IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 July 2025, accepted 17 August 2025, date of publication 26 August 2025, date of current version 5 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3603007

RESEARCH ARTICLE

Enhanced Voice Phishing Detection Using an LLM-Based Framework for Data Augmentation and Classification

HYUNGHEE PARK[1], **JIWON LEE**[1], **SANGHYUN HAN**[2], **AND HAEWON BYUN**[3]
[1]Department of Convergence Technology Engineering, Sungshin Women's University, Seoul 02844, South Korea
[2]SK Telecom, Seoul 02844, South Korea
[3]School of AI Convergence, Sungshin Women's University, Seoul 02844, South Korea

Corresponding author: Haewon Byun (hyewon@sungshin.ac.kr)

This work was supported in part by SK Telecom (SKT) through the 2024 SKT AI Fellowship Program and the collaborative research conducted within the scope of the fellowship, and in part by a Korea Institute for Advancement of Technology (KIAT) grant funded by the Korean government (Ministry of Trade, Industry and Energy (MOTIE) of the Republic of Korea) (Human Resource Development (HRD) Program for Industrial Innovation) under Grant No. RS-2025-02263062.

ABSTRACT Existing voice phishing detection models based on call transcripts often suffer from limited generalizability due to insufficient scenario diversity and the absence of ambiguous samples in data. To address these challenges, we propose an integrated framework that leverages GPT-4o, a large language model (LLM), to generate realistic call transcripts from actual fraud cases and to build an expert-guided phishing detection model. Using case reports from the Financial Supervisory Service (FSS) and Korean call transcripts from the KorCCVi dataset, we generate Korean phishing call transcripts that capture underrepresented fraud tactics. In addition, we generate non-phishing call transcripts that retain phishing-like linguistic patterns by removing or attenuating core fraudulent cues, thereby enabling training on ambiguous cases. The generated data are quantitatively evaluated for linguistic naturalness, scenario diversity, and detection difficulty. To assess the sample-level detection difficulty in semantic space, we introduce the metric Class Centroid Distance Variability (CCDV). We also propose the Domain Expert LLM, implemented using GPT-4o, a prompt-engineered detection model that incorporates six analytical criteria validated by a domain expert. The model not only improves detection performance but also produces structured analytical reports that enhance interpretability. In experiments, the Domain Expert LLM achieves an F1 score of 0.9686 on previously unseen and ambiguous transcripts, substantially outperforming conventional models such as RandomForest and KoBERT, which yield an average F1 score of approximately 0.70.

INDEX TERMS Expert-guided prompting, large language models, synthetic data evaluation, voice phishing detection.

I. INTRODUCTION

Voice phishing is an increasingly sophisticated form of financial fraud in which perpetrators deceive victims by phone to induce unauthorized transfers or the disclosure of sensitive information. Common tactics include impersonating bank officials to offer fictitious loans and posing as law-enforcement agents to issue fabricated legal threats. Recently, schemes have expanded to exploit tax-refund and subsidy programs and to use artificial intelligence (AI), particularly generative AI, to impersonate victims' family members during fabricated crises such as kidnappings or medical emergencies, leading to unauthorized transfers [1]. This trend is particularly concerning in South Korea, where advanced financial infrastructure, such as open banking and real-time remittance systems, can be exploited for these scams.

In response, machine learning (ML) [2], [3], [4] and deep learning (DL) [5], [6], [7] models have been developed for automated voice phishing detection from transcribed conversations. However, ML models relying on fixed feature patterns struggle with evolving fraud tactics. Although

The associate editor coordinating the review of this manuscript and approving it for publication was Senthil Kumar.

DL models such as BERT show improved contextual understanding, their performance significantly degrades in out-of-distribution (OOD) scenarios involving unfamiliar or ambiguous attack patterns.

These limitations are closely tied to the constraints of existing datasets. The widely used KorCCVi dataset [5], for example, focuses primarily on institutional impersonation and loan fraud, offering limited coverage of emerging schemes such as family impersonation and refund scams. Moreover, the stark lexical and syntactic contrast between phishing and non-phishing samples in this dataset simplifies the classification task during training, but undermines the model's ability to generalize to ambiguous real-world cases, often resulting in misclassifications.

To mitigate these limitations, Yu et al. [7] utilized financial consultation transcripts from AI Hub as non-phishing data. These transcripts include domain-specific vocabulary and syntactic patterns that closely resemble those found in voice phishing, making them more challenging to distinguish from actual phishing cases. However, the dataset still suffers from limited coverage of diverse fraud scenarios.

Recently, large language models (LLMs) have been proposed as a promising approach to overcome structural limitations in conventional voice phishing detection [8]. Unlike traditional ML and DL models, LLMs offer greater flexibility in adapting to novel phishing tactics through large-scale pretraining on diverse text corpora. However, general-purpose LLMs often lack the capability for fine-grained analysis, especially in high-stakes phishing detection tasks.

To address these issues, we propose an integrated framework that leverages GPT-4o, a large language model (LLM), for both case-informed call transcript generation and expert-guided phishing detection. Our approach synthesizes high-fidelity phishing transcripts based on actual fraud case reports from the Financial Supervisory Service (FSS) and verified call transcripts from the KorCCVi dataset. These generated transcripts incorporate core phishing tactics, such as fabricated legal threats, urgent financial demands, and impersonation of trusted entities. In parallel, we generate borderline non-phishing transcripts by attenuating or removing key fraudulent elements, enabling the model to learn from subtle linguistic cues and improve robustness in ambiguous scenarios.

To evaluate whether the generated data capture the linguistic and contextual characteristics of real-world call transcripts, we conduct a quantitative evaluation based on criteria such as naturalness, scenario diversity, and detection difficulty. In particular, we introduce a new metric, Class Centroid Distance Variability (CCDV), which quantifies classification difficulty by measuring the semantic distance of each sample from class centroids in the embedding space.

Furthermore, to address the limited generalization performance of conventional models, we present Domain Expert LLM, implemented using GPT-4o, a detection model that integrates domain knowledge through expert-guided prompt engineering. Specifically, the prompts incorporate six analytical criteria and their linguistic characteristics, which were identified by a domain expert. The model automatically generates structured analysis reports based on these criteria for each call transcript. This improves detection performance and enhances the interpretability and practical applicability of the results in real-world decision-making for financial and investigative institutions.

The main contributions of this paper are as follows.

- We propose a dataset construction method that leverages GPT-4o to generate realistic and diverse phishing transcripts based on actual fraud cases and validated call transcripts.
- We introduce a novel evaluation metric, Class Centroid Distance Variability (CCDV), to measure detection difficulty at the sample level, and combine it with additional criteria to assess data quality.
- We develop the Domain Expert LLM, an interpretable detection model that incorporates expert-guided analytical criteria to improve both classification accuracy and transparency. The model also generates structured case-level reports to support real-world investigative and decision-making processes.

This study extends our previous conference work [9] by incorporating a rigorous and multidimensional evaluation of data quality, including the CCDV metric, and embedding domain expertise into the LLM-based classifier through expert-guided prompt engineering and automatic report generation.

The remainder of this paper is organized as follows. Section II reviews related work. Section III details the methodology, including LLM-based script generation, evaluation metrics, and expert-guided model development. Section IV presents the experimental setting, and Section V discusses the results and implications. Section VI concludes the study with future directions.

II. RELATED WORKS

A. VOICE PHISHING DATASETS CONSTRUCTED FROM CALL TRANSCRIPT DATA

The construction of high-quality datasets is essential for improving the performance of voice phishing detection models. However, the acquisition or disclosure of real-world phone call data is often restricted due to privacy regulations and wiretapping laws. To overcome these limitations, various dataset construction strategies have been proposed.

Derakhshan et al. [10] constructed a phishing dataset by instructing 15 graduate students to generate 75 simulated scam call scenarios across five types of fraud. These were combined with 140 non-phishing scripts from the CallHome corpus, resulting in a dataset consisting of 215 manually scripted and transcribed conversations. Although this approach helped alleviate data scarcity, the scripted nature of dialogues may not fully reflect the spontaneity and interactional nuances of authentic phone conversations.

Zhao et al. [11] curated phishing-related textual samples from social media platforms (Sina Weibo) and news portals (Baidu), applying machine learning–based filtering to collect 647 and 1,443 cases, respectively. Although this method leveraged publicly available incident data, the absence of audio and conversational context limited its ability to capture prosodic and pragmatic features inherent to real-time voice exchanges.

Li et al. [12] developed a telecom fraud dataset containing 12,506 samples, sourced from an anti-fraud platform operated by the Public Security Bureau. The dataset spans four major scam types and is categorized into five labels. However, as most recordings included only the victim's voice, the perpetrator's speech was reconstructed using extracted keywords, which may not accurately represent the original script content.

For Korean-language phishing detection, the KorCCVi v2 dataset [5] is one of the most widely used resources. It was compiled from victim call recordings provided by the Financial Supervisory Service (FSS) and general non-fraudulent conversations from the National Institute of Korean Language (NIKL). Although it reflects realistic fraud scenarios, the data set exhibits significant class imbalance - phishing cases account for only 23.7% - and is largely limited to loan fraud and institutional impersonation schemes, restricting its coverage of more recent or subtle types of scams.

To address these limitations, this study proposes a framework that goes beyond basic data augmentation. By integrating real-world case features into LLM-generated call transcripts and introducing a multifaceted data quality assessment framework, our approach supports the development of robust detection models that generalize effectively to diverse and evolving voice phishing scenarios.

B. LLM-BASED DATA AUGMENTATION

Data augmentation techniques aim to increase the size and quality of training data. Traditional approaches focus primarily on manipulating the surface structure of text, such as word insertion, deletion or replacement, and back-translation. With the advent of LLMs, which are capable of understanding complex contexts and generating human-like text, the scope of data augmentation has expanded. Recent studies include rewriting existing data and generating new samples under a variety of specific conditions. They have shown that these LLM-based approaches effectively improve minority class representation, contextual diversity, and model generalization performance [13], [14], [15], [16].

Zhao et al. [14] compared two data augmentation techniques using ChatGPT (GPT-3.5 turbo): (1) rewriting existing samples and (2) generating entirely new samples. The results showed that both approaches contributed to model performance, with further improvements observed when the synthetic data from both methods were combined. In particular, the rewriting approach preserves existing information and the generation approach supplements it with new content, indicating that a balance between information retention and supplementation can effectively enhance model performance.

Li et al. [15] investigated the impact of data augmentation using GPT-3.5 turbo and task subjectivity on the performance of classification models. Across ten diverse classification tasks, they compared models trained on three data types: zero-shot synthetic data, few-shot synthetic data, and original data. Few-shot methods consistently outperformed zero-shot methods, which was attributed to the richer context and linguistic patterns in human-written text. In low subjectivity tasks, such as spam detection and topic classification, models trained on synthetic data performed comparably to those trained on original data. The results demonstrate the applicability of LLM-based data augmentation for voice phishing detection.

Gopali et al. [16] evaluated several approaches to mitigate class imbalance in the Myers-Briggs Type Indicator (MBTI) dataset, including data-level and algorithm-level methods (e.g., Random Over Sampler, L2 regularization) and two forms of data augmentation: word-level and LLM-based augmentation through GPT-3.5-turbo with few-shot prompting. LLM-based augmentation outperformed other methods in both minority class representation and overall evaluation metrics.

Despite recent advances, research on voice phishing remains limited. In particular, data generation techniques for ambiguous call transcripts near class boundaries have been relatively underexplored. In this study, we propose LLM-based data augmentation for phishing detection with two main objectives. First, we aim to enhance the representation of the phishing class by augmenting existing data and generating new phishing data based on real-world phishing cases. Second, we generate non-phishing data based on the same cases to reflect topics and contexts similar to phishing, thereby enabling robustness evaluation near class boundaries and improving model performance.

C. EVALUATION METHODS FOR LLM-GENERATED DATA

Recent studies on LLM-based data generation and augmentation have adopted a wide range of intrinsic evaluation metrics to assess the quality, utility, and naturalness of generated text. Commonly used criteria include lexical diversity, distributional similarity, semantic consistency, and stylistic fidelity, which collectively aim to ensure that synthetic data approximate the characteristics of human-authored text. Divekar and Durrett [17] introduced SynthesizRR, a framework that combines LLMs with retrieval augmentation to generate synthetic datasets for tasks such as news classification and sentiment analysis. The generated data were evaluated along two primary dimensions: lexical diversity, measured via Self-BLEU, and distributional similarity, evaluated using MAUVE scores based on GPT2-XL embeddings. Chim et al. [18] generated synthetic text grounded in real social media posts and evaluated it across three dimensions: style preservation, semantic consistency, and privacy protection. Style was quantified using Idiolect Embeddings and Part-of-Speech (POS) Distance, and semantic consistency was measured using BERTScore and Fréchet BERT Distance (FBD). Privacy retention was assessed by computing (1 – Self-BLEU), indicating how lexically dissimilar the generated content was from the source. Shakil et al. [19] proposed a set of metrics to evaluate LLM-generated news summaries. The evaluation included conciseness, relevance, contextual coherence, and readability, which were assessed by compression ratio, ROUGE, Latent Semantic Analysis (LSA) similarity, and the Flesch–Kincaid score, respectively. Their experiments demonstrated that these metrics strongly correlate with GPT-3.5–based human evaluation scores.

Whereas previous studies have tailored evaluation metrics to the objectives of specific applications, many rely on surface-level or distributional comparisons. In domains such as voice phishing detection, which often involve highly ambiguous cases, conventional criteria may fall short. It is therefore essential to assess whether synthetic data can capture the nuanced difficulty of real-world detection tasks. To this end, we propose a multifaceted framework that evaluates the naturalness, diversity, and detection difficulty of synthetic data for voice phishing detection. For detection difficulty, we introduce a novel sample-level metric, Class Centroid Distance Variability (CCDV), which quantifies detection difficulty in semantic space. CCDV complements existing metrics by explicitly capturing ambiguity between phishing and non-phishing classes. This task-specific criterion improves our ability to evaluate the robustness and practical utility of synthetic voice phishing datasets.

D. DOMAIN KNOWLEDGE-GUIDED PROMPT ENGINEERING

Zero-shot and few-shot prompting techniques often fail to deliver sufficient accuracy and reliability in domain-specific tasks, particularly when addressing subtle distinctions or highly contextualized language patterns. To overcome these limitations, recent research has explored prompt engineering strategies that incorporate domain knowledge, either through expert feedback or structured data sources such as knowledge graphs (KGs).

Karayanni et al. [20] proposed a prompt optimization framework for clinical note classification that integrates expert feedback. Their method identifies low-confidence predictions and iteratively refines prompts based on domain expert input. This approach compensates for the limitations of existing automated prompting systems (e.g., DSPy), which typically lack domain-specific reasoning capabilities. Their results demonstrated that even minimal expert involvement can significantly enhance classification performance, offering potential applicability to high-stakes fields such as law and finance.

Domain-informed prompting has also been applied using structured scientific databases and KGs. For example, Liu et al. [21] leveraged the PubChem, UniProt, and Materials Project databases to design prompts enriched with scientific terminology and logical inference steps. These enhancements improved the factual accuracy and trustworthiness of LLM-generated answers in scientific question-answering tasks, and expanded the model's utility in the Science, Technology, Engineering, and Mathematics (STEM) domains.

In the political domain, Mou et al. [22] introduced the Political Experts through Knowledge Graph Integration (PEG) framework, which enables LLMs to reason over real-time political knowledge. PEG constructs a multi-view political knowledge graph (MVPKG) that incorporates the positions of politicians, legislative actions, and public opinions. By linking localized evidence to broader political contexts, the system supports evidence-based zero-shot question answering without additional fine-tuning.

In education, Abu-Rasheed et al. [23] proposed a GPT-4–based recommendation system that uses a KG aligned with educational taxonomies and a Markov decision process (MDP)-based learning path algorithm to improve the quality of explanation. Expert intervention during prompt design ensured pedagogical validity and empirical results showed improved trustworthiness and accuracy in generated explanations, as measured by ROUGE-based metrics.

Although domain-informed prompt engineering has demonstrated strong potential in various NLP tasks, its application to security-critical domains such as voice phishing detection remains largely unexplored. To bridge this gap, we define six expert-guided criteria informed by input from a linguist affiliated with the Supreme Prosecutor's Office of Korea, who contributed in a non-official capacity. These criteria are incorporated into prompt design to improve both detection accuracy and practical applicability in real-world scenarios.

III. METHOD

This section presents the components and implementation process of the proposed LLM-based integrated framework. As illustrated in Fig. 1, the workflow consists of three main stages. 1) Generating and augmenting both phishing and non-phishing call transcripts using LLMs, 2) Evaluating the quality of the generated data and building the final dataset, 3) Detecting voice phishing using the Domain Expert LLM.

A. LLM-BASED GENERATION AND AUGMENTATION OF REALISTIC PHISHING AND NON-PHISHING CALL SCENARIOS

1) CASE-INFORMED GENERATION OF CALL TRANSCRIPTS

To enhance the realism and diversity of the phishing data, we generated call transcripts based on 143 real-world case reports obtained from the Financial Supervisory Service (FSS) [24]. These cases cover various types of fraud,

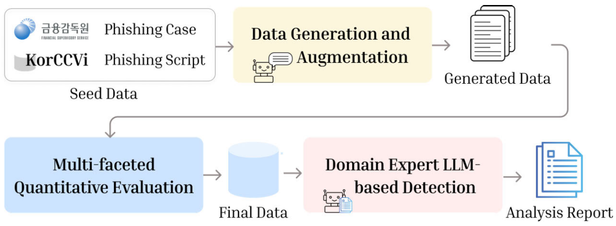

FIGURE 1. Overall workflow of the proposed voice phishing detection framework.

TABLE 1. Representative voice phishing case reported by the Financial Supervisory Service (FSS).

The suspect called the victim multiple times from May 24 at 12:51 PM to 4:00 PM, impersonating Prosecutor Kim Su-jin from the Seoul Central District Prosecutor's Office. The suspect falsely claimed that a criminal had used the victim's bank account and that in order to prove the victim's non-involvement in the crime, the victim needed to transfer a certain amount of money from their own account and purchase gift cards. Deceived by this claim, the victim transferred 1.3 million won from their bank account and purchased Culture Land and Shinsegae gift cards worth 2 million won. The victim then sent the gift card details via KakaoTalk, allowing the suspect to gain financial benefits.

including employment scams and family impersonation, and offer rich contextual information such as dialogue flow, impersonation tactics, and psychological manipulation involving urgency or emotional pressure.

During generation, situational attributes such as time, location, and impersonated entities were naturally incorporated to reflect real-world contexts. To preserve narrative realism while ensuring privacy protection, personally identifiable information was replaced with contextually appropriate synthetic content.

To improve the robustness of the model, we also generate non-phishing call transcripts that retain the overall dialogue structure and linguistic patterns of phishing calls, but omit or soften key fraudulent components such as explicit monetary requests. These borderline samples, often underrepresented in public datasets, are essential for training models to detect subtle linguistic cues and make accurate predictions in ambiguous scenarios. Table 1 summarizes a representative case of voice phishing reported by the Financial Supervisory Service (FSS). The case exemplifies a prototypical phishing scheme that combines impersonation of legal authorities with psychological manipulation tactics. An example of a case-informed call transcript is provided in Appendix A.

2) TRANSCRIPT-BASED AUGMENTATION OF PHISHING CALL TRANSCRIPTS

To address class imbalance in existing datasets, we perform data augmentation using authentic voice phishing scripts. Traditional word-replacement methods fail to reproduce the natural conversational flow typical of spoken interactions.

To overcome these limitations, we introduce an LLM-based data augmentation strategy that preserves core fraud strategies from real scripts while systematically varying contextual attributes such as time, location, and the identity of impersonated entities. This approach produces linguistically natural and contextually realistic call dialogues, thereby improving both class distribution and scenario diversity in the data.

3) LLM PROMPT OPTIMIZATION FOR REALISTIC CALL SCRIPT GENERATION

The initial call scripts generated using a baseline prompt exhibited several unnatural characteristics: (1) uniform sentence lengths, (2) overly structured and grammatically complete sentences lacking the natural variability typical of spontaneous speech, and (3) inconsistent or contextually inappropriate use of honorifics and anonymized entities, reducing overall realism. Appendix B illustrates these problems with representative excerpts from the dialogue.

To address these limitations, we refined the prompts to explicitly enforce variation in dialogue length and response style based on scenario context. For example, in urgent request situations, brief and decisive question–answer exchanges were used to convey immediacy. In contrast, explanatory scenarios incorporated longer and more detailed utterances to enhance conversational flow. This strategy resulted in a more diverse and naturalistic distribution of sentence lengths across dialogues.

All personally identifiable information (e.g., names, phone numbers) was replaced with synthetic placeholders. Additionally, prompt templates were revised with linguistic guidelines to ensure consistent and contextually appropriate use of honorifics and speech registers. Table 2 shows a prompt excerpt reflecting these refinements, and Appendix C presents example scripts generated with the optimized prompt.

TABLE 2. Prompt template for generating realistic and anonymized call dialogues.

- Individuals from financial or public institutions (such as banks, universities, or prosecutors' offices) should use a relatively formal tone and speech style.
- In conversations with acquaintances (such as family or friends), use a more familiar and informal tone and speech style.
- When simulating phone conversations with financial or public institutions (e.g., banks, universities, prosecutors' offices, workplaces), ensure that the communication follows the actual contact methods, procedures, and behaviors typically employed by these institutions.
- Each conversation should have a duration ranging between 5 to 15 minutes.
- The dialogue should reflect the natural linguistic style used in real conversations. Incorporate brief responses and swift conversational transitions to mimic the natural flow of a real-life phone call, making the dialogue concrete and realistic.
- Scenarios that involve urgent information requests should be structured with short, tense exchanges to heighten the sense of urgency, whereas situations requiring educational explanations should consist of longer, more descriptive dialogues.
- Any personal information should be represented in an appropriately anonymized or placeholder format instead of real details.

B. QUANTITATIVE EVALUATION OF NATURALNESS, DIVERSITY, AND DETECTION DIFFICULTY

To assess the quality of the generated data, we employ three key evaluation criteria: naturalness, diversity, and detection difficulty. This multifaceted evaluation ensures that the constructed dataset mirrors more closely real-world conditions, thereby improving its utility for training and evaluating voice phishing detection models.

1) NATURALNESS

The naturalness of generated scripts was evaluated using the MAUVE score [25], to quantify distributional similarity between real and synthetic scripts in embedding space.

Unlike traditional pairwise metrics, MAUVE analyzes the global embedding distribution of entire conversations, enabling a contextual assessment of linguistic fidelity. To better reflect Korean language characteristics, we substituted the original GPT-2 embeddings with 768-dimensional KoBERT embeddings.

The embeddings were clustered using k-means ($k = 500$), and the final MAUVE score was calculated using Kullback–Leibler divergence. The score ranges from 0 to 1, with higher values indicating greater alignment with the distribution of real-world dialogues.

2) DIVERSITY

We evaluated the lexical and structural diversity of the generated data using two complementary metrics, Self-BLEU and Compression Ratio (CR) [26]. Self-BLEU measures local diversity by analyzing n-gram overlap across samples, and CR quantifies global redundancy within the dataset.

Self-BLEU calculates the BLEU [27] score for each generated sample using all other samples in the dataset as references. By averaging these pairwise scores, Self-BLEU captures the degree of repetitiveness across the generated corpus. Lower Self-BLEU values indicate higher diversity of texts, suggesting reduced redundancy and greater variation in wording.

Compression Ratio (CR) evaluates global repetitiveness using a gZip-based compression method. Specifically, CR is defined as the ratio of the compressed file size $|\text{compress}(X)|$ to the original file size $|X|$. A lower CR implies greater diversity, as it reflects fewer redundancies in recurring sequences.

3) DETECTION DIFFICULTY

In real-world voice phishing detection, distinguishing between non-phishing and phishing calls is often ambiguous. However, existing datasets typically display clear lexical differences between these two classes, making detection easier than in actual scenarios. To ensure that the generated data include sufficient ambiguous cases, Fisher's Discriminant Ratio (FDR) [28] and the Silhouette Score (SS) [29] were applied to quantitatively evaluate detection difficulty.

Additionally, this paper proposes Class Centroid Distance Variability (CCDV) as a evaluation metric. Unlike conventional methods that analyze only a single class relationship or measure the average distance between classes, CCDV evaluates the distribution of distances between each data point and all class centroids. This approach provides a more nuanced understanding of how closely each instance aligns with multiple class centers, offering deeper insights into detection complexity.

Fisher's Discriminant Ratio (FDR) quantifies class separability by measuring the ratio of the between-class mean difference to the within-class variance, as defined in (1). A higher FDR value indicates greater separation between classes and lower intra-class variance, making detection more straightforward, where μ_1 and μ_2 are the mean vectors of the two classes, and σ_1^2 and σ_2^2 denote their respective within-class variances.

$$\text{FDR} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{1}$$

Silhouette Score (SS) evaluates how well a sample fits within its own cluster and measures its separation from other clusters. Values near 1 indicate well-separated classes (easier detection), whereas values approaching -1 suggest substantial overlap (more challenging detection), where a_i represents the average distance between sample i and other samples within the same cluster (cohesion), b_i denotes the average distance between sample i and the nearest neighboring cluster (separation). The SS is defined in (2):

$$\text{SS}(i) = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{2}$$

Class Centroid Distances Variability (CCDV) is a quantitative metric designed to measure how uniformly an individual data point is positioned relative to all class centroids, while also accounting for the relative proximity to the opposite class centroid. A larger CCDV value indicates a more challenging instance for detection models.

After converting all call transcripts in the KorCCVi dataset into KoBERT embeddings, representative centroids $\mathbf{C}_k$ are computed for both phishing and non-phishing data. During this process, the top and bottom 5% of values in each dimension are excluded, and a truncated mean is applied to mitigate outlier effects, ensuring a stable centroid.

The distance $\text{D}(\mathbf{x}, \mathbf{C}_k)$ between a generated data point $\mathbf{x}$ and each class centroid $\mathbf{C}_k$ is then measured by cosine distance, which effectively captures semantic similarity in high-dimensional embedding spaces. This distance is defined in (3):

$$\text{D}(\mathbf{x}, \mathbf{C}_k) = 1 - \frac{\mathbf{x} \cdot \mathbf{C}_k}{\|\mathbf{x}\| \, \|\mathbf{C}_k\|} \tag{3}$$

Next, CCDV defines a probability distribution based on distances to each class centroid and employs Shannon Entropy to measure how evenly the data point is distributed across all centroids. This allows for a quantitative assessment of whether $\mathbf{x}$ is strongly associated with a single centroid or equidistant from multiple centroids. Specifically, the probability p_k for centroid $\mathbf{C}_k$ is defined in (4):

$$p_k = \frac{e^{-D(\mathbf{x},\mathbf{C}_k)}}{\sum_j e^{-D(\mathbf{x},\mathbf{C}_j)}} \tag{4}$$

where j indexes each class centroid. The set of all p_k forms a probability distribution, from which the Shannon entropy is used to compute the CCDV value. The entropy-based component of CCDV is defined in (5):

$$\mathrm{CCDV}_{\mathrm{entropy}}(\mathbf{x}) = -\sum_k p_k \log p_k \tag{5}$$

Additionally, to increase detection difficulty for data points that are closer to the opposite class centroid, the Relative Distance Ratio (RDR) is introduced. It compares the distances to the true class centroid ($\mathbf{C}_{\mathrm{true}}$) and the opposite class centroid ($\mathbf{C}_{\mathrm{opp}}$), as defined in (6):

$$\mathrm{RDR} = \frac{D(\mathbf{x}, \mathbf{C}_{\mathrm{opp}})}{D(\mathbf{x}, \mathbf{C}_{\mathrm{true}}) + \epsilon} \tag{6}$$

where ϵ is a small constant added to prevent division by zero. A higher ratio implies that $\mathbf{x}$ is relatively closer to the opposite class centroid, suggesting increased difficulty for detection models.

Finally, the overall CCDV value is computed by summing the entropy-based component and the Relative Distance Ratio, as shown in (7):

$$\mathrm{CCDV}(\mathbf{x}) = \mathrm{CCDV}_{\mathrm{entropy}}(\mathbf{x}) + \frac{D(\mathbf{x}, \mathbf{C}_{\mathrm{opp}})}{D(\mathbf{x}, \mathbf{C}_{\mathrm{true}}) + \epsilon} \tag{7}$$

C. DOMAIN-KNOWLEDGE EMBEDDED PROMPT ENGINEERING FOR VOICE PHISHING DETECTION

In this section, we propose Domain Expert LLM, an LLM-based detection model that integrates expert knowledge to enhance both the generalization performance and the trustworthiness of voice phishing detection. Fig. 2 illustrates the overall implementation process.

In Phase 1, the evaluation criteria for the call transcripts were developed in consultation with an expert. These criteria were embedded in the Domain Expert LLM prompts to enable multifaceted analysis. In Phase 2, the prompts were iteratively refined based on expert feedback and the initial output of the model, further improving its reliability and effectiveness in detecting a wide range of phishing scenarios.

1) PHASE 1: EXPERT-GUIDED ANALYSIS CRITERIA ESTABLISHMENT

To improve the consistency and trustworthiness of voice phishing detection, we consulted with the expert to define six key analysis criteria, as summarized in Table 3. These criteria were selected for their relevance to phishing detection and their importance and priority were systematically determined. In addition, specific fraud tactics corresponding to each criterion were explicitly outlined to enable more precise analyzes. Less effective criteria were removed and higher priority criteria were placed at the top of the evaluation hierarchy, ensuring comprehensive coverage of a wide range of phishing scenarios.

TABLE 3. Phishing analysis criteria and selected examples of specific tactics.

Priority	Analysis Criterion	Examples of Specific Tactics
1	Information Theft Patterns	Installation of a fake app after deleting the mobile banking app.
2	Monetary Demand Methods	Mentioning terms such as "temporary account" or "secure account" and promising to restore funds.
3	Creation of Urgency / Pressure	Legal threats such as arrest and detention.
4	Complex Deceptive Tactics	Simultaneously leveraging multiple communication channels (telephone, SMS, email).
5	Victim Behavior Control and Isolation	Inducing victims to proceed to designated locations.
6	Trust Building Tactics	Utilizing victims' personal information (e.g., from social media or data breaches) to establish credibility.

2) PHASE 2: EXPERT FEEDBACK-BASED PROMPT REFINEMENT

The initial prompt for Domain Expert LLM was developed according to the analysis criteria established in Phase 1. To validate the trustworthiness of the detection process and results, the expert examined the reports generated by the model. Specifically, 26 low similarity cases were selected and individual analysis reports were generated for each. These reports were reviewed by the expert, whose feedback informed further refinement of the prompt.

An important insight gained during this process was the need to clearly distinguish legitimate institutional responses from phishing tactics, particularly in cases involving contact with individuals. Consequently, these responses were explicitly specified, helping the model to distinguish more effectively between legitimate and phishing intent. Table 4 shows representative examples of institutional responses, based on expert consultation.

TABLE 4. Institutional responses for phishing detection.

Normal Institutional Responses
Financial Supervisory Service can trace accounts without a warrant because it already has detailed account information, eliminating the need for direct requests.
Prosecutors focus solely on fact verification and do not instruct or guide victims to take any actions.
Upon detecting abnormal overseas transactions or repeated withdrawals, banks notify customers and advise account protection (e.g., via payment blocking) rather than instructing fund transfers to third-party accounts.

The finalized Domain Expert LLM prompt, illustrated in Fig. 3, includes explicitly defined expert roles and objectives, guidelines for phishing analysis criteria and legitimate

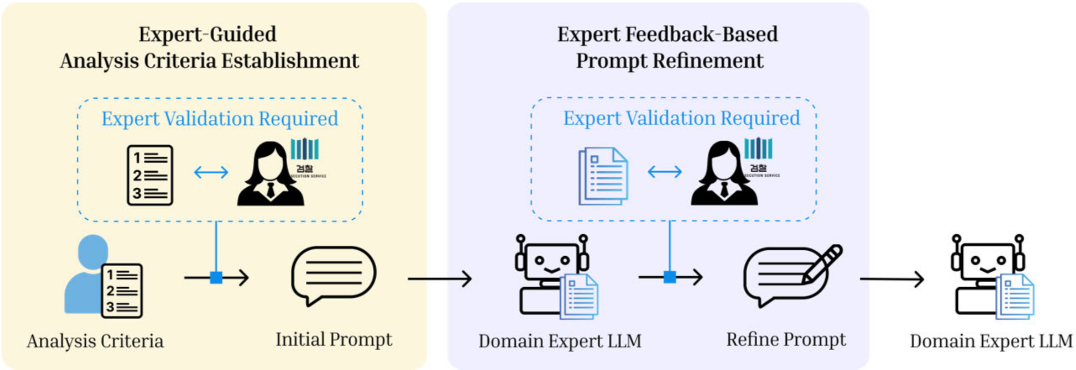

FIGURE 2. Overview of the Domain Expert LLM implementation process.

Role and Objective

You are a financial fraud expert. Your role is to analyze the provided call transcript to identify and assess potential signs of voice phishing scams. Based on recent trends in voice phishing techniques, write an analysis report following the format below. Maintain objectivity and only analyze elements that are directly observable in the call transcript.

Voice Phishing Scam Analysis Criteria (Ranked by Priority)

1. Information Request / Data Exfiltration Patterns
a) Direct Information Requests
   Requests for financial transaction details
   Requests for personally identifiable information (PII)
b) Gradual Information Extraction
   Begins with less sensitive information, progressively requesting more critical details
c) Technical Manipulation for Data Theft ...

Consideration of Genuine Institutional Call Procedures

1. Detect voice phishing by comparing it to actual institutional call procedures.
- Legitimate institutions (banks, universities, prosecutors' offices, etc.) often use official applications or require in-person visits when necessary. If such measures are present, the call is more likely to be genuine rather than a phishing attempt.
- Prosecutors only verify facts and do not instruct or guide the recipient to take specific actions. ...

Comprehensive Evaluation

- Describe the overall patterns and combinations of observed scam techniques.
- Highlight the top 3–5 most prominent or dangerous scam tactics.
- Summarize the overall risks the potential victim may face (financial loss, data compromise, etc.).
- Identify any unique or novel scam techniques observed in this case.
- Assess the likelihood of this call being a voice phishing attempt on a five-point scale:
  Very Low / Low / Medium / High / Very High.
- Note that a Medium rating still indicates a potential phishing attempt.

FIGURE 3. Excerpts from the Domain Expert LLM system prompt.

TABLE 5. Expanded KorCCVi dataset description.

Class	Source	Samples	Percentage
Phishing	KorCCVi Generated	695 1,305	13% 24.9%
Non-Phishing	KorCCVi AI Hub	2,232 1,000	42.7% 19.1%
Total		5,232	100%

TABLE 6. Additional evaluation dataset description.

Class	Source	Samples	Percentage
Phishing	Generated	143	50%
Non-Phishing	Generated	143	50%
Total		286	100%

institution responses, and a structured evaluation report format. This comprehensive approach ensures consistent and trustworthy voice phishing assessments from the perspective of financial fraud experts.

The generated analysis reports offer transparent detection rationales that can be used to improve model performance, refine analysis criteria, and serve as practical resources for law enforcement agencies and financial institutions. These reports can aid in phishing prevention and internal security training. Appendix D presents examples of analysis reports generated by Domain Expert LLM.

IV. EXPERIMENTAL SETUP

A. DATASETS

1) EXPANDED KorCCVi

We constructed the Expanded KorCCVi dataset by augmenting the original KorCCVi dataset with additional voice phishing scripts and financial/insurance consultation dialogues from AI Hub [30]. Specifically, the number of voice phishing cases increased from 695 to 2,000 through the inclusion of 1,305 newly generated phishing scripts, and the number of non-phishing cases increased from 2,232 to 3,232 through the addition of 1,000 general conversation scripts. This augmentation mitigates class imbalance and encompasses a broader range of conversational scenarios, thereby enhancing the realism and representativeness of the dataset.

2) ADDITIONAL EVALUATION DATASET

To assess the model's capability to handle various phishing cases and ambiguous non-phishing conversations, a separate evaluation dataset was constructed. It contains 286 cases: 143 generated phishing scripts based on real phishing scenarios and 143 non-phishing scripts that closely mimic phishing dialogues. By challenging the model to differentiate between actual phishing and these ambiguous non- phishing cases, this evaluation dataset provides a rigorous test of the model's discrimination ability.

B. METRICS

1) DETECTION PERFORMANCE METRICS

The performance of the Domain Expert LLM was evaluated using accuracy, precision, recall, and F1-score. Accuracy measures the overall proportion of correctly classified instances. Precision measures the proportion of correctly identified phishing samples among all instances predicted as phishing, whereas recall represents the proportion of phishing samples correctly detected out of all actual phishing instances in the dataset. The F1-score, a harmonic means of precision and recall, offers a balanced assessment of the detection performance of the model.

2) ANALYSIS REPORT QUALITY ASSESSMENT

The quality of the generated analysis reports was evaluated using a five-point scale based on four key criteria: Accuracy, which measures how faithfully the report reflects the actual facts of the original conversation; Comprehensiveness, which assesses whether all key points and details are included; Clarity, evaluating the ease of understanding and logical structure; and Depth of Analysis, determining if the report goes beyond simple summarization to offer deeper insights. For this evaluation, a reference-free LLM-based assessment method [31] was adopted, and the prompt used for evaluation is provided in Appendix E.

C. IMPLEMENTATION DETAILS

For call script generation and the Domain Expert LLM implementation, we employed GPT-4o (version: gpt-4o-2024-08-06) released by OpenAI. GPT-4o was the most advanced model available at the time, capable of generating up to 16,384 tokens per request—making it well-suited to handle long-context tasks such as phishing dialogues and non-phishing conversations. Compared to GPT-3.5, GPT-4o exhibited more refined reasoning abilities, enabling more realistic reconstructions of voice phishing scenarios. In terms of hyperparameter settings, we used default values for top-p, frequency penalty, and presence penalty, while setting the temperature to 0 to ensure stable and reproducible outputs for identical inputs.

V. RESULTS AND ANALYSIS

A. GENERATED DATA

1) NATURALNESS

According to the original MAUVE study, generated text with a MAUVE score below 0.6 (typically in the 0.37 to 0.59 range) was rated low quality in human evaluations, and scores above 0.8 were associated with text closely resembling human writing in terms of sentence structure and word choice. Based on this insight, a MAUVE score of 0.8 was adopted in this study as the threshold to evaluate text quality and distributional similarity. If the generated call scripts exceed this threshold, they can be considered meaningfully similar to real-world call scripts. Table 7 presents the MAUVE scores that compare real call scripts and generated call scripts.

TABLE 7. MAUVE score comparison between real and generated call scripts.

p_{text}	q_{text}	MAUVE
Real Call Script	Real Call Script	0.9626
Generated Call Script	Generated Call Script	0.9635
Real Call Script	Generated Call Script	0.8345

TABLE 8. POS (Part-of-Speech) tags in Mecab-ko used for diversity evaluation.

Tag	Description
NNG	General noun
NNP	Proper noun
NNB	Dependent noun
NNBC	Counter-like noun
NR	Numeral
NP	Pronoun
VV	Verb
VA	Adjective
VX	Auxiliary verb

Key findings show that the Real vs. Real comparison yielded a MAUVE score of 0.9626, indicating a high degree of internal consistency among actual call scripts. Similarly, Generated vs. Generated scored 0.9635, suggesting a coherent internal structure in the generated dataset. Although the Real vs. Generated comparison yielded a somewhat lower score of 0.8345, it still exceeds the 0.8 threshold. This outcome indicates that the generated data effectively capture realistic language patterns and maintain meaningful similarity to real-world text.

2) DIVERSITY

To evaluate whether the generated scripts cover a sufficiently diverse range of topics and vocabulary comparable to real scripts, three datasets were analyzed: (1) real call scripts from the KorCCVi dataset, (2) baseline-generated scripts produced by simply instructing the LLM to generate call dialogues without specific contextual input, and (3) scripts generated by the proposed method, where LLMs received voice phishing cases or real call scripts as input. To measure lexical diversity, nouns, verbs, and adjectives were extracted using Mecab-ko from KoNLPy [32] for morphological analysis, focusing on parts of speech that contribute primarily to meaning. Table 8 summarizes the part-of-speech (POS) tags used in the evaluation. To reduce length bias, 100 scripts per data type were randomly sampled and all text lengths were normalized to 300 tokens.

Table 9 presents the Self-BLEU and CR scores for real and generated call scripts. Lower Self-BLEU and CR scores indicate reduced lexical redundancy and increased vocabulary diversity.

Generated scripts generally exhibited higher Self-BLEU and CR scores than real call scripts, suggesting a greater tendency to repetitive expressions. Among the tested methods, the baseline generation produced the highest self-BLEU score (0.424 for phishing and 0.113 for non-phishing) indicating the most pronounced lexical repetition, and thus

TABLE 9. Diversity measures (Self-BLEU and CR) for real and generated scripts.

Data Type	Class	Self-BLEU (↓)	CR (↓)
Real Call Script	Phishing	0.055	2.858
	Non-phishing	0.058	2.96
(Baseline) Generated Call Script	Phishing	0.424	4.199
	Non-phishing	0.113	2.858
(Proposed Method) Generated Call Script	Phishing	0.161	3.174
	Non-phishing	0.107	2.89

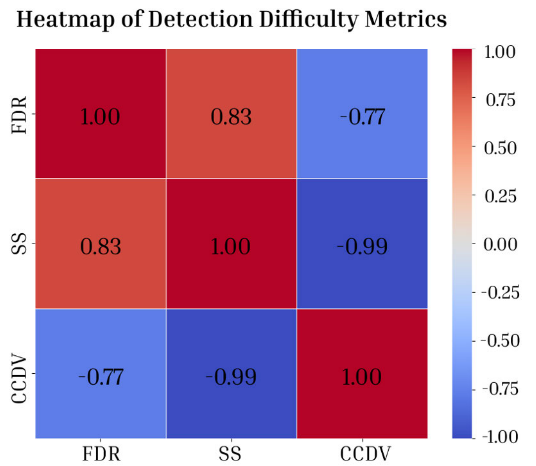

FIGURE 4. Correlations between detection difficulty metrics.

the lowest diversity. Additionally, the highest CR implies that repeated expressions rendered the text more compressible, confirming the challenge of generating varied expressions. The proposed method, which incorporates real voice phishing cases or transcripts, considerably improved lexical diversity. For example, the Self-BLEU score for phishing scripts decreased to 0.161, reflecting a wider range of vocabulary compared to the baseline. The lower CR of the proposed method also shows less redundancy and more natural variations.

3) DETECTION DIFFICULTY

To validate whether the proposed CCDV metric effectively captures the detection difficulty, FDR and SS were compared with CCDV using both real and generated call scripts. Fig. 4 presents a heatmap visualization showing the correlation between FDR, SS, and CCDV.

The results indicate a strong correlation between CCDV and these established difficulty measures: CCDV versus FDR at -0.99 (very strong negative correlation) and CCDV versus SS at -0.77 (strong negative correlation). Consequently, CCDV produces evaluation results consistent with existing metrics and adds the unique advantage of quantifying difficulty at the individual data point level.

Fig. 5 visually demonstrates the concept and impact of CCDV. One high-CCDV sample (top 5%) and one low-CCDV sample (bottom 5%) from the generated data were randomly selected and plotted, along with real voice phishing

TABLE 10. Detection difficulty analysis using different evaluation metrics.

Data Type	FDR (↓)	SS (↓)	CCDV (↑)
Real Call Script	38.452	0.857	0.516
Generated Phishing Call Script	17.028	0.836	0.794
Generated Non-phishing Call Script	6.254	0.693	15.616

and non-phishing call, in a two-dimensional embedding space.

High-CCDV samples tend to be farther from their respective class centroids—or closer to the opposite class centroid, indicating greater ambiguity and difficulty for detection models. In contrast, low-CCDV samples cluster near their own class centroids, making them easier to classify. These findings confirm that CCDV effectively pinpoints and examines cases that pose challenges to detection models. Moreover, CCDV complements existing detection difficulty metrics by providing a more granular evaluation of ambiguous cases, thus improving efforts to refine and strengthen phishing detection models.

Experimental results show that generated call scripts exhibited lower FDR and SS but higher CCDV, thereby indicating increased detection difficulty and a more ambiguous boundary between phishing and non-phishing conversations. Notably, the CCDV value rose significantly in generated non-phishing scripts, suggesting that detection models are more likely to misclassify generated non-phishing calls as phishing. This occurs because many of these samples lie closer to the phishing data in the embedding space, rather than aligning with typical non-phishing conversations. Such a trend heightens the risk of false positives, where genuinely non-phishing calls are mistakenly flagged as phishing, potentially reducing overall detection accuracy.

However, the elevated number of ambiguous cases introduced by the generated dataset can strengthen the robustness of the model by exposing detection algorithms to complex and ambiguous real-world scenarios.

B. DOMAIN EXPERT LLM DETECTION PERFORMANCE

The detection performance of various models on both the Expanded KorCCVi test set and the Additional Evaluation dataset is presented in Table 11. All models were trained on the Expanded KorCCVi dataset before being evaluated on these test sets.

1) PERFORMANCE ON THE EXPANDED KorCCVi TEST SET

In the Expanded KorCCVi dataset, traditional ML and DL models exhibited high detection performance, with F1 scores exceeding 0.98. Notably, KoBERT achieved the highest performance, recording an accuracy of 0.9981 and an F1-score of 0.9975. Other traditional machine learning models, such as RandomForest and XGBoost, also maintained F1 scores above 0.99, demonstrating stable detection performance.

By contrast, Vanilla LLM and Domain Expert LLM (Phase 1, Phase 2) showed relatively lower F1 scores, ranging from 0.95 to 0.97. Analysis revealed that LLM-based detection models exhibited reduced performance when key indicators of fraudulent tactics were missing or ambiguous. In particular, when phishing scripts lacked explicit financial demands or identity verification requests, the models tended to underestimate associated risk, leading to misclassification as non-phishing calls (Appendix F).

Phase 2 was designed as a stricter detection model, incorporating standard operational procedures from financial and legal institutions. As a result, non-phishing call detection performance improved, reducing false positives (FP) and elevating Precision to 0.9947, compared to 0.9693 for Vanilla LLM and Phase 1. However, stricter detection criteria also raised the likelihood that modified phishing patterns—particularly those resembling normal calls—would be classified as non-phishing, highlighting a trade-off between gains in precision and reductions in recall.

2) PERFORMANCE ON THE ADDITIONAL EVALUATION DATASET

The performance of traditional ML and DL models, including Attention-based CNN-BiLSTM, decreased considerably when tested on the Additional Evaluation dataset, which featured previously unseen voice phishing cases and ambiguous cases. The precision of KoBERT decreased to 0.6041, while the F1 scores of other traditional classification models decreased to around 0.70, underscoring their limited adaptability to novel phishing tactics.

In contrast, both Vanilla LLM and Domain Expert LLM (Phase 1, Phase 2) maintained F1 scores above 0.95, demonstrating greater robustness in detecting unfamiliar phishing patterns. Notably, Phase 2 used stricter classification criteria, thereby improving non-phishing call detection and reducing false positives (Precision: 0.9787, versus 0.9524 for Vanilla LLM and 0.9693 for Phase 1). However, this stricter approach led to a reduction in recall (0.9650), as some modified phishing cases were more likely misclassified as non-phishing calls.

3) CRITERION-BASED EVALUATION OF THE DOMAIN EXPERT LLM-GENERATED REPORT

To assess the effect of expert knowledge on the quality of the detection report, we compared reports generated by Vanilla LLM (no expert insights) with those produced by Domain Expert LLM (incorporating expert knowledge). This comparison aimed to determine whether the integration of expert knowledge genuinely improves the trustworthiness of detection. The evaluation was conducted on 200 sample scripts from the Expanded KorCCVi dataset. Table 12 summarizes the quality assessment of detection reports, distinguishing whether expert knowledge was applied. The results show that Domain Expert LLM outperformed Vanilla LLM overall, particularly in accuracy (Faithfulness) and analytical depth, indicating that voice phishing analysis criteria (i.e., expert knowledge) contributed to greater trustworthiness and analytical rigor in the detection reports.

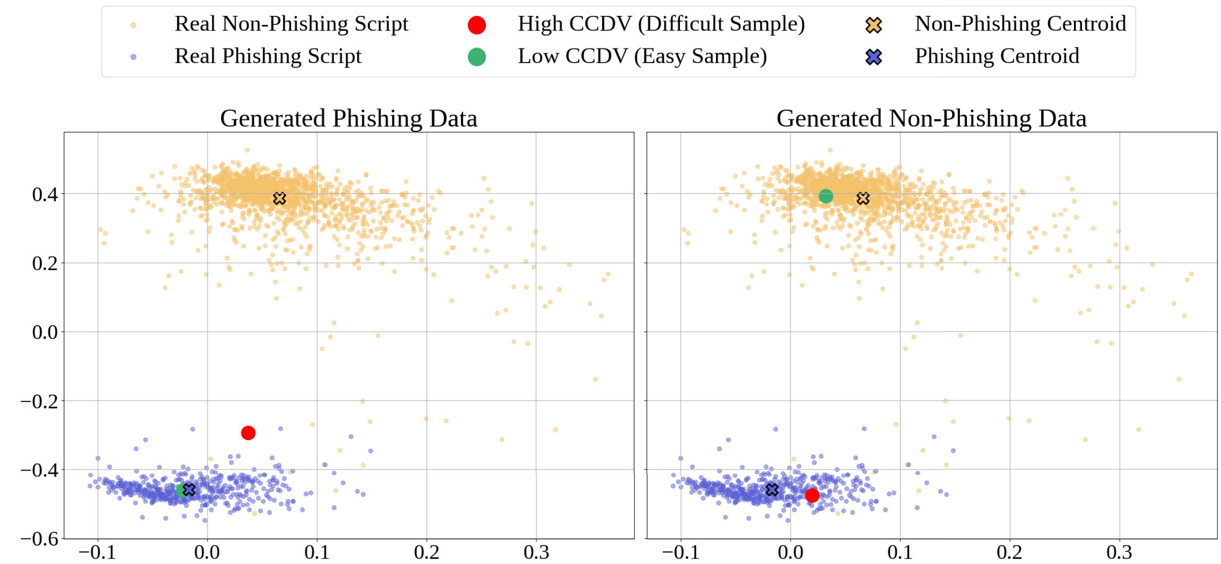

FIGURE 5. 2D embedding visualization of generated phishing (left) and non-phishing (right) scripts highlighting high and low CCDV samples.

TABLE 11. Detection performance.

Model \ Dataset	Expanded KorCCVi Dataset				Additional Evaluation Dataset			
	Acc (↑)	F1 (↑)	Recall (↑)	Precision (↑)	Acc (↑)	F1 (↑)	Recall (↑)	Precision (↑)
Logistic Regression	0.9942	0.9925	0.9925	0.9925	0.6013	0.7091	0.9720	0.5582
DecisionTree	0.9598	0.9481	0.9481	0.9481	0.5664	0.6884	0.9580	0.5372
RandomForest	0.9923	0.9900	0.9876	0.9925	0.5419	0.6858	1.0	0.5218
AdaBoost	0.9885	0.9852	0.9876	0.9828	0.5629	0.6913	0.9790	0.5343
GradientBoosting	0.9904	0.9876	0.9876	0.9876	0.5489	0.6861	0.9860	0.5261
HistGradientBoosting	0.9923	0.9901	0.9901	0.9901	0.5454	0.6875	1.0	0.5238
XGB	0.9933	0.9913	0.9901	0.9925	0.5559	0.6924	1.0	0.5296
LGBM	0.9904	0.9876	0.9851	0.9900	0.5524	0.6908	1.0	0.5276
CatBoost	0.9942	0.9925	0.9901	0.9950	0.5559	0.6909	0.9930	0.5298
LinearSVC	0.9933	0.9913	0.9876	0.9950	0.6013	0.7121	0.9860	0.5573
KoBERT	0.9981	0.9975	0.9950	1.0	0.6041	0.7064	0.9510	0.5619
Attention based CNN-BiLSTM	0.9100	0.9087	0.9100	0.9120	0.7543	0.7501	0.7544	0.7741
Vanilla LLM	0.9837	0.9790	0.9802	0.9778	0.9650	0.9655	0.9790	0.9524
Domain Expert LLM (Phase 1)	0.9636	0.9523	0.9358	0.9693	0.9720	0.9722	0.9790	0.9655
Expert-Guided LLM (Phase 2)	0.9703	0.9604	0.9284	0.9947	0.9720	0.9718	0.9650	0.9787

TABLE 12. G-EVAL-4's scores for quality assessment of detection reports with and without expert knowledge.

Model \ Dataset	Faithfulness		Completeness		Analytical Depth		Clarity	
	Non-Phishing	Phishing	Non-Phishing	Phishing	Non-Phishing	Phishing	Non-Phishing	Phishing
Vanilla LLM	2.311	4.331	1.162	4.44	1.661	4.092	4.033	4.582
Proposed	4.016	4.4	1.196	4.538	1.562	4.538	4.391	4.505

TABLE 13. Example of an actual voice phishing case.

Ms. K (female, in her 50s) received a call in the morning from a scammer claiming to be from the school administration office, stating that her child had suffered a head injury and required immediate medical treatment, demanding a deposit of 10 million KRW. As a result, she transferred 8 million KRW and fell victim to the scam. The scammer accurately knew her child's name, mobile phone number, and school name.

For non-phishing conversation reports, Vanilla LLM recorded a score of 2.311, whereas Domain Expert LLM achieved 4.016, indicating a substantial improvement. A similar trend was observed for phishing call scripts, where the score increased from 4.331 to 4.4, further suggesting that expert knowledge improves the accuracy of the report.

Domain Expert LLM also received higher scores for analytical depth. In phishing detection reports, Vanilla LLM scored 4.092 and Domain Expert LLM scored 4.538, demonstrating that integrating voice phishing analysis criteria improves the depth of the analysis. Conversely, the analytical depth for non-phishing conversation reports showed a tendency to decrease, as non-phishing scenarios often do not require elaborate analysis, thus yielding more concise reports. This simplicity appeared to improve the clarity score.

TABLE 14. Generated voice phishing call script based on the real case.

Hello, this is the administration office at Seoul National University. Am I speaking with Kim Min-seo's mother? Yes, that's me. Who is this? Ah, yes. I'm sorry to call like this, but this is urgent. Kim Min-seo just had a bad fall down the stairs at school and hit her head hard. What? Min-seo? What happened? Is she okay? She's on the way to the emergency room now. She's conscious, but her CT scan shows a possible brain hemorrhage, so she needs emergency surgery right away. Oh my god… No… I need to go. I'll go to the hospital now. Ma'am, they're preparing for surgery right now, so coming to the hospital won't change anything at this point. Also, Min-seo specifically asked us not to contact you because she didn't want you to worry. What? No, I still need to be there. I understand, but the most urgent problem right now is the surgery cost. This procedure isn't covered by insurance, so you'll need to pay 10 million won upfront before we can proceed. Ten million won? That's… a huge amount… Yes, because it's brain surgery, the cost is high. The longer we wait, the more dangerous it gets. We need the payment as soon as possible. I… I don't think I can get 10 million won right away… If you can transfer at least 8 million won now, we can start the surgery. You can take care of the rest later. Okay… I'll go to the bank now. Where do I send it? Thank you. Here's the account number. NongHyup Bank, account holder: Seoul National University College of Medicine. Okay, I wrote it down. I'll send the money now. Please, save my Min-seo. Don't worry. We're doing everything we can. As soon as the payment is confirmed, we'll begin the surgery. Yes, okay. I'll transfer it now. Alright, we'll be standing by. And please refrain from contacting Min-seo—she was very clear that she didn't want you to worry. Okay… I'll send it right now.

TABLE 15. Generated voice phishing call script based on the real case.

Hello, is this Soyeon's mother? Yes, this is. May I ask who's calling? Ah, hello. I'm Soyeon's homeroom teacher. I'm calling because something urgent came up. Oh, I see. What's going on? Well… during P.E. class today, Soyeon had a minor accident. It's nothing too serious, but she bumped her head a little, so we took her to the hospital for a check-up. Oh no, really? Is she okay? What did the doctor say? Yes, thankfully, it doesn't seem like a serious injury. The doctor just wants to run a few more tests, but she's calm and doing fine. So far, there doesn't seem to be anything to worry about. Oh, that's a relief… but was she really scared? She was a bit startled at first, but she's doing much better now. The doctor wants to monitor her condition a little longer, but they don't expect any major problems. If you can, it might be best to come to the hospital and talk to the doctor yourself. Of course! I'll come right away. Which hospital is she at? We're at Donga Hospital near the school. We took her to the closest hospital just to be safe. Got it! I'll head there right now. Thank you for letting me know, teacher. Sure, just come as soon as you can and be safe. When you arrive, just let them know, and I'll be with Soyeon. Okay, see you soon. Yes, see you soon.

However, the clarity score for the phishing detection reports decreased slightly. As expert knowledge was incorporated, the detection criteria became more detailed, leading to longer reports and a modest reduction in clarity. Consequently, an optimization strategy is required to preserve the complexity of the detection criteria while improving clarity.

TABLE 16. Initial system prompt.

You are an AI designed to generate general call scripts with similar conversation topics by referencing actual voice phishing cases. • Based on real voice phishing cases, simulate the conversations and create specific and realistic call scripts. • Ensure consistency in speech styles and appropriate titles based on roles to construct detailed and lifelike voice phishing call scenarios. • Write only the spoken content, as in STT (Speech-to-Text) format.

TABLE 17. Call script generated using the initial system prompt.

This is **Prosecutor OOO** from the Prosecutor's Office. I am urgently requesting your cooperation regarding a criminal case associated with your name. To protect your assets, you need to make an immediate transfer. A transfer? What do you mean? Are you saying I'm involved in a crime? We have confirmed that your personal information has been illegally used. To prevent further damage, you must immediately transfer funds to a specific account. Prosecutor, I'm not sure if this is the right thing to do. I really don't know what to do. I understand. But now, there is no more time to waste. I will now provide you with the account number. **It is –***. Understood. I'll make the transfer right away.

TABLE 18. Call script generated using the revised system prompt.

This is Prosecutor **Ji-hoon Park** from the Seoul Central District Prosecutor's Office. We have detected that personal information registered under your name, **Mr. Min-Soo Kim**, has been leaked and misused in multiple fraud cases. Have you recently noticed any suspicious activity involving your personal information? … For a precise verification, can you access your banking application while we are on the call? Please log into your banking app and navigate to the **"Safe Transfer"** menu. To finalize this process, you must transfer a designated amount to a temporary secure account that we have set up for verification purposes. I'm feeling a bit uneasy… Are you sure this is a legitimate procedure? Mr. Kim, I completely understand your concern. However, this is a standard security procedure to ensure your account is protected. We are conducting a real-time check for any irregular transactions, so we need your **immediate cooperation**.

VI. CONCLUSION

This study proposed an LLM-based integrated framework to improve both the quality of existing voice phishing datasets and the trustworthiness of detection models. To address critical problems, such as limited scenario diversity, the challenge of distinguishing phishing from non-phishing conversations, and class imbalance, an LLM-based method was employed to generate and augment call scripts, combined with a multifaceted quantitative data evaluation. To ensure model reliability, we established phishing detection criteria in consultation with an expert from the Supreme Prosecutor's

TABLE 19. Call script analysis report generated by the Domain Expert LLM.

1. Information Request / Data Exfiltration Patterns

a) Direct Information Requests: There is no explicit request for financial or personally identifiable information in the conversation. However, the mention of "transfer card" suggests an indirect intent to obtain financial details.

b) Gradual Information Extraction: The conversation appears to confuse the customer, potentially leading to a gradual disclosure of additional information. For instance, repeated misstatements of amounts and terminology could cause the customer to inadvertently provide their financial details.

c) Technical Manipulation for Data Theft: No clear technical methods for data theft are observed. However, phrases like "reissuing the transfer card" may be used to create a situation where the victim feels compelled to provide information.

2. Monetary Solicitation Strategies

a) Direct Requests: The conversation includes explicit monetary demands, such as requesting a "transfer card issuance fee" and specifying an amount (e.g., 1.5 million KRW) to be paid.

b) Indirect Requests: The scammer employs an indirect approach, disguising the payment request as part of a "fraudulent payment resolution."

c) Incremental Requests: The initial amount of 4 million KRW is mentioned, followed by an additional demand for 1.5 million KRW, gradually increasing the required payment.

d) Safety Assurance as a Bait: The conversation uses assurances such as "we will ensure that there is no fraudulent transaction record" to reinforce the legitimacy of the monetary request.

...

Comprehensive Evaluation

This conversation involves a fraudulent entity impersonating an authoritative institution, referencing internal regulations while demanding payments. The scammer employs time pressure and legal threats to manipulate the victim. By controlling the victim's actions and requesting payments for multiple reasons in a staged manner, confusion is induced.

Most Notable and Risky Scam Tactics:

- Creating urgency through time pressure, forcing the victim into immediate decision-making.
- Indirect but specific monetary requests, disguised under terms like "fraudulent payment resolution."
- Involvement of multiple roles to confuse the victim further and undermine trust.

Potential Risks to the Victim:

- Financial Loss: Large sums of money may be demanded under the pretext of loan processing or fraud resolution.
- Psychological Distress: The use of threats and pressure tactics may cause anxiety and emotional distress. Unique Fraud Techniques Observed:
- The use of "fraudulent payment resolution" as a novel pretext to justify monetary requests.

Likelihood of Voice Phishing: Very High

TABLE 20. LLM-based analysis report faithfulness evaluation prompt.

You will be given a call analysis report generated by Domain Expert LLM based on a call transcript. Your task is to rate the report's analysis on one metric: Faithfulness (1-5)
Please make sure you read and understand these instructions carefully.
Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
Faithfulness (1-5)

- Assess whether the analysis report accurately reflects the content of the call transcript.
- Factual Alignment: Does the report strictly rely on information explicitly present in the call transcript?
- Hallucinated Content: Does the report introduce or distort any details that are not found in the call transcript?
- Accuracy of Interpretation: Does the report correctly interpret and represent the meaning of the call transcript?

Evaluation Steps:

1) Carefully read the call transcript to understand its key content.
2) Read the analysis report and verify whether it is factually accurate and fully based on the call transcript.
3) Check if the report contains any information that is incorrect, hallucinated, or misinterpreted.
4) Assign a Faithfulness score on a scale of 1 to 5. (1 = Highly Inaccurate, 5 = Highly Accurate)

Example:
Conversation History: {Call Transcript}
Response (Analysis Report): {Analysis Report}
Evaluation Form (scores ONLY):
● Faithfulness

TABLE 21. Identity verification attempt without explicit fraud request.

Hello, this is Jung Ho-soo from the Daebang Planning Team in the financial sector. We have identified an identity theft case registered in your name, so I'm calling you. Are you available to talk?
Yes, I can talk. What is this regarding?
First, I'd like to verify your identity. Is your name Jang Bo-ri?
Yes, that's correct.
Have you recently had any items that might expose your personal information like your ID or mobile phone stolen or lost? No, I haven't.
Understood. While investigating fraudulent bank accounts and illegal credit card activities, our team discovered an account registered in your name. Have you ever transferred or sold an account to someone else?
No, I have never done that.
In that case, it appears you are the victim, Jang Bo-ri.
...

Office (in a personal capacity) and embedded them in the Domain Expert LLM prompt.

Experimental results show that the generated data effectively mitigated the limitations of conventional datasets, such as insufficient scenario diversity, similarity between phishing and non-phishing calls, and class imbalance. Moreover, the expert-guided call analysis criteria played a key role in aligning the model's detection capabilities with real-world operational needs.

This study contributes notable academic and practical insight to the field of voice phishing detection. From an academic perspective, using LLMs to generate realistic and diverse voice phishing data tackles existing dataset problems such as scenario bias and class imbalance. This approach establishes a stronger data foundation for voice phishing detection research. Additionally, by defining multifaceted evaluation criteria—covering naturalness, diversity, and detection difficulty—and introducing Class Centroid Distances Variability (CCDV) for objective per-data-point detection difficulty measurement, this work presents a systematic method for verifying how faithfully the generated data mirror genuine voice phishing scenarios. Consequently, it holds significant academic value for future voice phishing detection studies. From a practical standpoint, expert consultation led

TABLE 22. Risk assessment report in the absence of strong fraud indicators.

1. Information Request / Data Theft Patterns a) Direct Information Requests: There were **no explicit demands for financial transaction details** during the call. However, an **attempt was made to verify personal identification information**. The caller confirmed the victim's name, seemingly preparing to request more sensitive data. … 2. Monetary Demands **Neither direct nor indirect demands for money were observed**. At this stage, the focus appears to be on gathering information. … **Evaluation of the Likelihood that this Call is a Voice Phishing Attempt (on a 5-point scale): Moderate** There are early signs of various scam techniques, but **no direct monetary demands or pressure are present at the moment**. However, depending on subsequent actions, the situation has a high potential to escalate into a scam. While the initial conversation provides limited information to definitively determine fraud, it may lay the groundwork for a long-term scam plan.

to the development of call analysis criteria applicable to a variety of voice phishing cases, and expert validation helped ensure high trustworthiness in real-world settings. Moreover, the automatic generation of detailed analysis reports offers tangible evidence for subsequent research or operational tasks (e.g., refining model performance or analysis criteria). These reports can be utilized broadly by investigative agencies and financial institutions for prevention efforts or internal security training. Furthermore, by leveraging LLM to create ''unrealized yet plausible'' phishing types and augment them as training data, this approach enables the earlier detection of new attack vectors, offering substantial practical value in quickly responding to emerging voice phishing threats.

Despite these achievements, several limitations remain. First, the lack of sufficient non-phishing conversation scripts from various institutions limited the effective learning and validation of subtle differences between actual institutional calls and impersonation attempts. Second, the potential impact of transcription errors on detection outcomes was not adequately considered, potentially causing performance degradation in the detection model. Third, due to constraints in optimizing the Domain Expert LLM prompt, the detection criteria were set too strictly, resulting in some phishing cases going undetected. This underscores the need for additional optimization to balance sensitivity and specificity in real-world applications.

Future studies may acquire additional data, including public-sector and financial institution consultation scripts and on-site manuals, to better reflect the nuances of authentic institutional calls in both data generation and analysis criteria. Further assessments on a systematic quantitative analysis of the effects of conversation length and transcription errors on model performance should be undertaken to develop error-correction methods. Lastly, an automated optimization strategy for detection criteria could be pursued to further improve the balance between sensitivity and specificity in practical deployments.

APPENDIX A
EXAMPLES OF GENERATED DATA

See Tables 13 –15.

APPENDIX B
INITIAL SYSTEM PROMPT AND GENERATED SCRIPT EXAMPLE

See Tables 16 and 17.

APPENDIX C
ENHANCED RESULTS THROUGH PROMPT REVISION

See Table 18.

APPENDIX D
CALL SCRIPT ANALYSIS REPORT BY DOMAIN EXPERT LLM

To support the interpretability and trustworthiness of the model's decisions, we established evaluation criteria for analyzing call transcripts. These criteria were developed with reference to expert input, provided in a personal capacity, from a practitioner affiliated with the Supreme Prosecutor's Office.

APPENDIX E
LLM-BASED ANALYSIS REPORT QUALITY EVALUATION PROMPT

See Table 20.

APPENDIX F
IMPACT OF CONTENT AMBIGUITY ON VOICE PHISHING DETECTION

See Tables 21 and 22.

ACKNOWLEDGMENT

The authors would like to thank Kyung-Wha Kim from the Forensic Science Division, Supreme Prosecutor's Office in Seoul, South Korea, for her guidance on defining evaluation criteria for the analysis of phishing calls and on validating reports produced by the model.

REFERENCES

[1] T. F. Blauth, O. J. Gstrein, and A. Zwitter, "Artificial intelligence crime: An overview of malicious use and abuse of AI," *IEEE Access*, vol. 10, pp. 77110–77122, 2022, doi: 10.1109/ACCESS.2022.3191790.

[2] M. Lee and E. Park, "Real-time Korean voice phishing detection based on machine learning approaches," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 7, pp. 8173–8184, Jul. 2023, doi: 10.1007/s12652-021-03587-x.

[3] M. K. M. Boussougou and D.-J. Park, "A real-time efficient detection technique of voice phishing with AI," in *Proc. Korea Comput. Congr. (KCC)*, Jun. 2021, pp. 768–770.

[4] M. K. M. Boussougou, S. Jin, D. Chang, and D.-J. Park, "Korean voice phishing text classification performance analysis using machine learning techniques," in *Proc. Annu. Conf. Korea Inf. Process. Soc.*, Nov. 2021, pp. 297–299.

[5] M. K. M. Boussougou and D.-J. Park, "Attention-based 1D CNN-BiLSTM hybrid model enhanced with FastText word embedding for Korean voice phishing detection," *Mathematics*, vol. 11, no. 14, p. 3217, Jul. 2023, doi: 10.3390/math11143217.

[6] M. K. M. Boussougou and D.-J. Park, "Exploiting Korean language model to improve Korean voice phishing detection," *KIPS Trans. Softw. Data Eng.*, vol. 11, pp. 437–446, Oct. 2022, doi: 10.3745/KTSDE.2022.11.10.437.

[7] S. Yu, Y. Kwon, M. Kim, and K. Lee, "Korean voice phishing detection applying NER with key tags and sentence-level N-gram," *IEEE Access*, vol. 12, pp. 52951–52962, 2024, doi: 10.1109/ACCESS.2024.3387027.

[8] Z. Shen, S. Yan, Y. Zhang, X. Luo, G. Ngai, and E. Y. Fu, "'It warned me just at the right moment': Exploring LLM-based real-time detection of phone scams," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2025, pp. 1–7, doi: 10.1145/3706599.3720263.

[9] H. Park, J. Lee, S. Choi, N. Yoon, and S. Han, "Voice phishing detection improvement by conversation analysis and LLM-based scenario augmentation," in *Proc. Fall Conf. Korean Soc. Internet Inf.*, Jeju, South Korea, 2024, pp. 343–344.

[10] A. Derakhshan, I. G. Harris, and M. Behzadi, "Detecting telephone-based social engineering attacks using scam signatures," in *Proc. ACM Workshop Secur. Privacy Analytics*, Apr. 2021, pp. 67–73, doi: 10.1145/3445970.3451152.

[11] Q. Zhao, K. Chen, T. Li, Y. Yang, and X. Wang, "Detecting telecommunication fraud by understanding the contents of a call," *Cybersecurity*, vol. 1, no. 1, Dec. 2018, Art. no. 8, doi: 10.1186/s42400-018-0008-5.

[12] J. Li, C. Zhang, and L. Jiang, "Innovative telecom fraud detection: A new dataset and an advanced model with RoBERTa and dual loss functions," *Appl. Sci.*, vol. 14, no. 24, p. 11628, Dec. 2024, doi: 10.3390/app142411628.

[13] A. K. P. Venkata and L. Gudala, "The potential and limitations of large language models for text classification through synthetic data generation," *Int. Res. J. Eng. Appl. Sci.*, vol. 12, no. 2, pp. 9–15, Apr. 2024, doi: 10.55083/irjeas.2024.v12i02002.

[14] H. Zhao, H. Chen, T. A. Ruggles, Y. Feng, D. Singh, and H.-J. Yoon, "Improving text classification with large language model-based data augmentation," *Electronics*, vol. 13, no. 13, p. 2535, 2024, doi: 10.3390/electronics13132535.

[15] Z. Li, H. Zhu, Z. Lu, and M. Yin, "Synthetic data generation with large language models for text classification: Potential and limitations," in *Proc. Conf. Empirical Methods Natural Lang. Process., Assoc. Comput. Linguistics (ACL)*, 2023, pp. 10443–10461, doi: 10.18653/v1/2023.emnlp-main.647.

[16] S. Gopali, F. Abri, A. S. Namin, and K. S. Jones, "The applicability of llms in generating textual samples for analysis of imbalanced datasets," *IEEE Access*, vol. 12, pp. 136451–136465, 2024, doi: 10.1109/ACCESS.2024.3463400.

[17] A. Divekar and G. Durrett, "SynthesizRR: Generating diverse datasets with retrieval augmentation," 2024, *arXiv:2405.10040*.

[18] J. Chim, J. Ive, and M. Liakata, "Evaluating synthetic data generation from user generated text," *Comput. Linguistics*, vol. 51, no. 1, pp. 191–233, Mar. 2025.

[19] H. Shakil, A. M. Mahi, P. Nguyen, Z. Ortiz, J. Kalita, and M. T. Mardini, "Evaluating text summaries generated by large language models using OpenAI's GPT," in *Proc. 23rd Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2024, pp. 951–956, doi: 10.1109/ICMLA61862.2024.00140.

[20] N. Karayanni, A. Awwad, C.-L. Hsiao, and S. P Shanmugam, "Keeping experts in the loop: Expert-guided optimization for clinical data classification using large language models," 2024, *arXiv:2412.02173*.

[21] H. Liu, H. Yin, Z. Luo, and X. Wang, "Integrating chemistry knowledge in large language models via prompt engineering," *Synth. Syst. Biotechnol.*, vol. 10, no. 1, pp. 23–38, Mar. 2025, doi: 10.1016/j.synbio.2024.07.004.

[22] X. Mou, Z. Li, H. Lyu, J. Luo, and Z. Wei, "Unifying local and global knowledge: Empowering large language models as political experts with knowledge graphs," in *Proc. ACM Web Conf.*, Singapore, May 2024, pp. 2603–2614, doi: 10.1145/3589334.3645616.

[23] H. Abu-Rasheed, C. Weber, and M. Fathi, "Knowledge graphs as context sources for LLM-based explanations of learning recommendations," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Kos Island, Greece, May 2024, pp. 1–5.

[24] Financial Supervisory Service. (2023). *Voice Phishing Exhibition 2023*. Accessed: Apr. 1, 2025. [Online]. Available: https://www.fss.or.kr/fss/bbs/B0000203/list.do?menuNo=200686

[25] K. Pillutla, S. Swayamdipta, R. Zellers, J. Thickstun, S. Welleck, Y. Choi, and Z. Harchaoui, "MAUVE: Measuring the gap between neural text and human text using divergence frontiers," in *Proc. 35th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Red Hook, NY, USA, 2021, pp. 4816–4828.

[26] C. Shaib, J. Barrow, J. Sun, A. F. Siu, B. C. Wallace, and A. Nenkova, "Standardizing the measurement of text diversity: A tool and a comparative analysis of scores," 2024, *arXiv:2403.00553*.

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Philadelphia, PA, USA, 2002, pp. 311–318.

[28] R. A. FISHER, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936, doi: 10.1111/j.1469-1809.1936.tb02137.x.

[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[30] AI Hub. (2023). *Public Service (Call Center) Query-Response Data 2023*. National Information Society Agency, Daegu, South Korea. [Online]. Available: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=98

[31] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: NLG evaluation using GPT-4 with better human alignment," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Singapore, 2023, pp. 2511–2522.

[32] E. L. Park and S. Cho, "KoNLPy: Korean natural language processing in Python," in *Proc. Annu. Conf. Hum. Lang. Technol.*, Seoul, South Korea, Oct. 2014, pp. 133–136.

HYUNGHEE PARK received the B.Eng. degree in information systems engineering from Sungshin Women's University, Seoul, South Korea, in 2023, where she is currently pursuing the M.Eng. degree in future convergence technology engineering.

She actively involved in research projects focusing on AI-driven multimodal learning and has contributed to studies on generative AI applications. Her research interests include deep learning, multimodal learning, and generative AI.

JIWON LEE received the B.Eng. degree in information systems engineering from Sungshin Women's University, Seoul, South Korea, in 2023, and the M.Eng. degree in future convergence technology engineering in 2025.

Her research interests include web services, digital audio forensics, and generative AI.

SANGHYUN HAN received the B.S. degree in statistics from Sungkyunkwan University, Seoul, South Korea, and the M.S. degree in applied data science from the University of Michigan, Ann Arbor, MI, USA.

He is currently a Data Scientist at SK Telecom, Seoul. His research interests include applied data science and machine learning.

HAEWON BYUN received the B.Sc. degree in computer science from Yonsei University, Seoul, South Korea, in 1990, and the M.Eng. and Ph.D. degrees in computer science from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1992 and 2004, respectively.

Since 2006, she has been a Professor with the School of AI Convergence, Sungshin Women's University, Seoul. Her research interests include computer graphics, multimodal deep learning, and generative AI.