One Size Fits All? A Modular Adaptive Sanitization Kit (MASK) for Customizable Privacy-Preserving Phone Scam Detection

Kangzhong Wang
The Hong Kong Polytechnic University
Hong Kong, China
kangzhong.wang@connect.polyu.hk

Zitong Shen
The Hong Kong Polytechnic University
Hong Kong, China
esther.shen@connect.polyu.hk

Youqian Zhang*
The Hong Kong Polytechnic University
Hong Kong, China
you-qian.zhang@polyu.edu.hk

Michael MK Cheung
The Hong Kong Polytechnic University
Hong Kong, China
man-ki-michael-mk.cheung@polyu.edu.hk

Xiapu Luo
The Hong Kong Polytechnic University
Hong Kong, China
csxluo@comp.polyu.edu.hk

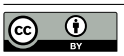
Grace Ngai
The Hong Kong Polytechnic University
Hong Kong, China
grace.ngai@polyu.edu.hk

Eugene Yujun Fu
The Education University of Hong Kong
Hong Kong, China
eugenefu@eduhk.hk

Abstract

Phone scams remain a pervasive threat to both personal safety and financial security worldwide. Recent advances in large language models (LLMs) have demonstrated strong potential in detecting fraudulent behavior by analyzing transcribed phone conversations. However, these capabilities introduce notable privacy risks, as such conversations frequently contain sensitive personal information that may be exposed to third-party service providers during processing. In this work, we explore how to harness LLMs for phone scam detection while preserving user privacy. We propose MASK (Modular Adaptive Sanitization Kit), a trainable and extensible framework that enables dynamic privacy adjustment based on individual preferences. MASK provides a pluggable architecture that accommodates diverse sanitization methods—from traditional keyword-based techniques for high-privacy users to sophisticated neural approaches for those prioritizing accuracy. We also discuss potential modeling approaches and loss function designs for future development, enabling the creation of truly personalized, privacy-aware LLM-based detection systems that balance user trust and detection effectiveness, even beyond phone scam context.

*Corresponding author.



MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758164

CCS Concepts

• **Human-centered computing**; • **Security and privacy**; • **Computing methodologies** → **Artificial intelligence**; • **Applied computing**;

Keywords

Sanitization, Phone Scam, Large Language Model, User-Centric Privacy, Modular Frameworks

ACM Reference Format:
Kangzhong Wang, Zitong Shen, Youqian Zhang, Michael MK Cheung, Xiapu Luo, Grace Ngai, and Eugene Yujun Fu. 2025. One Size Fits All? A Modular Adaptive Sanitization Kit (MASK) for Customizable Privacy-Preserving Phone Scam Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746027.3758164

1 Introduction

Phone scams remain a persistent and evolving threat to individuals and society. By impersonating authorities, family members, or service providers, scammers deceive victims over the phone for illegal financial gain. This form of fraud continues to cause significant economic damage: for instance, recent reports estimate that in 2024 alone, phone scams led to trillions of dollars in losses globally [31]. Moreover, victims may also suffer psychological trauma, anxiety, and long-term distrust in communication channels [6, 19, 20].

Efforts to prevent phone scams have been ongoing for years. Governments and organizations have adopted measures such as public awareness campaigns [5, 17, 21, 35] and technical solutions like caller ID verification [37]. However, as defensive mechanisms improve, so too do the tactics of scammers, who constantly find ways to bypass existing safeguards [27, 39]. In recent years, researchers have begun to shift their focus toward the content of

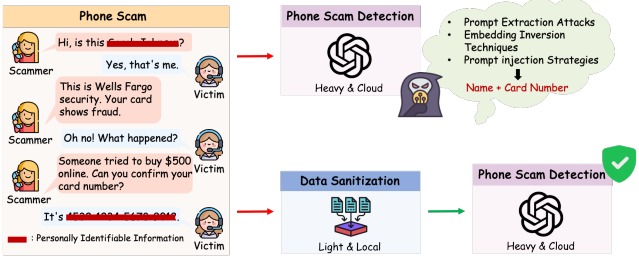


Figure 1: LLM-based phone scam detection approaches. Cloud-based models possibly risk privacy leakage, while data sanitization (at user/local devices) enables safe detection with cloud models.

scam communications [9, 10, 29, 32, 33, 38]. Since scammers must engage their targets in conversation, analyzing these interactions has emerged as a promising detection strategy. With the rapid advancement of large language models (LLMs), it has become increasingly feasible to analyze voice-to-text transcriptions of phone calls and detect scam patterns with high accuracy. This approach extends to processing multimedia content, leveraging the rich information embedded in audio and potentially other modalities. Several recent studies [11, 26, 32, 33] have successfully demonstrated the effectiveness of LLMs in identifying scam-related content in conversations (see more details in Section 2.1).

However, conversations between a caller and recipient often contain sensitive information, and the LLM-based approach introduces a serious concern, user privacy. As shown in Fig. 1, sending these conversations to a centralized LLM (e.g., a cloud-based server) for analysis can expose users to significant privacy risks: If such data were to be leaked or misused, the consequences could be severe, ranging from identity theft to further exploitation (more discussion in Section 2.2). It's crucial to note that traditional encryption does not fully solve the privacy problem, as encrypted data still needs to be decrypted before being processed by an LLM, during which the sensitive content is exposed.

To mitigate privacy risks in LLM-powered applications, recent work has proposed prompt sanitization techniques that remove or mask sensitive information before sending it to the model [1, 2, 8, 12, 34, 40]. However, the context of phone scam detection introduces unique challenges. Conversations in this setting are often highly sensitive, involving personal, financial, or emotional information, and the sensitivity of the content evolves dynamically throughout the conversation process. Moreover, real-time scam detection requires low-latency processing and minimal user intervention, making traditional interactive sanitization interfaces (e.g., highlight-and-adjust UIs) impractical. Further, users' expectations and preferences for privacy vary widely, influenced by their personal values, contexts, and risk perceptions, and could be shaped in real-time as the conversation unfolds. Prior research has shown that users' privacy needs in LLM-based applications are deeply subjective and context-dependent [24, 44, 45].

We argue that protecting privacy in phone scam detection should *center user autonomy and preference* and *go beyond static sanitization rules*, which most follow a "one-size-fits-all" paradigm, failing to account for the subjective, contextual, and evolving nature of privacy preferences. We propose the Modular Adaptive Sanitization Kit (MASK), a user-centric, configurable privacy framework designed to support diverse privacy needs in real-time. MASK marks a brave and novel departure from conventional thinking: Rather than enforcing a fixed sanitization pipeline, MASK embraces privacy pluralism by allowing users to dynamically configure how much and what kind of information gets sanitized across multimedia content, such as voice and text data. It respects the user's autonomy, adapts to their evolving context, and supports interactive, real-time privacy decisions, all while maintaining the utility needed for accurate phone scam detection.

2 Background

2.1 LLM-Based Phone Scam Detection

A few studies have explored the use of large language models (LLMs) for real-time detection of telecom fraud through the analysis of phone call transcripts. Shen et al. [33] proposed a framework that evaluates fraudulent intent in conversations and delivers immediate warnings to users. Their work highlights a key trade-off between detection accuracy and response timeliness but remains focused solely on textual input. In another work, Shen et al. [32] examined the strengths and limitations of LLMs in this domain, identifying challenges such as data bias, hallucinations, and low recall, and emphasizing the need for high-quality, diverse training datasets. In a follow-up work, Ma and Wang et al., [26] create a synthetic phone scam datasets to facilitate extensive studies in this area. Chang et al., [11] further investigated the vulnerability of LLMs to adversarial scam messages, constructing a dataset of original and manipulated texts to analyze misclassification rates.

2.2 Privacy Challenges in Cloud-Based LLMs

Deploying large language models (LLMs) in cloud environments raises significant privacy concerns, particularly when users submit sensitive data via conversational prompts. These interactions often include personally identifiable information (PII)—such as names, addresses, and financial details—which must be transmitted to LLM service providers for processing [23]. However, such data is typically logged and stored by the providers [23, 43], increasing the risk of privacy breaches. Stored prompts can be exploited through a range of emerging attack vectors, including prompt injection, embedding inversion, and prompt extraction [7, 25]. Prompt injection allows attackers to manipulate LLM behavior, potentially eliciting unauthorized disclosure of user inputs. Embedding inversion techniques can reconstruct original prompts from internal model embeddings, while prompt extraction attacks can retrieve user data directly from logged interactions [7, 25, 43]. These threats reveal a critical paradox in phone scam detection: the very AI tools designed to protect users may themselves become sources of privacy risk. This underscores the urgent need for privacy-preserving mechanisms that can mitigate these vulnerabilities while maintaining the effectiveness of LLM-based detection systems.

2.3 Data Sanitization

A critical assumption in privacy-preserving LLM applications is that private information can be efficiently identified and removed [3].

Existing sanitization techniques typically rely on rule-based filters, machine learning-based named entity recognition (NER), and local LLMs for topic identification. Rule-based methods redact predefined patterns of personally identifiable information (PII), while NER models detect and mask entities such as names and addresses [34]. More recently, lightweight local LLMs deployed on user devices have been used to identify sensitive topics, enabling selective sharing of non-sensitive content with cloud-based services [22, 34]. Despite their promise, these methods face challenges. Over-sanitization can degrade the utility of data for downstream tasks, including scam detection, while under-sanitization risks PII leakage due to the context-dependent subtleties of natural language [4]. Alternative approaches such as differential privacy, attempt to obscure PII by injecting statistical noise. However, they often incur high computational costs and struggle to preserve semantic fidelity in nuanced conversational contexts [4, 41].

3 Modular Adaptive Sanitization Kit (MASK)

To addresses the limitations of existing one-size-fits-all privacy-preserving approaches. This paper proposes and explores the design of the Modular Adaptive Sanitization Kit (MASK). At its core, MASK is a flexible, user-centric privacy-preserving framework that employs a pluggable architecture that allows users to dynamically select and configure sanitization modules based on their individual privacy preferences and contextual requirements. In particular, MASK consists of three primary components (Fig. 2 (b)): a privacy preference adapter, a modular sanitization layer, and an extensible plugin architecture.

3.1 Privacy Preference Adapter and Risk Tolerance Parameter

The privacy preference adapter serves as the central component in MASK that mapping user-defined privacy requirements to modular sanitzation strategies. The adapter supports fine-grained privacy control, particularly, through user-specific risk tolerance. The "*risk tolerance*" parameter functions analogously to the "*temperature*" setting in commercial LLMs that controls output randomness and creativity. In MASK, *risk tolerance* provides users with intuitive control over sanitization intensity, directly affecting the aggressiveness of privacy-preserving operations. A low *risk tolerance* (similar to low *temperature*) trigger the MASK to perform over-protection with more aggressive sanitization strategies and techniques, resulting in conservative, high-level of privacy-preserving prompts sent to the cloud server.

Fig. 2 (b) illustrates this concept using the input phrase "*Hello, is this Wang?*" as an example. Under low risk tolerance settings, the system might apply extreme sanitization by converting the entire input into a tokenized feature vector that only preserves structural information, such as [1, 1, 0, 0, ...] representing the count of different entity types (e.g., one name entity, one question marker, zero location entities, zero phone numbers). This approach maximizes privacy by eliminating all semantic content while retaining minimal structural signals for detection. In contrast, a higher risk tolerance setting would apply selective privacy entity masking, producing outputs like "*Hello, is this [Name]?*" that preserve the conversational structure and context while only masking explicit personal identifiers. We anticipate that users can specify their privacy requirements for MASK through intuitive interfaces that map their privacy preferences into concrete sanitization behaviors. This parameterization allows the same framework to serve privacy-conscious users who prioritize data protection and accuracy-prioritized users who need maximum detection performance.

Furthermore, we anticipate that the preference adapter can go beyond fixed approaches that rely on static mappings between risk tolerance and specific sanitizers. Instead, it supports more fine-grained privacy control through a multi-dimensional preference model that jointly considers multiple factors such as entity and data sensitivity levels, contextual information, and user risk tolerance in an integrated manner. A particular promising direction is to implement with neural network models that can dynamically balance the trade-off between risk tolerance (privacy preservation rate) and semantic maintenance. This sophisticated balancing can be implemented through joint modeling approaches that optimize both privacy preservation and semantic similarity as components of a unified loss function in future developments.

3.2 Modular Sanitization Layer Supporting Extensible Plugin

The modular sanitization layer implements a diverse collection of privacy-preserving methods which organized as interchangeable modules. Our MASK is initially implemented with four primary sanitization strategies, namely, (1) *TF-IDF Keyword Representation*, (2) *PII Statistical Representation*, (3) *PII Masking Anonymization*, and (4) *Transcript Summarization* (Details in Section 4). This offers the initial set of options matching users' different *risk tolerance* levels. For example, for users demanding maximum privacy protection, the traditional TF-IDF keyword-based and statistical methods that operate without deep semantic understanding, ensuring minimal information leakage at the cost of potential accuracy reduction.

To accommodate future developments and specialized requirements, MASK is implemented with an extensible plugin architecture that allows researchers and developers to integrate new sanitization techniques seamlessly. This forward-looking design includes well-defined APIs and standardized interfaces that enable the incorporation of emerging privacy-preserving methods, from rule-based detection methods to complex neural network approaches. Moreover, the plugin architecture also supports trainable components, allowing sanitization modules to adapt and improve their performance while maintaining user-specified privacy constraints by neural network with multi-task learning.

4 Privacy-Preserving Sanitization Strategies

In this section, we present four primary sanitization strategies that were initially implemented and examined within the Modular Sanitization Layer of MASK. It is important to note that these strategies are provided as examples, and the proposed framework offers flexibility for customization, as discussed earlier. In addition, note that in the following sections of this study, we utilize the datasets provided by [32], which consist of data collected from real scam and benign phone calls in Chinese, rather than synthetic

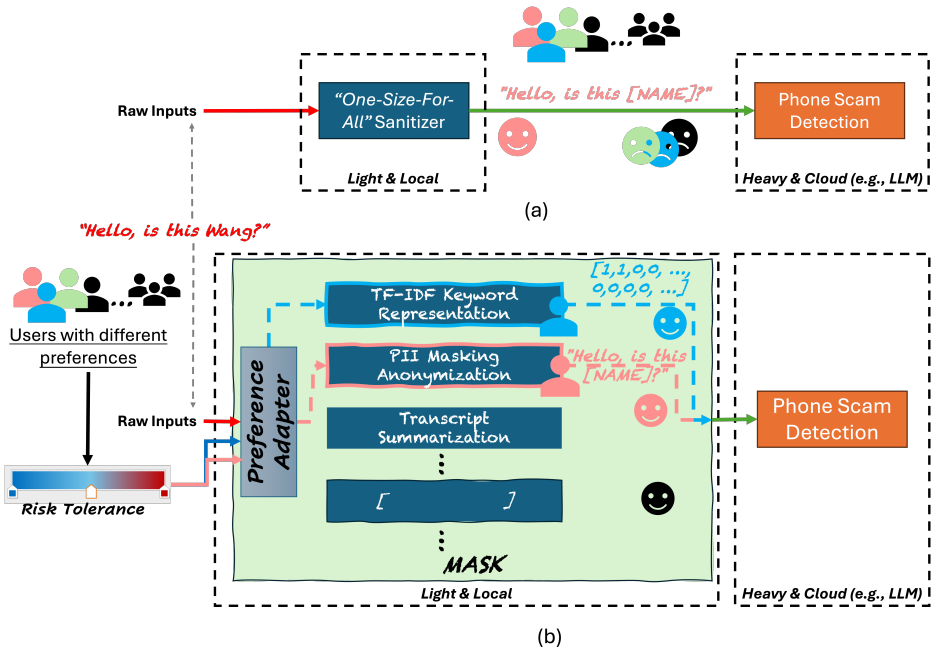


Figure 2: Overview of privacy-preserving frameworks: (a) Existing methods employ fixed (e.g., PII-like entity masking) with one-size-fits-all approaches, offering limited flexibility and no user preference customization. (b) The proposed Modular Adaptive Sanitization Kit (MASK) enables dynamic privacy adjustment through a pluggable architecture that accommodates various sanitization methods such as traditional keyword-based and PII-like methods for users with different privacy preferences. It also provides extensible plugin capabilities for future enhancements, particularly for sophisticated neural network approaches.

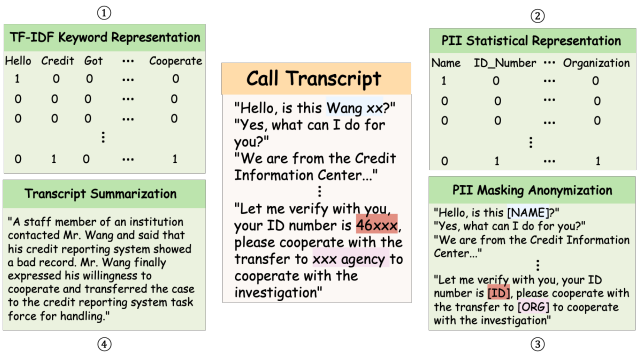


Figure 3: Overview of the four sanitization strategies: (1) TF-IDF keyword representation, (2) PII statistical representation, (3) PII masking anonymization, and (4) transcript summarization.

data. We believe this choice better reflects real-world scenarios and enhances the practical value of this work.

4.1 TF-IDF Keyword Representation

Term frequency-inverse document frequency (TF-IDF) [36] is a widely used statistical method in information retrieval and text mining that quantifies the importance of a word within a document relative to an entire corpus. In this study, the so-called "TF-IDF Keyword Representation" method converts each conversational transcript into a structured numeric vector format based on the frequency of selected discriminative keywords.

This method highlights words that are more indicative of scam-related content compared to normal conversations, thereby preserving critical semantic information while mitigating the risk of privacy leakage, as shown in Figure 3. Note that to identify the most discriminative features, the difference in mean TF-IDF scores between the two classes (scam v.s. normal phone calls) is calculated, and the top-N keywords with the largest absolute differences are selected. For each utterance in a transcript, a fixed-length numeric vector is generated, where each element represents the occurrence frequency of a selected keyword within that utterance. These numeric vectors serve as privacy-preserving representations for subsequent downstream scam detection tasks using cloud-based LLMs.

4.2 PII Statistical Representation and Masking Anonymization

PII refers to personally identifiable information. Both PII Statistical Representation and PII Masking Anonymization employ a two-step hybrid approach to in conversational transcripts. As shown in Figure 3, each transcript is processed using regular expression-based pattern matching to detect structured PII, such as identification numbers, phone numbers, account identifiers, emails, URLs,

bank card numbers, and dates, and a neural named entity recognition (NER) model [18] to capture unstructured PII, including personal names, organizations, and locations. All detected entities are replaced with standardized, category-specific placeholders (e.g., `[ID]`, `[PHONE]`, `[NAME]`, `[ORG]`, `[LOC]`), with unmatched numeric sequences masked as `[NUM]`.

The two strategies differ in how they represent the processed transcripts for downstream analysis:

- **PII Statistical Representation** transforms each utterance into a fixed-length vector, where each dimension reflects the count of a specific PII category within the utterance.
- **PII Masking Anonymization** outputs an anonymized transcript in which all identified PII entities are substituted by their category placeholders.

By distinguishing between statistical abstraction and structural anonymization, these methods provide flexible options for balancing privacy protection and semantic retention in language-based detection pipelines.

4.3 Transcript Summarization

Transcript Summarization is designed to maximize privacy protection and reduce input redundancy by condensing entire conversational transcripts into concise and high-level summaries. These summaries are constructed to retain only the core events and essential information, while systematically omitting extraneous details and all personally sensitive content. This method leverages local large language models (LLMs) to generate privacy-preserving, document-level representations prior to any cloud-based analysis, thereby minimizing the risk of sensitive data exposure. In this study,we select Qwen3-4b [42] as our local lightweight LLM for transcript summarization. Qwen3-4b combines a compact model size, which is suitable for efficient deployment on local or resource-constrained devices including smartphones, with robust generation capabilities and high summarization quality on Chinese conversational data. These characteristics make Qwen3-4b particularly well suited for secure, edge-based preprocessing in real-world scenarios. As shown in Figure 3, for each transcript, the complete textual content is provided as input to the local LLM. The model is prompted to produce a brief summary that distills the most salient events, actions, and key facts, while explicitly excluding personal names, identifiers, contact information, or subjective commentary. The output is a single, coherent summary statement for each transcript, ensuring that only non-sensitive, task-relevant information is utilized for downstream scam detection analysis.

4.4 Evaluation Metrics of Sanitization

We evaluate the effectiveness of our sanitization strategies along two primary dimensions: (1) privacy preservation and (2) semantic retention.

Privacy Preservation. To quantitatively assess the privacy-preserving capacity of each sanitization strategy, we introduce the *PII Removal Rate* (PRR). This metric measures the proportion of PII entities that are successfully removed or obfuscated from transcripts by the sanitization process. For each transcript, we employ a comprehensive PII detection pipeline, combining regular expression matching and NER to identify all types of PII in both the original and sanitized versions. The PRR is computed as:

$$\mathrm{PRR} = 1 - \frac{\sum_{i=1}^{N} |\mathrm{PII}^{(i)}_{\mathrm{sanitized}}|}{\sum_{i=1}^{N} |\mathrm{PII}^{(i)}_{\mathrm{raw}}|} \tag{1}$$

where $\mathrm{PII}^{(i)}_{\mathrm{raw}}$ and $\mathrm{PII}^{(i)}_{\mathrm{sanitized}}$ denote the sets of detected PII in the i-th transcript before and after sanitization, respectively, and N is the total number of evaluated transcripts. Higher PRR values indicate a greater degree of privacy protection.

Semantic Retention. It is crucial that sanitization methods do not excessively degrade the semantic content necessary for downstream LLM-based analysis. To this end, we assess the semantic retention rate (SRR) of each strategy by computing the average semantic similarity between the original and sanitized versions of each transcript:

$$\mathrm{SRR} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{Sim}(T_i, \hat{T}_i) \tag{2}$$

where T_i and $\hat{T}_i$ denote the original and sanitized versions of the i-th transcript, respectively, and $\mathrm{Sim}(\cdot,\cdot)$ represents a document-level semantic similarity function. In this study, we employ Sentence-BERT cosine similarity [30] to quantify the preservation of semantic content. Higher SR values indicate better retention of core semantic information, thereby allowing the LLM to more accurately interpret user intent and fulfill the intended task.

5 Experimental Results

We comprehensively evaluate the effectiveness of the proposed sanitization strategies for phone scam detection on conversational transcripts. Experiments are conducted using four advanced large language models (LLMs) that have demonstrated strong performance in Chinese understanding tasks [13]: ChatGPT-4o (or simply denoted as GPT)[28], Gemini-2.5-Flash (or Gemini) [15], Qwen2.5-MAX (or Qwen) [14], and DeepSeek-V3 (or DeepSeek)[16]. To systematically compare the privacy-utility trade-offs of each approach, all sanitization methods are applied to the dataset, and the processed transcripts are used as input for each LLM-based scam detector. Detection performance is reported using standard evaluation metrics, including accuracy (Acc.), precision (P.), recall (R.), and F1 score, across all combinations of data strategy and model. The results presented below highlight the comparative robustness of each LLM under the different sanitization methods.

5.1 Overall Evaluation

Table 1 presents a systematic comparison of the sanitization strategies in terms of their impact on both privacy protection and detection performance. The evaluation covers all four detection models and captures differences in semantic retention, privacy risk, and the preservation of critical conversational information.

The **PII Masking Anonymization** approach achieves a strong balance between privacy and detection effectiveness. By removing all identified sensitive information (PRR = 1.000) while largely retaining the semantic content of the original conversation (SRR = 0.861), this method maintains nearly the same F1 scores as the unprocessed baseline. The minimal decline in detection accuracy indicates that replacing explicit entities with category placeholders

Sanitization	Model	Acc.	P.	R.	F1	PRR	SRR
Original Text	GPT	0.97	**1.00**	0.94	0.97	0.000	1.000
	Gemini	0.96	0.96	0.96	**0.96**		
	Qwen	**0.98**	**1.00**	**0.96**	**0.98**		
	DeepSeek	**0.98**	**1.00**	**0.96**	**0.98**		
TF-IDF Keyword Representation	GPT	**0.83**	0.88	0.76	**0.82**	1.000	0.168
	Gemini	0.49	0.49	**0.98**	0.66		
	Qwen	0.67	**1.00**	0.32	0.48		
	DeepSeek	0.79	**1.00**	0.58	0.73		
PII Statistical Representation	GPT	0.27	0.21	**0.16**	**0.18**	1.000	0.085
	Gemini	**0.53**	**1.00**	0.06	0.11		
	Qwen	0.52	0.75	0.06	0.11		
	DeepSeek	**0.53**	**1.00**	0.06	0.11		
PII Masking Anonymization	GPT	0.96	**1.00**	0.92	0.96	1.000	0.861
	Gemini	**0.98**	**1.00**	0.96	**0.98**		
	Qwen	0.94	0.89	**1.00**	0.94		
	DeepSeek	**0.98**	**1.00**	0.96	**0.98**		
Transcript Summarization	GPT	0.88	**1.00**	0.76	0.86	0.903	0.412
	Gemini	**0.89**	**1.00**	**0.78**	**0.88**		
	Qwen	0.87	0.97	0.76	0.85		
	DeepSeek	0.86	**1.00**	0.72	0.84		

Table 1: Performance of different sanitization strategies and detection models. Metrics include accuracy, precision, recall, F1 score, privacy removal rate (PRR), and semantic retention rate (SRR).

is sufficient to prevent privacy leakage, while preserving the narrative and interaction cues necessary for reliable classification. This level of performance demonstrates that structured anonymization at the entity level supports both compliance with privacy standards and robust detection results.

By comparison, the **TF-IDF Keyword Representation** and **PII Statistical Representation** methods, although highly effective in removing sensitive details (PRR = 1.000), result in considerable degradation of semantic content. With much lower semantic retention rates (SRR = 0.168 and 0.085, respectively) and a pronounced drop in F1 scores, these strategies often discard critical context and cues , making it more difficult for the model to accurately distinguish scams from ordinary conversation.

The **Transcript Summarization** approach provides a moderate level of privacy protection (PRR = 0.903) and intermediate semantic retention (SRR = 0.412), with corresponding F1 scores situated between those of the entity masking and abstraction-based methods. Summarization condenses the transcript to its main points, omitting sensitive or redundant material, which can be advantageous for applications requiring smaller inputs or strict privacy requirements. Nevertheless, this reduction in detail sometimes leads to a loss of information relevant for accurate detection, and thus performance does not reach the level observed with entity-level masking.

These results illustrate the fundamental trade-off between privacy protection and semantic utility in the context of transcript sanitization. Entity-level masking stands out as the most effective strategy for balancing both aims, providing strong privacy safeguards with minimal compromise in downstream task performance. Methods based on greater abstraction or summarization, while sometimes necessary for stricter privacy or resource constraints, tend to increase the risk of missing critical context and reducing detection reliability. Careful selection of sanitization techniques is therefore essential, particularly in applications where both privacy and detection accuracy are priorities.

5.2 Impact of Sanitization on Precision and Recall

While the accuracy and F1 score offers an overall measure of detection performance, a more detailed analysis of precision and recall provides greater insight into the strengths and weaknesses of each sanitization strategy. Precision reflects the proportion of predicted scam cases that are indeed scams, highlighting the model's ability to avoid false positives. Recall, in contrast, measures the fraction of true scam cases that are correctly detected, indicating how well the model avoids missed detections. Both metrics are critical for practical deployment: high recall ensures effective identification of scams, while high precision reduces disruption for legitimate users.

Figure 4 visualizes the trade-off between privacy protection and utility for all sanitization methods. Each subplot displays the relationship between PII removal rate (PRR), semantic retention rate (SRR), and detection performance. In panel (a), circle size represents precision; in panel (b), circle size reflects recall. Colors are used to distinguish different sanitization approaches.

The results indicate that the **Original Text** and **PII Masking Anonymization** strategies both achieved high precision and recall, clustering in the upper part of each plot with large circle sizes. This suggests that entity-level masking preserves most detection capability while eliminating explicit identifiers. In contrast, both **TF-IDF Keyword Representation** and **PII Statistical Representation** achieve high PII removal but exhibit a substantial loss of semantic retention, which is reflected in much smaller circle sizes, especially for precision, indicating lower detection reliability. These methods tend to generate more false positives, largely due to the absence of conversational detail and context needed for robust discrimination between scam and non-scam transcripts.

The **Transcript Summarization** method occupies an intermediate position, with moderate semantic retention and detection performance. While summarization reduces input redundancy and enhances privacy, the moderate circle sizes indicate some loss of information needed for complete scam detection. The method is suitable in settings where privacy or input length constraints are strict, but users should be aware of the accompanying decline in recall and precision.

Overall, the visual analysis in Figure 4 underscores the need to maintain a balance between privacy protection and detection accuracy. Approaches that retain conversational structure, such as entity-level masking, offer the most reliable compromise. Excessive abstraction or compression, however, tends to compromise one or both critical aspects of detection.

5.3 Qualitative Comparison of LLM Responses Across Sanitization Methods

We further investigate how different sanitization strategies affect the reasoning and explanation provided by LLMs in scam detection. Representative model outputs were systematically reviewed to identify how varying levels of abstraction and information loss influence both the accuracy and interpretability of LLM-generated responses.

TF-IDF Keyword Representation. When the input is represented by TF-IDF keywords, the model primarily focuses on the

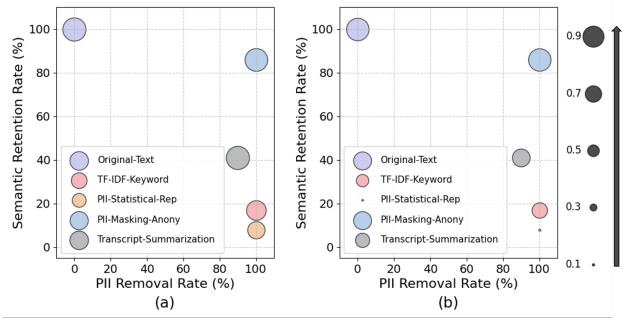


Figure 4: Trade-off between PII removal rate (PRR), semantic retention rate (SRR), and detection performance. (a) Circle size represents precision. (b) Circle size represents recall. Colors indicate different sanitization strategies.

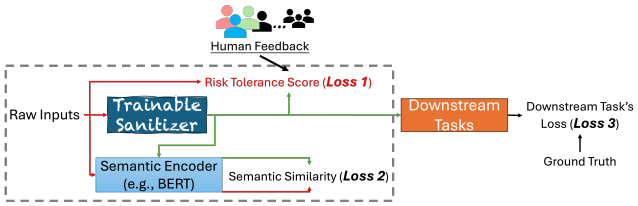


Figure 5: Trainable sanitization framework that jointly optimizes privacy, semantic retention, and task accuracy.

Local LLM	Accuracy	Precision	Recall	F1 Score
Gemma3-4b	0.62	0.57	0.98	0.72
Qwen3-4b-Think	0.79	0.70	1.00	0.83
Qwen3-4b-NoThink	0.70	0.83	0.50	0.62

Table 2: Performances of local LLMs for phone scam detection on original transcripts.

frequency and presence of specific terms associated with scam scenarios. The explanations tend to be formulaic, with decisions based on whether high-risk words such as "police", "identity card", or "account" are present. This method can sometimes identify scams that match expected vocabulary, but it lacks the ability to interpret context or conversational nuance. As a result, it often produces explanations that are repetitive and shallow, and it can miss less typical scams or produce false alarms when benign conversations contain similar keywords.

PII Statistical Representation. Using statistical features of personally identifiable information (PII) types, the model's judgments are mainly based on the diversity and frequency of sensitive entities, such as names, numbers, or time references. The LLM typically justifies its output by noting the absence of multiple or high-risk identifiers, for example, the lack of ID numbers, phone numbers, or bank card information. This method is effective for scams that depend on explicit information exchange, but it is limited in its ability to detect scenarios where scams rely more on manipulation or indirect cues. As a result, the explanations are narrow and sometimes overlook relevant behavioral patterns.

PII Masking Anonymization. When input transcripts are anonymized by replacing sensitive details with category placeholders, the LLM has access to the complete conversational structure minus explicit identifiers. In this setting, model explanations are more detailed and grounded in the sequence and nature of the interaction. The responses often discuss the logic of the conversation, including impersonation, urgency, or attempts to create psychological pressure. This approach enables the model to reference the development of the conversation, escalation, and behavioral indicators, supporting a more reliable and transparent rationale for its predictions.

Transcript Summarization. Summarized transcripts provide the model with a condensed version of the original conversation. The LLM's responses under this condition are typically brief, highlighting general risk factors such as requests for money or mention of urgency. However, these summaries frequently exclude the subtle cues and interactional context that are sometimes necessary to distinguish complex scams from legitimate disputes. The explanations are less specific, and the detection of nuanced or atypical scams becomes more difficult.

6 Discussion

6.1 Privacy Preferences and Trainable Sanitization

Our analysis highlights that the choice of sanitization strategy plays a critical role in balancing privacy protection with the preservation of semantic information essential for accurate downstream detection. This insight also suggests that users' privacy preferences, particularly their individual risk tolerance, are intrinsically linked to detection outcomes. Since privacy is inherently subjective and context-dependent, users differ in what information they are comfortable disclosing and how much potential exposure they are willing to accept. The analysis above demonstrates that MASK can provide a configurable framework that allows users to select sanitization strategies tailored to their personal privacy expectations. These strategies operate within a predefined spectrum of privacy-utility trade-offs (e.g., Figure 4), enabling users to make informed decisions that align with their sensitivity thresholds while maintaining detection effectiveness.

Further, as depicted in Figure 5, a trainable sanitization module can further enhance MASK's adaptive capability. By jointly training the sanitization module with a semantic encoder and a downstream detection model, the system can dynamically optimize for multiple objectives, including privacy risk minimization, semantic fidelity, and detection performance. This multi-objective optimization enables more nuanced control over the trade-off curve between privacy and utility.

6.2 On-Device Detection with Lightweight LLMs

While the MASK's "sanitize-then-detect" paradigm offers a modular and privacy-preserving approach, a natural and naive solution would be to perform scam detection directly on the user's device

using a compact LLM. This method would eliminate the need to transmit any user data to external servers, thus maximizing privacy by design.

However, as shown in Table 2, our experiments reveal that running a local lightweight LLM for detection remains challenging in practice. Current small models often lack the necessary accuracy and contextual understanding compared to larger, cloud-based models. Deploying them effectively for real-time phone scam detection, which often involves long, complex conversations, requires significant advances in several areas.

One possible direction is to fine-tune or distill larger models into smaller ones optimized for on-device usage. However, this approach requires access to large-scale, high-quality labeled datasets, which are currently scarce in the domain of phone scam detection. Moreover, the real-time, streaming nature of phone conversations adds further constraints on latency and memory usage, making the deployment of even moderately sized models challenging on consumer-grade mobile devices.

7 Conclusion

This work proposes a novel adjustable sanitization framework that incorporates human preferences in selecting sanitization strategies for privacy-preserving phone scam detection using cloud-based large language models. We explore common sanitization strategies and the results demonstrate an inherent trade-off between privacy protection and the preservation of semantic information required for accurate detection. The framework can be further extended with adaptive, trainable sanitizers to better balance privacy and utility. These findings offer practical guidance for deploying privacy-conscious language understanding systems in sensitive real-world applications.

8 Acknowledgements

This work was supported the Hong Kong Research Grant Council under Grant 15600219.

References

[1] Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023. Driving Context into Text-to-Text Privatization. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. Association for Computational Linguistics, Toronto, Canada, 15–25. doi:10.18653/v1/2023.trustnlp-1.2

[2] Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023. Guiding Text-to-Text Privatization by Syntax. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. Association for Computational Linguistics, Toronto, Canada, 151–162. doi:10.18653/v1/2023.trustnlp-1.14

[3] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What Does it Mean for a Language Model to Preserve Privacy? arXiv:2202.05520 [stat.ML] https://arxiv.org/abs/2202.05520

[4] Matthew Brown and Andrew Trask. 2022. What Does it Mean for a Language Model to Preserve Privacy? *arXiv preprint arXiv:2202.12345* (2022).

[5] Jeremy Burke, Christine Kieffer, Gary Mottola, and Francisco Perez-Arce. 2022. Can Educational Interventions Reduce Susceptibility to Financial Fraud?. In *Journal of Economic Behavior & Organization*, Vol. 198. 250–266. doi:10.1016/j.jebo.2022.03.028

[6] Johan HM Buse, Jonathan Ee, and Shilpi Tripathi. 2023. Unveiling the Unseen Wounds—A Qualitative Exploration of the Psychological Impact and Effects of Cyber Scams in Singapore. *Psychology* 14, 11 (2023), 1728–1742.

[7] Nicholas Carlini, Florian Tramèr, and Eric Wallace. 2021. Extracting Training Data from Large Language Models. *arXiv preprint arXiv:2012.07805* (2021).

[8] Ricardo Silva Carvalho, Theodore Vasiloudis, Oluwaseyi Feyisetan, and Ke Wang. 2023. TEM: High utility metric differential privacy on text. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 883–890.

[9] Isha Chadalavada, Tianhui Huang, and Jessica Staddon. 2024. Distinguishing Scams and Fraud with Ensemble Learning. *arXiv preprint arXiv:2412.08680* (2024). doi:10.48550/arXiv.2412.08680 arXiv:2412.08680 [cs.CR]

[10] Chen-Wei Chang, Shailik Sarkar, Shutonu Mitra, Qi Zhang, Hossein Salemi, Hemant Purohit, Fengxiu Zhang, Michin Hong, Jin-Hee Cho, and Chang-Tien Lu. 2024. Exposing LLM Vulnerabilities: Adversarial Scam Detection and Performance. *arXiv preprint arXiv:2412.00621* (2024). doi:10.48550/arXiv.2412.00621 arXiv:2412.00621 [cs.CR]

[11] Chen-Wei Chang, Shailik Sarkar, Shutonu Mitra, Qi Zhang, Hossein Salemi, Hemant Purohit, Fengxiu Zhang, Michin Hong, Jin-Hee Cho, and Chang-Tien Lu. 2024. Exposing LLM Vulnerabilities: Adversarial Scam Detection and Performance. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 3568–3571.

[12] Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A Customized Text Sanitization Mechanism with Differential Privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 5747–5758. doi:10.18653/v1/2023.findings-acl.355

[13] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

[14] Alibaba Cloud. 2025. *Qwen2.5-MAX: An Advanced Multimodal AI Model for General and Specialized Tasks*. Technical Report. Alibaba Cloud, Hangzhou, China. Available at https://www.alibabacloud.com/en/product/qwen.

[15] Google DeepMind. 2025. *Gemini-2.5-Flash: A High-Performance Language Model for Reasoning and Multimodal Tasks*. Technical Report. Google DeepMind, London, UK. Available at https://deepmind.google/technologies/gemini.

[16] DeepSeek. 2024. *DeepSeek-V3: A Large-Scale Open-Source Language Model for Specialized Applications*. Technical Report. DeepSeek, Beijing, China. Available at https://www.deepseek.com.

[17] Marguerite DeLiema, Martha Deevy, Annamaria Lusardi, and Olivia S. Mitchell. 2020. Financial Fraud Among Older Americans: Evidence and Implications. In *The Journals of Gerontology: Series B*, Vol. 75. 861–868. doi:10.1093/geronb/gby151

[18] Matthew Honnibal. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *(No Title)* (2017).

[19] Xinxin Hu, Hongchang Chen, Shuxin Liu, Haocong Jiang, Guanghan Chu, and Ran Li. 2022. BTG: A Bridge to Graph Machine Learning in Telecommunications Fraud Detection. *Future Generation Computer Systems* 137 (2022), 274–287. doi:10.1016/j.future.2022.07.020

[20] Xinxin Hu, Haotian Chen, Junjie Zhang, Hongchang Chen, Shuxin Liu, Xing Li, Yahui Wang, and Xiangyang Xue. 2024. GAT-COBO: Cost-Sensitive Graph Neural Network for Telecom Fraud Detection. *IEEE Transactions on Big Data* 10, 4 (2024), 528–542. doi:10.1109/TBDATA.2024.3352978

[21] Rasmus Ingemann Tuffveson Jensen, Julie Gerlings, and Joras Ferwerda. 2024. Do awareness campaigns reduce financial fraud? *European Journal on Criminal Policy and Research* (2024), 1–36.

[22] Alice Johnson and Bob Brown. 2024. DePrompt: Desensitization and Evaluation of Personal Identifiable Information in Large Language Model Prompts. *arXiv preprint arXiv:2404.09876* (2024).

[23] Haoran Li, Yulin Chen, Jinglong Luo, Jiecong Wang, Hao Peng, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, Zenglin Xu, Bryan Hooi, and Yangqiu Song. 2024. Privacy in Large Language Models: Attacks, Defenses and Future Directions. arXiv:2310.10383 [cs.CL] https://arxiv.org/abs/2310.10383

[24] Tianshi Li, Sauvik Das, Hao-Ping Lee, Dakuo Wang, Bingsheng Yao, and Zhiping Zhang. 2024. Human-centered privacy research in the age of large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–4.

[25] Kai Liu and Wei Chen. 2023. Prompt Injection Attacks on Large Language Models: A Survey. *arXiv preprint arXiv:2310.09876* (2023).

[26] Zhiming Ma, Peidong Wang, Minhua Huang, Jingpeng Wang, Kai Wu, Xiangzhao Lv, Yachun Pang, Yin Yang, Wenjie Tang, and Yuchen Kang. 2025. TeleAntiFraud-28k: An Audio-Text Slow-Thinking Dataset for Telecom Fraud Detection. *arXiv preprint arXiv:2503.24115* (2025).

[27] Hossen Mustafa, Wenyuan Xu, Ahmad-Reza Sadeghi, and Steffen Schulz. 2018. End-to-End Detection of Caller ID Spoofing Attacks. In *IEEE Transactions on Dependable and Secure Computing*, Vol. 15. 423–436. doi:10.1109/TDSC.2016.2580509

[28] OpenAI. 2024. *ChatGPT-4o: A Multimodal Large Language Model*. Technical Report. OpenAI, San Francisco, CA, USA. Available at https://openai.com/chatgpt.

[29] Lu Peng and Rongheng Lin. 2018. Fraud Phone Calls Analysis Based on Label Propagation Community Detection Algorithm. In *Proceedings of the 2018 IEEE World Congress on Services (SERVICES '18)*. 23–24. doi:10.1109/SERVICES.2018.00025

[30] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[31] Sam Rogers. 2024. *International Scammers Steal Over $1 Trillion in 12 Months in Global State of Scams Report 2024.* Global Anti-Scam Alliance. https://www.gasa.org/post/global-state-of-scams-report-2024-1-trillion-stolen-in-12-months-gasa-feedzai Accessed: 2025-05-30.

[32] Zitong Shen, Kangzhong Wang, Youqian Zhang, Grace Ngai, and Eugene Yujun Fu. 2025. Combating Phone Scams with LLM-based Detection: Where Do We Stand? (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 28 (Apr. 2025), 29487–29489. doi:10.1609/aaai.v39i28.35298

[33] Zitong Shen, Sineng Yan, Youqian Zhang, Xiapu Luo, Grace Ngai, and Eugene Yujun Fu. 2025. "It Warned Me Just at the Right Moment": Exploring LLM-based Real-time Detection of Phone Scams. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Article 18, 7 pages.

[34] John Smith and Jane Doe. 2024. Casper: Prompt Sanitization for Protecting User Privacy in Web-Based Large Language Models. *arXiv preprint arXiv:2405.12345* (2024).

[35] Russell G Smith and Tabor Akman. 2008. Raising public awareness of consumer fraud in Australia. *Trends & Issues in Crime & Criminal Justice* 349 (2008).

[36] Karen Sparck Jones. 1988. *A statistical interpretation of term specificity and its application in retrieval.* Taylor Graham Publishing, GBR, 132–142.

[37] True Software Scandinavia AB. 2024. Truecaller. https://www.truecaller.com. Mobile application software.

[38] Vincent S. Tseng, Jia-Ching Ying, Che-Wei Huang, Yimin Kao, and Kuan-Ta Chen. 2015. FrauDetector: A Graph-Mining-based Framework for Fraudulent Phone Call Detection. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. 2157–2166. doi:10.1145/2783258.2788623

[39] Huahong Tu, Adam Doupé, Ziming Zhao, and Gail-Joon Ahn. 2019. Users really do answer telephone scams. In *28th USENIX Security Symposium (USENIX Security 19)*. 1327–1340.

[40] Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. 2023. Locally Differentially Private Document Generation Using Zero Shot Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 8442–8457. doi:10.18653/v1/2023.findings-emnlp.566

[41] Li Wang and Ming Zhao. 2025. On Protecting the Data Privacy of Large Language Models (LLMs) and LLM Agents: A Literature Review. *arXiv preprint arXiv:2501.09876* (2025).

[42] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[43] Ye Zhang and Xinjian Li. 2024. Security and Privacy Challenges of Large Language Models: A Survey. *arXiv preprint arXiv:2402.12345* (2024).

[44] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.

[45] Jijie Zhou, Eryue Xu, Yaoyao Wu, and Tianshi Li. 2025. Rescriber: Smaller-LLM-Powered User-Led Data Minimization for LLM-Based Chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–28.