Tạp chí điện tử
Khoa học và Công nghệ Trường Đại học Công nghệ Đông Á
Website Tạp chí:https://vjai.org.vn

Anti-scam Application for Multi-level Marketing and Conferencing Using Artificial Intelligence

Le Trung Thuc[1*], Nguyen Anh Quan[1], Hoang Manh Cuong[1], Ngo Thuy Giang[1], Nguyen Huu Phuong[1]

[1] Faculty of Information Technology, East Asia University of Technology

ARTICLE INFO

Keywords:
artificial intelligence,
cheat,
fraud detection,
small language models (SLMs)

ABSTRACT

The rapid growth of online multi-level marketing (MLM) and virtual conferencing platforms has increased the risk of fraudulent activities. Scammers exploit persuasive communication and social engineering techniques to deceive users, causing financial losses and reducing trust. This paper proposes an AI-based anti-scam application for detecting fraudulent behaviors in MLM and online conferencing environments. The system integrates Small Language Models (SLMs) with Retrieval-Augmented Generation (RAG) to provide efficient, context-aware scam detection in real time. By combining retrieved knowledge from scam databases and regulatory sources with live text or speech analysis, the system identifies deceptive patterns such as exaggerated profit claims and coercive recruitment language. Experimental results demonstrate promising detection accuracy with low latency suitable for live conferencing applications. The proposed approach is scalable, adaptable, and capable of integration with existing online meeting and MLM management platforms, contributing to improved security and transparency in digital communication environments.

* Corresponding author
Email: thuclt@eaut.edu.vn
DOI: https://doi.org/10.65153/efccc665
Received: 19/04/2026; Received in revised form: 02/06/2026; Accepted: 06/06/2026
Available online: 09/06/2026
Published by: East Asia University of Technology

1. INTRODUCTION

The rapid expansion of digital communication platforms has significantly transformed the way multi-level marketing (MLM) activities and business interactions are conducted. Online conferencing tools enable geographically distributed participants to communicate, collaborate, and conduct business activities efficiently. However, these platforms also create new opportunities for fraudulent and scam-related behaviors. Prior studies have shown that AI-based cybersecurity and fraud prevention have become increasingly important as cyber threats and online scams continue to evolve in complexity [1], [2], [3]. In particular, MLM-related scams often exploit real-time conversations, persuasive narratives, and social engineering techniques to manipulate participants, resulting in financial losses and long-term damage to user trust.

Traditional anti-fraud solutions primarily rely on rule-based mechanisms or manual moderation. Although these approaches can be effective in certain controlled settings, they are often insufficient for dynamic and real-time communication environments. Rule-based systems may fail to capture evolving scam strategies, while manual moderation is typically limited in scalability and responsiveness. As a result, these conventional approaches are unable to provide timely intervention during live conversations. Recent advances in artificial intelligence, particularly in natural language processing and speech analysis, have demonstrated considerable potential for automated scam and fraud detection [4], [5], [6]. Nevertheless, Large Language Models (LLMs) generally require substantial computational resources and may introduce latency that is unsuitable for real-time conferencing scenarios.

To address these limitations, this paper proposes an AI-driven anti-scam application that leverages Small Language Models (SLMs) in combination with Retrieval-Augmented Generation (RAG). The use of SLMs enables efficient and low-latency inference, making the system suitable for real-time deployment, including edge and on-device environments [7]. Meanwhile, the RAG framework enhances contextual understanding by retrieving relevant information from curated scam knowledge bases, historical fraud cases, and regulatory documents [8]. This hybrid approach allows the system to dynamically adapt to emerging scam patterns without requiring frequent model retraining.

The main contributions of this work are as follows. First, the paper reviews related studies on scam detection, fraud detection, language model-based analysis, and real-time communication security. Second, it proposes a real-time anti-fraud architecture suitable for MLM and online conferencing environments. Third, it integrates SLMs and RAG to balance detection accuracy, interpretability, and computational efficiency. Finally, it presents

an empirical evaluation demonstrating the feasibility and effectiveness of the proposed system in detecting fraud-related communications. Overall, this research aims to improve security, transparency, and trust in the modern digital communication ecosystem.

2. RELATED WORK

Artificial intelligence has been widely applied to cybersecurity and fraud prevention in recent years. Prior studies have shown that AI-based systems can support threat monitoring, anomaly detection, and fraud prevention across cybersecurity, e-commerce, financial transactions, insurance, and anti-money laundering domains [1], [2], [3], [9], [10], [11]. These studies demonstrate the potential of machine learning, deep learning, and natural language processing in identifying suspicious patterns and reducing fraud-related risks. However, most existing approaches focus on structured transactions, offline analysis, or general cybersecurity threats rather than real-time scam detection in conversational environments.

With the development of natural language processing, textual and conversational data have become important sources for fraud detection. Tsaliki [4] explored the use of Large Language Models for fraud prevention in e-commerce, while Boulieris et al. [5] introduced FraudNLP, showing that fraud-related user behaviors can be modeled as language-like sequences. These works indicate that fraudulent activities often contain semantic and linguistic signals that can be detected by NLP-based methods. Nevertheless, LLM-based solutions usually require substantial computational resources, which may limit their use in real-time conferencing and on-device applications.

Audio-text scam detection has also received increasing attention. Ma et al. [6] introduced TeleAntiFraud-28K, an audio-text dataset for telecom fraud detection that provides speech-text pairs, fraud annotations, and reasoning cues. This dataset is relevant to scam detection because fraudulent communication often occurs through spoken interactions. However, telecom fraud differs from MLM and online conferencing scenarios, where deceptive behaviors may involve exaggerated profit claims, recruitment pressure, misleading promises, and persuasive social interaction.

Recent research on compact language models and retrieval-augmented generation provides a promising direction for efficient and context-aware scam detection. TinyLlama is a lightweight 1.1B-parameter Small Language Model suitable for low-resource and edge-based deployment [7]. Meanwhile, Retrieval-Augmented Generation enables language models to use external knowledge sources during inference, improving contextual grounding and interpretability [8]. Multilingual text embedding models such as multilingual-e5-base can further support semantic retrieval across language contexts [12].

Overall, existing studies confirm the effectiveness of AI, NLP, and language models in fraud detection. However, three gaps remain. First, limited attention has been given to real-time scam detection in MLM and online conferencing environments. Second, Large Language Models may not be suitable for low-latency and on-device deployment. Third, many fraud detection systems provide classification results without sufficient contextual explanation. To address these gaps, this study proposes an SLM–RAG-based anti-scam system designed to balance detection accuracy, interpretability, computational efficiency, and real-time deployability.

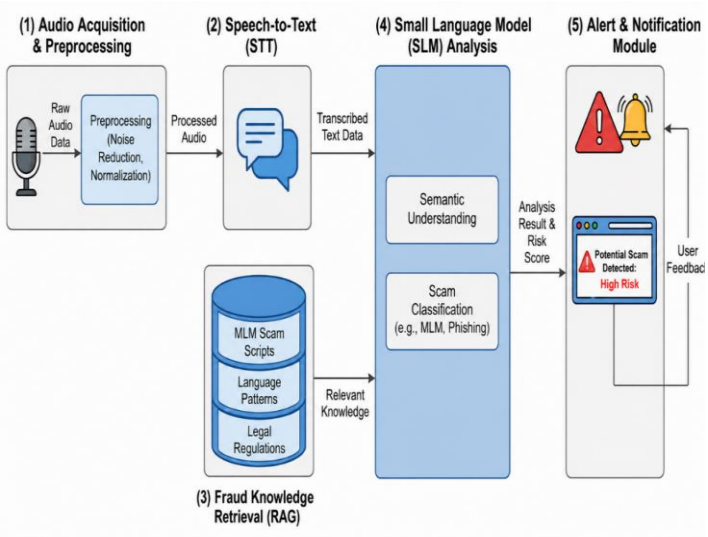

Figure 1. Overall architecture

3. ARCHITECTURAL DESIGN

This section presents the architecture of the proposed real-time anti-scam system designed for MLM and online conferencing environments. The system is developed to operate flexibly under different communication scenarios while ensuring low latency, privacy preservation, and reliable scam detection. To achieve these objectives, the architecture supports two complementary operating modes: post-meeting scam analysis and real-time scam monitoring.

3.1. Overall Architecture

The proposed architecture consists of five main components, as illustrated in Fig. 1: (1) audio acquisition and preprocessing, (2) speech-to-text transcription, (3) fraud knowledge retrieval based on Retrieval-Augmented Generation (RAG), (4) Small Language Model (SLM)-based analysis, and (5) alert and notification.

The first component captures raw audio data from online meetings or surrounding conversations and performs preprocessing operations such as noise reduction and normalization. The processed audio is then passed to the speech-to-text module, which converts spoken content into transcribed text data. In parallel, the fraud knowledge retrieval component retrieves relevant knowledge from predefined sources, including MLM scam scripts, language patterns, and legal regulations. This retrieved knowledge is used to enrich the analytical context of the SLM.

The SLM analysis component serves as the core reasoning module of the system. It performs semantic understanding and scam classification, such as identifying MLM-related fraud or phishing-like behaviors. By combining transcribed conversational data with retrieved fraud-related knowledge, the system produces an analysis result and a

corresponding risk score. Finally, the alert and notification module presents the detection outcome to the user through visual or auditory warnings. User feedback may also be collected to support future refinement of the system.

3.2. Post-Meeting Scam Analysis Mode

In the first operating mode, the system is applied to online meetings involving multiple participants. When a user initiates a meeting, the application records the entire audio session. The recording process automatically terminates when the meeting ends. To ensure privacy and user consent, the recorded data is processed only after the user explicitly authorizes the analysis.

Once permission is granted, the recorded audio is transcribed into text and segmented into conversational units. These segments are then analyzed by the SLM, which leverages the RAG module to retrieve relevant scam patterns, historical MLM fraud scenarios, and regulatory knowledge. Based on semantic similarity, persuasive intent, and behavioral indicators, the system evaluates the likelihood of scam-related activity within the conversation.

If the conversation is classified as potentially fraudulent, the system generates an alert and delivers it to the user. The alert may include a risk level, detected suspicious patterns of the reasons behind the warning. This post-meeting analysis mode is particularly useful for reviewing completed conversations, identifying potential manipulation attempts, and supporting users in making informed decisions after the meeting.

3.3. Real-Time Scam Detection Mode

The second operating mode focuses on real-time scam detection during live online meetings or in-person MLM presentations. In this mode, users can activate the application to continuously monitor the surrounding conversation or meeting audio.

The system processes incoming speech segments in near real time. The SLM continuously evaluates the semantic content of the conversation, while the RAG module dynamically retrieves relevant scam-related knowledge to update the contextual understanding of the system. When suspicious patterns are detected, such as exaggerated profit claims, coercive recruitment language, misleading promises, or repetitive manipulation strategies, the system immediately notifies the user through visual or auditory alerts.

3.4. Privacy and Ethical Considerations

Since the proposed system processes conversational data that may contain sensitive personal and financial information, privacy and ethical safeguards are essential. The system is designed to operate based on explicit user consent: recorded meetings are analyzed only after authorization, and real-time monitoring is activated manually by the user.

Privacy is further supported through data minimization and localized processing. By using Small Language Models (SLMs), the

system can perform on-device or edge-based inference, reducing the need to transmit raw audio or transcribed text to external servers. When RAG is used, anonymized or abstracted representations are preferred for knowledge retrieval, and recorded data can be deleted after analysis according to configurable retention policies.

The system is intended as a decision-support tool rather than an autonomous enforcement mechanism. Detection results are presented as risk assessments and warnings, not definitive accusations. In addition, the system does not support covert or unauthorized recording, thereby reducing the risk of surveillance misuse while supporting user safety, transparency, and responsible AI deployment.

4. Proposed method

To balance detection accuracy, interpretability, and computational efficiency, the proposed system integrates a Small Language Model (SLM) with a Retrieval-Augmented Generation (RAG) framework. This design addresses the limitations of conventional Large Language Models (LLMs), particularly their high computational requirements and latency, while maintaining reliable scam detection performance in real-time MLM and online conferencing environments.

Although SLM and RAG architectures are well-established, this study adapts and integrates them into a specialized online conference monitoring system for MLM-related scam detection. The proposed method provides several specific advantages.

First, the system is designed to identify MLM-specific scam language patterns. By combining retrieved contextual knowledge with real-time text or speech input, the system can detect deceptive patterns such as exaggerated profit claims, coercive recruitment language, misleading promises, and persuasive narratives commonly observed in MLM scams.

Second, the system supports real-time contextual adaptation. During live meetings, the SLM continuously analyzes incoming speech segments, while the RAG module dynamically retrieves relevant scam-related knowledge to update the system’s contextual understanding. This enables the model to interpret suspicious statements not only as isolated utterances but also in relation to known scam strategies and fraud scenarios.

Third, the system is optimized for low-latency deployment in conferencing environments. The use of lightweight SLMs, such as TinyLlama 1.1B or similar compact models, enables efficient inference suitable for edge-based or on-device execution. This design reduces dependence on large-scale cloud infrastructure and addresses the latency limitations of LLM-based approaches, which are often unsuitable for real-time online conferencing scenarios.

Fourth, the proposed method leverages a specialized anti-fraud knowledge base through the RAG module. In this study, the TeleAntiFraud-28K dataset is used as a specialized knowledge source for fraud-related communication patterns. TeleAntiFraud-28K is an open-source audio-text slow-thinking dataset designed for automated telecom fraud analysis. It contains 28,511 processed speech-text pairs and provides detailed annotations for fraud reasoning, making it suitable for studying models that can both detect fraudulent behavior and interpret the underlying fraud cues. The dataset is constructed using privacy-preserved transcriptions from anonymized call recordings, LLM-based self-instruction sampling, and multi-agent adversarial synthesis to improve scenario diversity. In the proposed system, retrieved knowledge from TeleAntiFraud-28K is provided to the SLM as auxiliary context, enabling more informed scam detection without increasing the model size or requiring frequent retraining [6].

Finally, the method incorporates privacy-aware design principles. The combination of SLM and RAG supports localized processing, reducing the need to transmit raw audio or transcribed text to external servers. In addition, the system is designed as a decision-support tool that provides risk warnings rather than definitive accusations, thereby reducing the risk of false judgments in sensitive conferencing environments.

Algorithm 1. Real-Time AI-Driven Anti-Scam Detection Algorithm

Input:

Capture audio *AU* input from the online conferencing session

Vector database *DB*

SLM

Embedding model *E*

Output:

Classification result O ∈ {0: potential fraud, 1: non-fraud}

Real-time alert A if suspicious activity is detected

1: Initialize the *SLM*, embedding model *E*, and vector database *DB*

2: Load the curated *TeleAntiFraud-28K* knowledge base into *DB*

3: **while** the live conferencing session is active **do**

4: Capture audio *AU* input from the meeting

5: Transcribe the audio into a conversational text segment *T*

6: Generate the query embedding $q = E(T)$

7: Retrieve the *top-k* relevant scam-related contexts R_k from *DB* using cosine similarity

8: Construct the analysis prompt P by integrating T with the retrieved knowledge R_k

9: Execute SLM inference: $O = SLM(P)$

10: Parse *O* to determine the scam likelihood, risk score S, and specific fraud tactics

11: **if** *O* indicates potential fraud **then**

12: Generate and deliver a real-time visual or auditory alert *A* to the user
13: **end if**
14: Log the analysis result *O* and alert *A* for post-meeting review
15: **end while**

Algorithm 1 describes the continuous real-time workflow of the proposed anti-scam detection system for live virtual conferencing sessions. The process begins by initializing the Small Language Model, the embedding model, and the ChromaDB vector database, which is preloaded with the curated *TeleAntiFraud-28K* knowledge base. During a live session, the system captures audio input and converts it into text segments using a speech-to-text module. Each text segment is transformed into an embedding vector and used to retrieve the most relevant scam-related contexts from the vector database.

The retrieved contexts are then combined with the current conversational segment to construct an analysis prompt for the SLM. Based on this prompt, the SLM evaluates whether the conversation contains deceptive language, coercive recruitment strategies, exaggerated profit claims, or other fraud-related indicators. If potential fraud is detected, the system generates a risk score and immediately notifies the user through visual or auditory alerts. The analysis results and alerts are also logged for post-meeting review, enabling users to examine suspicious interactions after the session has ended.

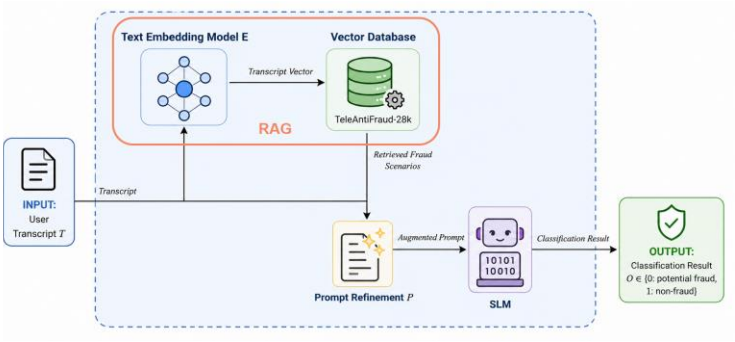

Figure 2. Proposed system architecture integrating SLM and RAG

Fig. 2 presents the proposed RAG-enhanced SLM architecture for the scam detection system. The architecture illustrates how the user transcript is transformed into a vector representation, enriched with retrieved fraud-related knowledge, and analyzed by the SLM to produce the final fraud decision. The main processing steps are described as follows:

1. **Input Ingestion:** The system receives the user dialogue transcript, denoted as T, which is generated from the live audio stream through the Speech-to-Text module.

2. **Query Vector Generation:** A text embedding model, namely *multilingual-e5-base* [12], encodes the transcript T into a dense query vector q = E(T).

3. **Context Retrieval:** The RAG module computes the semantic similarity between the query vector q and the reference embeddings stored in the TeleAntiFraud-28K vector database. Using top-k nearest neighbor retrieval, the system extracts the most relevant fraud scenarios and their associated metadata.

4. **Prompt Refinement:** The original transcript T, the retrieved fraud-related

contexts, and predefined system instructions are integrated to construct an augmented prompt P.

5. **SLM Inference and Analysis:** The augmented prompt P is forwarded to the Small Language Model (SLM), which analyzes the input context to identify deceptive tactics, generate a fraud-related reasoning summary, and provide defensive recommendations when necessary.

6. **Decision Processing:** The raw output of the SLM is parsed into a structured format, such as JSON. A threshold-based decision mechanism is then applied to the computed risk score to classify the final output O into one of two labels: potential fraud or non-fraud.

4.1. SLM Selection and Efficiency Considerations

The proposed system employs TinyLlama, a compact 1.1B-parameter Small Language Model (SLM), as the core semantic analysis engine [7]. Compared with large-scale language models, TinyLlama requires fewer computational resources, consumes less memory, and provides faster inference, making it suitable for real-time and edge-based deployment. Despite its compact architecture, the model retains sufficient language understanding capability to capture linguistic cues, persuasive intent, and conversational patterns related to scam behaviors. By using the SLM for primary inference, the system can maintain low latency in both post-meeting analysis and real-time monitoring scenarios.

4.2. RAG-Based Knowledge Augmentation

To compensate for the limited internal knowledge of the SLM, the system integrates a Retrieval-Augmented Generation (RAG) mechanism, which combines parametric model knowledge with retrieved external knowledge sources [8]. In the proposed system, the RAG module retrieves scam-related information from the TeleAntiFraud-28K dataset, a slow-thinking audio-text dataset designed for telecom fraud detection [6]. This dataset contains 28,511 processed audio-text pairs, totaling approximately 307 hours of audio, and provides fraud labels, reasoning cues, and diverse scam-related conversational patterns [6].

During inference, each conversational segment is encoded into a vector representation and matched against the vectorized knowledge base to retrieve the most relevant fraud-related contexts. For text representation, the multilingual-e5-base embedding model is used to generate dense multilingual embeddings, supporting effective semantic retrieval across language contexts [12]. ChromaDB is used as the vector database, and cosine similarity is applied to retrieve the top-k most relevant entries. In this study, top-k is set to 5 to balance contextual richness and the token constraints of the SLM.

The retrieved contexts are then integrated with the original transcript and system instructions to construct an augmented prompt. This prompt is provided to the SLM to support

fraud classification, risk scoring, reasoning summarization, and recommendation generation. This design enables the system to exploit external domain-specific knowledge without increasing the model size or requiring frequent retraining.

4.3. Interpretability and Balanced Decision-Making

The integration of SLM and RAG improves not only detection performance but also interpretability. Since each prediction is supported by retrieved fraud-related examples and contextual knowledge, the system can provide explanations indicating which linguistic patterns or fraud scenarios contribute to a high-risk assessment. This transparency is important in MLM and online conferencing environments, where misclassification and false accusations must be carefully avoided.

Overall, the proposed SLM–RAG integration provides a practical balance among detection accuracy, interpretability, and computational efficiency. By combining the lightweight inference capability of TinyLlama [7], external knowledge retrieval through RAG [8], multilingual semantic embedding with multilingual-e5-base [12], and domain-specific fraud knowledge from TeleAntiFraud-28K [6], the system is suitable for scalable and real-time anti-scam detection in MLM and online conferencing scenarios.

5. Evaluation

To evaluate the effectiveness and deployment feasibility of the proposed SLM–RAG architecture, this study uses both predictive performance metrics and system efficiency metrics. The evaluation is conducted on conversational data extracted from MLM-related meeting scenarios, with fraud-related knowledge supported by the TeleAntiFraud-28K dataset [6]. The proposed model is compared with two baseline architectures: an SLM-only model based on TinyLlama 1.1B [7] and an LLM-based model using LLaMA-2 7B. The embedding component is implemented using multilingual-e5-base for semantic representation and retrieval [12], while the knowledge augmentation mechanism follows the Retrieval-Augmented Generation paradigm [8].

The predictive performance is measured using Accuracy, Precision, Recall, and F1-score, which are commonly used in fraud detection and classification tasks [5], [9], [11]. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Accuracy measures the overall proportion of correctly classified samples and is defined as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

Precision measures the proportion of predicted fraud cases that are actually fraudulent:

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

Recall measures the proportion of actual fraud cases that are correctly detected:

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

F1-score is the harmonic mean of Precision and Recall, providing a balanced measure when both false positives and false negatives are important:

$$F1 - score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

In addition to predictive performance, the evaluation also considers system-level efficiency. Average inference latency measures the mean time required for the system to process one input segment and generate a prediction. It is computed as:

$$Average\ inference\ latency = \left(\frac{1}{N}\right) \times \sum t_i \quad (5)$$

where N is the number of tested samples and t_i is the inference time for the $i - th$ sample. Peak memory usage measures the maximum memory consumption observed during model inference and is defined as:

$$Peak\ memory\ usage = Max(m_i) \quad (6)$$

where m_i denotes the memory usage recorded during the $i - th$ inference process. These two metrics are important for assessing whether the system can be deployed in real-time and on-device environments.

Table 1. Quantitative comparison of different model architectures

Metric	**SLM-only** *(Tiny 1.1B)*	**SLM** *(Tiny 1.1B)* **– RAG** *(multilingual-e5-base)*	**LLM** *(LLaMA-2 7B)*
Accuracy (%)	81.4	**89.7**	90.8
Precision (%)	79.2	**88.1**	86.5
Recall (%)	83.6	**91.2**	92.4
F1-score (%)	81.3	**89.6**	89.4
Average inference latency (*ms*)	**38**	52	420
Peak memory usage (*GB*)	**2.1**	2.8	16.5
On-device deployability	Yes	**Yes**	No

As shown in Table 1, the proposed SLM–RAG model achieves strong detection performance while maintaining low computational cost. It obtains an accuracy of 89.70%, precision of 88.10%, recall of 91.20%, and F1-score of 89.60%. Compared with the SLM-only model, the proposed approach improves accuracy by 8.30 percentage points and F1-score by 8.30 percentage points. This improvement shows that integrating RAG with an SLM helps compensate for the limited internal knowledge of the compact model by providing relevant fraud-related contexts during inference.

Compared with the larger LLaMA-2 7B model, the proposed SLM–RAG architecture

achieves comparable predictive performance with substantially lower resource requirements. Although LLaMA-2 7B achieves slightly higher accuracy and recall, its average inference latency reaches 420 ms and its peak memory usage reaches 16.50 GB. In contrast, the proposed SLM–RAG model requires only 52 ms of average inference latency and 2.80 GB of peak memory usage. These results indicate that the proposed architecture is more suitable for real-time and on-device deployment scenarios.

Overall, the evaluation results demonstrate that the proposed SLM–RAG architecture provides a practical balance among detection accuracy, inference latency, memory efficiency, and on-device deployability.

6. Conclusions

This paper proposed an AI-driven anti-scam detection system that integrates a Small Language Model (SLM) with Retrieval-Augmented Generation (RAG) for real-time scam detection in MLM and online conferencing environments. The proposed architecture combines lightweight language model inference with external fraud-related knowledge to improve contextual understanding, detection accuracy, and interpretability while maintaining low latency and on-device deployability.

Experimental results demonstrate that the proposed SLM–RAG approach achieves a strong balance between predictive performance and computational efficiency. Compared with the SLM-only baseline, it provides better accuracy, recall, F1-score, and interpretability. Compared with the LLM-based architecture, it achieves comparable detection performance with substantially lower inference latency and memory usage, making it more suitable for real-time and edge-based deployment.

Overall, the proposed system provides a practical and scalable solution for enhancing security, transparency, and user awareness in MLM and online conferencing contexts. Future work will focus on expanding the fraud knowledge base, improving multilingual scam detection, incorporating user feedback, and evaluating the system in broader real-world communication scenarios.

REFERENCES

[1] R. M. Rajendran and B. Vyas, “Cyber security threat and its prevention through artificial intelligence technology,” International Journal for Multidisciplinary Research, vol. 5, no. 6, 2023.

[2] M. Z. Siddiqui, S. Yadav, and M. S. Husain, “Application of artificial intelligence in fighting against cyber crimes: A review,” International Journal of Advanced Research in Computer Science, vol. 9, no. 2, pp. 118–122, 2018.

[3] M. M. Mijwil, M. Aljanabi, and C. ChatGPT, “Towards artificial intelligence-based cybersecurity: The practices and ChatGPT generated ways to combat cybercrime,” Iraqi Journal for Computer

Science and Mathematics, vol. 4, no. 1, p. 8, 2023.

[4] K. C. Tsaliki, "Leveraging Large Language Models for fraud prevention in e-commerce," International Journal of Innovative Research in Science, Engineering and Technology, vol. 13, no. 8, 2024.

[5] P. Boulieris, J. Pavlopoulos, A. Xenos, and V. Vassalos, "Fraud detection with natural language processing," Machine Learning, vol. 113, no. 8, pp. 5087–5108, 2024.

[6] Z. Ma, P. Wang, M. Huang, J. Wang, K. Wu, X. Lv, Y. Pang, Y. Yang, W. Tang, and Y. Kang, "TeleAntiFraud-28K: An audio-text slow-thinking dataset for telecom fraud detection," 2025. [Online]. Available: https://arxiv.org/abs/2503.24115. [Accessed: Jun. 1, 2026].

[7] P. Zhang, G. Zeng, T. Wang, and W. Lu, "TinyLlama: An open-source Small Language Model," arXiv preprint arXiv:2401.02385, 2024.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[9] S. Lenka and R. Tiwari, "Real-time fraud prevention in digital wallet transactions using CNN-RNN hybrid networks," Cuestiones de Fisioterapia, vol. 54, no. 2, pp. 533–541, 2025.

[10] C. S. Pareek, "From detection to prevention: The evolution of fraud testing frameworks in insurance through AI," Journal of Artificial Intelligence, Machine Learning and Data Science, vol. 1, no. 2, pp. 1805–1812, 2023.

[11] O. Iguodala and A. Oyiborhoro, "AI-powered anti-money laundering (AML) and fraud detection: Enhancing financial security through intelligent fraud detection," World Journal of Advanced Research and Reviews, vol. 26, pp. 3702–3714, 2025.

[12] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 text embeddings: A technical report," arXiv preprint arXiv:2402.05672, 2024.