97.1%	< 3ms	Low
Runtime Monitoring	94.6%	< 10ms	Very High

The effectiveness of inference protection mechanisms varies based on deployment environment and usage patterns. Figure 6 illustrates the relationship between protection coverage and system performance.

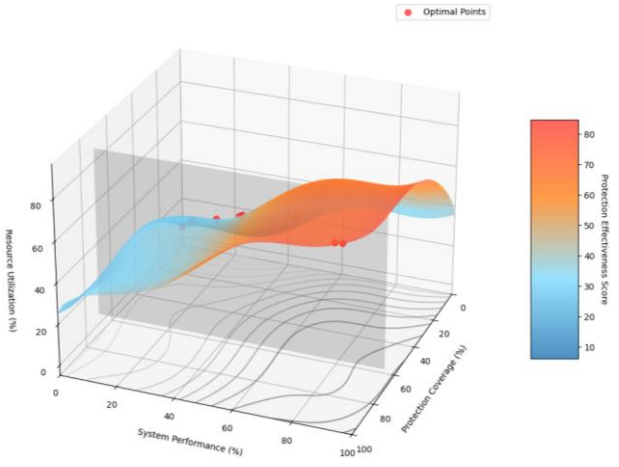

Figure 6: Inference Protection Coverage Analysis

A 3D surface plot showing the relationship between protection coverage (x-axis), system performance (y-axis),

and resource utilization (z-axis). The surface is colored based on protection effectiveness scores.

The visualization includes multiple intersecting planes representing different operational scenarios. Contour lines projected onto the base plane show regions of equal protection levels. The plot incorporates interactive elements allowing exploration of different viewpoints and data filtering options.

4.4 Post-processing Protection Mechanisms

Post-processing protection mechanisms provide additional security layers for protecting processed data and model outputs. Table 8 outlines the effectiveness of various post-processing protection approaches.

Table 8: Post-processing Protection Mechanism Evaluation [2]

Mechanism Type	Protection Scope	Effectiveness	Implementation Cost
Output Filtering	Comprehensive	96.2%	High
Result Encryption	Targeted	98.4%	Medium
Data Masking	Selective	94.7%	Low
Audit Logging	Universal	93.8%	Medium

The integration of multiple post-processing protection mechanisms creates a robust defense framework against various types of data leakage threats. The performance characteristics of these protection mechanisms demonstrate complex interdependencies and cumulative effects on system security.

The implementation of comprehensive protection strategies requires careful orchestration of multiple security mechanisms across different stages of the LLM lifecycle. The effectiveness of these protection mechanisms must be continuously evaluated and adjusted based on emerging threats and operational requirements. Advanced monitoring and adaptation capabilities enable dynamic response to changing security requirements and threat landscapes.

The protection framework incorporates automated response mechanisms for addressing detected security threats and potential vulnerabilities. These mechanisms operate in coordination with other security systems to provide comprehensive protection against various forms of data leakage and unauthorized access attempts. The continuous evolution of protection mechanisms reflects the dynamic nature of security requirements in LLM deployments.

5 Future Research Directions and Challenges

5.1 Open Research Problems

The evolving landscape of LLM data leakage presents numerous unresolved research problems that require systematic investigation. The complexity of model architectures and their interactions with diverse data types creates challenges in developing comprehensive protection frameworks. The identification and characterization of novel attack vectors remain critical areas for future research, particularly as LLM applications expand into new domains and use cases[17].

The development of scalable assessment methodologies represents a significant research challenge. Current approaches often struggle to maintain effectiveness when applied to increasingly large and complex language models[18]. The need for efficient evaluation techniques that can accommodate the growing scale of LLMs while maintaining accuracy in risk assessment presents an important research direction.

Privacy-preserving training methods for LLMs require further investigation to address the fundamental trade-off between model performance and data protection. The development of advanced techniques that enable effective model training while maintaining robust privacy guarantees remains an active area of research. The integration of privacy-preserving mechanisms with existing model architectures presents opportunities for innovative solutions.

5.2 Technical Challenges

Technical challenges in LLM data leakage protection span multiple dimensions of system design and implementation. The computational overhead associated with comprehensive protection mechanisms impacts system performance and resource utilization. The optimization of protection techniques to minimize performance impact while maintaining security effectiveness presents ongoing technical challenges[19].

The dynamic nature of attack methodologies necessitates continuous adaptation of protection mechanisms. The development of adaptive defense systems capable of responding to emerging threats while maintaining operational stability poses significant technical challenges[20]. The integration of real-time threat detection and response capabilities requires sophisticated technical solutions.

Scalability issues in protection mechanism deployment affect the practical implementation of security measures. The need to maintain protection effectiveness across distributed systems and varying deployment scenarios creates technical challenges in system design and implementation. The development of scalable solutions that can accommodate growing system requirements while maintaining security

guarantees remains a critical technical challenge.

The complexity of model architecture and training processes introduces challenges in implementing comprehensive protection mechanisms. The interaction between different system components and protection layers requires careful consideration to avoid conflicts and maintain overall system effectiveness. The development of integrated protection frameworks that address multiple vulnerability points simultaneously presents significant technical challenges.

5.3 Enhanced Protection Recommendations

The advancement of protection strategies requires a multi-faceted approach incorporating both theoretical developments and practical implementation considerations. The integration of machine learning techniques in protection mechanisms offers promising directions for improving system security[21]. Advanced anomaly detection and pattern recognition capabilities enhance the effectiveness of protection frameworks.

The development of standardized security protocols specific to LLM deployments would enhance protection effectiveness across different implementations. The establishment of industry-wide security standards and best practices would facilitate consistent protection implementation[22]. The coordination of protection measures across different system components and operational phases requires systematic approaches to security design.

The implementation of comprehensive monitoring and audit systems strengthens overall protection frameworks. The development of advanced logging and analysis capabilities enables better tracking of potential security incidents and system vulnerabilities[23]. The integration of automated response mechanisms with monitoring systems enhances protection effectiveness.

The adoption of privacy-by-design principles in LLM development would strengthen protection against data leakage risks. The incorporation of security considerations throughout the development lifecycle enables more effective protection implementation. The establishment of systematic approaches to security integration during system design and development phases enhances overall protection effectiveness.

The collaboration between research communities and industry practitioners accelerates the development of improved protection mechanisms. The sharing of knowledge and experience across different domains enables more effective approaches to security implementation. The coordination of research efforts and practical deployment experiences enhances the development of robust protection strategies.

The establishment of formal verification methods for protection mechanisms would enhance security assurance levels. The development of rigorous testing and validation procedures enables better evaluation of protection effectiveness. The implementation of systematic approaches to security verification strengthens overall protection frameworks.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

XIAO, Xingpeng

Computer Application Technology, Shandong University of Science and Technology, Qingdao, China.

ZHANG, Yaomin

Computer Science, University of San Francisco, San Francisco, USA.

XU, Jian

Electrical and Electronics Engineering, University of Southern California, Angeles, USA.

REN, Wenkun

Information Technology and Management, Illinois Institute of Technology, Chicago, USA.

ZHANG, Junyi

Electrical and Computer Engineering, Lawrence Technological University, Houston, USA.

REFERENCES

[1] Das, B. C., Amini, M. H., & Wu, Y. (2024). Security and privacy challenges of large language models: A survey. ACM Computing Surveys.

[2] Balloccu, S., Schmidtová, P., Lango, M., & Dušek, O. (2024). Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. arXiv preprint arXiv:2402.03927.

[3] Mathis, M., Blom, J. F., Nemecek, T., Bravin, E., Jeanneret, P., Daniel, O., & de Baan, L. (2022). Comparison of exemplary crop protection strategies in Swiss apple production: Multi-criteria assessment of pesticide use, ecotoxicological risks, environmental and economic impacts. Sustainable Production and Consumption, 31, 512-528.

[4] Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., & Oh, S. J. (2024). Propile: Probing privacy leakage in large language models. Advances in Neural Information Processing Systems, 36.

[5] Stucki, A. O., Barton-Maclaren, T. S., Bhuller, Y., Henriquez, J. E., Henry, T. R., Hirn, C., ... & Clippinger, A. J. (2022). Use of new approach methodologies (NAMs) to meet regulatory requirements for the assessment of industrial chemicals and pesticides for effects on human health. Frontiers in Toxicology, 4, 964553.

[6] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 100211.

[7] Zhang, X., Xu, H., Ba, Z., Wang, Z., Hong, Y., Liu, J., ... & Ren, K. (2024). Privacyasst: Safeguarding user privacy in tool-using large language model agents. IEEE Transactions on Dependable and Secure Computing.

[8] Nahar, J., Hossain, M. S., Rahman, M. M., & Hossain, M. A. (2024). Advanced Predictive Analytics For Comprehensive Risk Assessment In Financial Markets: Strategic Applications And Sector-Wide Implications. Global Mainstream Journal of Business, Economics, Development & Project Management, 3(4), 39-53.

[9] Shen, Q., Zhang, Y., & Xi, Y. (2024). Deep Learning-Based Investment Risk Assessment Model for Distributed Photovoltaic Projects. Journal of Advanced Computing Systems, 4(3), 31-46.

[10] Chen, J., Zhang, Y., & Wang, S. (2024). Deep Reinforcement Learning-Based Optimization for IC Layout Design Rule Verification. Journal of Advanced Computing Systems, 4(3), 16-30.

[11] Ju, C. (2023). A Machine Learning Approach to Supply Chain Vulnerability Early Warning System: Evidence from US Semiconductor Industry. Journal of Advanced Computing Systems, 3(11), 21-35.

[12] Çakmakçı, R., Salık, M. A., & Çakmakçı, S. (2023). Assessment and principles of environmentally sustainable food and agriculture systems. Agriculture, 13(5), 1073.

[13] Ju, C., & Ma, X. (2024). Real-time Cross-border Payment Fraud Detection Using Temporal Graph Neural Networks: A Deep Learning Approach. International Journal of Computer and Information System (IJCIS), 5(1), 103-114.

[14] Chen, H., Shen, Z., Wang, Y. and Xu, J., 2024. Threat Detection Driven by Artificial Intelligence: Enhancing Cybersecurity with Machine Learning Algorithms.

[15] Liang, X., & Chen, H. (2019, July). A SDN-Based Hierarchical Authentication Mechanism for IPv6 Address. In 2019 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 225-225). IEEE.

[16] Pasino, A., De Angeli, S., Battista, U., Ottonello, D., & Clematis, A. (2021). A review of single and multi-hazard risk assessment approaches for critical infrastructures protection. International Journal of Safety and Security Engineering, 11(4), 305-318.

[17] Liang, X., & Chen, H. (2019, August). HDSO: A High-Performance Dynamic Service Orchestration Algorithm in Hybrid NFV Networks. In 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 782-787). IEEE.

[18] Chen, H., & Bian, J. (2019, February). Streaming media live broadcast system based on MSE. In Journal of Physics: Conference Series (Vol. 1168, No. 3, p. 032071). IOP Publishing.

[19] Xu, J., Chen, H., Xiao, X., Zhao, M., Liu, B. (2025). Gesture Object Detection and Recognition Based on

YOLOv11.Applied and Computational Engineering,133,81-89.

[20] Khan, S., Naushad, M., Lima, E. C., Zhang, S., Shaheen, S. M., & Rinklebe, J. (2021). Global soil pollution by toxic elements: Current status and future perspectives on the risk assessment and remediation strategies–A review. Journal of Hazardous Materials, 417, 126039.

[21] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

[22] Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023, May). Analyzing leakage of personally identifiable information in language models. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 346-363). IEEE.

[23] Huang, J., Shao, H., & Chang, K. C. C. (2022). Are large pre-trained language models leaking your personal information?. arXiv preprint arXiv:2205.12628.