Adaptive Linguistic Prompting (ALP) Enhances Phishing Webpage Detection in Multimodal Large Language Models

Atharva Bhargude* **Ishan Gonehal** **Dave Yoon**
Kaustubh Vinnakota **Chandler Haney** **Aaron Sandoval†** **Kevin Zhu†**
Algoverse AI Research
atharvab167@gmail.com, kevin@algoverseacademy.com

Abstract

Phishing attacks represent a significant cybersecurity threat, necessitating adaptive detection techniques. This study explores few-shot Adaptive Linguistic Prompting (ALP) in detecting phishing webpages through the multimodal capabilities of state-of-the-art large language models (LLMs) such as GPT-4o and Gemini 1.5 Pro. ALP is a structured semantic reasoning method that guides LLMs to analyze textual deception by breaking down linguistic patterns, detecting urgency cues, and identifying manipulative diction commonly found in phishing content. By integrating textual, visual, and URL-based analysis, we propose a unified model capable of identifying sophisticated phishing attempts. Our experiments demonstrate that ALP significantly enhances phishing detection accuracy by guiding LLMs through structured reasoning and contextual analysis. The findings highlight the potential of ALP-integrated multimodal LLMs to advance phishing detection frameworks, achieving an F1-score of 0.93—surpassing traditional approaches. These results establish a foundation for more robust, interpretable, and adaptive linguistic-based phishing detection systems using LLMs.

1 Introduction

With over 1.2 million attempts blocked in 2024 alone (Acharya et al., 2024), phishing remains a persistent cybersecurity threat as attackers continuously refine their tactics to evade detection (Zara et al., 2024). Traditional detection methods, such as heuristic URL matching and brand verification, often falter against novel evasion tactics (Li et al., 2024). While earlier work (Lee et al., 2024) compared modality contributions using zero-shot brand + domain verification, we instead focus on how prompt engineering alone can enhance phishing detection performance using Multimodal Large Language Models (MLLMs).

The emergence of Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs), represents a paradigm shift in phishing detection (Tushkanov, 2023), offering fine-grained semantic analysis. These MLLMs, including GPT-4o and Gemini 1.5 Pro, enable joint reasoning over HTML, screenshots, and URLs (Touvron et al., 2023).

In this paper, we introduce **Adaptive Linguistic Prompting (ALP)**, a method inspired by few-shot prompting (Brown et al., 2020; Agrawal et al., 2022) and structured reasoning (Wei et al., 2023). It is a few-shot prompting framework that guides LLMs to perform structured, modality-specific reasoning using curated exemplars rather than zero-shot templates. Our contributions include: (1) a refined 8-shot prompting framework (Prompt-Enhanced ALP) for HTML + Screenshot and URL analysis, and (2) design insights—such as a "suspicious-first" URL heuristic—that improve F1 (0.91 vs. 0.93), demonstrating the impact of prompt tuning independent of model changes.

2 Related Works

Traditional phishing detection relies on heuristics like URL analysis, HTML structure analysis, and blacklisting (Li et al., 2024). While effective for known threats, these methods fail against zero-day attacks and sophisticated mimicry tactics (Kulkarni et al., 2024). Machine learning (ML)-based models extract statistical patterns to improve detection (Whittaker et al., 2010; Xiang et al., 2011), but adversarial perturbations and dynamic content limit their effectiveness (Lee et al., 2024).

Recent phishing detection methods use computer vision to spot brand imitation via logos and page layout (Abdelnabi et al., 2020; Lin et al., 2021; Ji et al. 2024), however these models require

*Lead Author
†Senior Author

Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI), pages 77–85
July 31, 2025

continuous retraining.

Lee et al., 2024 proposed a multimodal LLM approach integrating brand identification and domain verification to detect phishing inconsistencies. Their findings show LLMs outperform traditional methods in accuracy and robustness but face challenges like prompt injection and high computational costs (Divakaran and Peddinti 2024). Other models, like ChatSpamDetector (Koide et al., 2024), also leverage LLMs. Our research advances this field by integrating few-shot ALP prompting to refine linguistic feature extraction and reasoning, reducing dependence on continuous retraining while strengthening phishing detection.

3 Methodology

For the experiments, we selected the GPT-4o and the Gemini 1.5 Pro models for their advanced multi-modal analysis architectures, both of which effectively correlate verbose textual content and visual inputs (Yin et al., 2024). This is crucial for phishing detection, where identifying malicious sites often requires analyzing a combination of HTML content, syntactic patterns, and screenshots.

3.1 Data Collection

We use the curated Lee et al. (2024) dataset sourced from Lee (2024), containing 1607 Benign Brands and 289 Phishing Brands; each brand is represented by one screenshot and one JSON file containing HTML attributes. 311/1607 Benign Brands and all 289 Phishing Brands are retained through our filtering approach, to optimize computational efficiency.

3.2 Dataset Filtering

To optimize computational efficiency while preserving dataset diversity, the benign dataset is curated to 311 brands. This size mitigates token-based processing costs in GPT-4o while preserving analytical rigor. Our filtering approach (automated clustering paired with manual validation) prioritizes domain diversity, incorporating both high-reputation entities (e.g., "google.com", "amazon.com") and traffic-tiered domains identified through web-traffic metrics by (Howarth, 2025). Each brand is represented by two standardized instances—one screenshot and one JSON file containing HTML attributes—to streamline processing while capturing multimodal features critical for phishing detection. The selection criteria emphasizes minimizing redundancy without compromising domain heterogeneity, ensuring coverage of both commonly accessed and niche pages.

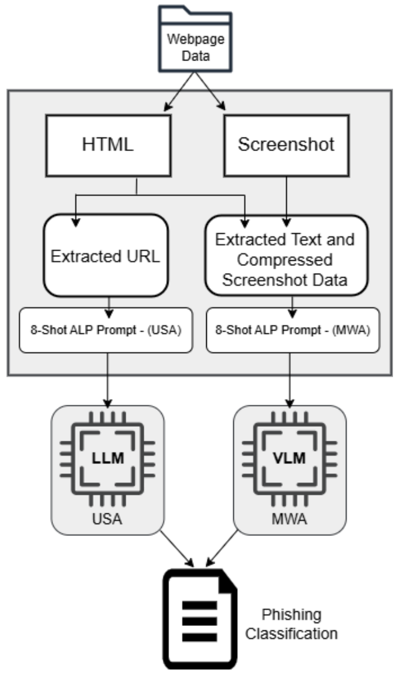

Figure 1: Overview of the Multimodal VLM & LLM Phishing Webpage Detection System.

3.3 Few-Shot ALP Prompting Methodology

This paper introduces a novel prompting methodology that integrates few-shot Adaptive Linguistic Prompting (ALP) to enhance phishing detection using LLMs for the analysis of multimodal data, including text, images, and URLs. We build on Lee et al.'s data splits and baseline prompts, reimplementing their zero-shot brand and domain setup, refining 8-shot prompts for HTML + Screenshot and URL analysis using held-out validation, and evaluating all methods on F1, precision, and recall.

Two distinct analyses were implemented to leverage the multi-modal capabilities of GPT-4o and Gemini 1.5 Pro for phishing detection. Both analyses utilized the curated dataset, ensuring consistent input modalities.

The first analysis, Multimodal Webpage Analysis (MWA), utilizes multi-modal capabilities by analyzing both the HTML textual content and the screenshot data of a given webpage. The diagram (Figure 1) illustrates the flow of MWA prompting, wherein HTML and screenshot data are extracted and processed through an 8-shot ALP reasoning prompt.

1. Multimodal Webpage Analysis (MWA) Instruc-

tion Prompt utilizing ALP

> You are a phishing detection assistant. Given a randomly selected phishing or benign URL, you will analyze its HTML and screenshot data. Use the provided few-shot examples as guides to assess phishing risk. Structure your analysis into brand recognition, diction and syntax, and screenshot-specific evaluations.
>
> (HTML and screenshot input)
> (8-shot ALP examples)
>
> Your response should follow this order:
> 1. Step-by-Step Reasoning:
> 2. Key Indicators:
> 3. Supporting Evidence:
> 4. Confidence Score 0-10:
> 5. Phishing Risk (Phishing/Benign):

The second analysis, URL Structure Analysis (USA), applies URL-matching techniques from existing literature, leveraging structured data from embedded URLs to classify phishing attempts. In this analysis, zipped files containing website data are processed systematically, with each HTML file being parsed to extract URLs. These URLS are then processed through an 8-shot ALP reasoning prompt and are then used to evaluate the model's ability to identify phishing through URL analysis.

2. URL Structure Analysis (USA) Instruction Prompt

> You are a phishing detection expert. Your task is to classify URLs as only either 'phishing' or 'benign' and provide a detailed explanation of why. Focus on key features like domain name, protocol, URL path, and potential phishing indicators.
>
> (URL input)
> (8-shot MWA examples)
>
> Your response should follow this order:
> URL:
> Features:
> Reasoning:
> Label: (Benign or Phishing)

3.4 Combining Analysis Results

We fuse MWA and USA outputs with a straightforward, risk-aware rule. When both analyses agree, we accept that label immediately. If they disagree, we label phishing whenever USA predicts it or when MWA's confidence exceeds 8.5—otherwise we label benign. The 8.5 cutoff was set a priori on a training split to balance precision and recall. USA, optimized for URL analysis, is adept at identifying domain-specific anomalies, while MWA is adept at capturing subtler phishing signals, such as content and visual mimicry. The combined analysis enhances detection accuracy, particularly in cases involving visually deceptive phishing tactics.

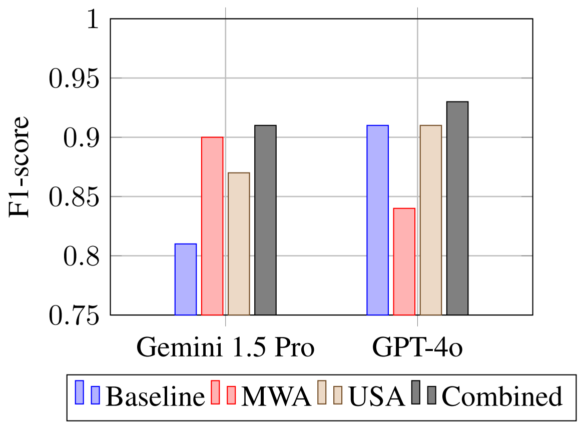

Figure 2: F1-score for Gemini 1.5 Pro and GPT-4o between Baseline, MWA, USA, & Combined Analysis

4 Results and Discussion

The results demonstrate the efficacy of multi-modal phishing detection modified by Adaptive Linguistic Prompting (ALP) over heuristic zero-shot prompting methods seen in our baseline (Lee et al., 2024).

The performance assessment revealed different strengths between GPT-4o and Gemini 1.5 Pro across MWA, USA, and the combined analysis. MWA outperformed static approaches by detecting advanced deception tactics, particularly brand mimicking. Gemini 1.5 Pro performed well in identifying brand representation anomalies due to its comprehensive analysis of visual elements and textual consistency markers.

USA demonstrated complementary strengths, with GPT-4o excelling in domain authenticity and URL pattern analysis, identifying suspicious structures and security protocol inconsistencies. The

Model	Approach	Precision	Recall	F1
GPT-4o	Baseline	0.91	0.91	0.91
	MWA	0.80	0.89	0.84
	USA	0.91	0.91	0.91
	Combined Analysis	0.91	0.94	0.93
Gemini 1.5 Pro	Baseline	0.76	0.85	0.81
	MWA	0.94	0.87	0.90
	USA	0.88	0.85	0.87
	Combined Analysis	0.91	0.92	0.91

Table 1: Performance comparison across baseline, MWA, USA, & Combined Analysis

combined analysis improved performance by integrating both approaches.

GPT-4o's combined analysis achieved a precision of 91.67%, recall of 94.12%, and an F1-score of 0.93, while Gemini 1.5 Pro achieved 91% precision, 92% recall, and an F1-score of 0.91. These results suggest an optimized phishing detection framework could integrate both models, with GPT-4o for USA and Gemini for MWA to maximize accuracy.

While testing USA, we incorporated a "suspicious-first" prompting strategy into USA, as seen below.

> **If you are unsure and feel that a link is suspicious of phishing activity, label it phishing** **Otherwise, if the link relates to an official domain, label it Benign**

This guided models to classify uncertain cases as potential threats, significantly improving USA detection accuracy while maintaining a practical balance between security and usability. This finding suggests prompt engineering could be as crucial as model architecture in developing robust detection systems as such a simple prompt increased USA's accuracy from 81% to 91% in GPT-4o by adopting a risk-averse approach.

5 Conclusion

Our study demonstrates that Adaptive Linguistic Prompting (ALP) enhances phishing detection by guiding multimodal LLMs to systematically analyze brand, linguistic, visual, and URL-based deception tactics. By integrating semantic reasoning and few-shot prompting, ALP addresses critical gaps in traditional heuristic methods, which struggle with nuanced phishing cues (examples in Appendix A.1). The combined analysis function achieved an F1-score of 0.93, a significant improvement over the baseline and past literature, highlighting ALP's ability to harmonize detection modalities while prioritizing risk-averse classification. Our findings establish ALP as a robust, interpretable, and scalable solution, reducing reliance on continuous retraining.

6 Limitations

With frameworks utilizing ALP to advance phishing detection, certain limitations merit acknowledgment. First, the dataset, though diverse, may not fully capture emerging phishing tactics or region-specific attacks, potentially affecting applicability. Second, reliance on proprietary LLMs like GPT-4o introduces scalability and cost barriers, limiting accessibility for broader deployment. Third, the framework's performance on non-English content and adversarial evasion strategies (e.g., context-aware paraphrasing) remains unexplored. Fourth, the "suspicious-first" strategy prioritizes URL analysis in conflicting MWA and USA situations, which could marginalize nuanced visual mimicry detected by multimodal analysis. A weighted probability combination function could potentially be explored. Finally, comparisons to machine learning architectures other than transformer-based LLMs are absent. Future research should optimize prompt engineering to lower token consumption and explore multi-LLM frameworks to leverage specific model strengths. Further, expanding datasets to cover non-English phishing tactics, integrating cost-effective open-source models, and rigorously testing adversarial evasion strategies could strengthen ALP and enhance phishing detection accuracy.

References

Sahar Abdelnabi, Katharina Krombholz, and Mario Fritz. 2020. Visualphishnet: Zero-day phishing website detection by visual similarity. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 1681–1698.

Acharya, Lazzaro, Cinà, and Holz. 2024. Pirates of charity: Exploring donation-based abuses in social media platforms. *Preprint*, arXiv:2412.15621.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *Preprint*, arXiv:2205.12689.

Atharva Bhargude. 2025. Adaptive-linguistic-prompting-alp-multimodal-llm-phishing-detection. Accessed: 2025-01-30.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Dinil Mon Divakaran and Sai Teja Peddinti. 2024. Llms for cyber security: New opportunities. *Preprint*, arXiv:2404.11338.

Josh Howarth. 2025. Most visited websites in the world (november 2024). Accessed: 2025-01-29.

Fujiao Ji, Kiho Lee, Hyungjoon Koo, Wenhao You, Euijin Choo, Hyoungshick Kim, and Doowon Kim. 2024. Evaluating the effectiveness and robustness of visual similarity-based phishing detection models. *Preprint*, arXiv:2405.19598.

Takashi Koide, Naoki Fukushi, Hiroki Nakano, and Daiki Chiba. 2024. Chatspamdetector: Leveraging large language models for effective phishing email detection. *Preprint*, arXiv:2402.18093v1.

Aditya Kulkarni, Vivek Balachandran, Dinil Mon Divakaran, and Tamal Das. 2024. From ml to llm: Evaluating the robustness of phishing webpage detection models against adversarial attacks. *Preprint*, arXiv:2407.20361.

Jehyun Lee. 2024. Multimodal llm phishing detection - github. Accessed: 2025-01-29.

Jehyun Lee, Peiyuan Lim, Bryan Hooi, and Dinil Mon Divakaran. 2024. Multimodal large language models for phishing webpage detection and identification. *Preprint*, arXiv:2408.05941.

Yuexin Li, Chengyu Huang, Shumin Deng, Mei Lin Lock, Tri Cao, Nay Oo, Hoon Wei Lim, and Bryan Hooi. 2024. Knowphish: Large language models meet multimodal knowledge graphs for enhancing reference-based phishing detection. *Preprint*, arXiv:2403.02253.

Yun Lin, Ruofan Liu, Dinil Mon Divakaran, Jun Yang Ng, Qing Zhou Chan, Yiwen Lu, Yuxuan Si, Fan Zhang, and Jin Song Dong. 2021. Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 3793–3810.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Vladislav Tushkanov. 2023. What does chatgpt know about phishing? Accessed: 2025-01-29.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Colin Whittaker, Brian Ryner, and Marria Nazif. 2010. Large-scale automatic classification of phishing pages. In *Ndss*, volume 10, page 2010.

Guang Xiang, Jason Hong, Carolyn Rosé, and Lorrie Cranor. 2011. Cantina+: A feature-rich machine learning framework for detecting phishing web sites. *ACM Trans. Inf. Syst. Secur.*, 14:21.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Ume Zara, Kashif Ayyub, Hikmat Ullah Khan, Ali Daud, Tariq Alsahfi, and Saima Gulzar Ahmad. 2024. Phishing website detection using deep learning models. *IEEE Access*, 12:167072–167087.

A Appendix

A.1 Few-Shot-Prompts

Below is a few-shot prompt for MWA

> Example 1
> Url: https://secure-appleid-login.com/
> HTML Data: title: Apple ID Login meta description: Sign in to your Apple ID account. favicon: https://www.apple.com/favicon.ico logo alt text: Apple Logo footer text: Not Found headers text: Home | Support | Account | Privacy nav bar content: Not Found paragraphs text: Enter your Apple ID and password to sign in. span text: Forgot Apple ID? | Forgot Password? | Sign Up
> Screenshot Info: The Apple logo is high-resolution and matches official branding. The login form includes fields for Apple ID and password. The "Sign In" button is styled with Apple's official colors. The footer lacks additional information and links typically found on Apple's official site.
> Brand Recognition Analysis: The page uses the official Apple logo with correct alt text. The URL ("secure-appleid-login.com") is not an official Apple domain. HTTPS protocol is absent, which is unusual for Apple's official login pages.
> Diction and Syntax Analysis: The language is straightforward without excessive capitalization. Instructions are clear and mimic official Apple communications. There is no urgent or threatening language detected. Screenshot Analysis: Visual elements largely mimic Apple's official login page. However, the absence of additional footer links and the domain discrepancy raise suspicion.
> Step-by-Step Reasoning:
> 1. Branding elements appear legitimate with a correct Apple logo.
> 2. The URL does not match Apple's official domain and lacks HTTPS.
> 3. Missing footer elements typical of Apple's official site suggest phishing.
> Classification: PHISHING

The Multimodal Webpage Analysis (MWA) few-shot prompt guides the model to analyze phishing risk by breaking down the evaluation into structured components: brand recognition, diction and syntax, and screenshot-specific analysis. It provides step-by-step reasoning, key indicators, supporting evidence, a confidence score, and a final phishing/benign classification. This approach ensures the model systematically analyzes both textual and visual elements, leveraging examples to identify subtle inconsistencies in branding, language, and design that may indicate phishing.

Below is a few-shot prompt for USA

> Example 1:
> Link: 'http://university-scholarship-app.us/confirm-details'
> Features: The domain includes terms such as "university" and "scholarship", HTTP protocol, path "confirm-details".
> Reasoning: This link exhibits several traits of a phishing attempt. The domain uses generic terms like "university" and "scholarship," which are unrelated to any well-known institution, making it highly suspicious. The use of HTTP instead of HTTPS suggests that the site does not encrypt user data, leaving it vulnerable to interception. The path, "confirm-details," is a classic phishing tactic designed to solicit sensitive personal or account-related information under the guise of legitimacy. These factors combined strongly indicate malicious intent.
> Classification: PHISHING

The URL Structure Analysis (USA) few-shot prompt focuses on classifying URLs as phishing or benign based on domain name, protocol, URL path, and phishing indicators. It requires the model to provide detailed reasoning for its classification, emphasizing features like HTTPS usage, domain legitimacy, and suspicious path structures. By incorporating examples of phishing URLs, the prompt trains the model to recognize patterns such as generic domain names, insecure protocols, and deceptive paths, enabling it to detect phishing attempts through URL analysis alone.

A.2 Model Outputs

The following focuses on four different model outputs over MWA and USA for GPT and Gemini. The webpage in focus is 'https://www.12voltdoesit.com/'

1. GPT-4o MWA Response

Folder Name: 12voltdoesit AI Response: 1. *Step-by-Step Reasoning*: - **Brand Recognition Analysis**: The HTML content mentions the website title as "Home | 12VOLTDOESIT - Automotive Electronics Services," indicating a focus on automotive electronics. However, the favicon and logo alt text are marked as "Not Found," which raises suspicion about the legitimacy of the brand. The footer text provides a service description but lacks well-known brand indicators or established company credentials. The URL "12voltdoesit.com" does not correspond to a widely recognized brand.
- **Diction and Syntax Analysis**: Phrases like "we would like to help" and "top-rated Automotive Electronic Services" suggest a friendly tone, but the overall structure seems cluttered with excessive product listings and pricing information, which could be seen as unprofessional. There is no urgent or alarming language, which is a positive sign.
- **Screenshot Analysis**: The screenshot URL leads to an image that contains the visual representation of the webpage. The design is amateurish, potentially indicating a phishing attempt. The lack of identifiable branding elements in the HTML also reflects poorly in the screenshot.
2. *Phishing Risk (Phishing/Benign)*: Classification: Benign (Confidence score of 6.5 suggests some legitimacy, but caution is warranted due to missing brand recognition elements)
3. *Key Indicators*: - Informal diction and promotional language - Cluttered presentation of services and products
4. *Confidence Score*: 6.50
5. *Supporting Evidence*: The analysis reveals that the absence of recognizable branding elements such as a logo or favicon raises red flags about its legitimacy. The informal nature of the text may indicate a lack of professionalism, which is common in phishing sites. However, the lack of urgent language and the coherent description of services suggest that it may not be an outright phishing attempt. The overall impression is that while there are indicators of concern, there is no definitive evidence to classify the site as phishing, leading to a benign classification.
6. *Provide URL*: URL: https://www.12voltdoesit.com/

2. Gemini MWA Response

Folder Name: 12voltdoesit AI Response: 1. *Step-by-Step Reasoning*: 1. The URL seems legitimate, belonging to a business named "12voltdoesit" and using HTTPS. 2. Brand Recognition: The branding is consistent throughout the website, though the lack of a logo image in the HTML is somewhat unusual for a business site. The footer copyright and contact information reinforce the legitimacy of the business. 3. Diction and Syntax Analysis: The language used on the site is generally clear and understandable, although there are some minor grammatical errors and repetitions, such as "Installation Include in price!". The language doesn't employ urgency or pressure tactics commonly found in phishing attempts. 4. Screenshot Analysis: The screenshot displays a professional-looking website with clear sections for services, products, and contact information. The screenshot matches the information provided in the HTML data, further reinforcing the website's authenticity.
2. *Phishing Risk*: Classification: BENIGN
3. *Key Indicators*: Repeated phrases "Installation Include in price!", some minor grammatical inconsistencies. Clear pricing for all advertised services. Presence of contact information, copyright information, and a physical address.
4. *Confidence Score*: 9.50
5. *Supporting Evidence*: The website exhibits several key characteristics of a legitimate business. It uses a proper domain name, secure HTTPS, and provides consistent branding across the HTML and the screenshot. While there are minor grammatical errors and repetitive phrases, these do not necessarily indicate a phishing attempt and could simply be due to oversight. The absence of common phishing indicators like urgent language, requests for personal information (outside of standard login/registration), and misleading links further support the benign classification. The overall presentation and information provided create a high degree of confidence in its legitimacy.
6. *Provide URL*: URL: `https://www.12voltdoesit.com/`

Regarding MWA comparisons, GPT-4o excels in identifying branding inconsistencies and linguistic informality, often flagging sites with missing favicons, unprofessional design, or cluttered content as suspicious. Its cautious approach leads to lower confidence scores (e.g., Confidence Score: 6.5) even when classifying sites as Benign, reflecting a focus on structural red flags. Gemini 1.5 Pro emphasizes branding coherence and visual authenticity, showing higher confidence in benign classifications (e.g., Confidence Score: 9.50) when branding elements align, even in the presence of minor linguistic errors. Its strength lies in recognizing legitimate visual and textual patterns, often overlooking subtle inconsistencies that GPT-4o flags.
3. GPT-4o USA Response

Link: 'https://www.12voltdoesit.com/' Features: The domain is "12voltdoesit.com". The protocol is HTTPS. Reasoning: The domain "12voltdoesit.com" does not immediately relate to any widely recognized brand or institution, which raises suspicion. However, it uses HTTPS, which ensures a secure connection. Without additional context or known associations, it's difficult to determine the legitimacy of this domain. Given the lack of recognizable branding, it should be approached with caution. Label: Phishing

4. Gemini USA Response

Link: 'https://www.12voltdoesit.com/' Features: Domain "12voltdoesit.com", HTTPS. Reasoning: The domain name appears legitimate for a business related to 12-volt electronics. HTTPS ensures secure communication. No suspicious elements are present. Label: Benign

For USA, GPT-4o benefits from the risk-averse approach, labeling unrecognized or niche domains (e.g., 12voltdoesit.com) as phishing due to lack of brand association, even when HTTPS is present. Gemini 1.5 Pro demonstrates contextual adaptability, revising its classification from Phishing to Benign upon recognizing domain relevance (e.g., 12-volt electronics). Its ability to infer legitimacy from niche or technical domains highlights its flexibility in URL analysis.

The marginally better post-combination performance of GPT-4o stems from its conservative URL Structure Analysis (USA), which aligns well with the risk-averse "suspicious-first" strategy. While Gemini excels in Multimodal Webpage analysis (MWA), GPT-4o's acute URL scrutiny provides a robust safety net, particularly in cases where visual mimicry is deceptive. This complementary dynamic enhances overall detection accuracy, as GPT-4o's precision in URL analysis compensates for edge cases where Gemini's visual analysis might falter. However, an evaluation function containing Gemini 1.5 Pro for MWA and GPT-4o for USA could leverage the strengths of both models for potentially improved accuracy.

A.3 Code

All code and few-shot prompts for the Multimodal Webpage Analysis (MWA) and URL Structure Analysis (USA) frameworks can be found in this GitHub (Bhargude, 2025).