Journal of Information Science & Automation Systems
JISASS, Vol. 1, Issue 1, July-August, 2026, 6-10

Intelligent Cybersecurity Framework for Detecting and Explaining Malicious Prompt Injection Attacks in Large Language Models

Wilson Bakasa[1] , Ranjan Arulanandham[2]

[1]Lecturer, Department: Computer Science, University of Botswana, Botswana, bakasaw@ub.ac.bw
2Head of Department, Information and Communication Technologies, Northrise University, ranjan.arulanandham@northrise.net

Article's Information

Received: 04.02.2026
Accepted: 02.05.2026
Published: 25.07.2026

Keywords:
Large Language Models ,
BERT,
classification,
MPDD dataset

Abstract

Large Language Models (LLMs) are increasingly vulnerable to malicious prompt injection attacks that manipulate intended instructions and compromise system security. This paper proposes an intelligent cybersecurity framework combining Modern BERT for contextual prompt classification with Integrated Gradients for explainable detection. The framework preprocesses prompts, generates contextual representations, performs malicious-prompt classification, applies threshold-based security decisions, and identifies influential tokens responsible for predictions. Using the MPDD dataset, the proposed Modern BERT model achieves 97.78% accuracy, demonstrating effective malicious prompt detection. The integration of explainable analysis improves transparency and supports informed security decisions, providing a practical approach for strengthening the protection of LLM-based applications against prompt injection attacks.

https://jisass.com/index.php/jisass/index
*Corresponding author: bakasaw@ub.ac.bw

1. Introduction

Large Language Models (LLMs) are now commonly applied in intelligent applications; however, the reliance on natural language inputs makes LLMs vulnerable to the prompt injection, when malicious inputs replace the intended instructions of the user and manipulate LLMs. It has been proved recently that prompt injection can be used to compromise not only traditional LLMs but also multimodal architectures [1],[2]. The existing works are devoted to hybrid detection methods involving rule-based and machine learning models (One-Class SVM) as well as lightweight BERT-based filtering and multiple security layers [3], [4]. Many approaches focus on detecting or preventing attacks without offering a consistent framework that would combine prompt classification, security decision-making, and token attribution [5]. To overcome the aforementioned deficiencies, the proposed research develops an intelligent cybersecurity framework combining ModernBERT for contextual malicious-prompt detection and Integrated Gradients for explanation. The framework is able to classify inputs as either malicious or benign, apply the security decision based on the specified threshold, and provide tokens that influence malicious predictions.

1.1 Key Contribution

- Propose a ModernBERT-based framework for detecting malicious prompt injections.
- Built a threshold based LLM security decision layer to allow, flag or block prompts.
- Integrated Gradients for token-level explanations to improve the interpretability of malicious prompt detection.

2. Literature Review

Zhuhadar [6] proposed PromptSentinel-X, a leakage-aware, context-aware screening framework for LLM-powered web agents, with 1,581 benchmark records and 30,015 transfer/stress-testing records. On 465 group-aware tests, it got 98.49% accuracy and 88.87% macro-F1. Performance degraded in multi-turn, RAG

and memory scenarios motivating broader multimodal and multilingual evaluation. Srinivasan et al. [7] tackled the problem of prompt injection in Hindi and Hinglish LLMs by creating a 4,000 prompt multilingual data set and testing Transformer, Distill Transformer, SVM, XGBoost, Random Forest, and Logistic Regression algorithms. In this work, they succeeded in creating a hybrid rule-based transformer algorithm that attained 99.70% accuracy, proving multilingual detection reliably. Wang et al. [8] proposed a Policy Enforcement Point for MCP-based LLM agents, using declarative rules, cross-step information-flow labels and SHA-256 hash-chained auditing. On a controlled four attack class dataset, It reduced attack success rate from 40.0% to 5.0%. False positives and rule coverage holes still exist and need to be evaluated on further deployment.

3. Methodology

The proposed architecture will detect any form of malicious prompt injection attacks targeted on LLMs using ModernBERT as the backbone detector model. The prompts are pre-processed, either classified as either malicious or benign and filtered before they get into the LLMs. The Integrated Gradients technique gives explanation through the important words affecting the decision making.

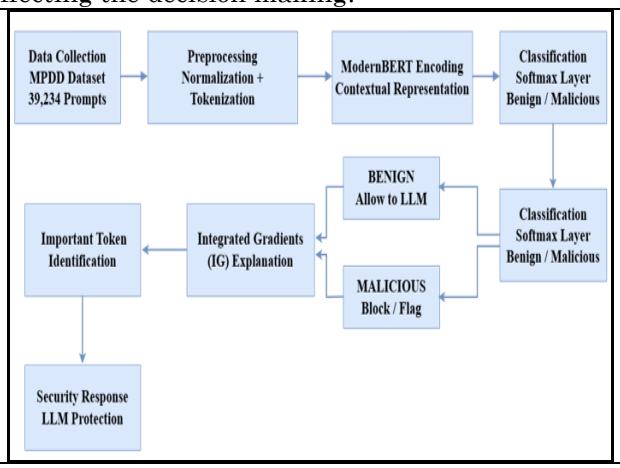

Figure 1. Proposed Workflow for Malicious Prompt Detection and Explainability.

Fig. 1 shows the approach used: MPDD preprocessing, ModernBERT classification, security decision, Integrated Gradients explanation, token detection, and LLM protection.

3.1 Data Collection:

The proposed framework for cybersecurity is based on the dataset MPDD [9], which has 39,234 prompts that are both benign and malicious. This variety of samples will be useful for training and evaluating the model for detecting attacks involving injection of malicious prompts aimed at LLMs.

3.2 Data Preprocessing:

The Prompts were collected and cleaned of duplicate, empty and irrelevant prompts, and security words and instructions were preserved. Normalization and tokenization of each prompt was done using a tokenizer from the ModernBERT model. The pre-processing process is represented as (1),

$$X = Tokenize\big(Normalize(P)\big) \quad \ldots(1)$$

The original prompt P and processed token sequence X. The data set was split into a training set, a validation set and a testing set by stratified sampling to maintain the class distribution.

3.3 ModernBERT-Based Prompt Detection:

The ModernBERT technology is capable of conducting contextual text classification by studying the incoming prompt prior to the processing of the large language model. It is defined in (2),

$$\boldsymbol{H = ModernBERT(X)} \quad \ldots(2)$$

When $\boldsymbol{X}$ is the tokenized input prompt, ModernBERT uses context to identify features, and $\boldsymbol{H}$ is the output from those features for classification processes.

3.4 Malicious Prompt Classification:

ModernBERT's contextual representation goes to a simple classification layer. The probability of the prompt being in the benign class is given by (3),

$$\hat{\boldsymbol{y}} = \boldsymbol{Softmax(WH + b)} \quad \ldots(3)$$

Where $\boldsymbol{W}$ is the classifier weights, $\boldsymbol{H}$ is the contextual features, $\boldsymbol{b}$ is the bias term, and $\hat{\boldsymbol{y}}$ is the predicted class probability.

3.5 LLM security Framework:

The proposed approach is a security layer that is placed between the user and the targeted large language model. The incoming prompt is first analysed by the ModernBERT detector, and then sent to the LLM. Then a security decision is applied based on a threshold (4),

$$\boldsymbol{S(P)} = \begin{cases} \boldsymbol{1, P(malicious) < \tau} \\ \boldsymbol{0, P(malicious) < \tau} \end{cases}$$

(4)

The predicted probability of being malicious and the detection threshold . If , the prompt continues on to LLM, otherwise, it is flagged/blocked. This helps to decrease the chances of malicious instructions reaching the target LLM.

3.6 Expalinable Malicious Prompt Detection:

To explain the predictions made by the ModernBERT detector, Integrated Gradients was added. Searches for the most discriminatory input tokens for classification. The value of token is computed as (5),

$$IG_i = (x_i - x_i') \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (5)$$

Where x is the prompt, and x^' is a baseline prompt. The resulting attribution scores highlight key words or phrases that helped to make the prediction. The higher the attribution value, the more influence on the model's decision.

The methodology involves pre-processing MPDD prompts, tokenization of inputs and their analysis by ModernBERT for categorizing malicious prompts. The security module takes threshold-based decisions on whether to stop or allow prompts, whereas the Influential Tokens component employs Integrated Gradients to find out the key tokens.

4. Results and Discussion

This proposed architecture takes into account MPDD for injection detection, prompt accuracy, decision of the LLM, characteristics of the dataset, and explainable predictions.

4.1 Dataset Characteristics

This section discusses the dataset MPDD [9], used for training and testing the suggested approach and their properties along with the distribution of classes. The balanced distribution of the benign and malicious prompts provides a good foundation for detecting malicious prompts in binary form.

Table 1. Characteristics of the MPDD Dataset

Characteristic	Value
Dataset	Malicious Prompt Detection Dataset (MPDD)
Data Type	Text
Total Prompts	39,234
Benign Prompts	19,617
Malicious Prompts	19,617
Benign Percentage	50.00%
Malicious Percentage	50.00%

4.2 ModernBERT Detection Performance

Training and validation results of ModernBERT show successful learning of contextual patterns relating to both malicious and legitimate prompts.

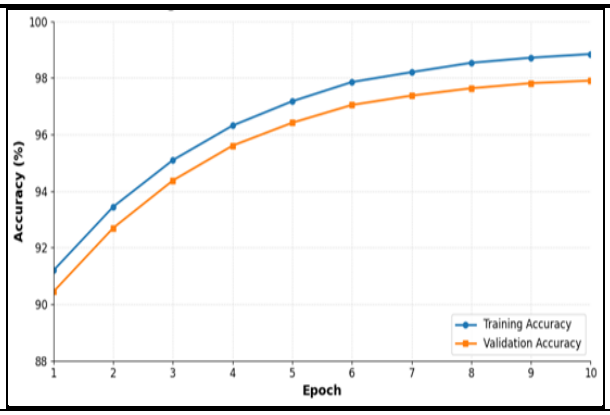

Figure 2. Training and Validation Performance of ModernBERT

Training and validation accuracy of ModernBERT continued to improve throughout ten epochs, achieving 98.85% training accuracy and 97.91% validation accuracy.

4.3 Malicious Prompt Classification

The classification ability of the proposed framework was evaluated by the confusion matrix and standard classification metrics. The results indicate the accuracy of the model in identifying malicious prompts from benign prompts.

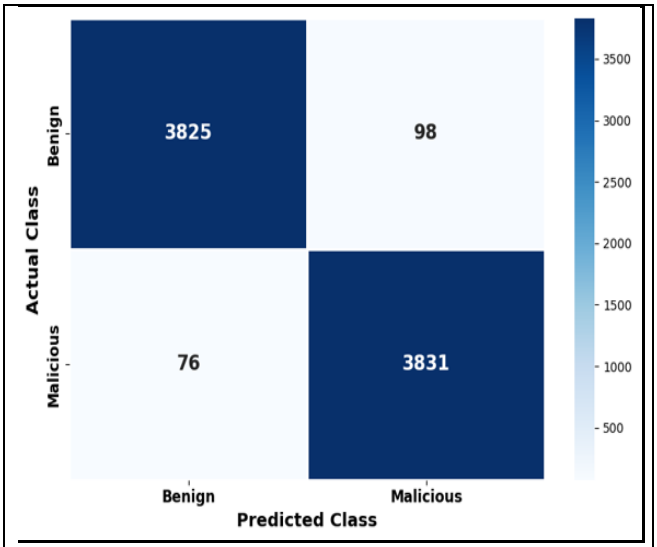

Figure 3. Confusion Matrix of Malicious Prompt Detection

The fig. 3 shows the accuracy of the ModernBERT classifier for benign and malicious prompts. In this regard, the model was able to classify 3,825 benign and 3,831 malicious prompts, having 98 false positives and 76 false negatives.

Table 1. Performance of the Proposed ModernBERT Model.

Metric	ModernBERT (%)
Accuracy	97.78
Precision	97.51
Recall	98.05
F1-Score	97.78
MCC	95.57
AUC	99.12

ModernBERT reached 97.78% accuracy, 97.51% precision, 98.05% recall, 97.78% F1-score, 95.57% MCC, and 99.12% AUC.

4.4 LLM Security Decision Analysis

The security layer evaluates the prompts before getting to the LLMs based on the probability of being malicious and the threshold.

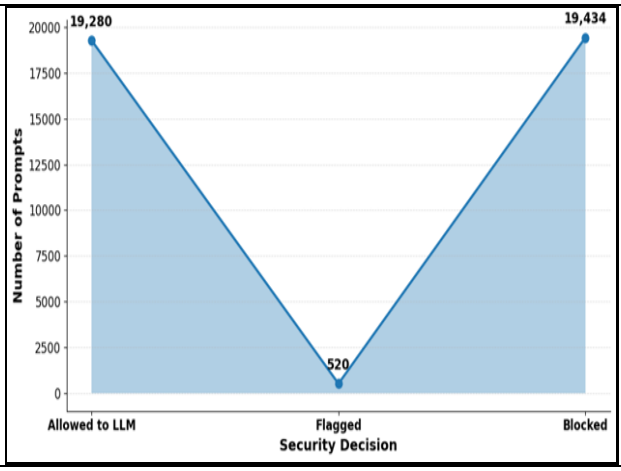

Figure 4. LLM Security Decision Analysis

Fig. 4 depicts security choices for 39,234 prompts: 19,280 accepted, 520 reviewed, and 19,434 rejected as potentially malicious.

4.4 Explainable Malicious Prompt Detection

The Integrated Gradients approach highlights influential tokens, thereby showing which tokens have the greatest impact on the classification by ModernBERT.

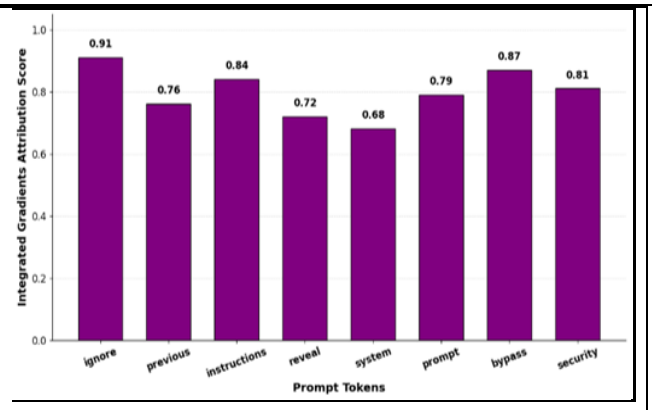

Figure 5. Integrated Gradients-Based Explainable of Malicious Prompt Detection

Fig. 5. Integrated Gradients identifies important tokens in the prompt. The most important tokens are ignore, bypass, and instructions.

4.5 Comparison with Existing Transformer-Based Models

The proposed architecture was compared to current Transformer architectures based on classification metrics used in the same experiment setup.

Table 1. Comparison with Existing Prompt Injection Detection Methods

Model	Accuracy
XGBoost [10]	90.79%
Fine-tuned Small Language [11]	95.83%
OpenAI text-embedding-3-small + XGBoost Classifier [12]	97.70%
Proposed ModernBERT	97.78%

Table III indicates the accuracy comparison between different models, where it can be seen that the proposed ModernBERT model performs better than others with an accuracy of 97.78%.

4.6 Discussion

The proposed ModernBERT model exhibits a high ability to recognize malicious prompts, with a recognition accuracy of 97.78%. The stability of training, high accuracy of validation, and few classification errors show good learning and recognition of benign and malicious prompts.

5. Conclusions

The proposed ModernBERT framework effectively detects malicious prompt injection attacks in LLMs, achieving 97.78% accuracy on the MPDD dataset. By combining contextual classification with Integrated Gradients, the framework identifies influential tokens and improves detection transparency. The approach provides an effective cybersecurity

mechanism for strengthening LLM protection against malicious prompts.

Future work focuses on larger multilingual datasets, adaptive attack scenarios, real-time deployment, and advanced explainability techniques, targeting improved robustness, scalability, generalization, interpretability, and cybersecurity protection across diverse LLM-based applications and evolving prompt-injection threats.

References

[1] J. Clusmann *et al.*, "Prompt injection attacks on vision language models in oncology," *Nat Commun*, vol. 16, no. 1, p. 1239, Feb. 2025, doi: 10.1038/s41467-024-55631-x.

[2] Y. Yang, Q. Jin, F. Huang, and Z. Lu, "Adversarial prompt and fine-tuning attacks threaten medical large language models," *Nat Commun*, vol. 16, no. 1, p. 9011, Oct. 2025, doi: 10.1038/s41467-025-64062-1.

[3] A. A. Alshammari and O. I. Alsaleh, "Detecting Prompt Injection Attacks in Generative AI Systems: A Hybrid SIEM and One-Class SVM Framework," *Electronics*, vol. 15, no. 11, p. 2242, May 2026, doi: 10.3390/electronics15112242.

[4] A. Alzahrani, "PromptGuard a structured framework for injection resilient language models," *Sci Rep*, vol. 16, no. 1, p. 1277, Jan. 2026, doi: 10.1038/s41598-025-31086-y.

[5] M. Podpora, M. Baranowski, M. Chopcian, L. Kwasniewicz, and W. Radziewicz, "LLM Firewall Using Validator Agent for Prevention Against Prompt Injection Attacks," *Applied Sciences*, vol. 16, no. 1, p. 85, Dec. 2025, doi: 10.3390/app16010085.

[6] L. P. Zhuhadar, "PromptSentinel-X: A Leakage-Aware and Context-Aware Framework for Prompt-Injection Detection in Large Language Model-Powered Web Agents," *Future Internet*, vol. 18, no. 7, p. 376, Jul. 2026, doi: 10.3390/fi18070376.

[7] J. Srinivasan, S. A. Regi, A. K. Anbarasan, S. A, V. T, and S. Venu, "Detection and analysis of prompt injection in indian multilingual large language models," *Sci Rep*, vol. 16, no. 1, p. 16208, Apr. 2026, doi: 10.1038/s41598-026-43883-0.

[8] S. Wang, S. Zhu, and R. Li, "Runtime Policy Enforcement for MCP-Based LLM Agents," *Electronics*, vol. 15, no. 13, p. 2829, Jun. 2026, doi: 10.3390/electronics15132829.

[9] "Malicious Prompt Detection Dataset (MPDD)." Accessed: Aug. 22, 2026. [Online]. Available: https://www.kaggle.com/datasets/mohammedaminejebbar/malicious-prompt-detection-dataset-mpdd

[10] M. Alamsabi, M. Tchuindjang, and S. Brohi, "Embedding-Based Detection of Indirect Prompt Injection Attacks in Large Language Models Using Semantic Context Analysis," *Algorithms*, vol. 19, no. 1, p. 92, Jan. 2026, doi: 10.3390/a19010092.

[11] A. A. Hsine and A. Arabo, "A Privacy-Preserving Middleware Architecture for Detecting Prompt Injection and Sensitive Data Exposure in Large-Language-Model Interactions," *Electronics*, vol. 15, no. 16, p. 3554, Aug. 2026, doi: 10.3390/electronics15163554.

[12] K. P. Vyas, M. Bhatt, D. Dudhatra, R. Gupta, and A. Sharma, "Malicious Prompt Classifier with Leave-One-Out Deletion Approach for Prompt Sanitization," *Procedia Computer Science*, vol. 282, pp. 3015–3024, 2026, doi: 10.1016/j.procs.2026.06.068.