KorSmishing Explainer: A Korean-centric LLM-based Framework for Smishing Detection and Explanation Generation

Yunseung Lee[*] and **Daehee Han**[*]
Division of Research and Development, KakaoBank Corp., Republic of Korea
{yun.lee, day.han}@lab.kakaobank.com

Abstract

To mitigate the annual financial losses caused by SMS phishing (smishing) in South Korea, we propose an explainable smishing detection framework that adapts to a Korean-centric large language model (LLM). Our framework not only classifies smishing attempts but also provides clear explanations, enabling users to identify and understand these threats. This end-to-end solution encompasses data collection, pseudo-label generation, and parameter-efficient task adaptation for models with fewer than five billion parameters. Our approach achieves a 15% improvement in accuracy over GPT-4 and generates high-quality explanatory text, as validated by seven automatic metrics and qualitative evaluation, including human assessments.

1 Introduction

Smishing, a form of financial fraud through SMS, has evolved into deceptive messages aimed at stealing personal information or coercing monetary transfers and has led to significant financial losses in South Korea (Kohilan et al., 2023). These losses amounted to approximately USD 100 million in 2021, affecting both individuals and financial institutions (Seo, 2022; Boukari et al., 2021).

Previous research on smishing detection has focused on a binary classification model that often achieves high accuracy (Sousa et al., 2021; Liu et al., 2021; Oswald et al., 2022). However, the lack of interpretability undermines user trust and practical applicability (Tenney et al., 2020; Rudin, 2019; Yuan et al., 2022).

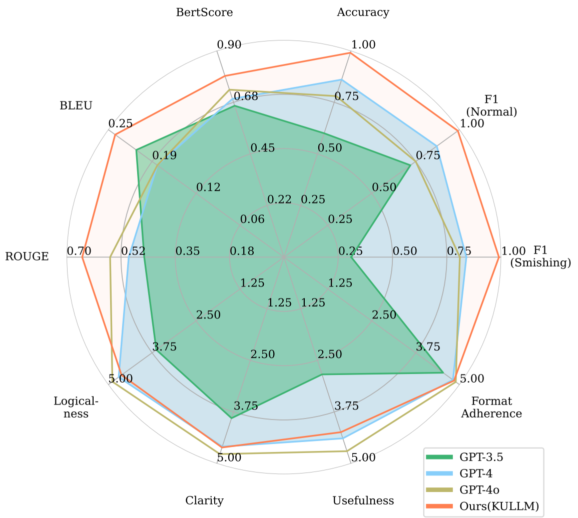


Figure 1: KULLM adapted with our proposed framework achieved significantly higher detection accuracy and produced explanations that were better than those of GPT-3.5, GPT-4, and GPT-4o with prompt engineering.

We introduce a novel framework for adapting a large language model (LLM) for explainable smishing detection in Korean. This framework enables the model to detect smishing and explain the results of its detection. Designed for integration into enterprise-level services, it aims to maintain robustness to minor text format variations and cost efficiency, surpassing OpenAI's GPT models with prompt engineering in performance. Our framework includes pseudo-label generation with a collected dataset and task-adaptive fine-tuning layers to optimize LLMs for both detection and explanation generation tasks, enhancing users' understanding of the model's results and supporting informed decision-making. Additionally, our framework ensures efficiency with models under five billion parameters, outperforming the accuracy and practicality of OpenAI's GPT models. The framework guarantees higher accuracy, practical utility, and cost efficiency, as illustrated in Figure 1.

To the best of our knowledge, this study is the first to propose an explainable NLP framework within the smishing detection. The model trained through this framework has significantly positive societal impacts in South Korea, highlighting its po-

---

[*] These authors contributed equally to this work.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 642–656
November 12-16, 2024 ©2024 Association for Computational Linguistics

tential to reduce financial fraud through improved detection and enhanced user comprehension.

2 Related Works

2.1 LLM Adaptation

Integrating LLMs into enterprise services requires a balance between high performance and cost efficiency (Touvron et al., 2023; Kwon et al., 2023). Although LLMs excel in natural language understanding and are general task solvers, prompt engineering can be both expensive and unstable owing to the need for detailed task descriptions and sensitivity to input changes (Chang et al., 2024; Wei et al., 2022; Kojima et al., 2022). Thus, optimizing LLMs for specific tasks through adaptation strategies is crucial for service-based applications (Brown et al., 2020). Parameter-efficient fine-tuning (PEFT) methods, such as Low-Rank Adapters (LoRA) and quantized LoRA (QLoRA), mitigate these challenges by updating only a portion of the model's parameters or by adding small adapter layers. These methods facilitate faster model adaptation, requiring significantly less computational power and storage space while maintaining performance (Hu et al., 2022; Dettmers et al., 2023).

2.2 Korean-Centric LLM

Most LLMs are primarily trained in English, which limits their effectiveness in handling Korean tasks. Even multilingual LLMs often encounter data imbalances that reduce their performance in non-English tasks (Jung and Plas, 2024; Lorandi and Belz, 2024; Sitaram et al., 2023). To develop effective Korean-centric LLMs, it is crucial to train on Korean data using techniques tailored to specific architectures.

The GPT-NeoX-based Korean Polyglot variants (Polyglot-ko), a Korean-centric LLM, demonstrate exceptional performance in specialized Korean tasks owing to its training on 863GB of Korean data. Polyglot-ko excels in the KOBEST dataset, which evaluates Korean understanding and reasoning (Ko et al., 2023). Additionally, the Korean University LLM (KULLM), an extension of the Polyglot-ko models, enhances performance through instruction tuning with Korean-translated datasets used for training models, such as Vicuna and Dolly LLMs (Lee et al., 2023a). Despite having under five billion parameters, these models slightly close the gap with GPT-4 in aspects such as fluency, coherence, and completeness, demonstrating the effectiveness of training on Korean data (Lee et al., 2023a).

2.3 Explainable Smishing Detection

Smishing detection, a subset of misinformation detection, primarily utilizes Transformer-based models to classify messages as smishing or normal (Kaddoura et al., 2020; Jiang et al., 2020; Oswald et al., 2022). These models leverage attention scores to identify influential tokens, enhancing interpretability (Letarte et al., 2018; Niu et al., 2019). However, practical applications in financial services require outputs that are more user-friendly than merely highlighting high attention tokens.

While existing studies on misinformation detection using LLMs focus on fact-checking or reasoning with knowledge databases, targeting areas such as fake news detection or Wikipedia-based fact-checking (Bang et al., 2023; Pelrine et al., 2023; Pan et al., 2023), smishing detection is less explored. Our approach advances smishing detection by incorporating detailed explanatory results through LLM adaptation, while previous work relies on prompt engineering to explain logical conflict of false statements (Cheng et al., 2023).

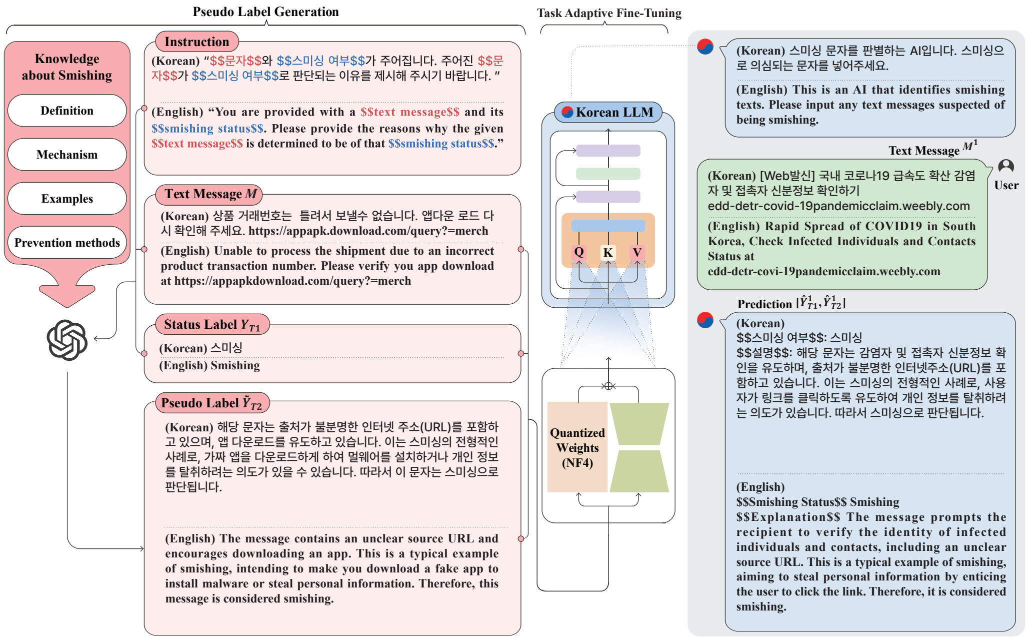
This method aims to enhance decision-making processes for service providers and users by providing clear explanations, thus improving the model's ability to distinguish smishing messages. This study introduces a novel end-to-end approach using LLMs for fraud prevention in the financial sector, extending the application of natural language processing techniques and enhancing security and trust within the financial services sector and beyond.

3 Methodology

Our framework is designed to adapt a Korean-centric LLM for smishing detection and to generate detailed explanations for the results. As illustrated in Figure 2, it consists of two core components: pseudo-label generation and task-adaptive fine-tuning. These elements work together to enhance the model's accuracy in classifying and explaining smishing attempts.

3.1 Pseudo-Label Generation

We assembled a dataset comprising text messages and their smishing status labels, structured as (M, Y_{T1}) pairs, where M represents the message content and Y_{T1} indicates the smishing status. Ini-



(Korean) "$$문자$$와 $$스미싱 여부$$가 주어집니다. 주어진 $$문자$$가 $$스미싱 여부$$로 판단되는 이유를 제시해 주시기 바랍니다."

(English) "You are provided with a $$text message$$ and its $$smishing status$$. Please provide the reasons why the given $$text message$$ is determined to be of that $$smishing status$$."

(Korean) 상품 거래번호는 틀려서 보낼수 없습니다. 앱다운 로드 다시 확인해 주세요. https://appapk.download.com/query?=merch

(English) Unable to process the shipment due to an incorrect product transaction number. Please verify you app download at https://appapkdownload.com/query?=merch

(Korean) 스미싱

(English) Smishing

(Korean) 해당 문자는 출처가 불분명한 인터넷 주소(URL)를 포함하고 있으며, 앱 다운로드를 유도하고 있습니다. 이는 스미싱의 전형적인 사례로, 가짜 앱을 다운로드하게 하여 멀웨어를 설치하거나 개인 정보를 탈취하려는 의도가 있을 수 있습니다. 따라서 이 문자는 스미싱으로 판단됩니다.

(English) The message contains an unclear source URL and encourages downloading an app. This is a typical example of smishing, intending to make you download a fake app to install malware or steal personal information. Therefore, this message is considered smishing.

(Korean) 스미싱 문자를 판별하는 AI입니다. 스미싱으로 의심되는 문자를 넣어주세요.

(English) This is an AI that identifies smishing texts. Please input any text messages suspected of being smishing.

(Korean) [Web발신] 국내 코로나19 급속도 확산 감염자 및 접촉자 신분정보 확인하기
edd-detr-covid-19pandemicclaim.weebly.com

(English) Rapid Spread of COVID19 in South Korea, Check Infected Individuals and Contacts Status at
edd-detr-covi-19pandemicclaim.weebly.com

(Korean)
$$스미싱 여부$$: 스미싱
$$설명$$: 해당 문자는 감염자 및 접촉자 신분정보 확인을 유도하며, 출처가 불분명한 인터넷주소(URL)를 포함하고 있습니다. 이는 스미싱의 전형적인 사례로, 사용자가 링크를 클릭하도록 유도하여 개인 정보를 탈취하려는 의도가 있습니다. 따라서 스미싱으로 판단됩니다.

(English)
$$Smishing Status$$ Smishing
$$Explanation$$ The message prompts the recipient to verify the identity of infected individuals and contacts, including an unclear source URL. This is a typical example of smishing, aiming to steal personal information by enticing the user to click the link. Therefore, it is considered smishing.

(a) Traning Phase

(b) Deployment

Figure 2: A Korean language-based smishing detection and explanation generation framework for enterprise-level services. (a) The LLM is adapted for explainable smishing detection using pseudo-label generation and task-adaptive fine-tuning to enable efficient training with minimal resources. (b) The chat-based user interface generates and displays the smishing status and explanations for input messages to the user.

tially, the dataset did not include Y_{T2} labels for explanation text.

To generate these labels, we leveraged GPT-4o's capabilities through prompt engineering, known for its excellence in natural language understanding (OpenAI, 2023). The generation of Y_{T2} pseudo-labels was guided by essential smishing-related knowledge and instructions, as shown in Equation (1). The process of constructing prompt templates is detailed in Appendix A.1.

$$\begin{aligned} &P(\tilde{Y}_{T2} | M, Y_{T1}, C), \\ &C = \{knowledge, instruction\} \end{aligned} \tag{1}$$

The pseudo-labels $\tilde{Y}_{T2}$, generated using GPT-4o under these conditions, are employed for the explanation generation task, resulting in a dataset of $(M, Y_{T1}, \tilde{Y}_{T2})$ pairs. Human verification ensured the quality of the explanation labels, with Cronbach's alpha at 0.89 and Cohen's kappa at 0.80, indicating high inter-rater agreement.

3.2 Task-Adaptive Fine-Tuning

To minimize costs without compromising service quality, we employed a parameter-efficient fine-tuning approach. This method integrates a rank-decomposition matrix into the quantized Korean LLM using the QLoRA technique, allowing fine-tuning with fewer parameters (Dettmers et al., 2023). Figure 2 shows that low-rank matrices, quantized to 4-bit normal float (NF4), are added to the query (Q), key (K), and value (V) in the attention layer of the Korean LLM.

The Korean LLM is fine-tuned to follow a structured response format: "$$Smishing Status$$ {*Smishing Status*}, $$Explanation$$ {*Explanation*}", accompanied by the instruction, "Please determine whether the given text message is a smishing message or a normal message." The model adapted from our framework is required to generate predicted smishing status label $\hat{Y}_{T1}$ and explanation $\hat{Y}_{T2}$ of the result. Fine-tuning a model with five billion parameters in our methodology takes approximately 9 hours on four RTX 3090 GPUs, as detailed in Appendix A.2.

4 Experimental Design

4.1 Dataset

Owing to the lack of a Korean dataset containing both smishing and normal text messages, we collected our own dataset, as detailed in Appendix B. This dataset comprises normal messages sourced through crowdsourcing, such as promotional texts, and smishing messages acquired from

	Smishing Detection			**Explanation Generation**							**Avg.**
Model	**F1 (Smishing)**	**F1 (Normal)**	**Accuracy**	**BertScore**	**BLEU**	**ROUGE**	**Logicalness**	**Clarity**	**Usefulness**	**Format Adherence**	**Rank**
KcBERT	0.99±0.00	0.99±0.00	0.99±0.00	-	-	-	-	-	-	-	-
GPT-3.5	0.31±0.01 (6)	0.72±0.01 (6)	0.60±0.00 (6)	0.66±0.00 (6)	0.21±0.00 (4)	0.45±0.00 (6)	3.63±0.00 (6)	3.90±0.00 (6)	2.84±0.01 (6)	4.53±0.00 (6)	5.8
GPT-4	0.84±0.01 (4)	0.87±0.00 (4)	0.86±0.01 (4)	0.69±0.00 (5)	0.18±0.00 (5)	0.50±0.00 (5)	4.70±0.00 (2)	4.59±0.00 (3)	4.39±0.00 (2)	4.81±0.00 (4)	3.8
GPT-4o	0.81±0.00 (5)	0.75±0.01 (5)	0.78±0.01 (5)	0.73±0.00 (4)	0.18±0.00 (5)	0.56±0.00 (4)	**4.88**±0.00 (1)	**4.77**±0.01 (1)	**4.70**±0.00 (1)	**4.89**±0.00 (1)	3.2
Korean LLM Adapted by Our Framework											
Polyglot(1B)	0.93±0.06 (3)	0.92±0.08 (3)	0.92±0.07 (3)	0.75±0.01 (3)	0.23±0.01 (3)	0.60±0.01 (3)	4.35±0.07 (5)	4.36±0.08 (5)	3.94±0.05 (5)	4.69±0.05 (5)	3.8
Polyglot (5B)	**0.99**±0.00 (1)	**0.99**±0.00 (1)	**0.99**±0.00 (1)	0.77±0.00 (2)	**0.25**±0.00 (1)	0.62±0.00 (2)	4.50±0.01 (4)	4.53±0.01 (4)	4.10±0.02 (4)	4.82±0.00 (3)	2.3
KULLM (5B)	**0.99**±0.00 (1)	**0.99**±0.00 (1)	**0.99**±0.00 (1)	**0.79**±0.00 (1)	0.24±0.00 (2)	**0.65**±0.00 (1)	4.62±0.01 (3)	4.61±0.01 (2)	4.24±0.02 (3)	4.84±0.01 (2)	1.7

Table 1: Quantitative performance on the test dataset for smishing detection and explanation generation tasks. The values represent the average performance over three runs for each seed, with ranks indicated in (). Overall performance is determined by the average ranking across multiple metrics.

a data vendor. To ensure privacy, all data was strictly anonymized, removing personal information.

Following the initial labeling, two in-house financial fraud detection experts reviewed the dataset to maintain high data quality. Messages were labeled as smishing (not normal) if they involved any form of deception or impersonation. In contrast, one-on-one communications (without impersonation) and official messages from legitimate institutions were classified as normal.

We placed significant emphasis on incorporating the essential linguistic features of the Korean language during the data collection process. The dataset followed a prescribed labeling format, and we verified that key linguistic elements, such as honorifics (including suffixes, nouns, and verbs) and relevant prefixes, were applied correctly during both the pseudo-label generation and human verification stages. This approach ensured that the dataset faithfully captured both the linguistic nuances and contextual aspects of the Korean language.

To prevent train-test bias (Lee et al., 2022), we removed duplicates and highly similar entries, such as those differing by only one or two characters. This was achieved by calculating the cosine similarity of texts at the embedding level using Korean-specific sentence-BERT (KR-SBERT) models and filtering out entries exceeding a certain similarity score (Park and Shin, 2021).

Through this process, we finalized a dataset comprising approximately 14,600 records—9,400 smishing texts and 5,200 normal messages. Additionally, explanation labels for smishing detection were generated using a pseudo-label generation layer, as illustrated in Figure 2. Finally, the dataset was divided into training, validation, and testing sets in a 6:2:2 ratio.

4.2 Models

For our comparison, we evaluated OpenAI's GPT-3.5-Turbo, GPT-4, and GPT-4o against several Korean language models fine-tuned within our framework. OpenAI's GPT models were chosen because of their near state-of-the-art adaptation and performance in various unseen NLP tasks in zero-shot settings via in-context learning (Brown et al., 2020). We adapted several Korean language models under five billion parameters as base models for our framework, specifically selecting Polyglot-ko and KULLM for their efficiency in Korean-centric tasks and their open-source licenses. Although BERT models cannot generate explanations for their results, we included the Korean comments BERT (KcBERT) with smishing datasets as a baseline solely for the smishing detection task (Lee, 2020).

4.3 Evaluation Metrics

To evaluate smishing detection, we utilized the F1 score, which balances precision and recall, as well as accuracy, calculating both for smishing and normal detection categories. Given the absence of a standard for evaluating explanation generation, we employed multiple metrics to assess the quality of the results. These included bilingual evaluation understudy (BLEU), recall-oriented understudy for gisting evaluation (ROUGE), and BertScore, which measure surface-level matching and semantic similarity between generated results and pseudo-label references (Papineni et al., 2002; Lin, 2004; Zhang et al., 2020).

We also employed an LLM-based reference-free metric for evaluating generated explanations (Liu et al., 2023; Lee et al., 2024). In LLM-based metrics, it is essential to define the aspects relevant to the task and formulate specific questions to use as prompts. For explainable smishing detection, we selected logicalness, clarity, usefulness, and format adherence as key aspects. Each question was evaluated on a scale from 1 to 5, as detailed in Appendix C. The scores for each aspect were averaged to assess the overall quality of the generated output. Additionally, qualitative results from human evaluations were incorporated into the analysis.

5 Results

5.1 Smishing Detection

Our framework's adapted Korean LLMs outperformed the F1 score and accuracy of GPT-3.5, GPT-4, and GPT-4o with prompt engineering, as shown in Table 1. Notably, the task-adapted KULLM and Polyglot models, each having five billion parameters, achieved an F1 score of 0.99 for both smishing and normal labels. Even with a one-billion-parameter Polyglot model, our adaptation enhanced accuracy to 6 percentage points better than that of GPT-4, the highest among the GPT models.

Although GPT-4o benefits from additional non-English training data, fine-tuning on a Korean dataset for specific tasks proves more efficient in enhancing performance. Furthermore, our proposed model eliminates the need for detailed prompt engineering, unlike OpenAI's GPT models, reducing application programming interface usage costs related to token processing. Consequently, our framework surpasses OpenAI's GPT models in both performance and operational efficiency.

5.2 Explanation Generation

We evaluated the performance of generated explanations across three dimensions: surface level, embedding level, and a reference-free metric. Based on surface-level metrics (BLEU, ROUGE) and embedding-level metrics (BertScore), the sentences generated by the model trained with our proposed method showed higher similarity to the reference sentences than those produced by GPT models. Specifically, the fine-tuned KULLM achieved a BertScore of 0.79 and a ROUGE score of 0.65, indicating that our framework's explanations are more likely to contain core keywords and are semantically more similar to the reference sentences.

Additionally, we assessed the generated outputs using a reference-free metric based on GPT-4o, covering logicalness, clarity, usefulness, and format adherence. Outputs generated by GPT-4o with prompt engineering performed best, while the KULLM-based framework and GPT-4 showed comparable performance, as demonstrated in Table 1. The KULLM-based model particularly excelled in format adherence, generating outputs that adhered well to the prescribed format and utilized honorifics, which is important in Korean (Hwang et al., 2024; Lee et al., 2023b). In terms of clarity, the responses were easy to understand and maintained consistent context within the output. Evaluating the quality of explanations requires considering multiple aspects, and using the average rank across different metrics provides a comprehensive performance overview.

Additionally, the KULLM adapted with our framework processes over 20 responses per second on a single Hopper 100 architecture GPU, with a 5-second timeout setting, making it suitable for real-time deployment. This measurement reflects the model's ability to complete full token generation for inference, not just the first token or inter-token latency.

In summary, the proposed framework based on Korean LLMs outperformed GPT models in average rank across smishing detection accuracy and explanation generation quality metrics. Considering detection accuracy, quality of explanations, and operational efficiency, our framework is the most suitable for smishing detection and explanation generation.

5.3 Qualitative Results

We conducted a qualitative evaluation on a test dataset of 2,036 instances. We anonymized the outputs of GPT-4o, the highest-performing GPT model, and the best-performing KULLM-based model from our task-adapted framework. Evaluators were asked to choose their preferred output including option for a tie. The results indicated a preference for the KULLM-based model. Inter-agreement among evaluators was assessed using

	Test Data	Reference	ChatGPT-4o	Fine-tuned KULLM
	(a) Normal message			
Kor	고객님 우체국입니다. 우체국택배를 오늘배달할 예정입니다. 등기번호 : 0000- 0000- 0000 발송인 : 김00 배달장소 : 전남 00군 00 우체국 000-0000-0000 박00 배달원	$$스미싱 여부$$: 정상 $$설명$$: 해당 문자는 우체국에서 발송한 택배에 대한 안내로 보입니다. 등기번호와 배달장소, 배달원 정보가 구체적으로 제공되어 있으며, 링크나 민감한 정보를 요구하는 내용이 포함되어 있지 않습니다. 따라서 정상 문자로 판단됩니다.	$$스미싱 여부$$: 스미싱 $$설명$$: 이 문자는 우체국을 사칭하여 발송된 것으로 보입니다. 우체국에서 발송하는 공식 문자는 등기번호와 같은 정보를 제공할 수 있지만, 발송인과 배달원의 이름, 전화번호를 포함하는 경우는 드뭅니다. 또한, 출처가 불분명한 인터넷주소(URL)나 전화번호를 포함한 문자는 스미싱일 가능성이 높습니다. 따라서 이 문자는 스미싱으로 판단됩니다.	$$스미싱 여부$$: 정상 $$설명$$: 해당 문자는 우체국에서 발송한 택배에 대한 안내로 보입니다. 발송인, 배달장소, 등기번호 등 구체적인 정보가 포함되어 있으며, 링크나 의심스러운 요청이 포함되어 있지 않습니다. 따라서 정상 문자로 판단됩니다.
Eng	Dear customer, this is the post office. Your parcel delivery is scheduled for today. Registered number: 0000-0000-0000 Sender: Kim 00 Delivery location: Jeonnam 00-gun 00 Post Office 000-0000-0000 Courier: Park 00	$$Smishing Status$$: Legitimate $$Explanation$$: This message appears to be a notification about a parcel sent by the post office. It provides specific details such as the registered number, delivery location, and courier information, without including any links or requests for sensitive information. Therefore, it is considered a legitimate message.	$$Smishing Status$$: Smishing $$Explanation$$: This message appears to be sent under the guise of the post office. While official messages from the post office may provide information such as a registered number, it is rare for them to include the sender's and courier's names and phone numbers. Additionally, messages containing unclear source URLs or phone numbers are likely to be smishing attempts. Therefore, this message is considered smishing.	$$Smishing Status$$: Legitimate $$Explanation$$: This message appears to be a notification about a parcel sent by the post office. It includes specific details such as the sender, delivery location, and registered number, without containing any links or suspicious requests. Therefore, it is considered a legitimate message.
	(b) Smishing message			
	[국제발신] 이00님 [은 행 공 동 인 증 서] 해외에서 발급되었습니다 본인 아닐시 문의 문의:****-****	$$스미싱 여부$$: 스미싱 $$설명$$: 이 문자는 은행에서 발송된 것처럼 보이지만, 실제로는 스미싱일 가능성이 높습니다. 은행은 보통 중요한 인증서 발급과 관련된 사항을 문자로 통보하지 않으며, 특히 국제발신으로 발송되는 경우는 드뭅니다. 또한, 문의 전화번호가 제공되어 있지만, 이는 사기꾼이 제공한 번호일 가능성이 큽니다. 따라서 이 문자는 스미싱으로 판단됩니다.	$$스미싱 여부$$: 스미싱 $$설명$$: 이 문자는 은행에서 발송된 것처럼 보이지만, 몇 가지 의심스러운 요소가 있습니다. 첫째, '국제발신'이라는 문구와 함께 발신자가 명확하지 않습니다. 둘째, '은행공동 인증서'와 같은 중요한 정보가 해외에서 발급되었다는 내용은 일반적으로 문자로 통지되지 않습니다. 셋째, 문의 전화번호가 별도로 제공되어 있는데, 이는 사기꾼들이 자주 사용하는 수법입니다. 따라서 이 문자는 스미싱일 가능성이 높습니다. 출처가 불분명한 인터넷주소(URL)는 클릭하지 않도록 주의하시기 바랍니다.	$$스미싱 여부$$: 스미싱 $$설명$$: 이 문자는 은행에서 발송된 것처럼 보이지만, 실제로는 스미싱일 가능성이 높습니다. 은행은 보통 중요한 인증서 발급과 관련된 사항을 문자로 통보하지 않으며, 특히 해외에서 발급된 인증서에 대해 문의를 요구하는 경우는 드뭅니다. 또한, 문의 전화번호가 제공되었지만, 이는 사기꾼이 제공한 번호일 가능성이 높습니다. 따라서 이 문자는 스미싱으로 판단됩니다.
	[International Sender] Lee 00, [Bank Cooperation Certificate] issued overseas. If this is not you, please contact us. Contact: ****-****	$$Smishing Status$$: Smishing $$Explanation$$: This message appears to be from a bank, but it is likely a smishing attempt. Banks typically do not notify about important certificate issuances via text message, especially not from international senders. Additionally, while a contact number is provided, it is likely to be a number given by scammers. Therefore, this message is considered smishing.	$$Smishing Status$$: Smishing $$Explanation$$: This message appears to be from a bank, but there are several suspicious elements. First, the term "international sender" and the lack of a clear sender are concerning. Second, important information like a "bank cooperation certificate" being issued overseas is not typically communicated via text message. Third, the inclusion of a separate contact number is a common tactic used by scammers. Therefore, this message is likely to be smishing. Please be cautious and avoid clicking on unclear source URLs.	$$Smishing Status$$: Smishing $$Explanation$$: This message appears to be from a bank, but it is likely a smishing attempt. Banks typically do not notify about the issuance of important certificates via text message, especially not requesting inquiries about certificates issued overseas. Additionally, while a contact number is provided, it is likely a number given by scammers. Therefore, this message is considered smishing.

Figure 3: Results of smishing detection and explanation generation using GPT-4o and the adapted KULLM from our framework. ‘Kor’ and ‘Eng’ denote ‘Korean’ and ‘English’ respectively. All English texts are translations of the original Korean texts. The reference represents the labeled test data used for the smishing detection and explanation generation tasks.

Cronbach’s alpha at 0.71 and Cohen’s kappa at 0.55, indicating a moderate level of consensus and reliability.

As illustrated in Figure 3, both models generated natural and plausible explanations. However, the responses from GPT-4o occasionally misclassified normal messages as smishing or included hallucinations, such as warnings about clicking on non-existent links in the messages. These reliability issues contributed to the higher qualitative evaluation scores for the adapted KULLM model.

6 Conclusion

This study introduces a framework for explainable smishing detection using a Korean LLM, designed for enterprise-level applications. By incorporating pseudo-label generation and task-adaptive fine-tuning, our framework improves the accuracy of smishing detection and generates clear, logical explanations comparable to those from GPT-4o. For service applicability, it ensures both efficiency and accuracy in training and inference with a model under five billion parameters. Furthermore, this research demonstrates the effectiveness of adapting non-English LLMs for explainable smishing detection. By providing more comprehensible explanations, it helps users more effectively identify smishing messages. The proposed framework contributes to the prevention of financial fraud and has a positive social impact.

7 Limitations

While conducting human evaluation of the generated explanations, we did not yet incorporate positive or negative feedback from service users through the user interface. Future research will aim to address this by creating an additional feedback loop that incorporates user feedback, enabling model updates and enhancing user satisfaction.

8 Ethical Considerations

User privacy is critical and mandatory for financial institutions. To ensure privacy, all text messages were anonymized before the LLMs were trained. Sensitive information, including the names of people, organization names, account numbers, and phone numbers, was either removed, anonymized, or synthesized during collection. While reviewing the collected dataset, the reviewers additionally anonymized the synthesized names and numbers to prevent coincidence with real names and numbers. Although our approach achieved significant performance in smishing detection and explanation generation, an intensive evaluation of safety measures, including guardrails and toxicity, is required before deployment. In the future, we plan to integrate these methods into the training phase to ensure that the LLMs follow safety measures before deployment.

Acknowledgement

We would like to express our deepest gratitude to KakaoBank Corp. for their generous support throughout this research. Special thanks go to Hyeji Jo for her work on the illustrations, which greatly enhanced the visual clarity of our paper. Additionally, we would like to thank the anonymous reviewers for their constructive comments, which significantly improved the quality of our work.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the* 13th *International Joint Conference on Natural Language Processing and the* 3rd *Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP) (Volume 1: Long Papers)*, pages 675–718.

Badr Eddine Boukari, Akshaya Ravi, and Mounira Msahli. 2021. Machine learning detection for smishing frauds. In 18th *IEEE Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–2.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Sijie Cheng, Zhiyong Wu, Jiangjie Chen, Zhixing Li, Yang Liu, and Lingpeng Kong. 2023. Unsupervised explanation generation via correct instantiations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12700–12708.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations (ICLR)*.

Yerin Hwang, Yongil Kim, Hyunkyung Bae, Jeesoo Bang, Hwanhee Lee, and Kyomin Jung. 2024. Kosmic: Korean text similarity metric reflecting honorific distinctions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 9954–9960.

Zhuoren Jiang, Zhe Gao, Yu Duan, Yangyang Kang, Changlong Sun, Qiong Zhang, and Xiaozhong Liu. 2020. Camouflaged chinese spam content detection with semi-supervised generative active learning. In *Proceedings of the* 58th *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3080–3085.

Vincent Jung and Lonneke" Plas. 2024. Understanding the effects of language-specific class imbalance in multilingual fine-tuning. In *Findings of the Association for Computational Linguistics (EACL)*, pages 2368–2376.

Sanaa Kaddoura, Omar Alfandi, and Nadia Dahmani. 2020. A spam email detection mechanism for english language text emails using deep learning approach. In 29th *IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 193–198.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Sungho Park, and Kyubyong Park. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models. *arXiv preprint arXiv:2306.02254*.

Rasenthiran Kohilan, Harsha Edirisinghe Warakagoda, Tharushi Thathsarani Kitulgoda, Nimalaprakasan Skandhakumar, and Nuwan Kuruwitaarachchi. 2023. A machine learning-based approach for detecting smishing attacks at end-user level. In *IEEE International Conference on e-Business Engineering (ICEBE)*, pages 149–154.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems (NeurIPS)*, 35:22199–22213.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the* 29th *Symposium on Operating Systems Principles*, pages 611–626.

Junbum Lee. 2020. Kcbert: Korean comments bert. In *Proceedings of the* 32nd *Annual Conference on Human and Cognitive Language Technology*, pages 437–440.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the* 60th *Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 8424–8445.

SeungJun Lee, Taemin Lee, Jeongwoo Lee, Yoona Jang, and Heuiseok Lim. 2023a. Kullm: Learning to construct korean instruction-following large language models. In *Annual Conference on Human and Language Technology*, pages 196–202. Human and Language Technology.

Seungjun Lee, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2023b. Improving formality-sensitive machine translation using data-centric approaches and prompt engineering. In *Proceedings of the* 20th *International Conference on Spoken Language Translation (IWSLT)*, pages 420–432.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, and Pilsung Kang. 2024. Checkeval: Robust evaluation framework using large language model via checklist. *arXiv preprint arXiv:2403.18771*.

Gaël Letarte, Frédérik Paradis, Philippe Giguère, and François Laviolette. 2018. Importance of self-attention for sentiment analysis. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP*, pages 267–275.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Xiaoxu Liu, Haoye Lu, and Amiya Nayak. 2021. A spam transformer model for SMS spam detection. *IEEE Access*, 9:80253–80263.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Michela Lorandi and Anya Belz. 2024. High-quality data-to-text generation for severely under-resourced languages with out-of-the-box large language models. In *Findings of the Association for Computational Linguistics (EACL)*, pages 1451–1461.

Guocheng Niu, Hengru Xu, Bolei He, Xinyan Xiao, Hua Wu, and Sheng Gao. 2019. Enhancing local feature extraction with global representation for neural text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the* 9th *international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 496–506.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

C. Oswald, Sona Elza Simon, and Arnab Bhattacharya. 2022. Spotspam: Intention analysis-driven SMS spam detection using BERT embeddings. *ACM Transactions on the Web*, 16(3):14:1–14:27.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 6981–7004.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, page 311–318.

Suzi Park and Hyopil Shin. 2021. Kr-sbert: A pre-trained korean-specific sentence-bert model. `https://github.com/snunlp/KR-SBERT`.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6399–6429.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Junbae Seo. 2022. Korean social trends 2022. `https://kostat.go.kr/boardDownload.es?bid=12312&list_no=422196&seq=2`.

Sunayana Sitaram, Monojit Choudhury, Barun Patra, Vishrav Chaudhary, Kabir Ahuja, and Kalika Bali. 2023. Everything you need to know about multilingual llms: Towards fair, performant and reliable models for languages of the world. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 21–26.

Gustavo Sousa, Daniel Carlos Guimarães Pedronette, João Paulo Papa, and Ivan Rizzo Guilherme. 2021. Sms spam detection through skip-gram embeddings and shallow networks. In *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 4193–4201.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*.

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2022. Explainability in graph neural networks: A taxonomic survey. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 45(5):5782–5799.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In 8th *International Conference on Learning Representations (ICLR)*.

A Implementation Details

A.1 Pseudo-Label Generation

We designed prompts for pseudo-label generation to create explanation labels using the OpenAI GPT-4o 2024-05-13 version of Azure. The prompts include a system template and a user template, corresponding to $knowledge$ and $instruction$ in Equation (1).

For the system template, domain knowledge related to smishing detection—such as definitions, mechanisms, examples, prevention methods, and recent cases—is organized into sections as illustrated in Figure 4. The user template, shown in Figure 5, provides the guidelines for the LLM, including persona, text input with smishing status, response guidelines, and answer format. In summary, both the system and user templates were used as knowledge and instructions for pseudo-labeling.

A.2 Task-Adaptive Fine-tuning

The implementation details of our framework are described in Table 2. Referring to QLoRA, we selected the appropriate number of epochs and batch sizes based on the training data size. The intrinsic rank of the matrix r, α, and dropout ratio were set experimentally. In our proposed framework, the QLoRA matrix was added to the query, key, and value components. Considering its deployment in services and the need to control the diversity of generated results, the temperature was set to 0. Finally, to constrain the diversity of the generated results for the input message, a low temperature and repetition penalty were set.

Fine-tuning	
epoch	7
batch size	4
lr	3e-5
QLoRA	
r	8
α	32
target module	'query_key_value'
dropout ratio	0.05
Generation	
temperature	0.0
top p	0.9
repetition penalty	1.1

Table 2: Configuration for task-adaptive fine-tuning and inference.

B Korean Smishing Dataset

A Korean dataset was developed to detect smishing using crowdsourcing. Representative types of smishing and normal messages are shown in Figure 6. Note that this classification is solely for the purpose of conveying information about the dataset, and the proposed model does not categorize messages by type.

Smishing messages are categorized into various types, such as the impersonation of financial institutions, government agencies, delivery fraud, payment fraud, child impersonation, and fake notifications of weddings or funerals. For example, despite the legal prohibition of financial institutions promoting loans via SMS in South Korea, impersonation messages often falsely offer benefits, such as low interest rates, to steal personal information. Government impersonation and payment fraud messages mimic legitimate services and direct recipients toward malicious websites or apps. New types of smishing in Korea, such as child impersonation and fake ceremony notifications, aim to extract personal information by pretending to be urgent calls from children or links related to personal events.

Normal messages that provide a realistic contrast for effective smishing detection training include legitimate notifications from financial institutions and government agencies, courier delivery updates, payment alerts, and promotional content.

C Experimental Details

C.1 Baselines

The baseline models used are the GPT-3.5-turbo 0613 version, GPT-4 0613 version, and GPT-4o 2024-05-13 version from Azure OpenAI. We designed system and user templates to specialize in smishing detection and explanation generation by leveraging the in-context learning capabilities of LLMs. The same system template used for pseudo-label generation was employed to sufficiently inject smishing-related knowledge into the model. Meanwhile, the user template included instructions such as persona and input text format, guiding the generation of responses in the same format as pseudo-label generation and fine-tuning. For more details, please refer to Figure 7.

C.2 Prompt Template for LLM-based Reference-Free Metric

For LLM-based evaluation metrics such as G-Eval, defining the measured aspects and creating specific questions to evaluate these aspects are crucial (Liu et al., 2023; Lee et al., 2024). To assess outputs for explanation generation, we selected logicalness, clarity, usefulness, and format adherence as the key aspects. We then listed detailed questions to measure each of these aspects. Each question is rated on a scale from 1 to 5 and the score for each aspect is the average of the scores for its related questions. Questions 1 and 2 measure logicalness. Questions 3 and 4 measure clarity. Question 5 measures usefulness. Questions 6 to 8 measure format adherence. Detailed information about the specific questions is given in Figure 8.

(Korean) 지금부터 당신은 스미싱 문자와 정상 문자를 구분하여 금융소비자들이 건전한 금융생활을 영위하고 사기로부터 보호받도록 도와주는 역할을 하게 됩니다. 앞으로 당신이 읽고서 판단해야 하는 부분은 $${SECTION}$$으로 시작하고 해당 부분이 끝났음을 나타내는 $${SECTION 끝}$$ 으로 표현됩니다.

(English) From now on, you will play a role in helping financial consumers lead a healthy financial life and protect them from fraud by distinguishing between smishing and normal messages. The sections you will need to read and judge start with $$SECTION$$ and end with $$SECTION END$$.

Definition

(Korean) $$스미싱의 정의$$ 스미싱은 사회공학적 기법 기반의 공격으로, 문자 내 링크 등을 클릭하게 만들어서 사람들을 속이고, 사람들로 하여금 멀웨어를 다운로드하거나 민감한 정보를 공유하거나 사이버 범죄자에게 송금하도록 만드는 사이버 범죄입니다. $$스미싱의 정의 끝 $$

(English) $$Definition of Smishing$$ Smishing is a cybercrime that uses social engineering-based attacks to deceive people into clicking on links in messages, leading them to download malware, share sensitive information, or transfer money to cybercriminals. $$End of Definition of Smishing$$

Mechanism

(Korean) $$스미싱의 작동원리$$ 사기꾼들은 사람들을 속여 휴대폰, 은행 계좌 또는 개인 데이터를 침해하기 위해 거짓 메시지와 악성 링크를 사용합니다. 사람들은 은행이나 특정 브랜드로부터 안내문자 및 홍보문자를 수신하는 것에 익숙하기 때문에 문자 메시지 내에 포함된 링크를 클릭할 가능성이 더 높습니다. 기술적으로 사기꾼들은 전화번호, 링크의 출처를 손쉽게 위장할 수 있습니다. $$스미싱의 작동원리 끝$$

(English) $$How Smishing Works$$ Scammers use fake messages and malicious links to deceive people into compromising their mobile phones, bank accounts, or personal data. People are accustomed to receiving informative and promotional messages from banks or specific brands, making them more likely to click on links contained within text messages. Technically, scammers can easily disguise the source of a phone number or link. $$End of How Smishing Works$$

Examples

(Korean) $$스미싱의 사례$$
- 금융 기관인 척 가장함: 사기꾼들은 계좌에 문제가 발생했다고 알려주는 피해자의 은행인 것처럼 가장할 수 있습니다.
- 정부 기관인 것처럼 가장함: 사기꾼들은 경찰관, IRS 담당관 또는 기타 정부 관료인 것처럼 가장할 수 있습니다. 이러한 스미싱 문자는 피해자에게 미납 벌금이 있다거나 정부 혜택을 받기 위해 청구를 해야 한다고 주장합니다. 예를 들면, COVID-19 팬데믹이 한창일 때, 세금 면제, 무료 COVID 검사, 그리고 기타 이와 유사한 서비스를 제안하는 스미싱 공격에 대해 경고했습니다.
- 고객 지원 팀인 것처럼 가장함: 공격자들은 Amazon, Microsoft와 같은 신뢰할 수 있는 브랜드 또는 심지어 피해자의 무선 서비스 제공업체의 고객 지원 상담원인 것처럼 가장합니다. 이들은 보통 피해자의 계정에 문제가 있거나 청구하지 않은 보상 또는 환불 건이 있다고 말합니다.
- 배송물 사기: 이 스미싱 메시지는 FedEx, UPS, 또는 미국 우체국(US Postal Service)에서 보낸 것처럼 가장합니다. 이러한 메시지는 표적에게 배송 과정에서 문제가 발생했다고 말합니다. 이러한 SMS 피싱 사기는 많은 사람들이 배송물을 기다리는 휴가 기간에 자주 발생합니다.
- 다단계 인증 사기: 해커는 피해자의 MFA 코드를 훔쳐 피해자의 소셜 미디어, 이메일 또는 은행 계좌로 침입하려고 합니다. 흔히 발생하는 MFA 사기 시나리오 중 하나에서는 해커가 피해자의 친구 중 한 명인척 가장합니다.
- 가짜 앱 다운로드: 피해자가 사실은 멀웨어 또는 랜섬웨어인 가짜 앱을 다운로드하도록 속이는 스미싱 사기도 있습니다. 멀웨어는 파일 관리 프로그램, 안티바이러스 앱 또는 대출 앱으로 가장하는 경우가 많습니다.
- 자녀 사칭 사기: 해커는 자녀인 것처럼 부모에게 문자를 보내고 답장을 요구하는 경우입니다. 주로 핸드폰이 고장나서 연락한다고 문자를 보내는 경우가 많습니다.

$$스미싱의 사례 끝$$

(English) $$Examples of Smishing$$
- Pretending to be a financial institution: Scammers can pretend to be the victim's bank, informing them of issues with their account.
- Pretending to be government agencies: Scammers can pose as police officers, IRS agents, or other government officials. These smishing messages claim that the victim has unpaid fines or needs to claim government benefits. For example, during the COVID-19 pandemic, warnings were issued about smishing attacks offering tax exemptions, free COVID tests, and similar services.
- Pretending to be customer support teams: Attackers pose as trusted brands like Amazon, Microsoft, or even the victim's wireless service provider's customer support agent. They usually claim there are issues with the victim's account or unclaimed compensation or refunds.
- Delivery fraud: These smishing messages pretend to be from FedEx, UPS, or the US Postal Service. They tell the target that there was a problem with their delivery. These SMS phishing scams often occur during the holiday season when many people are expecting packages.
- Multi-factor authentication (MFA) scams: Hackers try to steal the victim's MFA codes to break into their social media, email, or bank accounts. In one common MFA scam scenario, the hacker pretends to be one of the victim's friends.
- Fake app downloads: Smishing scams trick the victim into downloading fake apps, which are actually malware or ransomware. Malware often masquerades as file management programs, antivirus apps, or loan apps.
- Impersonating a child: Hackers send texts to parents pretending to be their child and asking for a response. Often, they claim to be contacting because their phone is broken.

$$End of Examples of Smishing$$

Prevention methods

(Korean) $$스미싱 피해 예방법$$
- 출처가 불분명한 인터넷주소(URL)는 클릭하지 않기
- 알 수 없는 출처의 앱은 설치하지 않기
- 백신 프로그램 설치하기
- 소액결제 주의하기
- 보안을 이유로 금융정보를 요구해도 절대 입력하지 않기
- 전자금융사기 예방서비스 가입하기
- 구글 플레이스토어·T스토어·올레마켓·LGU+앱스토어 등 공인된 마켓을 통해 앱 설치하기
- 한국의 금융기관은 대출을 받도록 유도할 수 없음을 알기. 대출광고 시에는 심의광고필 번호가 존재함. (ex. 준법감시인 심의필 제{YEAR}-광고-{5DIGITNUM}호)

$$스미싱 피해 예방법 끝$$

(English) $$Preventing Smishing Victims$$
- Do not click on internet addresses (URLs) from unknown sources
- Do not install apps from unknown sources
- Install antivirus programs
- Be cautious with microtransactions
- Never enter financial information when asked for security reasons
- Subscribe to electronic financial fraud prevention services
- Install apps through accredited markets such as Google Play Store, T Store, Olleh Market, LG U+ App Store, etc.
- Know that financial institutions in Korea cannot solicit loans. Loan advertisements will have a review advertisement required number (ex. Compliance Officer Review Required {YEAR}-Ad-{5DIGITNUM}).

$$End of Preventing Smishing Victims$$

Extra information

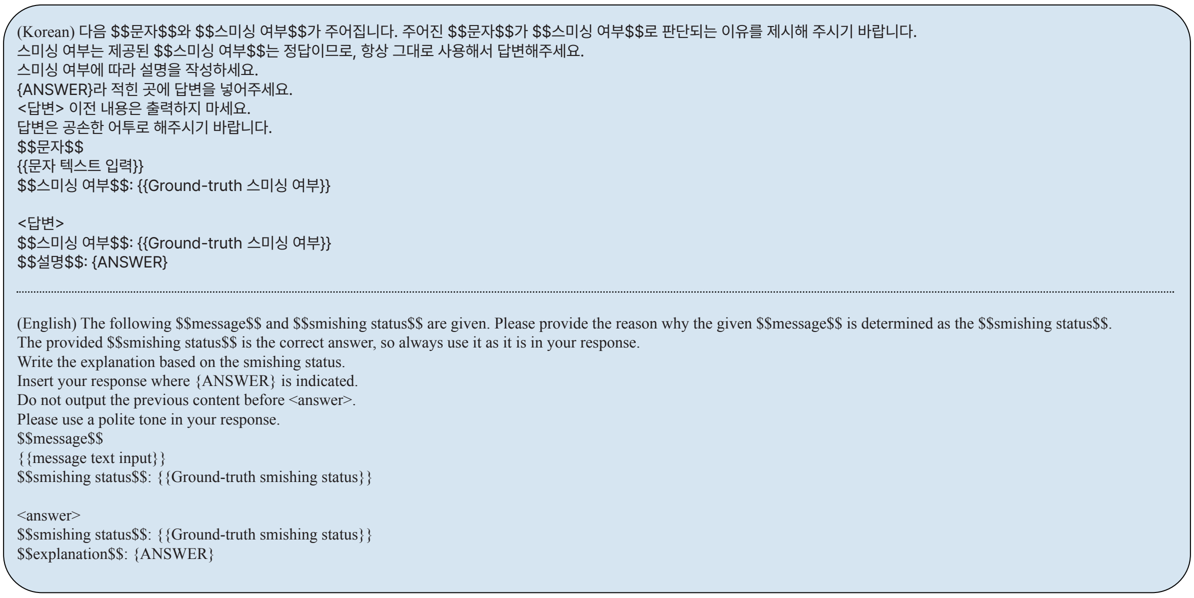
(Korean) $$신규 스미싱 유형 정보$$
- 은행 등 금융권에서는 전화나 문자를 통해 정부정책 대출을 포함하여 모든 종류의 대출신청을 권유하거나 광고하지 않습니다.
- 대출을 권유하는 경우 스미싱입니다.
- 스미싱 여부는 보수적으로 평가해주세요.

$$신규 스미싱 유형 정보 끝$$

(English) $$Information on New Types of Smishing$$
- Banks and financial institutions do not solicit or advertise any type of loan application, including government policy loans, via phone or text messages.
- It is smishing if there is a solicitation for loans.
- Please assess the possibility of smishing conservatively.

$$End of Information on New Types of Smishing$$

Figure 4: System template used in pseudo-labeling and prompt engineering of GPT models.



```
(Korean) 다음 $$문자$$와 $$스미싱 여부$$가 주어집니다. 주어진 $$문자$$가 $$스미싱 여부$$로 판단되는 이유를 제시해 주시기 바랍니다.
스미싱 여부는 제공된 $$스미싱 여부$$는 정답이므로, 항상 그대로 사용해서 답변해주세요.
스미싱 여부에 따라 설명을 작성하세요.
{ANSWER}라 적힌 곳에 답변을 넣어주세요.
<답변> 이전 내용은 출력하지 마세요.
답변은 공손한 어투로 해주시기 바랍니다.
$$문자$$
{{문자 텍스트 입력}}
$$스미싱 여부$$: {{Ground-truth 스미싱 여부}}

<답변>
$$스미싱 여부$$: {{Ground-truth 스미싱 여부}}
$$설명$$: {ANSWER}
```

```
(English) The following $$message$$ and $$smishing status$$ are given. Please provide the reason why the given $$message$$ is determined as the $$smishing status$$.
The provided $$smishing status$$ is the correct answer, so always use it as it is in your response.
Write the explanation based on the smishing status.
Insert your response where {ANSWER} is indicated.
Do not output the previous content before <answer>.
Please use a polite tone in your response.
$$message$$
{{message text input}}
$$smishing status$$: {{Ground-truth smishing status}}

<answer>
$$smishing status$$: {{Ground-truth smishing status}}
$$explanation$$: {ANSWER}
```

Figure 5: User template used in pseudo-labeling.

Smishing message	
Type 1: Posing as a Financial Institution	[Web발신] (광고) 카*오뱅크 *항상 고객님께 힘이 되어 드리는<ka**o Bank>가 되겠습니다. *5월 12일까지 접수시 적용되는 상품이 출시 되었습니다. +높은 승인률로 한도는 올리고 이자는 낮췄습니다. 최대 10년간 자유이용하세요. 상품내용 -접수기관: 카*오뱅크 -한도: 최소 300만 원~ 최대 2억 원 이내 -적용이자: 최초 1년 무이자, 연 2.3% ~3.5% 대 내외 (...) 신청/상담센터: 02-3473-**** •수신번호로 전화👉1번👉ARS에 따라 예약 신청
	[Sent via Web] (Advertisement) Ka**o Bank *We aim to always be a source of strength for our customers at <ka**o Bank>. *A new product has been launched applicable for applications received by May 12th. +With a high approval rate, the limit has been increased and the interest lowered. Enjoy freely for up to 10 years. Product Details: -Application Institution: Ka**o Bank -Limit: Minimum of 3 million KRW up to 200 million KRW -Interest Rate: First year interest-free, thereafter 2.3% ~ 3.5% per annum (...) Application/Consultation Center: 02-3473-**** •To apply, call the received number👉press 1👉reserve via ARS
Type 2: Posing as a Government Agency	[국제발신] (국세청) 2차 재난지원금 신속지급 즉시확인▼ ko.gl/z1j
	[International Message] (National Tax Service) Immediate confirmation for the expedited payment of the 2nd disaster relief fund▼ ko.gl/z1j
Type 3: Delivery Fraud	***고객님 상품이 발송될 예정입니다 배송조회☞ http://salewiz.co.kr/infobell_idx.html
	***Your product is scheduled to be shipped. Track delivery☞ http://salewiz.co.kr/infobell_idx.html
Type 4: Payment Fraud	[국제발신] Google play ***님 **/* **** ***,*** 승인완료 본인아닐시 문의:**-**** -****
	[International Message] Google Play ***Your transaction of **/* **** for ***,*** has been approved. If this was not you, please contact: **-**** -****
Type 5: Impersonating a Child	아빠나 통화중 폰 떨어뜨려서 액정이깨졌어.잠깐이번호 사용중...문자보면 여기로 답줘
	Dad, I dropped my phone while on a call, and the screen broke. I'm using this number for now... If you see this message, please reply here.
Type 6: Fake Notification of wedding or funeral	모바일청첩장결혼식일시: 4/08(토) 많이많이와주세요. https://c11.kr./1cnhj
	Mobile Wedding Invitation Wedding Date: April 8th (Saturday) Please come in large numbers. https://c11.kr./1cnhj

Normal message	
Type 1: Notification from Financial Institution	[Web발신] [카*오뱅크] ***님 카드(0***) 05월 후불교통 대금 135,050원이 06/12 출금될 예정입니다. 교통이용내역 조회 방법 [전체메뉴 > 내 카드 > 교통이용내역 조회] https://go.ka**obank.io (...)
	[Sent via Web] [Ka**o Bank] ***Your card (0***) will be debited with the postpaid transportation fee of 135,050 KRW on 06/12. To view your transportation usage details, go to [Main Menu > My Card > View Transportation Usage] https://go.ka**obank.io (...)
Type 2: Notification from Government Agency	[질병관리청] 귀국 시 큐코드 웹사이트 https://cov19ent.kdca.go.kr 접속하여 검역정보를 사전에 입력해주시기 바랍니다.
	[Disease Control and Prevention Agency] Upon returning to the country, please access the Q-code website https://cov19ent.kdca.go.kr and enter your quarantine information in advance.
Type 3: Delivery Notification	[Web발신] [C***통운 택배_배송출발] 반갑습니다, 고객님. 고객님의 소중한 상품이 배송 예정입니다. ・상품명 : 아이폰11슬림핏 - 1개 ・배송예정시간 : 12-14시 ※ 위탁장소 선택, 실시간 배송정보 https://mms.door**door.co.kr:8443/MMSPUSH/trust.do? (...)
	[Sent via Web] [C*** Express Delivery_Departure for Delivery] Greetings, customer. Your valuable product is scheduled for delivery. · Product Name: iPhone 11 Slim Fit - 1 piece · Expected Delivery Time: 12-14 hours · Choice of consignment location, real-time delivery information https://mms.door**door.co.kr:8443/MMSPUSH/trust.do? (...)
Type 4: Payment Notification	[K*] 휴대폰 소액결제 완료 안내 [Web발신] [멜*(자동)]110원 자동결제완료 결제일시: 03/07 14:00 잔여한도:499,890원 PG사명: 다* 아래 링크를 통해 한도를 상향하거나 하향 하실 수 있습니다.☞한도변경하기: https://ips.k*.com/o**eh/lu*in/myo**eh/limit/limit.do (...)
	[K*] Mobile Phone Micropayment Completion Notification [Sent via Web] [Mel*(Automatic)] 110 won automatic payment completed Payment Date and Time: 03/07 14:00 Remaining Limit: 499,890 won Payment Gateway Company Name: Da* You can increase or decrease your limit through the link below.☞Change Limit: https://ips.k*.com/o**eh/lu*in/myo**eh/limit/limit.do (...)
Type 5: Promotion/ Advertisement	(광고)라고 하기엔 혜택이 많은 [Web발신] 2019 최*우 <THE BRAIN> 고양공연 4명만 모여도 할인이 팡팡! 모이면 모일수록 커지는 할인혜택! 최대 30% 할인가로 대한민국 최고의 마술공연을 관람할 수 있는 기회 상식을 뒤집는 지상최대의 멘탈 매직콘서트 대한민국 No.1 마술사 최*우를 만나보세요! 1. 공연개요 일정: 2월 15일(금) ~ 17일(일) 장소: 고양**누리 **극장 관람등급: 만 5세 이상 2. 할인판매가 (8인 이상 예매시 30%할인 적용) R석: 53,900원 (...) 5. 문의 000-0000-0000 (자세한 내용은 홈페이지 참조: http://www.ticket****.co.kr/product/2***)
	(Advertisement) Too good to just call it an advertisement [Sent via Web] 2019's most outstanding <THE BRAIN> Goyang performance Group discounts galore! The more you gather, the bigger the discount! Seize the opportunity to watch Korea's best magic show at up to 30% off. Experience the greatest mental magic concert that will turn your common sense upside down, featuring Korea's No.1 magician Choi*Woo. 1. Performance Details Schedule: February 15th (Fri) to 17th (Sun) Venue: Goyang**Nuri **Theater Age Rating: 5 years and above 2. Discounted Sale Price (30% discount for bookings of 8 or more) R seats: 53,900 won (...) 5. Inquiries 000-0000-0000 (For more details, visit the website: http://www.ticket****.co.kr/product/2***)

Figure 6: Description of the collected Korean Smishing data. The left side of the table contains smishing messages, while the right side features normal message types. The dataset actually includes messages corresponding to each category, and the English-translated versions are also provided below.

```
(Korean) 다음 $$문자$$가 주어집니다. 먼저 $$문자$$를 보고 $$스미싱 여부$$를 판단하세요. 스미싱 여부는 스미싱 또는 정상 2단계로만 답변하세요.
다음으로 $$스미싱 여부$$로 판단한 이유를 제시해 주시기 바랍니다.
{ANSWER}라 적힌 곳에 답변을 넣어주세요.
<답변> 이전 내용은 출력하지 마세요.
답변은 공손한 어투로 해주시기 바랍니다.
$$문자$$
{{문자 텍스트 입력}}

<답변>
$$스미싱 여부$$: {ANSWER}
$$설명$$: {ANSWER}
```

```
(English) The following $$message$$ is given. First, determine the $$smishing status$$ of the $$message$$. Respond with either "smishing" or "legitimate".
Next, provide the reason for your determination of the $$smishing status$$.
Insert your response where {ANSWER} is indicated.
Do not output the previous content before <answer>.
Please use a polite tone in your response.
$$message$$
{{message text input}}

<answer>
$$smishing status$$: {ANSWER}
$$explanation$$: {ANSWER}
```

Figure 7: User template used for prompt engineering of GPT-3.5, GPT-4, and GPT-4o.

(Korean) 두 사람 간의 대화가 주어집니다. 다음의 입력(Input)과 입력에 대한 응답(Response)이 제시됩니다.
당신의 작업은 응답을 평가 단계에 따라 응답을 평가하는 것입니다.
이 평가 기준을 꼼꼼히 읽고 이해하는 것이 중요합니다. 평가하는 동안 이 문서를 계속 열어두고 필요할 때 참조해 주세요.
평가 외에 다른 설명을 추가하지 마세요.

평가 기준:
1. 입력 문자로부터 추론 가능한 정보로 답변이 구성되어있는가?
2. 답변이 논리적으로 구성되는가?
3. 답변이 간결하고 이해하기 쉬운가?
4. 답변 내의 정보가 일관된 맥락을 유지하는가?
5. 사용자가 스미싱 메시지를 인식하고 대응하는데에 답변이 실질적인 도움을 제공하는가?
6. 답변이 잘 읽히고 이해하기 쉬운 형식인가?
7. "$$스미싱 여부$$: ANSWER $$설명$$: ANSWER" 형식에 맞게 답변을 제공하는가?
8. 답변이 공손한 어조로 작성되었는가?

평가 단계:
1. Instruction, Input, 그리고 Response을 주의깊게 읽습니다.
2. 위의 평가 기준에 따라 Response을 평가합니다.

Input:
{{문자 텍스트 입력}}

Response:
{{평가를 위하여 모델이 생성한 응답 텍스트 입력}}

Result
1. 입력 문자로부터 추론 가능한 정보로 답변이 구성되어있는가? (1 - 5):
2. 답변이 논리적으로 구성되는가? (1 - 5):
3. 답변이 간결하고 이해하기 쉬운가? (1 - 5):
4. 답변 내의 정보가 일관된 맥락을 유지하는가? (1 - 5):
5. 사용자가 스미싱 메시지를 인식하고 대응하는데에 답변이 실질적인 도움을 제공하는가? (1 - 5):
6. 답변이 잘 읽히고 이해하기 쉬운 형식인가? (1 - 5):
7. '$$스미싱 여부$$: ANSWER $$설명$$: ANSWER' 형식에 맞게 답변을 제공하는가? (1 - 5):
8. 답변이 공손한 어조로 작성되었는가? (1 - 5):

……

(English) A conversation between two people is given. The following Input and Response are provided.
Your task is to evaluate the response according to the evaluation criteria.
It is important to read and understand these evaluation criteria thoroughly. Keep this document open during the evaluation process and refer to it as needed.
Do not add any additional explanations beyond the evaluation.

Evaluation Criteria:
1. Is the response composed of information that can be inferred from the input message?
2. Is the response logically structured?
3. Is the response concise and easy to understand?
4. Does the response maintain a consistent context throughout?
5. Does the response provide practical help for the user to recognize and respond to a smishing message?
6. Is the response written in a well-read and easy-to-understand format?
7. Does the response follow the format: "$$smishing status$$: ANSWER $$explanation$$: ANSWER"?
8. Is the response written in a polite tone?

Evaluation Steps:
1. Carefully read the Instruction, Input, and Response.
2. Evaluate the Response according to the above evaluation criteria.

Input:
{{Input text}}

Response:
{{Response text generated by the model for evaluation}}

Result
1. Is the response composed of information that can be inferred from the input message? (1 - 5):
2. Is the response logically structured? (1 - 5):
3. Is the response concise and easy to understand? (1 - 5):
4. Does the response maintain a consistent context throughout? (1 - 5):
5. Does the response provide practical help for the user to recognize and respond to a smishing message? (1 - 5):
6. Is the response written in a well-read and easy-to-understand format? (1 - 5):
7. Does the response follow the format: '$$smishing status$$: ANSWER $$explanation$$: ANSWER'? (1 - 5):
8. Is the response written in a polite tone? (1 - 5):

Figure 8: Prompt template used for LLM-based reference-free metric. Each aspect is measured with specific designated questions, each rated on a 5-point scale. Logicalness: Questions 1, 2. Clarity: Questions 3, 4. Usefulness: Question 5. Format Adherence: Questions 6, 7, 8.