"Is this Site Legit?": LLMs for Scam Website Detection

Yuan-Chen Chang(✉) and Esma Aïmeur

Department of Computer Science and Operations Research, University of Montreal, Montreal, Canada
yuan-chen.chang@umontreal.ca, aimeur@iro.umontreal.ca

Abstract. The proliferation of online scams has become a pressing concern in the digital age, exacerbated by the rise of Artificial Intelligence. Malicious actors now employ sophisticated techniques to create convincing fraudulent schemes, targeting vulnerable individuals through personalized approaches on social media. This paper addresses the challenges of scam website detection by leveraging the capabilities of Large Language Models (LLMs). While other papers have focused on fine-tuning LLMs, our research investigates if readily available LLMs can be directly applied to scam website detection. This paper explores text-based and screenshot-based methods, utilizing five prominent LLMs to analyze website content. The findings indicate that existing LLMs are effective in identifying scam websites and providing rapid, expert responses for assessing website legitimacy. A novel categorization of criteria is proposed based on the LLMs' decision-making processes. By comparing these models' performances, this paper aims to develop a more efficient and accessible solution for identifying fraudulent websites. This work contributes to enhancing cybersecurity measures, potentially reducing online scams and increasing user trust in digital interactions.

Keywords: Scam Detection · Large Language Models (LLMs) · Cybersecurity

1 Introduction

A scam is a fraudulent scheme or deceptive practice, usually aimed at obtaining money or personal information from individuals or organizations through dishonest means [4]. The problem of scams has been exacerbated by the rise of generative Artificial Intelligence (AI), which can be exploited by malicious actors to create more convincing and sophisticated fraudulent schemes [9,10,19].

The proliferation of social media and online activities has made it easier for scammers to profile potential victims and target them with personalized fraudulent ads. *Social engineering* techniques, which involve manipulating individuals' emotions to induce them to divulge confidential information or perform actions compromising their security, empower scammers to approach victims, establish

M. Barhamgi et al. (Eds.): WISE 2024, LNCS 15439, pp. 230–245, 2025.
https://doi.org/10.1007/978-981-96-0573-6_17

trust, and ultimately direct them to fraudulent websites. This not only results in financial loss but also causes significant emotional distress.

The dire consequences of scams are evident in real-life examples. For instance, a recent *CNN* report highlighted the tragic case of an elderly person committing suicide due to an online investment scam.[1] Additionally, some victims who have been scammed become targets of a secondary scam, where malicious actors offer fake legal service, claiming to recover their lost money. The scale of scams is also staggering. According to the *Canadian Anti-Fraud Centre*, a total of 569 million dollars was reported lost to fraud in 2023, representing a 48% increase compared to the 383 million dollars in 2021.[2]

Traditional detection systems have relied on blacklist-based approaches, where known phishing and scamming URLs are stored and used for comparison. However, these systems struggle to identify new and emerging phishing sites that have not yet been reported, leading to significant gaps in detection capabilities [21]. To address these challenges, this paper proposes leveraging *Large Language Models* (LLMs), which are advanced AI systems trained to understand and generate human-like language, for scam website detection. The advent and increased accessibility of LLMs provide an opportunity for real-time, expert-level responses.

The contributions of this paper are as follows:

1. A comparative analysis of two approaches for scam website detection: text-based and screenshot-based methods. This comparison evaluates the performance of each approach in identifying scam websites.
2. An extensive evaluation of five cutting-edge LLMs, including *GPT-4o*, *Copilot*, *Gemini*, *Perplexity*, and *Claude-3.5*. The study assesses and compares their performance in identifying scam websites.
3. A detailed analysis of the explanations provided by the LLMs for their decisions on website legitimacy. This analysis includes a proposed novel categorization of the criteria used by the LLMs in their decision-making process.

By leveraging existing AI models to counteract the sophisticated tactics employed by scammers, this study aims to provide an accessible solution for the general public to address the growing problem of online scams.

The rest of the paper is organized as follows: Section 2 presents an overview of related work. Section 3 outlines the methodology used in this paper, the experiment setup and the evaluation criteria. Section 4 showcases the results of the experiments. Section 5 discusses the findings and the limitations of the paper. Finally, Sect. 6 concludes the paper.

2 Related Work

Internet scams encompass a wide range of fraudulent activities, with phishing being a prominent subset. While this paper focuses on the broader category

[1] https://www.cnn.com/2024/06/17/asia/pig-butchering-scam-southeast-asia-dst-intl-hnk/index.html.

[2] https://antifraudcentre-centreantifraude.ca/index-eng.htm.

of scam websites, much of the existing research has concentrated on phishing detection, which is directly applicable to the wider context of online fraud.

The application of machine learning algorithms to distinguish between legitimate and fraudulent websites has been widely studied. For instance, Duan *et al.* [6] demonstrated that a random forest model, following website data preprocessing, is efficient in identifying phishing websites. Jha *et al.* [11] proposed a pipelined model integrating logistic regression to achieve a real-time application without significant hardware resources. Deep learning (DL) approaches have also shown promise in fraudulent website detection by capturing both the textual and visual cues of websites. Almousa *et al.* [1] focused on hyperparameter optimization in DL models for achieving high accuracy in phishing website detection. Gopali *et al.* [8] compared the performance of different DL models in identifying phishing URLs. Nonetheless, the performance of these models heavily depends on the configuration of the parameters and the size of the training data.

Recent research has explored the use of LLMs to enhance the identification of fraudulent websites. Jiang [12] utilized supervised learning techniques to fine-tune *GPT-3.5* and *GPT-4*, highlighting the potential application of LLMs in scam detection. Mahendre and Pandit [15] showed that fine-tuning *DeBERTa* achieved a better result than *GPT-4* in detecting phishing attempts. Saha Roy *et al.* [19] leveraged LLMs to generate phishing websites, which was then fed back to train a *BERT*-based model to detect phishing prompts. Comparing the performance prompt engineering to fine-tuning LLMs, Trad and Chehab [22] showed that even though fine-tuning LLMs resulted in better performance, with specifically crafted prompts, chat models like *ChatGPT* and *Claude* can identify phishing URLs most of the time using the URLs alone. Boumber *et al.* [3] proposed using a Retrieval Augmented Generation framework for few-shot learning in domain-agnostic settings to enhance the effectiveness of LLMs in detecting deception. Koide *et al.* [13] demonstrated the efficacy of *ChatGPT* in accurately identifying impersonated brands and social engineering techniques in phishing site detection.

While significant progress has been made in scam website detection, many existing approaches require specialized training, fine-tuning, or complex implementations. In contrast, we aim to investigate whether powerful, pre-trained, subscription-free LLMs can be directly applied to scam website detection using information that the general public can easily obtain from web pages. This approach explores the potential for offering the public immediate access to an effective scam detection solution by leveraging the capabilities of these readily available LLMs.

3 Methodology

This paper employs two methods to evaluate LLMs' capability in detecting scam websites, text-based and screenshot-based. During the initial data collection for the text-based method, it was found that some websites required *JavaScript* to load content, making the scraping method insufficient. Therefore, a screenshot-based detection approach was employed to capture the visual and structural

elements of these websites. This dual approach ensures a comprehensive evaluation of LLMs' capabilities in detecting scam websites. Fig 1 illustrates an overview of the methodology.

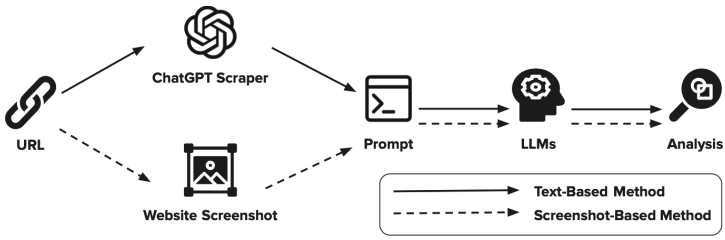

Fig. 1. An overview of the methodology used in this paper.

3.1 Data Collection

To construct a comprehensive dataset for our study on scam website detection, we sourced data from *PhishTank* and the *Global Anti-Scam Organization (GASO)*. *PhishTank* is a phishing intelligence source widely used in phishing detection research [5] and provides an **up-to-date** database of phishing websites. *GASO*, dedicated to combating online scams and fraud, provides an extensive list of scam websites.[3] These sources ensure the relevance and reliability of the collected data. To select legitimate websites, the *Tranco* list [14] generated on July 1, 2024, was utilized to ensure the inclusion of high-quality, trustworthy sites for comparison against scam sites. This list ranks the most popular websites and is widely used in research due to its reliability and comprehensive coverage of legitimate websites. To further enhance the dataset's relevance, the *Canadian Anti-Fraud Centre*'s Internet scam list[4] was cross-referenced to identify **recent** online scams. The taxonomy of fraud proposed by Button and Cross [4] was used to categorize the types of scams. The focus was specifically on two types:

1. **Investment Scam**: Scams that promise high returns on investments or low service fees, leading victims to provide sensitive information and invest money, which is then stolen. This includes fake investment opportunities, fraudulent digital wallets, and Ponzi schemes.
2. **Product and Service Scam**: Scams involving the offering of products or services fraudulently, often without intending to deliver the promised goods or services. This includes low-quality products and fake legal services.

[3] https://www.globalantiscam.org/scam-websites.
[4] https://antifraudcentre-centreantifraude.ca/scams-fraudes/medium-moyen-eng.htm.

A total of 400 websites were compiled, maintaining a 50–50 ratio of scam to legitimate sites. Scam websites were sourced equally from *PhishTank* and *GASO*.

For the text-based approach, the *Scraper* plugin[5] in *ChatGPT* was employed for its accessibility and ease of use. A simple prompt was used to query the model to collect real-time data from these sites, formulated as: `Extract text from {URL}`. For the screenshot-based approach, a screenshot of the first page each URL led to, along with the URL in the browser, was taken. Figure 2 provides an example of a screenshot of a scam website used for evaluation.

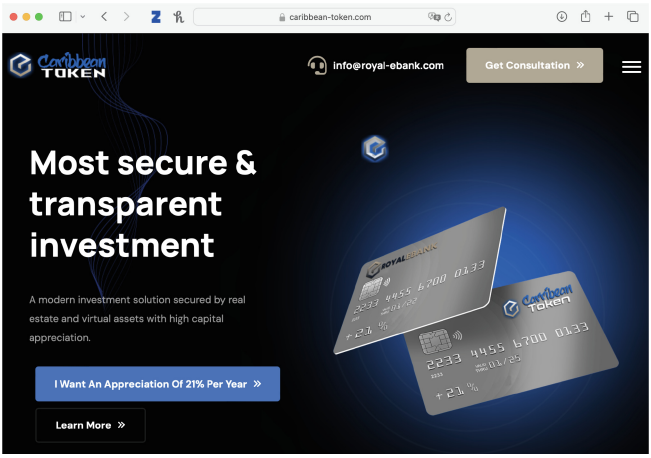

Fig. 2. A screenshot of an investment scam website.

3.2 Prompt Engineering

A simple, **generic prompt** was intentionally selected to evaluate the performance of the models with minimal guidance, thereby reflecting real-world scenarios where users may lack specialized expertise. For the text-based approach, the prompt template is as follows:

```
You're an expert in cybersecurity and cybercrime.
Decide whether the following website is a scam website
by analyzing the text extracted from the website and the URL.
The decision should be followed by a concise explanation.
URL = {url}
Text = {text}
```

This prompt guides the models in analyzing website content and determining its legitimacy, ensuring that LLMs provide concise yet informative

[5] https://chatgpt.com/g/g-870r1buu6-scraper.

explanations for their decisions, which is essential for transparency and user trust. For the screenshot-based approach, a similar prompt template was used, replacing `the text extracted from the website` with `a screenshot of the website`, and `Text = {text}` with the screenshot manually attached.

3.3 Model Selection

Five cutting-edge LLMs were selected to evaluate their performance in detecting scam websites: OpenAI's *GPT-4o* [17], Microsoft's *Copilot* [16], *Perplexity*'s default model [18], Google's *Gemini* [7], and Anthropic's *Claude-3.5* [2]. The selection was based on:

1. **State-of-the-Art Performance**: These models represent the latest advancements in natural language processing, demonstrating outstanding performance in various tasks.
2. **Accessibility**: Each model provides a user-friendly interface easily accessible to the public, ensuring the findings are relevant and applicable to a wide audience. All selected models are accessible without subscription services, making them practical scam detection tools for the general public.
3. **Image Processing Capabilities**: All the models chosen provide an interface for image processing, enabling users to analyze visual elements of websites in addition to text content.

3.4 Environment and Interfaces

All experiments were conducted on a laptop running macOS, using Safari as the browser. Evaluations were conducted exclusively using the publicly accessible interfaces provided by each LLM to demonstrate the feasibility and robustness of deploying existing chat models as scam detection tools in everyday scenarios.

3.5 Evaluation Metrics

The selected models were evaluated using the following metrics:

1. **Accuracy**, the proportion of correctly identified scam and legitimate websites.
2. **Precision**, the ratio of correctly identified scam websites to the total identified as scams.
3. **Recall**, the ratio of correctly identified scam websites to the total actual scams in the dataset.
4. **F1 Score**, the harmonic mean of precision and recall, a balance between the two metrics.

3.6 Categorization of Explanations

The explanations provided by the LLMs for their decisions were analyzed to understand their decision-making process. A categorization of the explanations was then developed based on the most frequently used factors by the LLMs to justify their decisions. The categories are as follows:

1. **Suspicious Domain or URL**: Examines unusual or suspicious domain names and misspellings in URLs.
2. **Format**: Assesses poor formatting or grammar. For the screenshot-based method, this also includes the unprofessional design of the web page.
3. **Generic and Unverifiable Claims**: Looks for vague language and unverifiable claims used to lure victims. For example, claiming a product or service to be the "Best and Most reliable" without detailed information or customer reviews.
4. **Suspicious Offers**: Identifies unrealistic promises, such as high returns, low prices, and quick profits.
5. **Transparency**: Evaluates the availability and authenticity of information establishing website legitimacy. This includes:
   (a) **Suspicious Contact Information**: Emails provided not aligning with domain names, or suspicious addresses.
   (b) **Impersonation**: Attempts to deceive by appearing legitimate or mimicking popular businesses.
   (c) **Lack of Transparency**: Absence of clear information about the operations, regulations, registrations, location, or legal details.
6. **Urgency**: Detects emotion manipulation and pressure tactics to rush victims into hasty decisions.
7. **Testimonials**: Analyzes the nature and presence of testimonials. This includes:
   (a) **Suspicious or lack of testimonials**: While genuine websites often have verifiable and positive testimonials, scam websites may have no testimonials or rely on fake testimonials to build a false image of happy consumers.
   (b) **Negative testimonials**: Some models have the ability to search through websites and use negative feedback or reports found online to determine the scamming nature of a website.

4 Results

4.1 Performance Metrics

Table 1 summarizes the performance metrics for the text-based and screenshot-based method.

Table 1. Performance metrics for the text-based/screenshot-based method. For each metric, the best performance is underlined for the text-based method and **bolded** for the screenshot-based method. The performance metrics for *Perplexity* using screenshots are omitted as *GPT-4o* is its default model for image processing.

	GPT-4o	*Copilot*	*Perplexity*	*Gemini*	*Claude-3.5*
Accuracy	0.929/0.940	0.881/0.931	0.866/-	0.780/0.915	0.940/**0.945**
Precision	0.983/**1.000**	0.990/**1.000**	1.000/-	0.952/0.977	0.992/**1.000**
Recall	0.873/0.880	0.769/0.865	0.731/-	0.590/0.850	0.888/**0.890**
F1	0.925/0.936	0.866/0.928	0.845/-	0.728/0.909	0.937/**0.942**

Performance Metrics for Text-Based Detection. Among the models, *Claude-3.5* demonstrated the highest overall accuracy at 0.94, followed closely by *GPT-4o* at 0.929. *Copilot* and *Perplexity* exhibited moderate performance, while *Gemini* showed the lowest accuracy at 0.78. In terms of precision, all models performed exceptionally well, with *Perplexity* achieving a perfect precision score of 1, indicating no false positives. Regarding recall, *Claude-3.5* again led the models with a score of 0.888, indicating its effectiveness in identifying actual scam websites. *Gemini* had the lowest recall at 0.59, suggesting it missed a significant number of scams. The F1 score, which balances precision and recall, was highest for *Claude-3.5* at 0.937, followed by *GPT-4o* at 0.925. These results indicate that *Claude-3.5* and *GPT-4o* offer the most balanced and effective performance for text-based detection.

Performance Metrics for Screenshot-Based Detection. Among the models, *Claude-3.5* exhibited the highest accuracy at 0.945, closely followed by *GPT-4o* at 0.94. Both *Copilot* and *Gemini* showed slightly lower accuracy, with scores of 0.931 and 0.915, respectively. In terms of precision, *GPT-4o*, *Copilot*, and *Claude-3.5* achieved perfect scores of 1, indicating that these models did not produce any false positives. *Gemini*, while slightly lower, still demonstrated strong precision with a score of 0.977. For recall, all models performed moderately well, with *Claude-3.5* leading at 0.89, indicating its superior ability to identify actual scam websites. The F1 scores were highest for *Claude-3.5* at 0.942, and lowest for *Gemini* at 0.909. These results indicate that while all models performed well in detecting scam websites using screenshots, *Claude-3.5* was slightly more effective overall.

4.2 Analysis of Explanations

Each explanation was examined to identify which of the criteria listed in the previous section were mentioned. To account for the different correct rates of each model, we used ratios to evaluate the identification of criteria in correctly classified cases and misclassified cases. This allows for a fair comparison of the models' performance.

Analysis of Criteria in Correctly Identified Scam Websites for Text-Based and Screenshot-Based Detection. Table 2 summarizes the analysis of criteria mentioned in the explanations provided by the LLMs for correctly identified scam websites.

Table 2. Frequency of criteria mentioned in LLM explanations for correctly identified scam websites using the text-based/screenshot-based method. For each criterion, the model that uses it the most frequently is underlined for the text-based method and **bolded** for the screenshot-based method. The frequency for *Perplexity* using screenshots is omitted as *GPT-4o* is its default model for image processing.

	GPT-4o	*Copilot*	*Perplexity*	*Gemini*	*Claude-3.5*
Domain/URL	0.828/0.742	0.755/0.372	0.705/-	0.577/0.268	0.714/**0.966**
Format	0.586/0.708	0.112/0.640	0.473/-	0.297/**0.784**	0.399/0.648
Generic Claims	0.787/0.640	0.483/0.256	0.616/-	0.712/0.670	0.667/**0.705**
Suspicious Offers	0.763/0.528	0.734/0.558	0.568/-	0.721/0.536	0.702/**0.739**
Lack of Transparency	0.917/0.933	0.790/0.616	0.904/-	0.856/0.825	0.940/**0.955**
Urgency	0.272/0.461	0.070/0.070	0.171/-	0.171/0.021	0.488/**0.545**
Testimonials	0.254/0.056	0.231/**0.523**	0.247/-	0.324/0.113	0.208/0.034

GPT-4o and *Claude-3.5* demonstrated the highest frequency of identifying **suspicious domains or URLs** in the text-based method (82.8%) and the screenshot-based method (96.6%) respectively. This indicates their strong capability to recognize domain-related red flags, which is essential for identifying fraudulent websites.

Regarding **grammatical and formatting issues**, *GPT-4o* (58.6%) was the most attentive in the text-based method, suggesting its adeptness at catching language anomalies indicative of scams. In the screenshot-based method, *Gemini* (78.4%) and *GPT-4o* (70.8%) led in identifying poorly designed scam websites, including the use of generic stock images.

GPT-4o also led in identifying **generic and unverifiable claims** in the text-based method (78.7%), followed by *Gemini* (71.2%) and *Claude-3.5* (66.7%). In the screenshot-based method, *Claude-3.5* (70.5%) used this criterion the most, followed by *Gemini* (67%) and *GPT-4o* (64%), indicating their ability to recognize misleading claims in both textual and visual content.

In identifying **suspicious offers**, except for *Perplexity* (56.8%), all models showed high effectiveness in the text-based method, with *GPT-4o* (76.3%) leading. In the screenshot-based method, *Claude-3.5* (73.9%) was the most effective, highlighting its sensitivity to offers that seem too good to be true. Other models identified suspicious offers around 50% of the time.

All models consistently identified the **lack of transparency**, a key indicator suggesting the site's legitimacy and authority. In the text-based method, *Claude-3.5* (94%), *GPT-4o* (91.7%) and *Perplexity* (90.4%) showed the highest frequencies. Similarly, in the screenshot-based method, *Claude-3.5* (95.5%)

and *GPT-4o* (93.3%) performed exceptionally well. Compared to other models, *Copilot* used this criterion less frequently, justifying less than 80% of the true positives in both methods.

For **urgency** or emotional manipulation, *Claude-3.5* was the most attuned in both methods, with frequencies of 48.8% in the text-based method and 54.5% in the screenshot-based method. This was followed by *GPT-4o*, which identified pressure tactics in 27.2% of the true positives with the text-based method and 46.1% in the screenshot-based method. *Copilot* (7%) was the least effective in identifying pressure tactics in the text-based method, while *Gemini* (2.1%) was the least effective in the screenshot-based method.

Gemini used **testimonials** in its justification the most in the text-based method (32.4%), utilizing both overly positive feedback and negative reviews found online to assess credibility. All the other models performed similarly, using testimonials as justification for around 20% of the true positives. In the screenshot-based method, *Copilot* (52.3%) frequently referred to negative reviews associated with the website or similar websites in the decision-making process, while other models relied significantly less on this criterion.

Analysis of Criteria in False Negatives for Text-Based and Screenshot-Based Detection. Table 3 summarizes the frequency of criteria mentioned in LLM explanations for scam websites incorrectly classified as legitimate.

Table 3. Frequency of criteria mentioned in LLM explanations for text-based/screenshot-based detection false negatives. For each criterion, the model that uses it the most frequently is underlined for the text-based method and **bolded** for the screenshot-based method. The frequency for *Perplexity* using screenshots is omitted as *GPT-4o* is its default model for image processing.

	GPT-4o	*Copilot*	*Perplexity*	*Gemini*	*Claude-3.5*
Domain/URL	0.286/**0.833**	0.907/0.500	0.667/-	0.236/0.333	0.737/0.800
Format	0.667/**1.000**	0.233/0.900	0.400/-	0.127/0.400	0.421/**1.000**
Detailed Information	0.905/0.750	0.512/0.300	0.711/-	0.291/0.133	0.789/**0.900**
No Suspicious Offers	0.095/0.083	0.256/0.100	0.156/-	0.109/0.067	0.368/**0.200**
Transparency	1.000/0.833	0.884/0.400	0.978/-	0.382/0.733	0.895/**0.900**
No Pressure Tactics	0.000/**0.667**	0.233/0.000	0.044/-	0.018/0.000	0.000/0.000
Testimonials	0.286/0.083	0.140/**0.200**	0.222/-	0.200/0.067	0.211/0.100

Claude-3.5 frequently cited **the absence of suspicious domain or URL** features as a justification for false negatives, with frequencies of 73.7% in the text-based method and 80% in the screenshot-based method. *Copilot* also relied heavily on this criterion, with frequencies of 90.7% for the text-based method, which is the highest among all, and 50% in the screenshot-based method. For the text-based method, *GPT-4o* mentioned this criterion less frequently (28.6%), but

in the screenshot-based method, the absence of suspicious URLs and domains was the most used criterion among all models (83.3%). *Gemini* used this criterion less often overall.

For **professional design**, including the absence of grammatical and formatting issues, *GPT-4o* (66.7%) was the most sensitive in the text-based method. In the screenshot-based method, professional design and the absence of formatting issues were significant justifications, with *GPT-4o* and *Claude-3.5* citing this factor 100% of the time. This highlights the importance of professional appearance in falsely identifying legitimate websites.

Detailed information, such as product specifications, was frequently used as a justification for false negatives by *GPT-4o* (90.5%) in the text-based method, followed by *Claude-3.5* (78.9%) and *Perplexity* (71.1%). In the screenshot-based method, *Claude-3.5* (90%) and *GPT-4o* (75%) frequently used detailed information as a justification for a website's legitimacy. This suggests that detailed content can mislead models into perceiving a website as legitimate.

The **absence of suspicious offers** was not a major justification for any of the models. In the text-based method, *Claude-3.5* cited this criterion at 36.8% of the false negatives, followed by *Copilot* at 25.6%. In the screenshot-based method, it was cited 20% of the time by *Claude-3.5*, while all other models mentioned this criterion less than 10% of the cases.

Transparency, such as openness and accessibility of contact information, was a major criterion for most models. In the text-based method, *GPT-4o* and *Perplexity* cited transparency 100% and 97.8%, respectively. Apart from *Gemini*, all models used transparency to justify the legitimacy of more than 80% of the false negatives. In the screenshot-based method, *Claude-3.5* (90%) and *GPT-4o* (83.3%) frequently cited this criterion, while *Gemini* used it 73.3% of the time. *Copilot* (40%) used this criterion less often.

The **absence of pressure** tactics was least cited across all models for the false negatives. While *Copilot* used this criterion to justify the legitimacy of 23.3% of the false negatives in the text-based method, *GPT-4o* and *Claude-3.5* did not mention it at all. In the screenshot-based method, only *GPT-4o* frequently cited this factor (66.7%).

Testimonials, such as positive testimonials listed on the web page and reviews found on the Internet, were a relatively low justification across all models. In the text-based method, *GPT-4o* (28.6%) mentioned testimonials most frequently, while *Copilot* (14%) used it the least. In the screenshot-based method, *Copilot* (20%) cited testimonials more often than the other models, which mentioned this criterion in less than 10% of the false negatives.

5 Discussion

5.1 Discussion on Model Performance in Text-Based and Screenshot-Based Methods

For both methods, *Claude-3.5* achieved the highest accuracy and F1 score, closely followed by *GPT-4o*. This suggests that these models are highly effec-

tive at analyzing both textual and visual content to identify scam websites. The screenshot-based method generally resulted in higher accuracy and F1 scores compared to the text-based method, which can be attributed to the additional context provided by visual elements. Specifically, *Gemini* achieved an F1 score higher than 0.9 with the screenshot-based method, marking a significant improvement from its performance of 0.73 in the text-based method. This improvement is likely due to *Gemini*'s difficulty in processing long text content, which was mitigated by the visual data. These results suggest that incorporating visual elements of websites provides additional context that aids in more accurate identification of scams. The structure and layout of a web page offer crucial information that complements textual analysis, leading to a more comprehensive and effective detection process. The consistent improvement across all models when using the screenshot-based method underscores the importance of a multimodal approach in enhancing the accuracy and reliability of scam detection systems. Overall, this observation suggests that for general users, publicly available pre-trained LLMs can be an effective tool to quickly verify the legitimacy of websites with just a screenshot, helping them avoid common scams such as phishing or fraudulent e-commerce sites. Moreover, it indicates that for cybersecurity experts, LLMs can be integrated into scam detection systems, offering a scalable and cost-effective solution to monitor and mitigate potential threats.

Another observation from the experiments is that *Gemini* sometimes had trouble processing images containing people. When screenshots included human faces, *Gemini* often stated, *"I can't analyze the security of the website based on the image you sent me since it contains people,"* and relied solely on the URL to determine the website's legitimacy. In contrast, *Copilot* blurred these faces when processing such screenshots, stating, *"Analyzing the image: Privacy blur hides faces from Copilot,"* which may have contributed to its better performance in these cases. This further highlights the need to consider ethical implications when processing visual data. The ability of models like *Copilot* to process images with privacy blurring demonstrates a commitment to protecting user privacy while leveraging visual information for scam detection. It is essential to ensure that AI systems respect user privacy and comply with ethical standards, especially when handling sensitive data such as images containing personal information. This highlights the need for continuous development and implementation of privacy-preserving techniques in AI-driven security solutions.

5.2 Discussion on Model Explanations

Our results demonstrated that different models exhibit varying strengths and weaknesses in detecting scam websites. *Claude-3.5* and *GPT-4o* consistently showed high values across most criteria, indicating their robust capability to consider multiple indicators to support their decisions. Notably, *Claude-3.5* excelled at recognizing emotional manipulation tactics on web pages, using this criterion more frequently than any other model. This suggests its advanced contextual understanding and sensitivity to psychological cues often employed in scam websites.

All models exhibited proficiency in identifying a lack of transparency and the presence of generic claims as indicators of a scam, highlighting these as reliable markers in both detection methods. For the text-based approach, the frequent use of URL and domain suspicions as justifications aligns with previous research suggesting that URL characteristics alone can serve as effective inputs for classification. In contrast, the screenshot-based approach shifts the focus from URL and domain characteristics to the design and structure of the web page. In this method, models prioritize visual elements, providing crucial contextual information absent in text analysis alone. This shift underscores the importance of a multimodal approach in enhancing scam detection accuracy.

Analyzing the false negatives reveals that detailed information and transparency often contribute to the misclassification of scam websites as legitimate. For example, one screenshot of a scam website classified as legitimate by all the models included the company's contact information, which featured a generic, obviously made-up phone number "*123-456-789*". None of the models identified this fraudulent contact information and instead justified the website's legitimacy by stating that it was transparent. This raises concerns, as scammers can fabricate details to deceive the models, or even exploit LLMs to generate more convincing and misleading information [9,10]. Furthermore, the presence of positive testimonials, either on the web page or found through the models' web search capabilities, also contributed to false negatives. This is problematic as scammers can easily create fake reviews to build a false image of legitimacy [20].

Implications for future model improvements are evident from these observations. There is a need to enhance models' ability to discern fabricated details and improve their sensitivity to subtle indicators of fraud. Additionally, ethical considerations must be addressed. The same technologies used for detecting scams can be exploited by malicious actors to generate informative and seemingly legitimate page content. This dual-use nature of AI necessitates stringent guidelines and ethical standards to prevent misuse.

5.3 Limitations

Several limitations in this study should be acknowledged. Firstly, the binary classification approach used to determine whether a website is a scam may introduce ambiguity and uncertainties. Enhancing accuracy could involve employing a scoring system to quantify the likelihood of a site being a scam. Secondly, this paper focused on assessing LLMs' ability to detect scam attempts using a general prompt. This may have caused the LLMs to rely on information from their training data and may not be up-to-date with the latest scam tactics. Future work can explore advanced techniques to enhance the models' capabilities in identifying emerging scam patterns and conduct a longitudinal study to examine the models' performance over time. Thirdly, the focus was on specific types of scams, leaving many other scam types unexplored. Incorporating different types of scams across various sectors, such as government grants, employment, romance, or phantom debt collection scams [4], could provide a more comprehensive understanding. Lastly, the paper relied on single screenshots for the screenshot-based method.

Incorporating multiple images in future studies could capture a broader range of scam indicators and improve detection accuracy. These limitations highlight the potential for further enhancements and expansions in this area of research.

6 Conclusion

This paper addressed the growing problem of online scams by leveraging the capabilities of readily available LLMs. A comprehensive comparison was conducted between text-based and screenshot-based methods on five state-of-the-art LLMs — *GPT-4o*, *Copilot*, *Gemini*, *Perplexity*, and *Claude-3.5* - for scam website detection. The results demonstrated that the screenshot-based approach outperforms the text-based method, underscoring the importance of incorporating visual elements in identifying fraudulent websites. Moreover, a detailed analysis of the LLMs' explanations resulted in the development of a novel categorization of the criteria. Our findings revealed while different models exhibit varying strengths in detecting specific scam indicators, transparency was identified as a critical factor for all LLMs in determining a website's legitimacy.

The analysis of the performance and explanations identified significant limitations in the current LLMs for scam detection. Models showcased different behaviours in processing images with potentially sensitive information, emphasizing the importance of implementing privacy-preserving techniques in AI-driven security solutions. Furthermore, detailed information and misleading testimonials, which could be fabricated by scammers or easily generated using AI technologies, often resulted in the misclassification of scam websites as legitimate. This further highlights the dual-edged nature of AI and the importance of developing ethical frameworks alongside technological advancements.

In conclusion, this paper, as a proof of concept, highlights the potential of leveraging accessible LLMs to counteract the sophisticated schemes employed by scammers, demonstrating a promising approach to real-time scam detection accessible to the general public. Future research should focus on addressing the identified limitations to enhance the effectiveness and reliability of scam detection systems, and incorporate user-centric evaluations to assess the practical applicability and user-friendliness of the proposed methods. By addressing these aspects, the proposed solution can offer a more effective defense against the evolving tactics of cybercriminals.

References

1. Almousa, M., Zhang, T., Sarrafzadeh, A., Anwar, M.: Phishing website detection: how effective are deep learning-based models and hyperparameter optimization? Secur. Priv. **5**(6), e256 (2022). https://doi.org/10.1002/spy2.256
2. Anthropic: Claude 3.5 sonnet (2024). https://www.anthropic.com/news/claude-3-5-sonnet. Accessed 9 Jul 2024

3. Boumber, D., Tuck, B.E., Verma, R.M., Qachfar, F.Z.: LLMs for explainable few-shot deception detection. In: Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics, pp. 37–47. IWSPA '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3643651.3659898
4. Button, M., Cross, C.: Cyber Frauds, Scams and Their Victims. Routledge, Taylor & Francis Group, London; New York (2017)
5. Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., Shukla, S.: Applications of deep learning for phishing detection: a systematic literature review. Knowl. Inf. Syst. **64**(6), 1457–1500 (2022). https://doi.org/10.1007/s10115-022-01672-x
6. Duan, C., Wang, M., Lu, X., Wang, J.: A phishing website detection system based on machine learning methods. Acad. J. Comput. Inf. Sci. **6**(5) (2023). https://doi.org/10.25236/AJCIS.2023.060512
7. Google: Gemini (2024). https://gemini.google.com. Accessed 9 Jul 2024
8. Gopali, S., Namin, A.S., Abri, F., Jones, K.S.: The performance of sequential deep learning models in detecting phishing websites using contextual features of URLs. In: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, pp. 1064–1066. ACM, Avila Spain (Apr 2024). https://doi.org/10.1145/3605098.3636164
9. Gressel, G., Pankajakshan, R., Mirsky, Y.: Discussion paper: Exploiting LLMs for scam automation: a looming threat. In: Proceedings of the 3rd ACM Workshop on the Security Implications of Deepfakes and Cheapfakes, pp. 20–24. ACM, Singapore Singapore (Jul 2024). https://doi.org/10.1145/3660354.3660356
10. Gupta, M., Akiri, C., Aryal, K., Parker, E., Praharaj, L.: From chatGPT to threatGPT: impact of generative AI in cybersecurity and privacy. IEEE Access **11**, 80218–80245 (2023). https://doi.org/10.1109/ACCESS.2023.3300381
11. Jha, A.K., Muthalagu, R., Pawar, P.M.: Intelligent phishing website detection using machine learning. Multimedia Tools Appl. **82**(19), 29431–29456 (2023). https://doi.org/10.1007/s11042-023-14731-4
12. Jiang, L.: Detecting scams using large language models (2024). https://arxiv.org/abs/2402.03147
13. Koide, T., Fukushi, N., Nakano, H., Chiba, D.: Detecting phishing sites using chatGPT (2024). https://arxiv.org/abs/2306.05816
14. Le Pochat, V., Van Goethem, T., Tajalizadehkhoob, S., Korczyński, M., Joosen, W.: Tranco: a research-oriented top sites ranking hardened against manipulation. In: Proceedings of the 26th Annual Network and Distributed System Security Symposium. NDSS 2019 (Feb 2019). https://doi.org/10.14722/ndss.2019.23386
15. Mahendru, S., Pandit, T.: Securenet: a comparative study of deBERTa and large language models for phishing detection (2024). https://arxiv.org/abs/2406.06663
16. Microsoft: Copilot (2024). https://www.microsoft.com/en-us/copilot. Accessed 9 Jul 2024
17. OpenAI: Hello gpt-4o (2024). https://openai.com/index/hello-gpt-4o/. Accessed 9 Jul 2024
18. Perplexity: perplexity AI (2024). https://www.perplexity.ai/hub/technical-faq/what-model-does-perplexity-use-and-what-is-the-perplexity-model. Accessed 9 Jul 2024
19. Roy, S.S., Thota, P., Naragam, K.V., Nilizadeh, S.: From chatbots to phishbots? – preventing phishing scams created using chatGPT, google bard and claude (2024). https://arxiv.org/abs/2310.19181

20. Salminen, J., Kandpal, C., Kamel, A.M., gyo Jung, S., Jansen, B.J.: Creating and detecting fake reviews of online products. J. Retail. Consum. Serv. **64**, 102771 (2022). https://doi.org/10.1016/j.jretconser.2021.102771
21. Tang, L., Mahmoud, Q.H.: A survey of machine learning-based solutions for phishing website detection. Mach. Learn. Knowl. Extr. **3**(3), 672–694 (2021). https://doi.org/10.3390/make3030034
22. Trad, F., Chehab, A.: Prompt engineering or fine-tuning? a case study on phishing detection with large language models. Mach. Learn. Knowl. Extr. **6**(1), 367–384 (2024). https://doi.org/10.3390/make6010018