SCAMNET: Toward Explainable Large Language Model-based Fraudulent Shopping Website Detection

Marzieh Bitaab[1], Alireza Karimi[1], Zhuoer Lyu[1], Ahmadreza Mosallanezhad[2], Adam Oest[3], Ruoyu Wang[1], Tiffany Bao[1], Yan Shoshitaishvili[1], Adam Doupé[1]

[1]Arizona State University, [2]NVIDIA, [3]Amazon
[1]{mbitaab,akarimi6,zlyu15,fishw,tbao,yans,adoupe1}@asu.edu,
[2]dmosallanezh@nvidia.com, [3]aoest@amazon.com

Abstract

Fraudulent shopping websites pose a significant threat to online consumers and legitimate businesses: in 2023, victims of such scams reported $392 million in losses to the Federal Trade Commission. This alarming trend not only impacts individuals but also erodes societal trust in e-commerce, necessitating urgent countermeasures. While previous studies have attempted to identify these fraudulent websites at scale, they face limitations such as potential bias in data collection, overreliance on easily manipulated features, and the lack of explainable results. This study explores the potential of Large Language Models (LLMs) in identifying fraudulent shopping websites, revealing that current LLMs underperform compared to existing machine learning models. To address this, we propose SCAMNET, a fine-tuned LLM for explainable fraudulent shopping website detection. Our experimental results on real-world datasets demonstrate a breakthrough in detection performance from 22.35% detection rate to 95.59%, particularly in identifying subtle deceptive tactics such as using a legitimate-looking website template. SCAMNET offers interpretable insights into its decision-making process, enhancing transparency and overcoming a key limitation of previous approaches.

Introduction

The proliferation of fraudulent e-commerce websites poses a significant threat to the society, rippling from end users, vendors, and financial departments. Fraudulent shopping websites, distinct from phishing sites, lure customers with enticing discounts on popular items, often resulting in the delivery of counterfeit goods or no products at all. These scams pose a broader risk to the society by indiscriminately targeting online consumers, unlike phishing attacks which impersonate only a specific brand. The Federal Trade Commission reports a surge in online shopping scams, ranking them the second most-reported fraud type in 2023, with $392 million in loss (Commission 2023).

Traditional approaches to identifying fraudulent websites rely heavily on manually crafted features, which can be time-consuming to create and potentially miss subtle indicators. Recent methods such as Scamdog Millionaire (Kotzias et al. 2023) and BEYOND PHISH (Bitaab et al. 2023) attempted to address this issue but each has limitations. Scamdog Millionaire extracts 111 features from collected domains to build a random forest-based classifier, yet its data collection process may introduce bias by relying on specific scam repositories and filtering out popular domains, potentially overlooking sophisticated scams masquerading as well-known entities. Moreover, the limited scope of manual verification in its real-world application raises concerns about the robustness of the validation process.

BEYOND PHISH, a state-of-the-art fraudulent shopping websites detection model, emphasizes URL-derived features such as length, presence of suspicious keywords, and unusual characters. This model detected fraudulent shopping websites with a high detection rate of 98.34% (Bitaab et al. 2023). However, scammers can easily manipulate URLs to appear legitimate, thus evading detection. Moreover, the reliance of this approach on a limited set of manually engineered features may struggle to keep pace with the evolving tactics employed by fraudsters (and our evaluation bears this out, as BEYOND PHISH's F-1 score drops from 90.19% on their original dataset to 27.14% on a similar dataset we collected two years later). BEYOND PHISH also lacks *explainability* which is a critical factor to the practice, since browsers that are responsible for blocking illicit websites cannot ban a website without a clear explaination. Therefore, explainability will reduce the legal risk for browser vendors such as Google and Microsoft.

Considering the limitations of traditional models and the unexplored potential of Large Language Models (LLMs) in identifying fraudulent shopping websites, two critical questions arise: (1) can LLMs effectively detect fraudulent shopping websites at scale using their inherent knowledge and capabilities and (2) what approach should we employ to maximize performance when using LLMs for explainable fraudulent shopping websites detection? To answer these questions, we examine LLMs' efficacy in identifying fraudulent shopping websites.

Our investigation involves empirical analysis to uncover LLMs' latent capabilities. We employ two distinct prompting strategies (zero-shot and few-shot) to classify websites as fraud or legitimate. These prompts incorporate the URL, website content, and WHOIS information. Surprisingly, our initial findings reveal that *even the most effective LLM-based approach falls short of the performance achieved by a recent*

open-source feed-forward neural network model (Bitaab et al. 2023). The results indicate that current LLMs can not serve as suitable replacements for task-specific trained models in identifying fraudulent shopping websites, and remains an untapped potential.

To augment LLMs for this task, we develop a multi-step approach. First, we created a refined dataset based on the BEYOND PHISH study, removing redundant data and websites with insufficient information such as empty website content or WHOIS information. Next, we used a two-phase fine-tuning process to build an explainable non-phishing detection model based on the Llama-3-8B-Instruct (Touvron et al. 2023) model. The initial phase used supervised fine-tuning to create a model capable of identifying non-phishing sites, while the second phase creates a small but precise explainable dataset that includes reasons a website is fraudulent or legitimate. We then use this dataset to fine-tune the model, resulting in an explainable fraudulent shopping websites detection system we call SCAMNET.

To assess SCAMNET's effectiveness, we perform extensive evaluations on real-world datasets. Our results indicate that SCAMNET performs better than traditional Machine Learning (ML) methods, even on temporal data, improving performance on recently gathered dataset from 22.35% detection rate to 95.59%. This suggests that fine-tuned LLMs can find meaningful patterns in website information. We also evaluated the explanations generated by our fine-tuned LLM and found that it can produce clear and insightful justifications. By using specific prompts to guide the LLM, we showed that these explanations significantly improve fraudulent shopping websites detection.

The main contributions of this paper are the following:

- We explore the effectiveness of using LLMs to detect fraudulent shopping websites, both out-of-the-box and fine-tuned. We demonstrate that LLMs can be fine-tuned to fraudulent shopping websites detection without manual feature extraction.
- We create a novel accurate dataset with explanations for fraudulent shopping websites detection.
- We conduct extensive experiments and ablation studies to show how SCAMNET works in real-world situations.

Background: LLMs for Classification

Large Language Models (LLMs) have emerged as powerful tools for various natural language processing tasks, including text classification (Kant et al. 2018; Menon and Vondrick 2022; Sun et al. 2023). These models have demonstrated remarkable capabilities in understanding and generating human-like text across various domains. Unlike traditional machine learning approaches that require extensive feature engineering for a given task, LLMs can be applied to classification tasks with minimal adaptation through techniques such as few-shot learning, unsupervised or supervised fine-tuning (Hegselmann et al. 2023).

The use of LLMs for text classification has gained significant attention due to their ability to capture complex linguistic patterns and contextual information. Studies have shown that LLMs can achieve competitive or superior performance compared to traditional supervised learning methods, especially in scenarios with limited labeled data (Lin et al. 2024). However, challenges remain in optimizing LLMs for specific classification tasks, particularly fine-tuning strategies. Recent research has explored various approaches to enhance LLM performance in text classification, including parameter-efficient fine-tuning (Ding et al. 2023), and the generation of synthetic training data (Chung, Kamar, and Amershi 2023). These advancements aim to leverage the pre-trained knowledge of LLMs while adapting them more effectively to specific classification tasks.

Model	Accuracy	TPR	FPR	F-1
GPT-3.5 - 0-shot	48.18%	99.19%	71.67%	51.75%
GPT-3.5 - 2-shot	29.73%	**100.00%**	97.62%	44.37%
GPT-4 - 0-shot	89.46%	91.26%	11.23%	82.91%
GPT-4 - 2-shot	90.21%	68.90%	**1.50%**	79.76%
BEYOND PHISH	**94.57%**	90.10%	3.71%	**90.19%**

Table 1: Performance metrics of BEYOND PHISH compared to state-of-the-art LLMs on the BEYOND PHISH evaluation dataset shows that fraudulent shopping websites detection is a new domain.

LLMs for Detecting Fraudulent Shopping Websites

·We first assess the capability of representative large language models (LLMs) to identify fraudulent shopping websites. Our evaluation employs two prompting approaches and benchmarks the results against the state-of-the-art BEYOND PHISH model. We selected BEYOND PHISH for comparison due to its open-source nature and our inability to access other similar models such as "Scamdog Millionaire" (Kotzias et al. 2023).

In this experiment, we use the dataset employed and released by (Bitaab et al. 2023) that evaluated the BEYOND PHISH classifier. The task involves classifying a website as legitimate or fraudulent based on the website's characteristics. To enhance the dataset quality of BEYOND PHISH's testing dataset we perform a de-duplication step.

We evaluate GPT-3.5-turbo and GPT-4, developed by OpenAI (Achiam et al. 2023). These LLMs can adapt to tasks through instructions or few-shot demonstrations, circumventing the impracticality of task-specific fine-tuning for such large models. We employ two prompting strategies:
0-Shot Prompting: This method constructs prompts containing only the task description and the given website information. To enhance response quality and reduce refusal rates, we optionally incorporate role-playing techniques in task descriptions.
2-Shot Prompting: This approach provides task-specific prompts along with two website info-label samples: one legitimate and one fraudulent website.

Through these methods, we aim to comprehensively assess the LLMs' potential in detecting fraudulent shopping websites and compare their performance against established machine learning models. Table 1 presents the results be-

tween GPT-3.5-turbo, GPT-4, and BEYOND PHISH on the testing set. The results reveal several key insights:

LLM Performance vs. Specialized Classifiers: Despite the general perception of LLMs as powerful tools, they underperformed compared to the machine learning-based classifier across both prompting approaches, which suggests that LLMs lack the task-specific knowledge required.

Impact of Prompting Approaches: Providing examples to an LLM for classification can improve its performance. Thus, we perform a two-shot approach which provides an example of a legitimate website and an example of a fraudulent non-phishing website. As we can select these examples randomly, we do this experiment five times and report the average results. There is minimal or no improvement between zero-shot and two-shot experiments. The only exception is GPT-4, where the two-shot prompting approach showed an improved (reduced) false positive rate.

Scalability and Privacy: Fine-tuning smaller models based on initial LLM outputs can balance performance and cost-effectiveness. This approach allows for secure, on-premise deployment and avoids potential privacy concerns associated with sending data to third-party API providers.

In conclusion, out-of-the-box LLMs fall short in replacing specialized ML models for tasks such as non-phishing detection. It seems that the combination of higher performance and lower costs makes traditional ML approaches more suitable for this application. However, fine-tuning smaller models may offer a path forward for balancing performance, cost, and scalability.

Proposed Method

In the previous section, we conducted an experiment to evaluate Large Language Models (LLMs) in detecting fraudulent shopping websites. Our findings indicate that commercial LLMs, in their current pre-trained state, are inadequate for the fraudulent shopping websites detection task. We believe this incapability is related to fraudulent shopping websites detection being a specialized task, which falls outside the typical training corpus of contemporary LLMs. This hypothesis is consistent with the challenges of applying general-purpose language models to highly specialized tasks without domain-specific adaptation (Hu et al. 2021). Our results underscore the importance of tailored model optimization and the potential need for curated datasets in detecting fraudulent shopping websites.

SCAMNET: An Explainable LLM-based Model for Fraudulent Shopping Websites Detection

Detecting potential fraudulent shopping websites necessitates the development of a robust classifier capable of differentiating between legitimate and fraudulent shopping websites. Previous research in this area is limited, as highlighted by (Bitaab et al. 2023), who underscore the scarcity of studies focused on detecting fraudulent shopping websites. The proposed open-source classifier, BEYOND PHISH, uses features derived from the website's content, WHOIS information, and the URL structure.

Inspired by this work and recent advancements in LLMs, we propose designing and optimizing an LLM specifically tailored for fraudulent shopping websites. The reason for employing an LLM lies in its ability to autonomously extract and process contextual features from textual data, potentially surpassing the efficacy of manual feature extraction methods traditionally used in classification tasks (Zhang et al. 2023). This approach aims to harness LLMs' pattern recognition and natural language processing capabilities to enhance the accuracy of phishing detection mechanisms.

Figure 1 shows the general tuning pipeline of SCAMNET. We employ the Llama-3-8B-Instruct model as the foundational LLM. We fine-tune Llama-3-8B-Instruct with a custom prompt designed to facilitate the detection task. Here is a refined and properly formatted version of our prompt template and its application:

```
# Information:
{input features}
# Pred:
{website's label}
```

In this prompt template:

- *input features* is populated with general textual features extracted from the website. These features include URL structure, website main page textual content, outgoing links (also called external links), and important WHOIS fields. We provide this information as follows:

  ```
  ## URL:
      website's URL
  ## Content:
      website's body text
  ## External Links:
      list of external links
  ## WHOIS Information:
      list of WHOIS information
  ```

  We prepare the input data to effectively use the Llama-3-8B-Instruct model. We begin by populating all fields in our data structure, except for the CONTENT field. These fields include metadata such as URL, WHOIS information, and external links. Then, the focus shifts to the *Content* field. Given the token limit, it is often necessary to truncate the website content to fit within the remaining token space after accounting for the other fields.
- *website's label* is the target output of the LLM, where it should predict whether the website is legitimate or fraudulent: The label is either $legit$ or $scam$.

In the process of adapting the Llama-3-8B-Instruct model to the specialized task of fraudulent shopping websites detection, we leverage the Low-Rank Adaptation (Hu et al. 2021) (LoRA) technique to incorporate task-specific information without extensively retraining the entire model. LoRA, a method that introduces additional trainable layers to the pre-existing architecture of Large Language Models (LLMs), selectively fine-tunes these layers while keeping the core model parameters intact (Hu et al. 2021). This selective adaptation strategy is important for several reasons. First, it significantly mitigates the risk of catastrophic forgetting, a common challenge in neural network training where

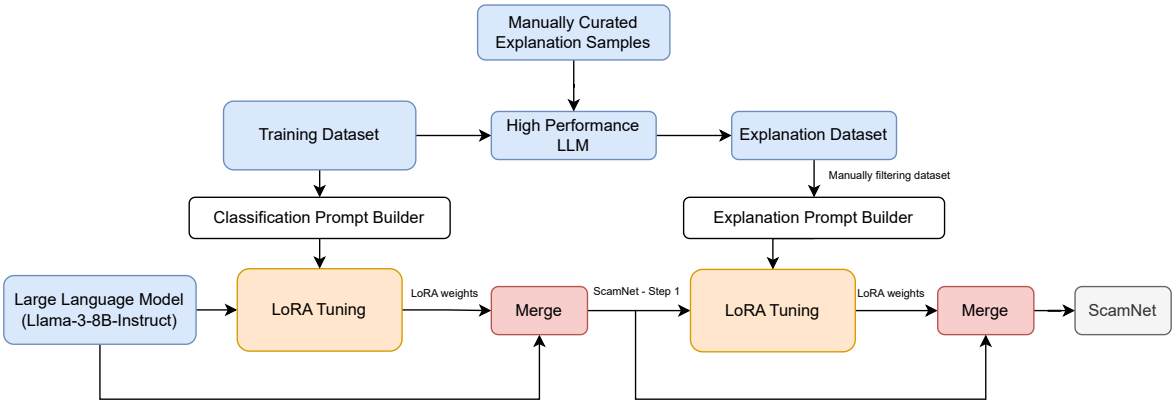

Figure 1: The proposed pipeline for creating an LLM-based fraudulent shopping websites detection. We use Llama-3-8B-Instruct as the base model to perform a two-step supervised fine-tuning.

Dataset	Model	Accuracy	TPR	FPR	F-1
BP Evaluation	BEYOND PHISH	94.57%	90.10%	3.71%	90.19%
	Llama-3-8B-Instruct	73.81%	89.35%	38.45%	75.06%
	SCAMNET	**99.47%**	**99.08%**	**0.22%**	**99.40%**
Real-World Evaluation	BEYOND PHISH	81.94%	22.35%	7.50%	27.14%
	Llama-3-8B-Instruct	75.55%	84.55%	26.05%	51.15%
	SCAMNET	**98.85%**	**95.59%**	**0.57%**	**96.15%**

Table 2: Performance metrics of BEYOND PHISH model compared to SCAMNET.

acquiring new knowledge leads to the erosion of previously learned information (Kirkpatrick et al. 2017). Furthermore, compared to traditional full-parameter fine-tuning methods, LoRA is more resource-efficient. It reduces the computational overhead and memory requirements associated with updating the vast parameter space of models such as Llama-3-8B-Instruct, making it a practical choice for scenarios with limited hardware capabilities. To ensure comparability in our study, we employed the deduplicated training and testing datasets used in the BEYOND PHISH methodology, which were obtained directly from the authors.

Building upon our LoRA-based fine-tuning approach, we create a pipeline to generate explainable reasons for the website classification as legitimate fraudulent. This process involves creation of a small and precise dataset of 200 explanations, showing the rationale behind categorizing a given website as either legitimate or fraudulent. Our curation methodology employs a multi-stage filtering process, leveraging commercial LLMs and expert human validation to filter-out low-quality generated data and improving the explanations to increase data quality.

Finally, we use this curated dataset to further fine-tune SCAMNET, enhancing it to not only classify websites but also provide justifications for its decisions. These justifications can be used by a human analyst to verify and validate the results of SCAMNET before applying a proactive measure, such as adding the website to a blocklist.

Dataset		# Legitimate	# Fraudulent	Collection Date
BP Data	Train Set - Step 1	10492	8430	Dec 2018 - Nov 2021
	Train Set - Step 2	100	100	Dec 2018 - Nov 2021
	Evaluation Set	2645	2086	Dec 2018 - Nov 2021
Real-World Data	Evaluation Set	1923	340	May 2024 -June 2024

Table 3: Dataset statistics.

Experiments and Results

We design experiments to measure the effectiveness of SCAMNET. We aim to answer three main research questions: (**Q1**) How does fine-tuning help in creating an LLM that can detect fraudulent shopping websites with explainable output? (**Q2**) How does the size of the tuning dataset and parameters affect the results? (**Q3**) How reliable are the generated explanations by our model?

To answer **Q1**, we evaluate our fine-tuned LLM after each step of tuning, before and after tuning the model on the given explainable dataset. Then, to answer **Q2**, we design an ablation study to show the effects of important fine-tuning parameters in the final model's outcome. Finally, to answer **Q3**, we perform a manual analysis on the generated reasoning by our model to assign a reliability score to them.

We compared several baselines in our experiments, including the state-of-the-art BEYOND PHISH method. Table 1 shows the results comparing the following baselines:

- **BEYOND PHISH** (Bitaab et al. 2023): This classifier uses manually extracted features from the website's content, WHOIS information, and URL to detect fraudulent shopping websites websites. This method is based on a feed-forward neural network model.
- **Llama-3-8B-Instruct** (Touvron et al. 2023): To evaluate the effectiveness of our tuning method, we also evaluate the base Llama-3-8B-Instruct model which is an instruction-tuned variant of Llama-3-8B.

Dataset

Training Datasets: In this section, we provide the details of the training datasets that we used for our two-step supervised fine-tuning approach:

Step 1 Dataset: For this step, we use the markdown-based template discussed previously. We use BEYOND PHISH's training set (Bitaab et al. 2023) to perform the first round of supervised fine-tuning.
Step 2 Dataset: This step involves creating an explainable dataset. We create a small but precise dataset that can be used to teach an LLM and provide explanations for its label. The process is as follows:

1. We first create several manually curated examples for randomly selected websites from our training set. For each website, we provide our reasons in three main categories. Below is a template we used to show the model's reasoning:

```
## WHOIS Information:
list of observations on website's
 ↪  WHOIS data

## Website's Content and External
 ↪  Links:
list of reasons why this website is
 ↪  legitimate or fraudulent based
 ↪  on its content and source code

## Miscellaneous:
list of reasons that do not belong
 ↪  to either of the above
 ↪  categories
```

2. After creating four reasoning examples (two legitimate and two fraudulent, following a four-shot strategy), we use a commercial LLM, GPT-4, to create the same type of reasoning for the rest of the dataset.
3. We then randomly select 200 samples, manually analyze and fix the reasoning, thus improving the dataset quality.

We then use this small and precise reasoning dataset to perform a second round of supervised fine-tuning.

Evaluation Datasets To assess SCAMNET, we consider two evaluation datasets. First, we evaluate the fine-tuned LLM on the same test set as BEYOND PHISH. This dataset is able to show the general capability of models on detecting fraudulent shopping websites.

To further improve our evaluation, we collect and perform evaluations on collected websites from real-world scenarios. We collect 2,267 websites from May 2024 to June 2024 by automatically searching the web using various shopping-related keywords, and our security experts manually assign labels to each website. Although our collected dataset is similar to BEYOND PHISH's evaluation dataset, it is two years more recent and thus representative of current fraudulent shopping websites. In addition, none of the models were trained on this dataset. The dataset statistics are shown in Table 3.

To show the importance of our newly collected dataset, we try to compare the dataset domains using t-distributed Stochastic Neighbor Embedding (t-SNE). Figure 2 shows the datasets' distribution in two dimensions for fraudulent shopping websites. Comparing the data points, we see that

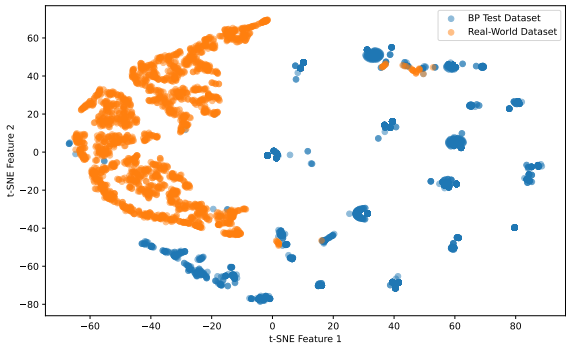

Figure 2: t-SNE plot of BP test dataset and our real-world collected dataset. This plot shows a gap between the two evaluation datasets, indicating a data shift that is happening in recent fraudulent shopping websites.

the two datasets do not have many data points in common. Finally, another important point from the t-SNE plot is that the new testing dataset is very diverse and does not belong to a single scam campaign. This makes the dataset suitable for assessing a model's robustness through time.

Optimization Algorithm

Our optimization process employs a two-step Supervised Fine-Tuning (SFT) approach using Low-Rank Adaptation (LoRA) to efficiently adapt the Llama-3-8B-Instruct model for fraudulent shopping websites detection while maintaining explainability.

Step 1: Initial Fine-Tuning

In the first step, we apply LoRA to fine-tune the model on our fraudulent shopping websites detection dataset derived from BEYOND PHISH's dataset (Bitaab et al. 2023), but with reduced noise. To maximize adaptation quality, we target all linear layers in the model, as opposed to only the attention blocks. This approach allows for a more comprehensive update of the model's knowledge while maintaining efficiency.

We perform the initial fine-tuning step on different values of learning rates and LoRA rank values. Finally, we select the best checkpoint according to the evaluation loss during tuning.

Step 2: Explainable Fine-Tuning

After the initial fine-tuning, we merge the LoRA weights into the base model. This step integrates the task-specific knowledge gained from the first fine-tuning phase directly into the model architecture.

We then perform a second round of SFT using our curated explainable dataset. This dataset contains detailed explanations for website classifications, enabling the model to generate coherent and insightful justifications for its decisions. Using a smaller and focused dataset aims to enhance the model's ability to provide clear explanations without overfitting or losing fraudulent shopping websites detection capabilities acquired in the first step.

Throughout both stages, we use the *AdamW* optimizer and a linear learning rate scheduler with *warmup*. This com-

Parameter	Value for Step 1	Value for Step 2
Learning rate	1e-4	1e-5
LoRA rank r	16	32
Number of epochs	2	2

Table 4: Hyperparameters for explainable fine-tuning process. Note that we always set the LoRA's parameter $\alpha = 2 \times r$.

bination helps manage the learning process effectively, particularly given the specialized nature of our task and the need to balance performance with explainability. We provide the hyperparameters used for each step in Table 4. We use four H100 GPUs to tune the models on $bf16$ precision.

Evaluation Results

As shown in Table 1, we explore the application of LLMs to the task of fraudulent shopping websites detection, a relatively uncharted area within the field. The results indicate a challenge in adapting general-purpose LLMs, such as GPT-4, to effectively identify fraudulent shopping websites. This finding is not unexpected, given the limited availability of robust datasets tailored for this task.

Table 2 shows the results of SCAMNET compared to other baselines on the two datasets. The BEYOND PHISH method, designed explicitly for fraudulent shopping websites detection, demonstrates superior performance on its own evaluation dataset, even better than state-of-the-art LLMs. However, it fails to reach the performance of SCAMNET, especially when it has a low false positive rate (**Q1**). Moreover, we observe a random behavior from non-tuned Llama-3-8B-Instruct model as it predicts the fraudulent label for almost all of the given websites. Comparing the results of SCAMNET to non-tuned Llama-3-8B-Instruct indicates that fraudulent shopping websites detection is a new domain for LLMs, and our proposed approach improves the results by adopting the domain of pre-trained Llama to fraudulent shopping websites detection.

Further, we perform the same evaluation on the new in-the-wild dataset, including the dataset collected over two years after than the training set. The results indicate that BEYOND PHISH does not perform well on newer fraudulent shopping websites. The performance degrades from a high TPR of 90.10% to a very low TPR of 22.35%, while SCAMNET shows promising performance. We believe that the reason the explainable model keeps its good performance on the real-world dataset is that we used a curated representative training set to improve its reasoning capabilities. This finding aligns with recent research that shows that using a high-quality dataset, although small in size, can improve the results by a significant margin (Zhou et al. 2024).

Ablation Study

Next, we aim to answer research question **Q2** by performing an analysis on important model tuning parameters. One important aspect of our model tuning is using LoRA, in which two important values affect the outcome, LoRA rank r_{LoRA} and scale value α_{LoRA}. We apply our approach on different LoRA rank values $r_{\text{LoRA}} \in \{8, 16, 32\}$ which keeping the scale value $\alpha = 2 \times r_{\text{LoRA}}$. Figure 3 shows the performance results for each experiment. Moreover, to show the impact of using our curated dataset for a second round of supervised fine-tuning, we include the results of SCAMNET after step 1 of SFT.

Algorithm 1: Two-Step LoRA Fine-Tuning for Explainable fraudulent shopping websites Detection.

Require: Pre-trained model M, fraudulent shopping websites detection dataset D_{np}, explainable dataset D_{exp}
Ensure: Fine-tuned explainable fraudulent shopping websites detection model M_{exp}

Step 1: Initial Fine-Tuning
Initialize LoRA adapters L_1 for all linear layers in M
Set hyperparameters: learning rate lr_1, LoRA rank r_1, LoRA alpha α_1, batch size b_1, epochs e_1
for epoch $= 1$ to e_1 **do**
    for batch in D_{np} **do**
        Update L_1 using AdamW optimizer and linear learning rate scheduler
    end for
end for
$M_1 \leftarrow$ Merge L_1 into M
Step 2: Explainable Fine-Tuning
Initialize new LoRA adapters L_2 for all linear layers in M_1
Set hyperparameters: learning rate lr_2, LoRA rank r_2, LoRA alpha α_2, batch size b_2, epochs e_2
for epoch $= 1$ to e_2 **do**
    for batch in D_{exp} **do**
        Update L_2 using AdamW optimizer and linear learning rate scheduler
    end for
end for
$M_{exp} \leftarrow$ Merge L_2 into M_1
return M_{exp}

The experimental results show an improvement in the performance of SCAMNET after the two-step supervised fine-tuning process. This enhancement was particularly evident after the second round of fine-tuning, which used our curated explainable dataset. The performance improvement after the second fine-tuning step suggests that the model became better adapted to the target domain's features. This highlights the value of iterative fine-tuning and domain-specific datasets in optimizing LLMs for specialized tasks.

Analysis of LoRA rank r variations shows distinct performance trends across two rounds of SFT. In the first round, performance improved as r increased to 16 but declined at $r = 32$, suggesting potential overfitting. Consequently, the checkpoint with $r = 16$ was selected for the second round. The second round shows improved performance at $r = 32$, contrasting with the first-round results. These findings underscore the complex relationship between LoRA rank and dataset properties, highlighting the necessity for iterative tuning and careful parameter selection in model optimiza-

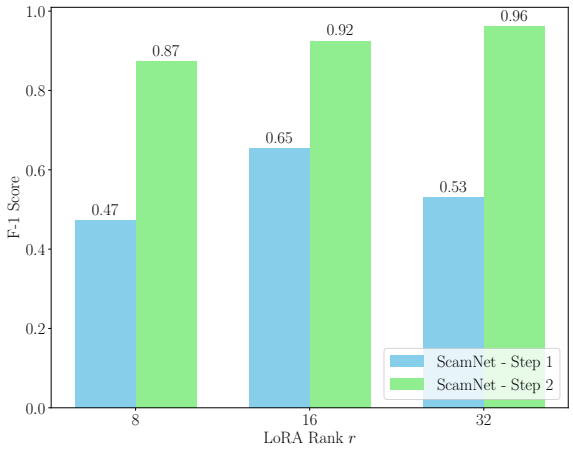

Figure 3: F-1 score of SCAMNET after step 1 and step 2 of supervised fine-tuning on the real-world evaluation set. The results indicate a performance improvement after performing the second round of tuning on our curated dataset.

	Score		
Label	Bad (0)	OK (1)	Good (2)
Fraudulent	0	76	20
Legitimate	4	4	92
Total	4	80	112

Table 5: Statistics of manually reviewed SCAMNET's generated explanations on the real-world evaluation dataset.

tion.

Reasoning Reliability

We design a human study to answer research question **Q3** to study the quality of generated explanations. We start by creating a simple scoring guideline to assign a general evaluation score to each of the generated reasoning. The scoring guideline is as follows:

- If the generated reasoning is not related to the given label or the provided website's information, the score is **0 - BAD**.
- If most of the generated reasons are related to the given label and the provided website's information, the score is **1 - OK**.
- If all of the generated reasons are related to the given label and the provided website's information, the score is **2 - GOOD**.

We assign 200 websites, where the label has been detected correctly, to three security experts to score based on the guidelines. We then aggregate their results by averaging the scores for each model's output.

Table 5 shows the statistics of the assigned scores to SCAMNET's generated explanations. The overall average score reported is 1.8, reflecting a **Good** reasoning quality. This result verifies that our model can identify important points when assigning a legitimate or fraudulent label to a website. We further show several examples of our model's explainable output in the supplementary materials.

Related Work

Fraudulent Shopping Websites Detection: Several studies explored approaches to identify deceptive online shops (Carpineto and Romano 2017). These approaches include classifying fraudulent websites based on their source code structure similarity (Beltzung et al. 2020) or using Doc2Vec for vectorization of HTML to detect fake Japanese shopping sites (Sakai et al. 2023). Other research focuses on detecting fraudulent shopping websites by analyzing user reviews (Manek et al. 2016). Two recent methods for detecting fraudulent online shopping websites are a random forest-based classifier (Kotzias et al. 2023) and a neural network-based approach (Bitaab et al. 2023). Both methods use feature extraction to assess website legitimacy but face challenges with evolving fraudulent techniques.

LLMs for classification: Recent studies have revealed significant limitations in ChatGPT's commonsense reasoning abilities compared to fine-tuned models (Qin et al. 2023; Laskar et al. 2023). While some research on GPT-4 showed promise in real-world physical reasoning tasks (Bubeck et al. 2023), overall evaluations indicate that ChatGPT under-performs in various logical reasoning challenges. Despite displaying high confidence, ChatGPT demonstrates an inability to maintain consistent beliefs about truth across diverse reasoning tasks (Wang, Yue, and Sun 2023). These findings highlight the need for more comprehensive assessments of LLMs' reasoning capabilities.

Conclusion

Our investigation into the efficacy of Large Language Models (LLMs) for fraudulent shopping websites detection revealed that commercial out-of-the-box LLMs under-perform compared to traditional ML-based models. However, fine-tuning a Llama-3-8B model using the same dataset as traditional ML models yielded surprising results. The fine-tuned LLM not only outperformed the ML-based model on the testing set but also demonstrated superior performance *data collected two years later*. These findings highlight a crucial insight: while LLMs possess sophisticated analytical capabilities, they may struggle to fully leverage their internal knowledge for specialized tasks without proper fine-tuning. Our research suggests that "mining" the potential of LLMs through targeted fine-tuning on the appropriate dataset, including samples with reasoning behind labels, can lead to significant performance improvements. Notably, the achieved False Positive Rate of 0.57% is particularly promising for practical security applications, as it minimizes the risk of legitimate websites being incorrectly flagged as fraudulent.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This work was supported in part by the Advanced Research Projects Agency for Health (ARPA-H) under Contract No. SP4701-23-C-0074, National Science Foundation

(NSF) Grant No. 2232915, Department of Navy award N00014-23-1-2563 and N00014-24-1-2193 issued by the Office of Naval Research. We gratefully acknowledge their support.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Beltzung, L.; Lindley, A.; Dinica, O.; Hermann, N.; and Lindner, R. 2020. Real-time detection of fake-shops through machine learning. In *2020 IEEE International Conference on Big Data (Big Data)*, 2254–2263. IEEE.

Bitaab, M.; Cho, H.; Oest, A.; Lyu, Z.; Wang, W.; Abraham, J.; Wang, R.; Bao, T.; Shoshitaishvili, Y.; and Doupé, A. 2023. Beyond Phish: Toward Detecting Fraudulent e-Commerce Websites at Scale. In *2023 IEEE Symposium on Security and Privacy (SP)*, 2566–2583. IEEE.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Carpineto, C.; and Romano, G. 2017. Learning to detect and measure fake ecommerce websites in search-engine results. In *Proceedings of the international conference on web intelligence*, 403–410.

Chung, J. J. Y.; Kamar, E.; and Amershi, S. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. *arXiv preprint arXiv:2306.04140*.

Commission, F. T. 2023. FTC Data Book 2023. https://www.ftc.gov/system/files/ftc_gov/pdf/CSN-Annual-Data-Book-2023.pdf.

Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.-M.; Chen, W.; et al. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3): 220–235.

Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; and Sontag, D. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, 5549–5581. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kant, N.; Puri, R.; Yakovenko, N.; and Catanzaro, B. 2018. Practical text classification with large pre-trained language models. *arXiv preprint arXiv:1812.01207*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Kotzias, P.; Roundy, K.; Pachilakis, M.; Sanchez-Rola, I.; and Bilge, L. 2023. Scamdog Millionaire: Detecting E-commerce Scams in the Wild. In *Proceedings of the 39th Annual Computer Security Applications Conference*, 29–43.

Laskar, M. T. R.; Bari, M. S.; Rahman, M.; Bhuiyan, M. A. H.; Joty, S.; and Huang, J. X. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. *arXiv preprint arXiv:2305.18486*.

Lin, X.; Wang, W.; Li, Y.; Yang, S.; Feng, F.; Wei, Y.; and Chua, T.-S. 2024. Data-efficient Fine-tuning for LLM-based Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 365–374.

Manek, A. S.; Shenoy, P. D.; Mohan, M. C.; and Venugopal, K. 2016. Detection of fraudulent and malicious websites by analysing user reviews for online shopping websites. *International Journal of Knowledge and Web Intelligence*, 5(3): 171–189.

Menon, S.; and Vondrick, C. 2022. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*.

Qin, C.; Zhang, A.; Zhang, Z.; Chen, J.; Yasunaga, M.; and Yang, D. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Sakai, K.; Takeshige, K.; Kato, K.; Kurihara, N.; Ono, K.; and Hashimoto, M. 2023. An Automatic Detection System for Fake Japanese Shopping Sites Using fastText and LightGBM. *IEEE Access*.

Sun, X.; Li, X.; Li, J.; Wu, F.; Guo, S.; Zhang, T.; and Wang, G. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, B.; Yue, X.; and Sun, H. 2023. Can chatgpt defend the truth? automatic dialectical evaluation elicits llms' deficiencies in reasoning. *arXiv preprint arXiv:2305.13160*.

Zhang, R.; Han, J.; Liu, C.; Gao, P.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; and Qiao, Y. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.