

Detecting Phishing Sites Using ChatGPT

목차

1. introduction
2. background
3. method
4. dataset
5. evaluation
6. limitations
7. conclusion

Introduction

ChatGPT 연구는 악의적인 웹 콘텐츠, 특히 피싱 사이트의 분석 및 탐지에 모델을 활용하는 데는 거의 주의를 기울이지 않음.

소셜 엔지니어링 기법(SE)을 사용하는 피싱 사이트는 합법적인 플랫폼으로 가장하여 민감한 정보를 노출하거나 금전적 피해를 입히도록 속임으로써 인터넷 사용자에게 심각한 위협이 된다.

LLM에 의해 촉진되는 자동화된 사이버 공격의 증가하는 위협에 효과적으로 대처하려면 **악성 웹 콘텐츠 탐지를 자동화하는 것이 중요하다**. 이를 위해서는 **LLM의 힘을 활용하여 피싱 사이트를 분석하고 분류하는 방법론을 탐구**해야 한다.

<contribution>

- **ChatGPT를 사용하여 피싱 사이트를 탐지하는 새로운 방법을 제안**한다. 웹 크롤링 기술을 활용하여 웹 사이트에서 정보를 수집하고 프롬프트를 생성함으로써 ChatGPT는 웹 사이트를 피싱으로 분류할 수 있다.
- 평가에서 **98.3%의 precision**와 **98.4%의 recall**로 높은 성능을 보여주며 피싱 사이트를 식별하는 LLM의 성능을 보여준다.
- GPT-3.5와 GPT-4 사이에 비교 분석을 수행했고 피싱 사이트를 식별하는 GPT-4의 능력, 특히 FN을 최소화하는 데 있어 눈에 띄는 향상을 발견했다. 의심스러운 도메인을 식별하고, 웹 콘텐츠에서 SE 기법을 탐지하며, 여러 요소를 고려해 포괄적인 피싱 탐지 기능을 제공하는 데 탁월했다.

Background

피싱 사이트 : 사용자를 심리적으로 유인하여 개인 정보나 돈을 훔치거나 악성 프로그램 감염을 유발하는 것을 목적으로 하는 사기성 웹 사이트

피싱 사이트 구성

1. 도메인 이름 모방과 공식 로고 및 브랜딩을 사용하여 사용자가 합법적인 서비스와 상호 작용하고 있다고 믿게 하는 것
2. SE 기술을 사용하여 사용자를 속이고 행동을 조작하는 것

기존 피싱 사이트 탐지 연구의 한계점

1. 피싱 사이트에 따라 대상 브랜드를 학습하고 알고리즘을 수정해야 한다.
2. 사회공학이 유발하는 심리적 조작의 맥락을 상세하게 분석할 수 없다.

Method

광범위한 텍스트 데이터에 대한 교육을 받은 ChatGPT를 사용하여 다양한 피싱 사이트를 탐지할 수 있다.

- **웹 크롤러**를 사용하여 입력 URL에 액세스하고 방문한 웹 사이트에서 스크린샷 이미지, HTML, URL 등의 정보를 얻습니다.
- 이 정보는 ChatGPT에 입력하는 **프롬프트 생성**하는 데 사용되며, 사이트가 피싱 사이트인지 아니면 비피싱 사이트인지 확인합니다

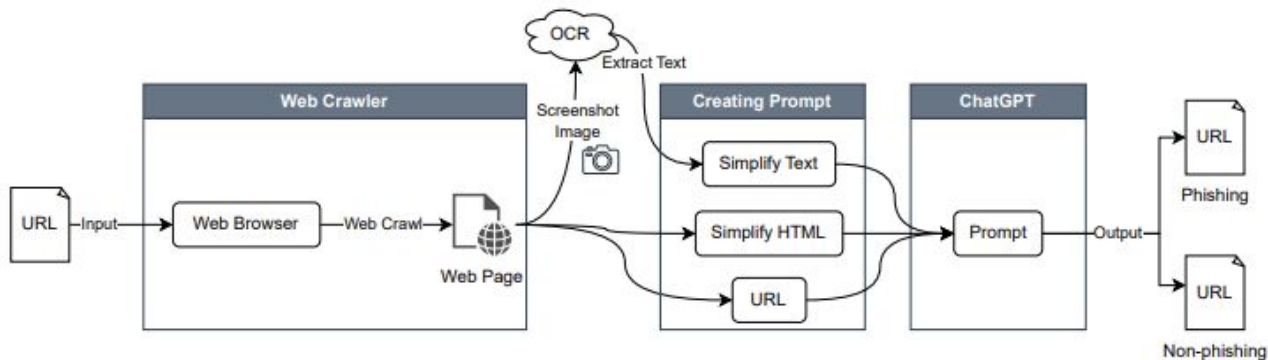


Figure 1: Overview of Proposed Method.

Method

1) Web Crawling

Chrome DevTools Protocol 로 **Google Chrome**을 자동화하여 웹 사이트를 방문하고 정보를 수집하는 웹 크롤러를 구현

- 입력 URL이 지정되면 웹 크롤러는 도달한 웹 페이지(리디렉션 후 최종 대상)의 URL, JavaScript 실행 후 HTML을 검색하고 스크린샷 이미지를 캡처한다.
 - javascript 를 실행 후 html 을 검색하는 이유 : 난독화된 JavaScript를 사용하여 DOM 요소를 생성하여 분석을 회피하기 위한 수단으로 사용하는 피싱 사이트가 있기 때문!
- Windows + Chrome 및 iPhone + Safari 의 두가지 환경을 emulate 하도록 크롤러 구성

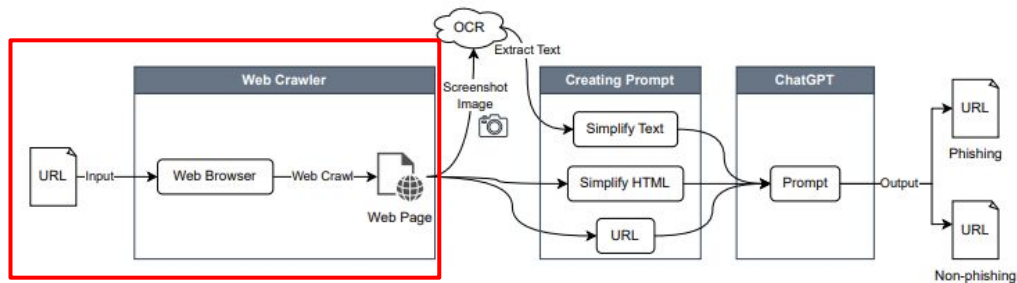


Figure 1: Overview of Proposed Method.

Method

2) Prompt

- 프롬프트의 목적 : ChatGPT에 웹 사이트 정보를 제공하고 해당 사이트가 피싱 사이트인지 여부를 확인하는 것
- Chain of Thought (CoT) prompting technique 을 기반으로 프롬프트 템플릿을 만듦.
 - CoT : LLM 이 그들의 추론을 설명하도록 장려하게 만드는 프롬프팅 기술, 다양한 추론 작업에서 성능을 향상시키는 데 효과적
- OCR(Optical Character Recognition) 사용

HTML에서 추출할 수 없는 텍스트를 프롬프트에 입력하기 위해 텍스트 추출을 위해 OCR을 사용하여 웹 사이트 정보를 프롬프트에 제공

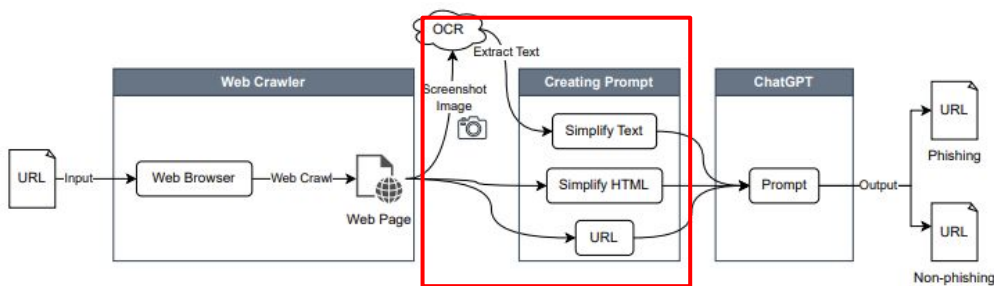


Figure 1: Overview of Proposed Method.

Method

2) Prompt

피싱 사이트 탐지 작업

1. 웹사이트에 사용자를 속이거나 유인하는 SE 기법이 포함되어 있는지 분석

- 상금, 가짜 악성 프로그램 감염 경고, 계정 문제 및 우편 소포 문제와 같이 피싱 사이트에서 일반적으로 사용되는 일반적인 **SE 기술을 제공**
- 비로그인 상태에서 컨텍스트가 없는 이러한 구문이 있으면 피싱 사이트의 가능성이 높다는 것을 나타냅니다.

2. 웹 사이트의 브랜드 이름을 추출

- HTML만으로는 정품 여부를 판단하기 어렵습니다. 따라서 LLM은 URL이 합법적인 사이트에 해당하는지 확인하라는 메시지를 표시합니다. → 브랜드 이름을 추출하여 해당 사이트의 URL 이 정규 도메인인지 확인하기 위한 단계

3. 사이트가 피싱 사이트인지 여부를 확인하고 그 이유를 명시적으로 설명한다.

4. JSON 형식으로 출력을 생성

suspected_domain 값을 true 로 출력, phishing_score 를 0-10 까지 생성

Prompt Template 1

You are a web programmer and security expert tasked with examining a web page to determine if it is a phishing site or a legitimate site. To complete this task, follow these sub-tasks:

1. Analyze the HTML, URL, and OCR-extracted text for any SE techniques often used in phishing attacks. Point out any suspicious elements found in the HTML, URL, or text.
2. Identify the brand name. If the HTML appears to resemble a legitimate web page, verify if the URL matches the legitimate domain name associated with the brand, if known.
3. State your conclusion on whether the site is a phishing site or a legitimate one, and explain your reasoning. If there is insufficient evidence to make a determination, answer "unknown".
4. Submit your findings as JSON-formatted output with the following keys:
 - phishing_score: int (indicates phishing risk on a scale of 0 to 10)

- brand: str (identified brand name or 'None' if not applicable)
- phishing: boolean (whether the site is a phishing site or a legitimate site)
- suspicious_domain: boolean (whether the domain name is suspected to be not legitimate)

Limitations:

- The HTML may be shortened and simplified.
- The OCR-extracted text may not always be accurate.

Examples of social engineering techniques:

- Alerting the user to a problem with their account
- Offering unexpected rewards
- Informing the user of a missing package or additional payment required
- Displaying fake security warnings

URL:

{URL}

HTML:

...

{Browser-rendered HTML}

...

Text extracted using OCR:

...

{OCR-extracted text}

...

Method

2) Prompt

- token : ChatGPT의 토큰은 문자나 단어를 나타내며 LLM에 의한 처리 및 분석에 사용되는 개별 텍스트 단위를 나타냄

GPT-3.5 token limit : 4096 , GPT-4 token limit : 8000

=> 본 연구에서는 HTML 및 OCR 추출 텍스트를 4,096개 토큰 내에 맞도록 단순화시킴.

Table 1: Token Counts for Prompt Template, HTML, OCR-extracted Text, and URL

Data	Maximum Token Count
Prompt Template	362 (fix)
HTML	3,000
OCR-extracted Text	500
URL	234
Total	4,096

프롬프트 템플릿은 362개의 토큰을 사용하며 HTML의 경우 최대 3,000개의 토큰, OCR 추출 텍스트의 경우 최대 500개의 토큰을 사용합니다. URL 은 단순화하지 않고 활용.

Method

3) Simplifying HTML & OCR-extracted Text

- **HMTL 단순화**

HTML 를 단순화하기 위한 프로세스에 대해 설명 피싱 탐지 및 브랜드 식별을 위한 필수 정보를 유지하면서 **HTML**을 단순화

HTML의 맨 위에 위치한 요소(제목, 메타 설명 요소, 로그인 양식 페이지 텍스트)의 정보의 길이를 최소화하면서 정보의 최대 보존을 달성하기 위해 알고리즘 1에 설명된 단계를 따름.

- **OCR-extracted text 단순화**

OCR에서 식별된 텍스트의 글꼴 크기를 검사하고 가장 작은 글꼴 크기부터 시작하는 문장을 제거

텍스트의 토큰 수가 500개 미만으로 떨어질 때까지 반복

Method

3) Simplifying HTML & OCR-extracted Text

Algorithm 1 Simplifying HTML

Require: inputHTML: Input HTML

Ensure: processedHTML: Processed HTML

```
1: function SIMPLIFYHTML(inputHTML)
2:   Remove style, script, and comment elements from
   inputHTML
3:   processedHTML  $\leftarrow$  Result after removal
4:   if lengthToken(processedHTML) < 3000 then
5:     return processedHTML
6:   end if
7:   Unwrap elements except for important tags
8:   Remove elements without text content
9:   Shorten href in a tags and src in img tags
10:  processedHTML  $\leftarrow$  Result after removal
11:  if lengthToken(processedHTML) < 3000 then
12:    return processedHTML
13:  end if
14:  while lengthToken(processedHTML) > 3000 do
15:    Remove an HTML element from the midpoint of the
    processedHTML
16:    processedHTML  $\leftarrow$  Result after removal
17:  end while
18:  return processedHTML
19: end function
```

<알고리즘 1>

1. 스타일, 스크립트, 주석 태그 요소 제거 → 중요하지 않은 요소 제거
2. 헤드, 제목, 메타, 본문, h1, h2, h3, h4, h5, h6, p, strong, a, img, hr, tbody, tr, td, ol, ul, li, ruby, label로 나열된 **중요 태그 이외의 HTML 요소를 푼다.** → 하위 요소 상위 요소에 통합됨.
3. 태그를 포함하지 않는 HTML 요소를 제거
4. **base64**로 인코딩된 **src** 속성 **img** 요소와 긴 URL을 포함하는 요소의 **href** 속성을 줄인다.
5. 토큰 카운트가 **3000** 미만으로 떨어질 때까지 중간 HTML 요소를 반복적으로 제거

Dataset

실험에 활용된 데이터 세트에는 각각 1,000개의 피싱 사이트와 비피싱 사이트가 포함됩니다.

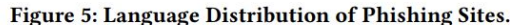
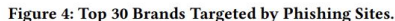
1) Phishing sites

피싱 사이트를 수집하기 위해 phishing intelligence source(**OpenPhish, PhishTank**) 와 CrowdCanary 사용

- CrowdCanary : 머신러닝 모델을 사용하여 **트위터에서 피싱 관련 게시물을 추출하는 방법**
 - '피싱', '스캠' 등 **키워드 기반 필터를 활용**해 피해자가 접하고 전문가가 관찰한 피싱 신고와 관련된 **게시물에서 URL을 추출**
- 2023년 3월부터 4월까지 **URL**을 수집하여 시드 **URL**로 사용
- 앞의 웹 크롤러를 사용하여 시드 **URL**에 액세스
- 두가지 요소에 집중하여 웹사이트의 스크린샷 이미지에 대한 철저한 조사를 실시
 - 잘 알려진 서비스와 관련된 로고 또는 브랜드 이름을 표시하거나 합법적인 사이트와 시각적으로 유사한 웹 사이트
 - 긴급성 또는 관심을 유발하기 위해 거짓 정보(예: 가짜 보상, 바이러스 감염 경고, 계정 문제)를 제시하는 것과 같은 **SE** 기술을 사용하는 웹 사이트.
- 수집된 웹 사이트 정보에서 다음 조건을 충족하는 웹 사이트는 제외
 - ChatGPT의 콘텐츠 필터와 충돌 가능성으로 인해 **R18**(도박, 도박, 데이트) 콘텐츠가 포함된 웹 사이트
 - 추출된 피싱 사이트 중 유사한 스크린샷이 발견된 경우는 **1개**를 제외하고 모두 제외

⇒ 위의 과정을 거쳐 **고유한 1,000개의 FQDN(정규화된 도메인 이름)**을 가진 **1,000개의 피싱 사이트 데이터** 세트를 얻음!

1) Phishing sites

[illegible]

데이터셋 내에서 피싱 사이트가 많이 사용하는 상위 브랜드 30

- 언어 탐지 라이브러리인 langdetect 를 사용하여 OCR 추출 텍스트의 언어를 식별하여 1,000개의 피싱 사이트가 22개의 다른 언어로 분산되었음을 밝힘.

Dataset

2) non-phishing sites

피싱 사이트와 동일한 수의 웹 사이트를 수집. (1000개)

- **seed URL** : 피싱 사이트가 대상으로 하는 **153개** 브랜드 웹 사이트, 트랜스코 목록의 상위 **2,000개** 도메인 이름으로 구성. → **236** 개
- 웹 크롤러를 사용하여 시드 URL에 액세스 ⇒ **1,661개**의 URL에 성공적으로 접속
 - 접속된 URL 중 포르노 **28**개, 도박 **3**개, 불법 **9**개 사이트는 제외
- **764**개는 트랜코 상위 사이트에서 무작위로 선택

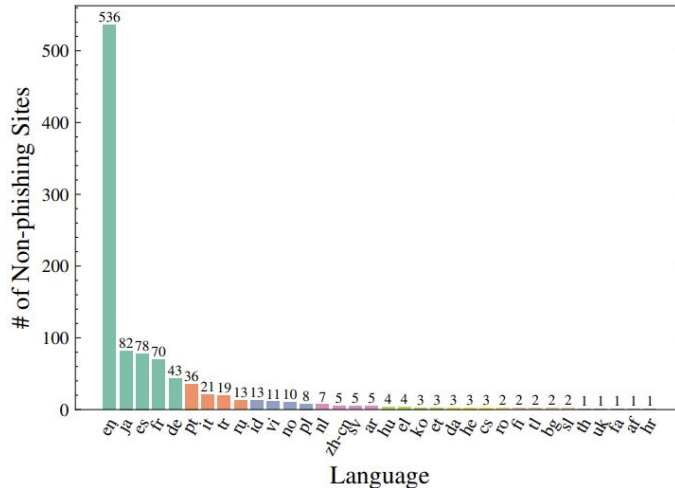


Figure 6: Language Distribution of Non-phishing Sites.

비피싱 사이트들은 **34**개 언어에 걸쳐 분포되어 있음

Dataset

3) Simplifying HTML and OCR-Extracted Text

앞에서 설명한 과정을 이용하여 피싱 사이트와 비피싱 사이트의 수집된 HTML과 OCR 추출 텍스트를 단순화했다.

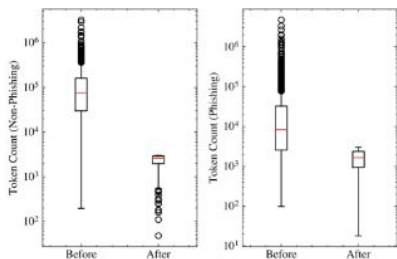


Figure 2: Number of Tokens Before and After Simplified HTML

HTML 단순화 전, 후의 토큰 수 변화
(왼쪽 non-phishing, 오른쪽 phishing)

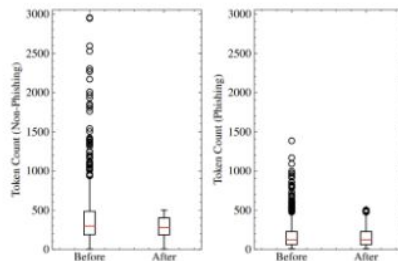


Figure 3: Number of Tokens Before and After Simplified OCR-extracted Text

OCR 추출 텍스트 단순화 전, 후의 토큰 수 변화

비피싱 사이트의 토큰 수가 더 많은 이유 :

정상 non-phishing 사이트는 복잡한 자바스크립트 코드를 통해 다양한 정교한 기능을 구현하거나 대형 플랫폼을 사용하는 경우가 많은 반면, 피싱 사이트는 사용자를 속이거나 민감한 정보를 훔치는 데 필요한 최소한의 기능만 가지고 있기 때문

Evaluation

GPT-3.5 와 GPT-4 두 개의 모델을 활용하여 평가함.

탐지 정확도를 비교하기 위해 동일한 프롬프트를 사용하여 데이터 세트의 각 URL에 대한 두 모델의 응답을 분석했습니다.

1) Summary of result

GPT 의 출력에서 JSON 응답의 **phishing key** 또는 **suspected_domain key** 값이 참인 경우 응답을 피싱으로 정의함, 두 키가 모두 거짓이면 비피싱으로 분류

Table 2: Confusion Matrix for GPT-3.5

		Predicted	
		Phishing	Non-phishing
Actual	Phishing	TP 867	FN 133
	Non-phishing	FP 15	TN 985

Table 3: Confusion Matrix for GPT-4

		Predicted	
		Phishing	Non-phishing
Actual	Phishing	TP 984	FN 16
	Non-phishing	FP 17	TN 983

Table 4: Performance Metrics Comparison

	GPT-3.5	GPT-4
Precision	98.3%	98.3%
Recall	86.7%	98.4%
Accuracy	92.6%	98.4%
F-measure	92.1%	98.4%

Evaluation

2) Examples of ChatGPT's Responses

< Phishing Example A >

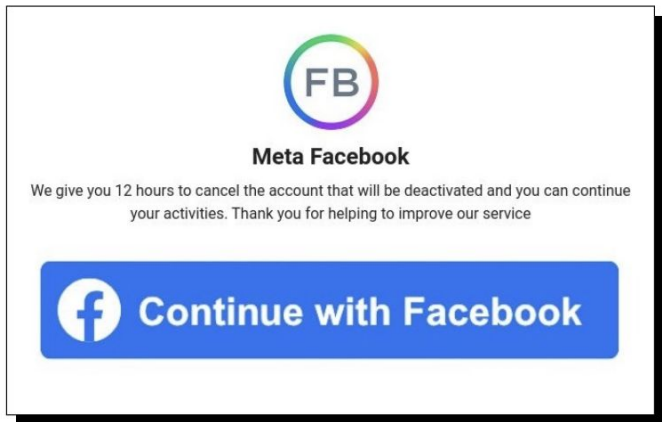


Figure 8: An Example of Facebook Phishing Site.

이 웹사이트는 페이스북을 사칭하고 사용자의 계정 문제에 대한 절박감을 조성하려고 함.

Response 1

1. Suspicious elements:

- The URL contains a subdomain "securitypages", which is unrelated to the legitimate Facebook domain (www.facebook[.]com).
- The extracted text alerts the user about having 12 hours to cancel the account. This is a social engineering technique often used in phishing attacks.
- Furthermore, the link provided redirects to a suspicious, non-legitimate URL: "http://centrenotificationvice.s.duckdns[.]org/centre.html".

2. The brand name identified is Meta Facebook.

3. The site appears to be a phishing site because of the suspicious URL and social engineering tactics in the text.

4. JSON-formatted output:

```
...
{
  "phishing_score": 9,
  "brands": "Meta Facebook",
  "phishing": true,
  "suspicious_domain": true
}
...
```

- ChatGPT는 이 웹 사이트의 브랜드 이름을 "Meta Facebook"로 식별하고 도메인 이름이 합법적인 Facebook이 아니라고 판단한다.
- ChatGPT는 SE 기법의 존재를 인식하고 HTML 내에 의심스러운 링크의 존재를 강조한다.
- ChatGPT는 phishing_score를 9로 설정하여 phishing 과 suspicious_domain을 모두 true로 표시하였다.

Evaluation

2) Examples of ChatGPT's Responses

< Phishing Example B >

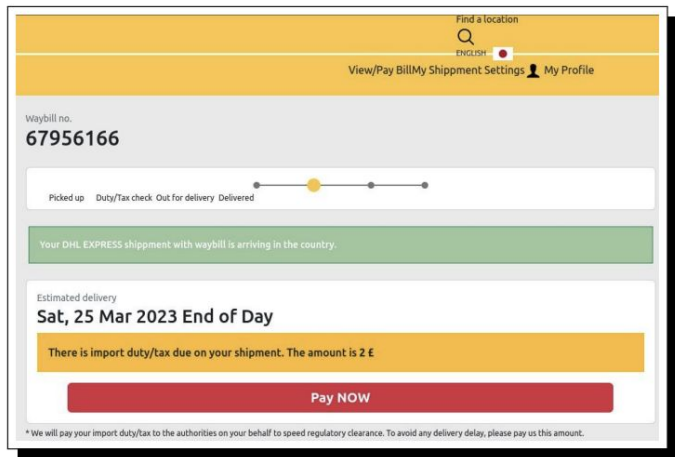


Figure 9: An Example of DHL Phishing Site.

DHL을 사칭한 피싱 사이트를 보여주며, 사용자를 속여서 금융 정보를 제공하도록 하는 경우를 보여줍니다.

Response 2

Based on my analysis of the HTML, URL, and OCR-extracted text, here are my findings:

1. Suspicious elements in the HTML and OCR-extracted text:

a. The site prompts the user to pay an import duty/tax amounting to 2€ on their shipment, which could be a social engineering technique to trick the user into providing their payment information.

b. The HTML structure has a card data input form where the user is expected to enter their card details, such as card number, expiry date, CVC number, and cardholder's name. This could be an attempt to steal sensitive card information.

2. The brand name identified is "DHL EXPRESS".

3. The URL ([https://siges.com\[.\]br/service/](https://siges.com[.]br/service/)) does not match the legitimate domain name associated with the DHL brand, which is "dhl.com". This mismatch is a red flag as it is not coming from the legitimate source.

Considering my findings, I conclude that this site is a phishing site attempting to steal user payment information by posing as a DHL EXPRESS page.

Here's the JSON-formatted output of my findings:

```
...
{
  "phishing_score": 9,
  "brands": "DHL EXPRESS",
  "phishing": true,
  "suspicious_domain": true
}
...
```

- ChatGPT 는 브랜드명을 DHL로 정확하게 인식하고 수입관세/세금비용 지불을 요청하는 SE기법을 정확하게 파악한다.
- 도메인 이름이 합법적인 DHL 웹사이트에 속하지 않는다고 판단한다.
- ChatGPT는 HTML 구조를 분석하여 카드 번호, CVC 번호 등 신용카드 정보를 도용하려는 피싱 사이트의 구조에 대한 상세한 통찰력을 제공한다.

Evaluation

3) Evidence for Identifying Phishing Sites

피싱 사이트를 탐지하는 데 사용되는 **7가지 증거 유형** 각각에 대해 해당 **SE 기술의 특성과 ChatGPT가 식별한 방법**을 설명한다.

1. **도메인 이름이 합법적이지 않음** : 검사를 받은 웹 사이트가 합법적인 브랜드를 사칭하지만 공식 도메인 이름과 일치하지 않는 도메인 이름을 가지고 있는 경우 => ChatGPT에 HTML 또는 OCR 추출 텍스트에서 브랜드 이름을 추출하여 합법적인 도메인 이름과 비교하도록 지시
2. **가짜 바이러스 감염 경고**
3. **가짜 계정 이슈** : 로그인 계정을 도용하기 위해 사용자 계정에 문제가 있다 경고를 표시하는 경우 => OCR을 사용하여 추출된 데이터에 사용자에게 계정 문제를 경고하는 텍스트가 있는지 확인
4. **급한 지불 요청** => 도메인 이름의 불일치와 '배달 중단 및 추가 요금' 같은 문구의 존재를 근거로 이를 피싱 사이트로 식별
5. **가짜 로그인 에러** : 입력 안했는데도 로그인 에러 뜨거나 잘못된 신용카드 정보를 입력했다는 경고가 뜨는 경우 => 'wrong credential' 이나 'invalid username or password' 같은 문구의 존재를 근거로 식별
6. **가짜 보상** => "get reward up to 600,000 RUB." 와 같은 문구의 존재를 근거로 식별
7. **SMS 인증 요청**

Evaluation

4) Comparative Analysis of Phishing Site Detection: GPT-4 vs. GPT-3.5

피싱사이트를 식별하는 성능에 대해 GPT-4와 GPT-3.5의 기능을 비교 분석함.

GPT-3.5 이 탐지하지 못한 사이트 중 GPT-4가 정확하게 파악한 피싱 사이트는 총 121개

이 사이트들을 분석한 결과 GPT-3.5에 비해 GPT-4가 우수한 성능을 보이는 세 가지 기능을 발견했다:

- 도메인 이름의 신뢰성을 확인하는 기능.
- SE 기술을 인식하는 능력.
- 여러 요소를 활용하여 피싱 사이트를 포괄적으로 탐지하는 기능

이유 :

- ChatGPT-3.5 의 훈련 자료가 2021년 9월로 거슬러 올라가는 데다 올해에 대한 지식이 부족해 부정확한 결과가 나올 수 있기 때문이다.
- GPT-4는 때때로 관련이 없는 관측치를 제공하지만 이를 보다 신뢰할 수 있는 정보와 결합하여 전체적으로 정확한 결정을 내리는 경향이 있다.

Limitations

- ChatGPT는 출력이 확률적으로 결정되기 때문에 유연하고 다양한 반응을 제공합니다. 그러나 이는 실험에 따라 탐지 결과가 달라질 수 있음을 의미하기도 한다
- 제안된 접근법을 포함하여 LLM을 사용하는 방법은 원래 프롬프트를 덮어쓰고 악성 콘텐츠가 응답에 삽입되는 프롬프트 인젝션에 취약할 수 있다.
- ChatGPT는 2021년 9월까지 데이터 교육을 받았기 때문에 해당 날짜 이후 생성된 서비스와 관련된 피싱 사이트와 비피싱 사이트를 정확하게 분류하지 못할 수 있다.

Conclusion

- 본 연구에서는 ChatGPT를 사용하여 피싱 사이트를 탐지하는 새로운 방법을 제시했다.
- 98.3%의 정밀도와 98.4%의 리콜로 놀라운 성능을 달성하여 피싱 사이트를 효율적으로 탐지하고 소셜 엔지니어링 기술을 발견할 수 있는 LLM의 잠재력을 보여주었다.
- 또한 GPT-3.5와 GPT-4의 비교 분석을 통해 거짓 음성을 최소화하는 GPT-4의 능력이 크게 향상된 것으로 나타났다.
- 본 연구는 다양한 사이버 보안 영역에서 LLM의 적용을 탐색하고 악성 웹 콘텐츠를 분석 및 탐지하기 위한 기능을 더욱 최적화하기 위한 새로운 방향을 제시한다.