Towards Safer Calls for Everyone: Designing a Benchmark Dataset for Evaluating Voice Phishing Detection Models

Joeun Kang[1], Gyuri Choi[1], Chanhyuk Yoon[2], Yongbin Jeong[2], Younggyun Hahm[2], Shea Husband[1], Hansaem Kim[1†]

[1]Yonsei University, Seoul, Republic of Korea
[2]Teddysum, Seoul, Republic of Korea
{j0eun, gyuri1345, shusband, khss}@yonsei.ac.kr,
{chyoon, ybjeong, hahmyg}@teddysum.ai

Abstract

Voice phishing is an evolving form of social engineering crime and requires the continuous advancement of detection technologies. We introduce a benchmark dataset designed to evaluate the practical performance of AI-based voice phishing detection models. The dataset includes diverse voice conversation scenarios and supports four evaluation tasks to assess open-source language models. Experimental results show that while some large-scale models demonstrate stable performance across multiple tasks, accuracy remains low in topic classification and dialogue structure recognition, regardless of model size. These findings highlight the complexity of voice phishing detection, which demands contextual reasoning and dialogue structure understanding beyond simple sentence-level comprehension. The proposed benchmark dataset provides a foundation for more robust evaluation and development of AI systems capable of detecting deceptive voice interactions, contributing to safer and more trustworthy communication environments.

Keywords: voice phishing, AI detection models, safety evaluations, benchmark dataset

1. Introduction

Voice phishing is a criminal act in which people are deceived through cellular communication, often leading to financial exploitation and causing severe social harm on a global scale. In response, governments and companies worldwide are actively exploring the use of artificial intelligence (AI) technologies with the aim of preventing such crimes and maintaining social order. For instance, the American cybersecurity firm Pindrop provides services that detect voice phishing and deepfake audio, enhancing contact center security across various industries. In addition, a European Union research report (October 2025) pointed out that the advancement of generative AI has increased the efficiency of fraudulent schemes, warning of the growing seriousness of scam call crimes and emphasizing the need for both technological and institutional countermeasures.

In South Korea, national agencies such as the Ministry of Science and ICT and Financial Services Commission, have established AI and data-driven memoranda of understanding to support the development of crime prevention technologies, while major telecommunication companies like KT have introduced technical measures and mobile applications to help prevent voice phishing. Despite these collective efforts, voice phishing methods continue to evolve, employing increasingly sophisticated linguistic and psychological deception tactics to mislead victims (European Parliament, 2025).

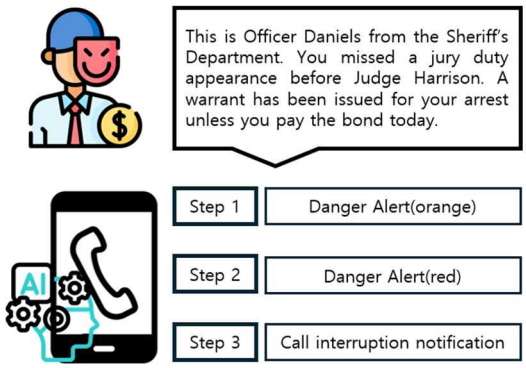

Figure 1: *Example of an AI-based voice phishing detection service (KT 후후(Whowho) app)*

Thus, voice phishing detection and identification technologies must evolve based on continuous monitoring and adaptability to emerging criminal patterns. In response to this issue, the present study aims to construct a benchmark dataset for objectively evaluating the practical effectiveness of voice phishing detection technologies. Designed using publicly available data, this dataset enables multifaceted verification through various task designs, ultimately contributing to assessing whether advances in AI-driven detection systems can lead to tangible reductions in real-world crime and financial loss.

This study is guided by the following research questions (RQs):

† Corresponding author

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7391–7404
11–16 May 2026.

- ✓ RQ1. How can a standardized benchmark dataset be built to evaluate the linguistic and contextual dimensions of voice phishing detection?
- ✓ RQ2. What do the comparative results of open-source AI models across four evaluation tasks reveal about the limitations and future directions of current detection technologies?

These research questions aim to establish a foundation for measuring and enhancing the reliability and applicability of AI-driven voice phishing prevention systems.

2. Related Works

Voice phishing detection models can be classified as domain-specific AI systems designed for specific industries or purposes. Developing such models requires both the construction of high-quality training datasets and the establishment of robust evaluation frameworks to verify detection performance. However, benchmark studies capable of precisely comparing detection capabilities across models remain limited.

The KorCCVi (Korean Call Content Vishing) dataset is a representative study (Boussougou et al., 2021a). KorCCVi is the first Korean-language voice phishing dataset and was created by combining telephone recordings collected from the Financial Supervisory Service's "Voice Phishing Keeper" website with general conversation data from AI Hub, comprising a total of 1,218 dialogues. The study evaluated detection performance using traditional machine learning (e.g., Random Forest, LGBM) and deep learning (e.g., RNN, BiLSTM), with the LGBM model achieving approximately 99% accuracy (Boussougou et al., 2021b).

Subsequently, Gao et al. (2024) developed a Chinese voice phishing dataset (ZhCCVi) based on KorCCVi to compare detection performance across languages, while Boussougou et al. (2022, 2023) applied models such as CNN-BiLSTM, KoBERT, and RoBERTa to examine potential improvements in classification accuracy. In addition, Kim et al. (2021) compared SVM and logistic regression models using datasets from the Financial Supervisory Service and the National Institute of Korean Language, and Lee and Park (2023) demonstrated that applying Doc2Vec embeddings to transcribed voice data yielded more efficient detection than speech-based approaches.

Most existing studies have focused on model training and individual performance enhancement, without sufficiently addressing more multilayered aspects of voice phishing detection such as conversational context, speaker roles, and dialogue structure. To address these limitations, this study constructs the first Korean benchmark dataset for voice phishing detection and designs multiple evaluation tasks to systematically assess the detection capabilities of language models.

3. Voice Phishing Benchmark Dataset Construction

We constructed a reliable benchmark dataset for voice phishing detection, going beyond simple data collection through meticulous qualitative inspection and refinement to ensure high data fidelity and contextual accuracy.

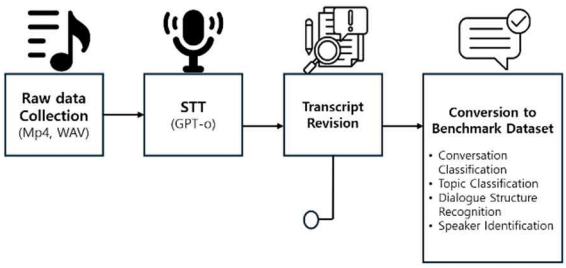

Figure 2: *Process for Construction of the Voice Phishing Benchmark Dataset*

Raw Data Collection The benchmark dataset was built using publicly available voice phishing dialogues and financial consultation data. First, 94 audio files were collected from the Financial Supervisory Service's "Voice Phishing Experience Center" website[2], which provides real-world voice phishing cases accumulated between 2015 and 2023. The data consist of short recordings of less than one minute to longer samples of up to four minutes, primarily in an MP4 format. In addition, a comparative dataset was sourced from AI Hub's Call Center Question–Answer Dataset (comprising a total of 101,501 dialogues)[3]. This dataset covers various domains, including shopping, public health, finance/insurance, and the 다산(Dasan) Contact Center. For this study, 80 dialogues from the finance/insurance domain, which were thematically similar to voice phishing conversations, were selected as general conversation samples.

Speech to Text The collected voice phishing audio files were converted into text using OpenAI Whisper and GPT-4o models. Whisper was first used for initial transcription, and the results were compared with GPT-4o's STT outputs to

[2] https://www.fss.or.kr/fss/bbs/B0000203/list.do?menuNo=200686

[3] https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=98

Task		Evaluation Unit	Prompt Type	Evaluation Method
Conversation Classification	Binary Classification of Voice Phishing Conversations and General Conversations	Dialogue	Closed-ended question	Accuracy
Topic Classification	Classification of Conversation Topic(s)	Dialogue	Closed-ended question	Accuracy
Dialogue Structure Recognition	Summary and Segmentation of Conversation Content	Dialogue +Sentence (Utterance)	Open-ended question	Accuracy, F1-Score
Speaker Identification	Classification of Speech into Speaker Roles: "Counsellor, Client, Other"	Sentence (Utterance)	Closed-ended question	Accuracy

Table 1. Task Description and Evaluation Method

qualitatively assess transcription accuracy and linguistic naturalness.

Transcription Revision Based on qualitative evaluation, the GPT-4o transcriptions were selected and further refined. During this stage, personally identifiable information (PII) such as names, addresses, bank names, and institution names that were not anonymized were annotated and masked. The conversational structure was also revised to clearly distinguish among speakers and to ensure logical flow throughout each dialogue.

Conversion to Benchmark Dataset After refinement, a total of 239 dialogues were constructed, of which 80 dialogues involving institutional impersonation and containing more than seven conversational turns were selected as the final evaluation samples. In combination with the 80 general dialogues selected from the finance/insurance domain of AI Hub's dataset, the final benchmark dataset comprised 160 dialogues in total.

The dataset was designed not only for simple classification tasks but also for comprehensive evaluation of the models' ability to understand both the global context and local utterance structures of voice phishing conversations. To this end, four evaluation tasks were defined as follows:

Conversation Classification (T1) determines whether an entire dialogue constitutes a voice phishing conversation or a normal one.

Topic Classification (T2) categorizes each dialogue into one of four thematic domains based on its overall context, assessing the model's understanding of topical intent.

Dialogue Structure Recognition (T3) summarizes and segments dialogues into semantic units, evaluating the model's comprehension of conversation flow and structural relationships.

Speaker Identification (T4) aims to classify the role of each speaker using only utterance-level textual information. Speaker roles are defined as *Counsellor* and *Client* so that both general counseling conversations and voice phishing conversations can be consistently represented.

These four tasks form a multi-layered and multilabel framework, in which each dialogue instance can be evaluated at the dialogue, sentence, and utterance levels to capture both macro-level context and micro-level patterns.

As a result, the proposed Korean Voice Phishing Benchmark Dataset serves as a foundation for evaluating multidimensional language understanding and provides an experimental basis for advancing AI-based voice phishing detection systems.

4. Experimental Settings

The objective of this study was to design a benchmark dataset for evaluating the performance of voice phishing detection models and to verify its practical applicability.

When the detection capability of language models is structured around classification tasks, it is essential to clearly define the evaluation units and methods for each task.

This study conducted classification-based evaluation experiments using a benchmark dataset consisting of 160 dialogues. We analysed the differences in task-level difficulty and accuracy arising from distinct evaluation units and examined variations in model responses based on prompt types (open-ended vs. closed-ended).

4.1 Task Design

The benchmark dataset was designed not only to measure classification accuracy but also to evaluate language understanding and the contextual reasoning capabilities of models.

T1 This task determines whether an entire dialogue constitutes a voice phishing conversation or not ("Yes" / "No"). Prompts were designed as closed-ended questions (e.g. "Is the following conversation a case of voice phishing scam?")

Model	T1: Acc	T2: Acc	T3: Boundary F1	T3: B-Cubed F1	T3: Similarity Accuracy	T4: Acc
Bllossom_253b (reasoning off)	83.12	50	34.99	72.99	69.17	87.12
Bllossom_3.1_70b	81.88	47.5	28.67	71.37	70.59	85.45
Bllossom_3_8b	53.12	46.25	19.14	58.86	44.83	40.47
Bllossom_3.2_3b	58.13	33.75	15.07	51.23	71.79	39.84
Gemma-3-27b-instruct	90.62	52.5	2.53	1.25	69.17	84.05
Gemma-3-12b-instruct	83.75	48.75	1.28	0.62	70.75	66.72
Gemma-3-4b-instruct	51.88	46.25	0	0	74.47	30.3
Gemma-3-1b-instruct	38.12	26.25	0	0	69.23	55.05
Gemma-3-270m-instruct	49.38	25	0	0	0	39.86
Llama_3.1_8B_Instruct	50	48.75	3.12	3.12	61.4	40.09
Llama_3.2_3B_Instruct	44.38	43.75	0	0	63.83	40.42
Qwen2.5-72B-Instruct	97.5	51.25	5.56	3.75	65.69	89.24
Qwen2.5-7B-Instruct	83.12	51.25	1.31	0.62	60.61	58.44
Qwen3-4B-Instruct-2507	56.25	52.5	0.44	0.62	71.23	68.74

Table 2. Experimental Results

and included few-shot examples to ensure stable model output. The few-shot setting was implemented as a one-shot example, which served as a mechanism to maintain consistency across the outputs of different models.

T2 Due to the limitation of public datasets, which contain only institutional impersonation-type phishing cases, this task was applied to general dialogues as well. The general conversations were categorized into three topics— (1) accident and compensation inquiries, (2) product subscription and cancellation, and (3) transfer, withdrawal, and loan services—and were collected from the aforementioned AI-Hub's dataset's finance and insurance domains, which are thematically similar to voice phishing conversations.

T3 This task required models to summarize a dialogue and then segment it into meaningful units based on that summary, thereby evaluating their understanding of dialogue structure and flow. It was formulated as an open-ended question task, in which the model's free-form summary and segmentation results were compared with gold standards.

T4 This task identifies the roles of individual speakers, either a counsellor (e.g., scammer) and or client (e.g., victim) based on each utterance within the dialogue.

4.2 Model and Evaluation Metrics

To ensure reproducibility and reliability, representative open-source large language models were selected for the experiments. Specifically, models from the Blossom, Gemma, LLaMA, and Qwen families were included, allowing for comparison across various model architectures and parameter scales.

Accuracy was used as the primary evaluation metric for most tasks, while Boundary F1 and B-Cubed F1 were employed for the Dialogue Structure Recognition task to assess segmentation quality. Specifically, Boundary F1 measures how accurately the model identifies segment boundaries within a dialogue, whereas B-Cubed F1 evaluates the degree of overlap between the model-generated and reference segments, reflecting how well the segmented content aligns with the gold standard.

All experiments were conducted exclusively with open-source models, enabling replication under identical conditions, thereby ensuring the objectivity and applicability of the experimental results.

5. Results and Analysis

5.1 Detailed Accuracy Result Across Tasks and Models

In T1, Qwen2.5-72B-Instruct achieved the highest level of accuracy with 97.5%, followed by Gemma-3-27B-Instruct (90.6%) and Bllossom_253b (83.1%). In contrast, smaller models such as Gemma-3-1B-Instruct (38.1%) and Gemma-3-270M-Instruct (49.4%) exhibited relatively low performance.

In T2, Qwen2.5-72B-Instruct again recorded the highest accuracy of 51.25%, while bllossom_253b (50%) and Qwen3-4B-Instruct (52.5%) exhibited similar levels of performance. However, since

most models achieved accuracies in the 30–40% range, it was confirmed that topic classification is a highly challenging task.

T3 was evaluated using Boundary F1, B-Cubed F1, and Similarity Accuracy. Experimental results revealed generally low performance across models, with many achieving only 0–5% in Boundary and B-Cubed F1 scores. However, bllossom_253b performed relatively well, with Boundary F1 = 34.99 and B-Cubed F1 = 72.99, suggesting that larger models may possess relative strengths in dialogue summarization and structural segmentation.

In T4, Qwen2.5-72B-Instruct achieved the highest accuracy of 89.24%, followed by bllossom_253b (87.12%) and Gemma-3-27B-Instruct (84.05%), with all of them performing in a stable manner. In contrast, smaller models recorded notably lower results across all subtasks.

5.2 Quantitative Analysis of Model Size and Task Difficulty

The experimental results depict a trend in which performance improves with model size. Qwen2.5-72B-Instruct and bllossom_253b demonstrated outstanding performance across most tasks, indicating that larger models are relatively better suited for voice phishing detection.

However, except for T1, no model achieved an accuracy above 90%, even among large-scale LLMs, indicating that while model size contributes significantly to performance improvement, voice phishing detection remains a high-difficulty task due to its complex contextual nature.

In T2 and T3, overall performance remained low regardless of model size, likely because voice phishing categories are not standardized, and the definition of gold-standard answers for dialogue structure recognition remains ambiguous. Additionally, several small models performed poorly even in T1, suggesting that voice phishing detection requires not only sentence-level understanding but also contextual reasoning and pattern recognition across utterances.

In summary, these findings indicate that model performance in voice phishing detection is influenced not only by model scale but also by the characteristics of training data and the definition of task objectives.

5.3 Error Analysis of Dialogue Structure Recognition (T3)

To identify the causes of low performance in T3, qualitative analysis was conducted on model-generated outputs from the open-ended evaluation format. For dialogue summarization, models tended to interpret specific utterances through keyword-based heuristics, often generating summaries that diverged from the actual conversational context.

For example, in dialogues discussing the verification of a suspect's involvement in a fraud case, the model instead summarized it as "an explanation of account closure and illicit fund amount," focusing on surface-level terms rather than actual meaning. In some cases, the model overinterpreted utterances—for instance, paraphrasing "I will help you receive a non-indictment decision" as "a promise to assist in proving victimhood and securing a non-indictment ruling."

Models also exhibited a tendency to compress dialogues into a single event, ignoring temporal or causal flow. Unlike human annotators who segmented dialogues into three or more meaningful stages, models frequently produced flat or overly condensed structures. For segmentation, overlapping and redundant divisions were also observed.

For instance, "details of the scam" and "explanation of how a fake account was used" were treated as separate segments, despite the latter being a subset of the former. Models also failed to distinguish between core and auxiliary utterances, segmenting filler phrases such as "Please cooperate actively" or discourse markers like "and" and "then" as independent segments.

The models failed to capture the finer flow and thematic transitions across utterances, often simplifying dialogues into single-topic narratives or generating excessive segmentation.

Consequently, whereas human annotators summarized each dialogue into three or four semantic units, distinguishing between core and peripheral utterances, models either collapsed dialogues into one topic or added unnecessary splits, thus failing to reflect the full structure and meaning of conversations.

5.4 Comparison of Human and Model Topic Classification (T2)

All voice phishing dialogues in our dataset are annotated as institutional impersonation. Unlike the general conversation dataset, which contains multiple topics, the voice phishing data represent a single category, making direct type classification between the two datasets difficult. To address this limitation, we conducted topic classification for general conversations using closed-ended question prompts. For voice phishing dialogues, we employed open-ended prompts and analysed the differences between the generated responses and the institutional impersonation category.

The analysis showed that human evaluators consistently classified most cases as impersonation of public institutions, aligning with the existing annotation scheme. This reflects a decision grounded in legal and administrative classification frameworks, ensuring clear criteria and annotation consistency. Such explicit annotation standards facilitate quantitative evaluation and improve the reproducibility of the results.

In contrast, the model performed context-based classification without relying on the predefined category of “impersonation.” Instead, it combined institutional entities (e.g., Prosecutor’s Office, the Financial Supervisory Service, banks, etc.) with fraud tactics (e.g., account freezing, identity theft, loan inducement, etc.) to infer more granular contextual patterns.

While this demonstrates strong contextual interpretation, it also risks reducing evaluation consistency, as models may apply differing contextual logic to the same input. Hence, the model’s detailed contextual classification cannot be regarded as inherently superior. From an evaluation design perspective, the more complex the annotation scheme becomes, the more finely the gold-standard criteria must be defined—leading to inevitable discrepancies between model outputs and human references, and consequently, lower accuracy scores.

Thus, context-sensitive classification is not always appropriate for evaluation purposes, where measurability and objectivity must take precedence.

6. Conclusions

We constructed a benchmark dataset for voice phishing detection and designed an evaluation framework to analyse the performance of various open-source language models. Across four evaluation tasks, the results showed a general trend of improved performance with increasing model size, with Qwen2.5-72B-Instruct and bllossom_253b demonstrating relatively superior results across most tasks.

However, Tasks 2 and 3 yielded consistently low accuracy scores regardless of model type or scale, indicating that these tasks require not only sentence-level understanding but also contextual reasoning and discourse structure recognition, which are inherently more challenging.

The qualitative analysis revealed that while the models exhibited more fine-grained, context-based classification than human annotators, their classification criteria lacked consistency, thereby reducing the overall evaluation stability. In particular, the error analysis of Task 3 showed that models often failed to capture the temporal flow and causal relationships of dialogues. Furthermore, models also showed a tendency to over-compress or redundantly segment units of meaning and thus did not fully reflect the structural complexity of conversations.

These findings confirm that voice phishing detection is a linguistically and cognitively complex task. Future research should move beyond isolated single-task experiments and focus on building scenario-based sequential tasks that more closely resemble real-world voice phishing interactions. Additionally, the refinement of phishing-type taxonomies, the clarification of gold-standard annotation criteria, and the construction of an expanded dataset encompassing both phishing and general dialogues are essential next steps.

We contribute to the first stage of benchmarking and comparative evaluation for voice phishing detection models. Moving forward, establishing a more realistic and scalable dataset and a scenario-driven evaluation framework will be crucial for verifying model performance in real-world contexts.

7. Limitations

This study is meaningful in that it systematically evaluated AI technologies for voice phishing detection by building a benchmark dataset and testing open-source language models.

However, several limitations remain. First, from the perspective of dataset design, the study did not fully capture the multimodal nature of voice phishing. Because the current dataset is text-based, it does not include non-textual cues that play a crucial role in real phishing interactions—such as vocal features (intonation, emotion, tone) or visual information (documents, links, transfer screenshots, etc.). Moreover, as with previous studies, the benchmark dataset was developed using publicly available data, which inherently limits the data scale and diversity and introduces potential data contamination or learnability issues for large language models.

Future work should focus on multimodal extensions that integrate audio and visual elements, as well as scenario-based data augmentation and content variation across different crime types to enhance realism and coverage.

Second, there exists a gap between the analytical evaluation framework used in this study and real-world service environments. Our evaluation focused on post-hoc performance comparison and did not incorporate real-time detection or continuous monitoring capabilities that are essential for operational systems.

Consequently, future work should extend the current evaluation framework toward real-time detection by developing scenario-driven evaluation tasks that simulate real conversations and user interactions, allowing the assessment of both detection speed and practical deployment feasibility in realistic service environments.

Future research should therefore develop scenario-driven evaluation tasks that simulate real conversations and user interactions to assess the feasibility and applicability of model deployment in realistic service conditions.

Finally, as this study represents an initial exploration into benchmark design for voice

phishing detection, the qualitative analysis remains at a case-study level. Future work should therefore adopt pragmatic and discourse-oriented approaches to better understand complex fraud strategies and evaluate models for real-world crime prevention. The datasets constructed in this study will be released in stages after completing modality-specific datasets (text, audio, and image), which is expected to improve reproducibility and comparability in voice phishing detection research and support diverse model development.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2025-02215142, Development of Pseudonymization Technology for Suspected Criminal Information).

References

Boussougou, MKM, & Park, DJ (2021). A real-time efficient detection technique of voice phishing with AI. *Proceedings of the Korean Information Science Society*, 768-770.

Boussougou, M. K. M., Jin, S., Chang, D., & Park, D. J. (2021). Korean voice phishing text classification performance analysis using machine learning techniques. In Annual Conference of KIPS (p. 297-299). Korea Information Processing Society.

Gao, P., Zheng, J., Shuai, C., & Zhang, L. (2024). A Hierarchical Dual-Role Interaction Network for Telephone Conversation Fraud Detection. IEEE Access.

Boussougou, MKM, Park, MG, & Park, DJ (2022). An Attention-Based CNN-BiLSTM Model for Korean Voice Phishing Detection. Proceedings of the Korean Information Science Society, 1139-1141.

Moussavou Boussougou, M. K., & Park, D. J. (2023). Attention-based 1D CNN-BiLSTM hybrid model enhanced with fasttext word embedding for Korean voice phishing detection. Mathematics, 11(14), 3217.

Boussougou, M. K. M., & Park, D. J. (2022). Exploiting Korean Language Model to Improve Korean Voice Phishing Detection. KIPS Transactions on Software and Data Engineering, 11(10), 437-446.

Boussougou, MKM, Hong, CK, Hong, S., & Park, DJ (2023). Towards Privacy-Preserving Korean Voice Phishing Detection: A Federated Learning Approach with RoBERTA. Proceedings of the Korean Information Science Society, pp. 626-628.

Kim Eun-jeong. (2022). Analysis of the criminal process of face-to-face voice phishing: Focusing on crime script analysis. Journal of Criminal Investigation, 31-54.

Kim, J. W., Hong, G. W., & Chang, H. (2021). Voice recognition and document classification-based data analysis for voice phishing detection. Human-centric Computing and Information Sciences, 11.

Lee, M., & Park, E. (2023). Real-time Korean voice phishing detection based on machine learning approaches. Journal of Ambient Intelligence and Humanized Computing, 14(7), 8173-8184.

Lee, Y., & Han, D. (2024, November). KorSmishing explainer: A Korean-centric LLM-based framework for smishing detection and explanation generation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track (pp. 642-656).

Shin, Seong-won. (2022). A study on the current state of voice phishing and countermeasures. Korean Journal of Public Security Administration, 19 (4), 165-185.

Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., ... & Huang, M. (2023). Safetybench: Evaluating the safety of large language models. arXiv preprint arXiv:2309.07045.

Appendix

A. Benchmark Task Design

The benchmark dataset constructed in this study consists of four tasks designed to evaluate not only classification accuracy but also dialogue understanding and contextual reasoning across different linguistic levels.

Each task targets a different unit of analysis. Conversation Classification operates at the conversation level, determining whether a dialogue constitutes a case of voice phishing. Topic Classification requires models to infer the main topic based on conversational and sentence-level semantic information. Dialogue Structure Recognition focuses on identifying the structural flow and meaningful segments of a conversation at the dialogue and sentence levels. Speaker Identification operates at the utterance level, classifying the role of each speaker.

This multi-level task design enables analysis of how models understand overall conversational context, semantic cues, and interaction structures. Among the tasks, Topic Classification and Dialogue Structure Recognition require relatively higher levels of language understanding and contextual reasoning, as they involve interpreting both the semantic context and structural flow of conversations.

1) Conversation Classification (T1)

The Conversation Classification task is formulated as a binary classification problem that determines whether a given dialogue is a voice phishing conversation or a normal conversation.

Voice phishing dialogues were constructed from audio recordings released by the Voice Phishing Experience Center. The recordings were first transcribed using a language-model-based STT system and then manually reviewed to correct transcription errors. The resulting transcripts were used as the voice phishing dialogue dataset. General conversation data were collected from financial and insurance customer service dialogues.

This task evaluates whether models can identify voice phishing conversations by considering the overall context of the dialogue.

2) Topic Classification (T2)

The Topic Classification task evaluates a model's ability to infer the main topic of a conversation.

Because publicly available voice phishing datasets are predominantly annotated under a single category—institutional impersonation—it is difficult to construct a multi-topic classification task using only phishing dialogues. To address this limitation, the topic classification task was primarily designed using general conversation data.

General conversations were categorized into the following finance-related topics:

- accident and compensation inquiries
- product subscription and cancellation
- balance and transaction history inquiries
- transfer, withdrawal, and loan services

In the closed-ended setting, models were asked to select the correct topic from the predefined options, enabling quantitative evaluation through multiple-choice classification.

In the open-ended setting, voice phishing dialogues were used, and models were asked to freely generate the inferred topic. The responses generated were then compared with the annotated category of institutional impersonation. Although this setup is not a strict classification benchmark, it provides useful insights into how models interpret and describe both general and phishing conversations.

3) Dialogue Structure Recognition (T3)

The Dialogue Structure Recognition task evaluates whether a model can understand the structural organization and semantic flow of a conversation.

Human annotators first analysed each dialogue and constructed a golden standard by summarizing the key content and segmenting the dialogue into meaningful units. Models were then asked to perform the following two tasks:

- segment the conversation into meaningful units
- summarize the dialogue content

The summaries generated and segmentation results were qualitatively compared with the human-annotated golden standard. All prompts in this task were designed as open-ended questions.

Unlike simple classification tasks, this task requires models to understand conversational flow, speech act progression, and contextual transitions. It therefore provides insights into how models interpret structural characteristics and contextual relationships within dialogues.

4) Speaker Identification (T4)

The Speaker Identification task classifies the role of each speaker at the utterance level.

Each utterance is labeled as either Counsellor (or perpetrator) or Client (or victim). These role categories were unified as *Counsellor* and *Client* so that the task can be consistently applied to both general service conversations and voice phishing dialogues.

This task evaluates whether models can infer speaker roles using linguistic cues present in individual utterances.

B. Annotation Process

The benchmark dataset constructed in this study was annotated by eight graduate students in computational linguistics. All annotators are native speakers of Korean and have had prior experience in both natural language processing and dialogue analysis. All annotators are native speakers of Korean and have had prior experience in both natural language processing and dialogue analysis. Before participating in the dataset construction, they received training on the annotation guidelines and procedures for each task and conducted sample annotation exercises followed by review sessions. The annotators were compensated through research support for their participation in the project.

The annotation process varies depending on the characteristics of each task.

1) Conversation Classification (T1)

The Conversation Classification task did not require additional manual annotation because the ground-truth labels were determined by the data sources. Voice phishing dialogues were constructed from audio materials released by the Voice Phishing Experience Center, while general conversations were obtained from financial and insurance dialogue datasets provided by AI Hub. Accordingly, dialogues were labeled as either voice phishing or general conversation based on their source.

2) Topic Classification (T2)

The Topic Classification task also utilized existing annotations from the original datasets without additional manual labeling. Voice phishing dialogues were labeled under the category of institutional impersonation, which is the primary annotation type provided in the public dataset. General conversations used topic annotations from financial service dialogues, including categories such as accident and compensation inquiries, product subscription and cancellation, and transfer, withdrawal, and loan services.

3) Dialogue Structure Recognition(T3)

The Dialogue Structure Recognition task involved manual analysis of dialogue structure. Based on previous studies on voice phishing dialogue scripts (Lee et al., 2024; Shin, 2022; Kim, 2022), the conversation structure was categorized into three stages:

- Approach
- Information Request
- Information Collection

Annotators segmented each dialogue into meaningful units and assigned each segment to one of the three stages. They also produced a brief descriptive summary for each segment to capture the key content of the interaction.

4) Speaker Identification

For the Speaker Identification task, speech recordings were first transcribed using an STT system. Annotators then reviewed the transcripts while referring to the original audio recordings to determine speaker boundaries and identify the speaker for each utterance.

Quality Control

Some tasks in the benchmark did not require additional manual annotation because the labels were already determined by the data sources or existing public dataset annotations. For example, in Conversation Classification, labels were assigned based on the source of the dialogue, and Topic Classification relied on pre-existing topic annotations. In these cases, inter-annotator agreement (IAA) was not calculated.

For tasks involving human interpretation, such as Dialogue Structure Recognition and Speaker Identification, a cross-checking procedure was applied to ensure data quality. Annotators reviewed each other's results and verified segmentation criteria and speaker assignments. When disagreements occurred, the final labels were determined through discussion and consensus.

C. Prompt Design for Experiments

To ensure consistent evaluation across different language models, the same system prompt was used for all experiments. The system prompt was standardized as follows:

- *You are a helpful AI assistant. Please answer the user's questions kindly.*

Each task used either a closed-ended or open-ended prompt depending on the evaluation objective.

1) Conversation Classification

The Conversation Classification task was conducted using a closed-ended prompt. Models were given dialogue and asked to determine whether the conversation corresponds to a voice phishing case. The model was required to return one of the predefined answer options.

[Task]

당신의 임무는 주어진 대화가 보이스피싱인지 아닌지를 [Options]에서 선택하여, **선택한 옵션의 값을 수정 없이 그대로, 오직 하나만 반환**하는 것입니다.

절대로 다른 설명, 괄호, 번호 등 부가적인 텍스트를 포함해서는 안 됩니다.

*Your task is to determine whether a given conversation is a voice phishing scam by selecting [Options] and **returning only one value for the selected option, without modification**. It must **never** include any additional text, such as other descriptions, parentheses, or numbers.*

[Example]

- Input:

[Question] 다음 대화가 보이스피싱인가요?

A: 서울중앙지검 첨단범죄수사팀입니다. 김민준 씨 명의로 대포통장이 개설되었습니다.

B: 네? 저는 그런 적 없는데요.

Is the following conversation a voice phishing scam?

A: This is the Advanced Crime Investigation Team of the Seoul Central District Prosecutors' Office. A fake bank account has been opened in the name of Kim Min-jun.

B: Huh? I've never done anything like that.

[Options] (A) Yes (B) No

- Output: Yes

[Question]

다음 대화가 보이스피싱인가요? *Is the following conversation a voice phishing scam?*

{대화}

[Options]

(A) Yes

(B) No

[Output Form]

[Options]의 선택지 중 하나를 선택하여 단어만 반환하세요. (예: Yes)

Select one of the options in [Options] to return only words (e.g., Yes).

2) Topic Classification

The Topic Classification task also used a closed-ended prompt. Models were asked to read the dialogue and select the main topic of the conversation from the provided options.

[Task]

당신의 임무는 주어진 대화의 주제를 [Options]에서 선택하여, **선택한 옵션의 값을 수정 없이 그대로, 오직 하나만 반환**하는 것입니다.

절대로 다른 설명, 괄호, 번호 등 부가적인 텍스트를 포함해서는 안 됩니다.

*Your task is to select a topic for a given conversation from [Options] and **return only one value for that selected option, without modification**. It should **never** include any additional text, such as other descriptions, parentheses, or numbers.*

[Example]

- Input:

  [Question] 다음 대화의 주제로 가장 적절한 것을 고르세요.

  A: 예탁금 통장 거래내역 조회해 주세요

  B: 본인확인후 안내드리겠습니다.

  [Options] (A) 사고 및 보상 문의 (B) 상품 가입 및 해지 (C) 잔고 및 거래내역 (D) 이체, 출금, 대출서비스

  [Question] Choose the most appropriate topic for the following conversation.

  A: Please check your deposit account transaction history.

  B: I will provide guidance after verifying your identity.

  [Options] (A) Accident and compensation inquiries (B) Product subscription and cancellation (C) Balance and transaction history (D) Transfer, withdrawal, and loan services

- Output: 잔고 및 거래내역 *Balance and transaction history*

[Question]

다음 대화의 주제로 가장 적절한 것을 고르세요.

{대화}

Choose the most appropriate topic for the following conversation.

{Conversation}

[Options]

(A) 사고 및 보상 문의 *Accident and compensation inquiries*

(B) 상품 가입 및 해지 *Product subscription and cancellation*

(C) 잔고 및 거래내역 *Balance and transaction history*

(D) 이체, 출금, 대출서비스 *Transfer, withdrawal, and loan services*

[Output Form]

[Options]의 선택지 중 하나를 선택하여 단어만 반환하세요. (예: 잔고 및 거래내역)

Select one of the options in [Options] to return only words (e.g., Balance and Transaction History)

3) Dialogue Structure Recognition

The Dialogue Structure Recognition task used an open-ended prompt. Models were asked to read the dialogue, summarize the conversation in several stages, and segment the dialogue according to its structural flow.

[Task]

분류 작업을 수행하려 합니다. Question(질문)에 대해 답변하세요.

대화의 흐름을 반드시 3 개에서 5 개 사이의 구간으로 나누고, 각 구간의 핵심 내용을 요약하여 아래 [Output Form]에 명시된 JSON 배열 형식으로만 반환해야 합니다.

- 각 구간의 요약('value')은 10 자에서 30 자 사이로 작성하세요.

- 당신의 답변은 반드시 '['로 시작하고 ']'로 끝나야 합니다. 다른 어떤 설명도 포함하지 마세요.

You are attempting to perform a classification task. Please answer the Question.

The conversation flow must be divided into 3 to 5 segments, and the key points of each segment must be summarized and returned only in the JSON array format specified in the [Output Form] below.

- The summary ('value') of each segment must be between 10 and 30 characters.

- Your answer must begin with '[' and end with ']'. Do not include any other explanation.

[Example]

- Input:

speaker_id 1: 여보세요?

speaker_id 2: 네, 서울중앙지검 첨단범죄수사팀의 김민준 수사관입니다. 본인 명의로 대포통장이 개설되어 연락드렸습니다.

speaker_id 3: 네? 저는 그런 적 없는데요.

speaker_id 4: 본인도 모르게 명의가 도용된 것 같습니다. 피해 사실을 접수하고 계좌를 보호해야 합니다.

speaker_id 1: Hello?

speaker_id 2: Yes, this is investigator Kim Min-jun from the Seoul Central District Prosecutors' Office's High-Tech Crime Investigation Team. I'm contacting you because a fake bank account has been opened in my name.

speaker_id 3: Huh? I've never done anything like that.

speaker_id 4: It seems my identity has been stolen without my knowledge. You should report the damage and protect your account.

- Output:

```
[
 {
   "target": [1, 2],
   "value": "신원 확인 및 연락 목적 설명" Identification and purpose of contact
 },
 {
   "target": [3, 4],
```

"value": "피해 사실 부인 및 명의 도용 가능성 제기" *Denial of damage and raising the possibility of identity theft*

}

]

[Question]

다음 대화를 읽고 3-5 단계로 요약하며 각 구간을 어떻게 구분할 수 있나요?

{대화순서번호}

{대화}

Read the following conversation and summarize it in 3-5 steps. How would you differentiate between each section?

{Conversation Number}

{Conversation}

[Output Form]

[

{

"target": [1, 2, 3, ...],

"value": "첫 번째 구간 요약" *Summary of the first section*

},

{

"target": [10, 11, 12, ...],

"value": "두 번째 구간 요약" *Second section summary*

}

]

This task was designed to analyse how models understand the semantic progression and structural organization of dialogues. The output generated was evaluated qualitatively by comparing them with the dialogue structure analysis created by human annotators.

4) Speaker Identification

The Speaker Identification task used a closed-ended prompt. Models were given a dialogue and asked to identify whether each utterance was spoken by the Counsellor or the client based on the provided options.

[Task]

당신의 임무는 주어진 [Options]에서 가장 적절한 역할 하나를 선택하여, **선택한 옵션의 값을 수정 없이 그대로, 오직 하나만 반환**하는 것입니다.

절대로 다른 설명, 괄호, 번호 등 부가적인 텍스트를 포함해서는 안 됩니다.

*Your task is to select the most appropriate role from the given [Options] and **return only one value for that option, without modification**. It should **never** include any additional text, such as descriptions, parentheses, or numbers.*

[Example]

- Input:

[Question] 'speaker_id 2'의 역할은 무엇입니까?

speaker_id 1: 안녕하세요, 고객님.

speaker_id 2: 네, 안녕하세요. 대출 문의드립니다.

[Options] (A) 상담사 (B) 내담자 (C) 기타

- Output: 내담자

[Question] What is the role of 'speaker_id 2'?

speaker_id 1: Hello, customer.

speaker_id 2: Yes, hello. I would like to inquire about a loan.

[Options] (A) Counsellor (B) Client (C) Other

- Output: Client

[Role Definition]

- 상담사: 대화를 주도하거나 정보를 제공하는 주체 (예: 검찰, 금융기관 직원, 상담원).

- 내담자: 정보를 받거나 요청하는 대상 (예: 일반 시민, 고객).

- 기타: 명확한 제 3 자, 자동응답시스템 등 위 두 역할에 해당하지 않는 경우에만 선택.

- Counsellor: The entity leading the conversation or providing information (e.g., prosecutor, financial institution employee, Counsellor).

- Client: The entity receiving or requesting information (e.g., general public, customer).

- Other: Select only if the role does not fall into the above two categories, such as a clear third party or automated response system.

[Question]

'speaker_id N'의 역할은 무엇입니까?

{대화}

What is the role of 'speaker_id N'?

{Conversation}

[Options]

(A) 상담사 *Counsellor*

(B) 고객 *Client*

(C) 기타 *Other*

[Output Form]

[Options]의 선택지 중 하나를 선택하여 단어만 반환하세요. (예: 내담자)

Select one of the options in [Options] to return only words (e.g., Client).

D. Computational Resources

All experiments were conducted on a single server equipped with eight NVIDIA A100-SXM4-80GB GPUs. Depending on the model size, inference was performed using 1–4 GPUs.

All evaluated models were used in their pretrained form, and no additional fine-tuning or training was conducted. Bllossom_253b was evaluated using int8 quantization, while all other models were evaluated with bfloat16 precision.

The benchmark was executed through a unified evaluation pipeline consisting of four tasks. All tasks were performed via inference only. The total computational cost of the experiments amounts to approximately 422.7 GPU-hours