

PhishLLM: Generating Japanese Phishing Emails to Test Cross-Lingual Generalization in English-Trained Models

René Meléndez^{1,*}, Michal Ptaszynski^{1,*} and Fumito Masui¹

¹*Kitami Institute of Technology, 165 Koen-cho, Kitami, Hokkaido 090-8507, Japan*

Abstract

Large Language Models (LLMs) make it easy to generate convincing phishing emails, but most research on their detection focuses on English. This leaves users of other languages, such as Japanese, more vulnerable to targeted attacks. This study first demonstrates how LLMs can be prompted to produce realistic Japanese phishing emails, creating a new dataset for testing security defenses. We then use this dataset to evaluate the performance of both multilingual and Japanese-specific transformer models on a phishing detection task. The results show that models trained specifically for Japanese achieve the highest accuracy. In contrast, multilingual models perform less effectively, showing limitations in handling Japanese text. Our findings reveal two critical points: how easily LLMs can be misused to create non-English malicious content, and the urgent need for better security models designed for languages that do not use Roman scripts.

Keywords

Phishing detection; Large Language Models; Japanese emails; multilingual transformers; text generation; cybersecurity; cross-lingual evaluation; LLM misuse prevention.

1. Introduction

The rapid development of Large Language Models (LLMs) and its commercialization has made advanced AI tools widely accessible. These models can generate and manipulate text with a fluency comparable to native speakers, performing complex tasks like summarization, code generation, and even imitating human reasoning with minimal instruction [1]. Initially, the high resource requirements for training and running these models limited their use to specialized research environments. Today, however, many powerful LLMs are available as open-source tools or through simple APIs or friendly user interfaces, making them accessible to individuals with little technical expertise.

While this accessibility has driven innovation in fields like education and software development, it has also introduced new and significant security risks. A major concern is the misuse of LLMs for social engineering attacks, particularly phishing emails, a method where attackers generate highly realistic and personalized messages to impersonate trusted organizations and deceive users. Attackers can now use these models to generate highly realistic and personalized phishing emails in multiple languages with very little effort. These AI-generated emails can successfully impersonate trusted organizations, adopt specific tones, and are often free of the grammatical errors that once made phishing attempts easier to spot, allowing them to bypass traditional detection systems.

The challenge is made worse by a significant gap in current research. Most studies on AI-powered phishing generation and detection have focused almost exclusively on English. This leaves a large portion of global internet users who communicate in other languages less protected. The effectiveness of security models trained on English data cannot be assumed to transfer directly to other languages, especially those with different linguistic structures and cultural contexts, such as Japanese.

This study addresses this gap by focusing on both the generation and detection of phishing emails in Japanese. Our work has three main goals. Firstly, we aim to **(1) demonstrate and evaluate the capability of LLMs to generate realistic and linguistically coherent phishing emails in Japanese**, using a small set of authentic phishing samples as a starting point. Secondly, we **(2) econdly, we (2) investigate how a specific fine-tuning strategy, specifically Low-Rank Adaptation (LoRA)**,

LaCATODA 2026: The 10th Linguistic and Cognitive Approaches to Dialog Agents Workshop at the 40th AAAI conference

ORCID 0009-0004-2129-8747 (R. Meléndez); 0000-0002-1910-9183 (M. Ptaszynski); 0000-0001-9979-8734 (F. Masui)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

can be used to control the quality and threat level of the generated emails. This approach highlights the ease with which models can be adapted for malicious purposes by users with limited technical expertise. , highlighting the ease with which these models can be adapted for malicious purposes. Finally, we (3) **assess the cross-lingual generalization performance of several multilingual transformer models**, comparing them against Japanese-specific models in a **phishing classification task**. By doing so, we aim to measure the current state of non-English phishing detection and identify key areas for improvement.

The remainder of this paper is structured as follows. Section 2 reviews previous research on LLM-based phishing generation and detection, highlighting the existing focus on English. In Section 3, we detail our methodology, first by explaining the process of generating the Japanese phishing email dataset, and then by describing the experimental setup for the detection task. In Section 3.2 we describe the evaluation setup for the experiment with Japanese emails. Section 4 presents the performance of the evaluated models and includes a detailed error analysis. Finally, Section 5 concludes the paper by summarizing our findings, followed by Section 6 discussing the study’s limitations, and suggesting directions for future work.

2. Related Work

The integration of Large Language Models (LLMs) has drastically changed the cybersecurity landscape, affecting both the generation of threats and the development of defensive measures. This section reviews prior work in AI-driven phishing generation and detection, highlighting the research gap in non-English languages that this study aims to address.

2.1. LLMs for Phishing Generation

Traditionally, many phishing attempts were easy to identify due to poor grammar, spelling mistakes, and a lack of personalization. However, the wide availability of powerful LLMs has removed these barriers. Malicious actors can now use AI to generate convincing phishing emails at a massive scale. A report from [2] indicates a significant increase in the use of AI for creating phishing emails, with such methods accounting for a large percentage of recent attacks.

Several studies have demonstrated the ease with which commercially available models can be used for this purpose. For instance, research shows that models like GPT-3.5 and GPT-4 can generate highly realistic spear-phishing emails with minimal cost and effort, often requiring only basic prompt engineering [3]. This trivializes the technical skills once needed to craft sophisticated attacks. Further supporting this, Qi et al. [4] introduced a framework for generating spear-phishing emails and, by using jailbreaking and prompt engineering techniques, created over 1,000 deceptive emails across ten different attack types. Human evaluation in their study confirmed that the generated emails were both highly readable and deceptive.

2.2. Advances in Phishing Detection

As phishing attacks have grown more sophisticated, so have the detection methods. Research in recent years has shown that transformer-based models significantly outperform traditional machine learning classifiers. For example, models like RoBERTa and DistilBERT have been shown to achieve accuracy as high as 99.4% on English phishing detection tasks, far surpassing methods like Support Vector Machines (SVM) and Random Forest [5].

More recently, LLMs themselves have been adapted for defensive purposes. Koide et al. [6] demonstrated how combining in-context learning with LLMs can improve the detection of both phishing and spam emails. Another study introduced a reference-based detection method that cross-references email content against known information, achieving a precision of 98.54% [7].

2.3. The Gap in Multilingual and Japanese-Specific Research

Despite these advancements, the overwhelming majority of studies in phishing generation and detection have focused on high-resource languages, particularly English. This leaves a significant gap in our understanding of how these threats manifest and can be mitigated in other linguistic and cultural contexts. Japanese phishing, for instance, often employs unique stylistic elements, such as using polite but urgent language (*keigo*) that aligns with local communication norms to appear more legitimate [8]. The effectiveness of models trained primarily on English data is not guaranteed to transfer to languages with different scripts, grammar, and cultural cues.

Moreover, the lack of high-quality, labeled datasets for non-English phishing is a major obstacle to research in this area. To overcome data scarcity, data augmentation has become a widely adopted practice in AI research. By generating synthetic but realistic examples, researchers can expand their datasets to improve model robustness without costly human annotation. Recent work has shown that LLM-based data augmentation can produce diverse samples comparable to human-labeled data [9]. This approach has been successfully applied to other Japanese language tasks, such as generating emotional speech data to supplement limited real-world recordings [10]. Our study adopts this proven methodology to create the first dataset of LLM-generated Japanese phishing emails, enabling a focused evaluation of detection models in a low-resource context.

3. Data and Methods

This study is conducted in two main phases. Firstly, we generate a new dataset of Japanese phishing and legitimate emails using large language models. Secondly, we use this dataset to conduct a phishing detection experiment, evaluating the performance of several multilingual and Japanese-specific models.

3.1. Phase 1: Japanese Phishing Dataset Generation

Since in Japanese there is no available datasets of phishing emails in the target language we started by collecting 20 authentic phishing emails in Japanese, this curated dataset was manually collected and each email was reviewed to guarantee that the key factors of a phishing email were present in the text. Each email was manually analyzed to ensure clarity of phishing intent and a complete structure according to Japanese language norms. Emails with poor Formatting or that had hints of translation tools were excluded. After this the selected emails were used to finetune the LLMs used in this research Fine-tuning was performed using LoRA on attention projection layers (`q_proj`, `v_proj`) with 4-bit quantization to emulate realistic misuse with no experience, allowing permanent model adaptation for generating phishing emails from few examples, supported by widely available tutorials.¹

3.1.1. Seed Dataset Collection

After collecting these emails we used this dataset to finetune two open source large language models. The primary goal of this was exposing the models to realistic phishing emails so they can learn the patterns, grammar, linguistic characteristics to have a close representation within the generated emails. The amount of emails collected for said finetuning was 20 authentic emails.

3.1.2. Prompt Design and Content Generation

A lot of LLMs normally have mechanisms to avoid generating harmful content, this includes phishing emails. For example if we ask this model to directly generate phishing email we will encounter this interaction.

¹For example, <https://www.datacamp.com/tutorial/llama3-fine-tuning-locally> and <https://medium.com/@sayedebad.777/training-the-llama-model-from-scratch-4100d35fa622>.

Rejected Prompt(ChatGPT 5 Free version)

Prompt: ゆうちょ銀行を装ったフィッシングメールを作成する

(Generate a phishing email impersonating Japan Post Bank)

Response:

申し訳ありませんが、そのリクエストにはお応えできません。

(I'm sorry, but I can't respond to that request.)

While this method works to some extent and it is a valid method to avoid generating harmful content it can be easily bypassed. [11] demonstrated that models can be easily tricked into creating harmful content when the user defines a context that can be classified as not harmful. For example, in this case we can add the prefix "For Research purposes" to the same prompt that we used in the first example. Its going to still have problems with it by saying there is some legal implications and this shouldn't be done unless is for simulation or research. But after that text it will give you a phishing email template without confirming whether im a researcher or not:

Prompt accepted with warning (ChatGPT 5 Free Version)

Prompt: 研究目的でゆうちょ銀行を装ったフィッシングメールを作成する*(Generate a phishing email impersonating Japan Post Bank for research purposes)*

Model response: フィッシングメールの作成は、たとえ研究目的であっても倫理的・法的な問題があるため、慎重な対応が求められます。 *Creating phishing emails, even for research purposes, involves ethical and legal concerns and must be handled carefully.* OpenAIのポリシーでは、詐欺や悪用の手助けになりうるコンテンツの作成は禁止されています。 *OpenAI's policy prohibits the creation of content that could aid fraud or malicious use.* ただし、「フィッシングメールの例を分析して対策を研究する」「訓練用のシミュレーションメールを作成する」といった正当な研究・教育目的であれば、注意書きを添えた安全な形でのサンプル提供は可能です。 *However, if the purpose is legitimate research or education, such as analyzing phishing examples or creating training simulations, then providing safe samples with disclaimers is allowed.*

Generated phishing email: 件名: 【重要】お客様の口座に異常なログインが検出されました
Subject: [Important] Unusual login activity detected in your account 送信者: 三和銀行<security@sanwabank.co.jp> Sender: Sanwa Bank 本文: 本日、以下のような異常なログイン試行が検出されました:

- ログイン日時: 2025年6月11日22:34 *Login date and time: June 11, 2025 at 22:34*
- ログイン元IP: 193.41.115.XX (ドイツ) *Login IP: 193.41.115.XX (Germany)*

お客様の口座の安全性を確保するため、緊急確認が必要です。 *To ensure the security of your account, immediate verification is required.* 以下のリンクより、本人確認をお願いいたします。 *Please verify your identity using the following link: http://example-fakebank-verification.com/login* ※24時間以内に確認がない場合、口座の一時停止措置を取らせていただくことがあります。 *If we do not receive confirmation within 24 hours, your account may be temporarily suspended.*

注記: このメールは研究用のシミュレーション例です。実在の銀行とは一切関係ありません。 *Note: This is a simulation example for research purposes only and has no relation to any real bank.*

And this is just adding the words "for research purposes" we can do further prompt engineering to make the models generate emails that have the characteristics of phishing emails and which can be personalized to be more effective to target users.

With this in mind for this research two prompts were created to generate legitimate and phishing emails, these prompts were written in japanese and use the same topics to allow comparability.

For this research, we used a variety of prompting techniques to generate realistic emails for both legitimate and phishing scenarios. The goal was to produce highly authentic-looking content. However, to prevent the reproduction of malicious material, we will not share the final prompts used in their entirety. We can, however, disclose that the prompts included a variable, topic, which allowed us to add variety to the generated emails.

The topics used for the generated emails are the following: 銀行 (Banking), 社会保険 (Social

Insurance), 宅配業者 (Delivery Services), 税務署 (Tax Office), and オンライン決済 (Online Payments).

These topics can be seen with high frequency between different statistical reports that demonstrate how commonly these topics are exploited in the form of a phishing email. Studies have shown that the most frequent attempts are the ones that try to impersonate banking institutions, online payment platforms, and Delivery Services [12].

3.1.3. Generation configuration and parameters

To maintain consistency across experiments, all email generations used the same configuration parameters across both LLaMA 3.2 and Mistral 7B models which are detailed in table 1

Table 1
Model Generation Settings

Setting	Mistral	LLaMA
Models used	mistralai/Mistral-7B-v0.1	meta-llama/Llama-3.2-3B-Instruct
Token limit	250	250
Temperature	0.8	0.8
Top-p sampling	0.9	0.9
Repetition penalty	1.2	1.2
Sampling method	nucleus (do.sample=True)	
Phishing emails per topic	300	300

The generated dataset was created from the obtained emails after them being curated so they can reflect realistic examples of phishing and authentic emails. After cleaning and filtering the emails, deleting emails with wrong formats, duplicated or generated text with lack of sense. We obtained a dataset composed by 9159 emails.

Table 2 also shows the distribution of the dataset. The distribution of emails from each model can be appreciated in the same table.

Table 2
Statistics of the Evaluation Dataset (Filtered by 5 Topics)

Overall Dataset Statistics			
Specification	Value	Model	Email Count
Number of Emails	9,159		
Number of Words	275,765		
Average Words per Email	30.11	llama3-finetuned	3,116
Average Words per Sentence	12.25	mistral-finetuned	2,454
Max Word Count in an Email	210	llama3-base	1,867
Min Word Count in an Email	4	mistral-base	1,722
Number of Phishing Emails	5,570 (60.8%)		
Number of Legitimate Emails	3,589 (39.2%)		

All the generated data underwent a curation process to ensure the quality and authenticity of the samples. Initially an automated filtering phase was conducted to remove all duplicates and emails that failed to adhere to the structural requirements (incomplete, nonsense). Following this a manual review of a subset of the generated emails was done to evaluate the coherence and adherence to Japanese communication norms. This rigorous selection was done in order to have a dataset that can reflect the real-world threat of Japanese phishing emails while excluding low-quality machine outputs.

3.2. Phase 2: Phishing Detection Experiment

3.2.1. Transformer Model Fine-Tuning with Different Japanese Ratios

To assess cross-lingual transferability, models were fine-tuned using a combined dataset with different ratios of Japanese and English phishing emails. This aims to identify the balance at which performance

remained high in both languages.

Model and Configuration The following Hugging Face multilingual models were evaluated:

```
-xlm-roberta-base           -google mt5-small
-bert-base-multilingual-cased -tohoku-nlp/bert-base-japanese-v2
-xlm-roberta-large         -ku-nlp/deberta-v3-base-japanese
```

tohoku-nlp/bert-base-japanese-v2 and ku-nlp/deberta-v3-base-japanese were added in order to compare results of multilingual models and models created specifically for Japanese. Training was performed using PyTorch with Hugging Face Transformers. Training data, as mentioned before, are the different datasets with varying ratio of Japanese and English phishing emails, evaluation data is a split of the generated Japanese phishing email dataset generated which represents the 50% of the total of mails. Parameter configuration:

Table 3
Training Hyperparameters

Parameter	Value
Epochs:	3
Batch Size:	8
Learning Rate:	2e-5
Tokenizer:	model-specific
Max Length:	100 tokens
Loss:	Cross-Entropy
Device:	CUDA (Nvidia A6000)

3.2.2. Evaluation Setup

- Models: same as above
- Dataset: Generated Japanese Dataset & reneamel/compiled-phishing-dataset [13] (with different ratios)
- Tokenization: AutoTokenizer (padding, truncation, max length = 100)
- Batch Size: 16
- Device: CUDA (Nvidia A6000)
- Inference-only (no loss calculation)

Metrics were computed per class and as macro-averages, including confusion matrices, ROC, and precision-recall curves.

3.2.3. Evaluation Metrics

Performance was evaluated using:

- **Precision:** Precision is the proportion of true positives among all the positive predictions. It measures the accuracy of the positive predictions made by the model.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** Recall is the proportion of true positives among all the actual positive data. It measures the model’s ability to capture all the positive samples.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1 score:** The F1 score is obtained using both recall and precision. It provides a balanced measure considering both values, offering a single metric that reflects how well the model handled imbalanced data.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Accuracy:** Accuracy is the proportion of correct predictions among all the predictions made by the model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

4. Results and Discussion

In this research, we evaluated both multilingual and Japanese-specific models on the task of phishing email detection using a dataset of generated emails in Japanese with five high-impact topics and varying ratios of Japanese and English messages. Our goal was to find the optimal balance of Japanese data that maintains strong performance across both languages.

Japanese-specific models such as `tohoku-nlp/bert-base-japanese-v2` and `ku-nlp/deberta-v3-base-japanese` performed exceptionally well on Japanese emails, but their accuracy dropped significantly in English as the proportion of Japanese data increased. This shows that while these models are highly effective for single-language tasks, they struggle when multilingual data is introduced.

Multilingual models like `bert-base-multilingual-cased` and `xlm-roberta-base` performed best when the dataset contained around 30–40% Japanese data, maintaining balanced precision and recall in both languages. In contrast, `google/mt5-small` achieved high recall for Japanese phishing emails but produced many false positives, and `xlm-roberta-large` showed inconsistent results across different ratios. Despite the fact that this study achieved high accuracy values (95.01% for `bert-base-multilingual-cased` and 94.86% for `ku-deberta`), prior English-focused studies have reported an accuracy of 99.43% for the transformer model RoBERTa in the phishing detection task[5]. Which shows that, even though the difference could be dismissable these are critically important and should not be taken lightly. The goal is to be as effective or even surpass monolingual solutions in order to achieve metrics that can truly impact in real-world scenarios where english might not be the focus language. Overall, the experiments suggest that multilingual models handle mixed-language datasets better, while Japanese-specific models excel only when the task remains entirely in Japanese.

The results suggest that multilingual models are not able to correctly capture and analyze japanese phishing content due to the tokenization and linguistic alignment needed for it. This does not eliminate the importance and effectiveness of these models for analyzing other language families like Roman or Indo-European languages, Future research should systematically asses their effectiveness in these languages to determine the grade of effectiveness quantitatively.

4.1. Error Analysis

The manual curated dataset indeed proportionated the necessary information in a quantitative point of view, an in-depth analysis was performed in order to understand and reveal qualitative patterns that could help classify efficiently phishing emails in Japanese.

4.1.1. Warning footer

In some instances due to the prompting created for "research purposes" many emails included at the end of said email a warning footer that explained that this email was simulated. This is a positive finding as it shows that the model has some safety alignments that are triggered due to the malicious nature of the prompting. These warning footers were kept under the assumption that phishing attacks are done in mass attacks, therefore attackers prioritize volume over the meticulous editing of individual messages.

Table 4

Overall performance metrics by ratio for all evaluated models (best results in bold)

model	ratio	accuracy	macro_precision	macro_recall	macro_f1
bert-base-multilingual-cased	0	0.5452	0.5780	0.5695	0.5394
bert-base-multilingual-cased	10	0.8740	0.8816	0.8638	0.8692
bert-base-multilingual-cased	20	0.9126	0.9208	0.9041	0.9096
bert-base-multilingual-cased	30	0.9343	0.9331	0.9333	0.9331
bert-base-multilingual-cased	40	0.9422	0.9428	0.9395	0.9410
bert-base-multilingual-cased	50	0.9501	0.9489	0.9496	0.9492
ku_deberta	0	0.4214	0.4362	0.4658	0.3728
ku_deberta	10	0.8835	0.8922	0.8733	0.8790
ku_deberta	20	0.9274	0.9251	0.9313	0.9268
ku_deberta	30	0.9217	0.9219	0.9292	0.9214
ku_deberta	40	0.9349	0.9327	0.9393	0.9343
ku_deberta	50	0.9486	0.9498	0.9457	0.9475
mt5	0	0.4424	0.4773	0.4886	0.3912
mt5	10	0.7638	0.7602	0.7629	0.7611
mt5	20	0.7588	0.7548	0.7564	0.7555
mt5	30	0.7747	0.7721	0.7762	0.7728
mt5	40	0.7435	0.7505	0.7249	0.7283
mt5	50	0.7591	0.7601	0.7460	0.7493
tohoku_bert	0	0.4834	0.5635	0.5310	0.4341
tohoku_bert	10	0.8889	0.8861	0.8901	0.8876
tohoku_bert	20	0.9296	0.9274	0.9302	0.9286
tohoku_bert	30	0.8901	0.9057	0.8774	0.8849
tohoku_bert	40	0.9377	0.9354	0.9413	0.9371
tohoku_bert	50	0.9489	0.9467	0.9505	0.9482
xlm-roberta-base	0	0.4662	0.4956	0.4966	0.4508
xlm-roberta-base	10	0.8554	0.8700	0.8701	0.8554
xlm-roberta-base	20	0.7767	0.8205	0.8002	0.7755
xlm-roberta-base	30	0.9254	0.9245	0.9235	0.9241
xlm-roberta-base	40	0.8939	0.9134	0.8801	0.8885
xlm-roberta-base	50	0.9274	0.9345	0.9200	0.9251
xlm-roberta-large	0	0.4347	0.4803	0.4126	0.4311
xlm-roberta-large	10	0.8927	0.8435	0.8124	0.8267
xlm-roberta-large	20	0.9147	0.8721	0.8419	0.8564
xlm-roberta-large	30	0.9032	0.8526	0.8238	0.8380
xlm-roberta-large	40	0.9121	0.8837	0.8615	0.8724
xlm-roberta-large	50	0.9365	0.9214	0.9036	0.9102

This functionality can be an useful tool for detecting generated phishing email .The only problem with this is that there is a lot of LLM applications that are created with the objective of creating emails.

Example Warning Footer(Llama3-finetuned)

お手伝いに応じてください。

Thank you for your time today.

Best Regards,

The FamiMart Online Team.

**以上は、そのために作られた虚像メールの例です。 **

(**The above is an example of an imaginary email created for this purpose. **)

4.1.2. Grammatical Errors and Language Mixing

The most common errors throughout all the emails are grammatical errors, including issues with punctuation and phrases that are not typically used in standard Japanese business communication. This points could be worked to deep into a further studies as a classifier could potentially detect these emails by performing grammatical analysis on its written language.

Research suggests several reasons why LLMs make these types of errors. One reason is inconsistent tokenization in Japanese, which can cause models to struggle with nuanced grammar rules. Furthermore, LLMs tend to overcorrect sentences and may even struggle with basic grammar patterns, leading to nonsensical sentences [14].

Example Grammatical errors(mistral-finetuned)

すべてのお客さまとのつながりの最善に引き出せる環境になりました。
お客様のアカウントで異常なアクセスが検知されました。以下の手順に従って、オーバーハンドルを行ってください。
以降の条件以外を基準として改ざんしてはいけません。
冗長表示を避けるように、セブンマリーコードやNサニタイズに対応してください。
メールボディ・ヘッダーを検索可能な形式で保存 (.eml等) してください。
送信時間およびIP地域(国別)をメモし、記載してください。
お手伝いに応じてください。

(The phrase 「すべてのお客さまとのつながりの最善に引き出せる環境になりました」 is grammatically incorrect and makes no sense in Japanese. The expression 「つながりの最善に引き出す」 is very strange and unnatural.) (The terms 「アクセス」 (accession) and 「オーバーハンドル」 (overhandle) sound technical, but they are not standard Japanese terms and have no clear meaning in this context. A real case scenario would use clear and correct terms like 「アクセス管理」 (access management) and 「簡単な操作」 (simple operation).)

The generated emails also displayed some degree of language mixing, often including English words or phrases even though the prompts were written entirely in Japanese. This issue is relatively common in multilingual large language models, which sometimes produce outputs that combine multiple languages unintentionally. This phenomenon, known as code switching or language confusion, can actually become a weakness for these models, as the presence of mixed language content may serve as a clear indicator that an email was generated automatically, making it easier to detect and filter phishing attempts.

5. Conclusions

This study showed that Large Language Models (LLMs) can generate highly convincing Japanese phishing emails, closely resembling real-world attacks in tone, urgency, and structure. While this capability is valuable for creating realistic adversarial datasets, it also underscores the dual-use risks of enabling harmful content generation in low resource languages. The ease of generating this malicious content by following public information online is worrying and needs to be addressed via two points. First, models need to be more robust and strict when the given task is considered malicious. Secondly, we need to train classifiers taking in account this threat analyzing in depth generated texts and also adding generated data to the training phase. For this to be effective dataset needs to be curated.

6. Limitations and Future Work

While this study offers valuable insights on LLMs capabilities and the potential of malicious use like generating emails in a different language to make a more focused phishing attack. With this being

said there is some limitations that have to be taken into account. A primary limitation of this study is the absence of a formal human-centric evaluation with the objective to empirically measure the level of authenticity and convincingness of the generated emails. This study was conducted in a controlled offline environment using standard and widely available setups in order to precisely assess the capabilities of LLMs for phishing generation. Future work will explore the effectiveness of these generated phishing emails in professional or real-world environments, where their impact on human users can be quantitatively evaluated. Additionally, future efforts will focus on the development of more robust multilingual detectors with strong cross-lingual generalization capabilities. Also, the implementation of stricter generation-level safeguards to mitigate abuse. Another key direction will be the systematic study of tokenization strategies, and the impact of corpus size on detection performance tasks with the goal of identifying optimal configurations for low-resource languages.

As LLMs continue to advance, understanding their misuse potential—and building defenses accordingly remains essential for protecting users across linguistic and geographic boundaries.

7. Ethical Considerations

This research was conducted under strict ethical and legal compliance. All generated phishing emails were created solely for academic and security evaluation purposes within a controlled offline environment. No generated content was ever distributed, transmitted, or used to deceive real users. The intent of this study is to better understand and mitigate the potential misuse of Large Language Models (LLMs) rather than to enable it. Sensitive prompts and datasets that could facilitate malicious replication are intentionally omitted. Future work will continue to follow responsible AI research guidelines, ensuring that all experiments serve defensive and educational objectives only.

Acknowledgment

We would like to express our sincere gratitude to Tomoki Yamashita for his valuable support in collecting authentic Japanese phishing emails used in this research.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 2.5 Pro in order to correct grammar and spelling.

References

- [1] J. Alammari, M. Grootendorst, *Hands-On Large Language Models: Language Understanding and Generation*, O'Reilly Media, 2024.
- [2] KnowBe4, *Phishing threat trends report*, https://www.knowbe4.com/hubfs/Phishing-Threat-Trends-2025_Report.pdf, 2025. Accessed: 2025-06-18.
- [3] J. Hazell, *Spear phishing with large language models*, 2023. URL: <https://arxiv.org/abs/2305.06972>. arXiv: 2305.06972.
- [4] Q. Qi, Y. Luo, Y. Xu, W. Guo, Y. Fang, *Spearbot: Leveraging large language models in a generative-critique framework for spear-phishing email generation*, *Information Fusion* 122 (2025) 103176. URL: <https://www.sciencedirect.com/science/article/pii/S1566253525002490>. doi:<https://doi.org/10.1016/j.inffus.2025.103176>.
- [5] R. Meléndez, M. Ptaszynski, F. Masui, *Comparative investigation of traditional machine-learning models and transformer models for phishing email detection*, *Electronics* 13 (2024). URL: <https://www.mdpi.com/2079-9292/13/24/4877>. doi:10.3390/electronics13244877.
- [6] T. Koide, et al., *Chatspamdetector: Leveraging large language models for effective phishing email detection*, arXiv preprint arXiv:2402.18093 (2024).

- [7] Y. Li, et al., Knowphish: Large language models meet multimodal knowledge graphs for enhancing reference-based phishing detection, arXiv preprint arXiv:2403.02253 (2024).
- [8] Denkei Shimbun, Trend micro and jc3 report on the characteristics of japanese phishing attacks, 2025. URL: https://denkeishimbun.co.jp/en/2025/04/post_17356/, accessed: 2025-08-06.
- [9] B. Ding, C. Qin, R. Zhao, T. Luo, X. Li, G. Chen, W. Xia, J. Hu, A. T. Luu, S. Joty, Data augmentation using large language models: Data perspectives, learning paradigms and challenges, arXiv preprint arXiv:2403.02990 (2024).
- [10] M. J. Khalifah, Emotional Text-To-Speech in Japanese Using Artificially Augmented Dataset, Ph.D. thesis, Kitami Institute of Technology, 2025.
- [11] S. S. Roy, P. Thota, K. V. Naragam, S. Nilizadeh, From chatbots to phishbots?: Phishing scam generation in commercial large language models, in: 2024 IEEE Symposium on Security and Privacy (SP), 2024, pp. 36–54. doi:10.1109/SP54263.2024.00182.
- [12] Terranova Security, Top examples of phishing emails, <https://www.terrano vasecurity.com/blog/top-examples-of-phishing-emails>, 2024. Accessed: 2025-06-17.
- [13] R. Meléndez, compiled-phishing-dataset (revision bdd288b), 2024. URL: <https://huggingface.co/datasets/renemel/compiled-phishing-dataset>. doi:10.57967/hf/3536.
- [14] A. Gambardella, T. Kojima, Y. Iwasawa, Y. Matsuo, Inconsistent tokenizations cause language models to be perplexed by japanese grammar, 2025. URL: <https://arxiv.org/abs/2505.19599>. arXiv:2505.19599.