

Small Language Models for Phishing Website Detection: Cost, Performance, and Privacy Trade-Offs

Georg Goldenits^{a,*}, Philip König^b, Sebastian Raubitzek^a, Andreas Ekelhart^{a,b}

^aSBA Research gGmbH, Floragasse 7/5.0G, Vienna, 1040, Vienna, Austria

^bUniversity of Vienna, Universitätsring 1, Vienna, 1010, Vienna, Austria

Abstract

Phishing websites pose a major cybersecurity threat, exploiting unsuspecting users and causing significant financial and organisational harm. Traditional machine learning approaches for phishing detection often require extensive feature engineering, continuous retraining, and costly infrastructure maintenance. At the same time, proprietary large language models (LLMs) have demonstrated strong performance in phishing-related classification tasks, but their operational costs and reliance on external providers limit their practical adoption in many business environments. This paper investigates the feasibility of small language models (SLMs) for detecting phishing websites using only their raw HTML code. A key advantage of these models is that they can be deployed on local infrastructure, providing organisations with greater control over data and operations. We systematically evaluate 15 commonly used **Small Language Models (SLMs)**, ranging from 1 billion to 70 billion parameters, benchmarking their classification accuracy, computational requirements, and cost-efficiency. Our results highlight the trade-offs between detection performance and resource consumption, demonstrating that while **SLMs** underperform compared to state-of-the-art proprietary LLMs, they can still provide a viable and scalable alternative to external LLM services. By presenting a comparative analysis of costs and benefits, this work lays the foundation for future research on the adaptation, fine-tuning, and deployment of **SLMs** in phishing detection systems, aiming to balance security effectiveness and economic practicality.

Keywords: Large Language Model, Small Language Model, LLM, SLM, Phishing Detection, Cybersecurity, Benchmarking

1. Introduction

Website phishing remains one of the most prevalent and damaging forms of online fraud, enabling attackers to steal user credentials, financial data, or other sensitive information by imitating legitimate websites [FBI IC3 \(2025\)](#); [Europol \(2024\)](#). The Anti-Phishing Working Group (APWG) reported over 1.1 million unique phishing websites detected in Q2 2025, representing a 13% increase compared to Q1 in 2025, continuing a multi-year upward trend in attack frequency [APWG \(2025\)](#). Unlike email phishing, web-based phishing exploits users' trust in domains, brands, and website structures, making it difficult to detect through superficial inspection [Chiew et al. \(2018\)](#). As phishing websites become more sophisticated, evasion tactics become more prevalent and campaigns emerge rapidly, automated and adaptive detection techniques are essential [Europol \(2024\)](#); [Mahboudi et al. \(2024\)](#).

Classical machine learning and deep learning approaches for website phishing detection, such as URL-based classifiers [Yang et al. \(2021\)](#); [Bahaghagh et al. \(2023\)](#) or visual similarity models based on **Convolutional Neural Networks (CNNs)** [Yerima & Alzaylae \(2020\)](#); [Abdelnabi et al. \(2020\)](#); [Lin et al. \(2021\)](#) have shown promising detection rates under controlled conditions. However, they face several limitations in practice. Mod-

els trained on static features are vulnerable to adversarial evasion, such as obfuscating malicious scripts or dynamically generating HTML content [Panum et al. \(2020\)](#); [Nahapetyan et al. \(2025\)](#); [Abuadbba et al. \(2022\)](#). Furthermore, **Machine Learning (ML)** classifiers require continuous retraining or feature updates to remain effective against rapidly evolving phishing tactics, leading to high operational costs. Challenges such as high false-positive rates, susceptibility to adversarial attacks, scalability issues, and data imbalances limit the usefulness of **ML**-based detection approaches [Kavya & Sumathi \(2024\)](#); [Bountakas et al. \(2023\)](#). These factors underscore the need for more adaptable and robust detection mechanisms.

Recent work has proposed using **Large Language Models (LLMs)** for phishing website detection by treating the HTML structure and embedded text as natural language input. Proprietary **LLMs**, such as GPT-4 or Gemini, have demonstrated strong performance in classifying phishing websites and even generating natural-language explanations of their reasoning [Liu et al. \(2024\)](#); [Koide et al. \(2024\)](#). However, these approaches are often impractical in commercial settings, as they depend on vendor-controlled APIs, incur high inference costs, and make fine-tuning or customisation expensive. They also pose data privacy risks, as sensitive information must be sent to external systems, and may suffer from availability issues due to varying request latencies that can range from few seconds to minutes [Barberá \(2025\)](#); [Irugalbandara et al. \(2024\)](#).

SLMs, which we define as models with $\leq 70\text{b}$ parameters,

*Corresponding author

Email address: ggoldenits@sba-research.org (Georg Goldenits)

present a promising alternative for phishing website detection, especially when they are open source and deployed locally. They offer several advantages, including significantly reduced operating costs, independence from vendor APIs, improved data privacy and confidentiality, easier compliance, and the ability to fine-tune models on-premise for domain-specific needs. However, to the best of our knowledge, no study outlines the utility, performance, costs, and benefits of using **SLMs** in comparison to large proprietary models.

In this study, we systematically benchmark a diverse set of 15 commonly used **SLMs**, ranging from 1 billion to approximately 70 billion parameters, for website phishing detection. The selected models include different sizes of Gemma¹, Deepseek², Qwen³ and Llama⁴, as well as models that are only available in one size, such as Mistral⁵, gpt-oss⁶, phi⁷, and dolphin⁸. We evaluate those models along three axes: time to analyse website HTML and produce results, phishing classification correctness, and output coherence. To generate this benchmark, each model classifies 1,000 websites (500 phishing and 500 benign) sampled from a freely available dataset Putra (2023).

To determine the suitability of local **SLMs** for phishing detection, we structure our approach along the following research questions:

- **RQ1:** Do **SLMs** deliver coherent and reliable outputs, including meaningful reasoning for phishing website detection?
- **RQ2:** How do the costs and benefits of local models compare to state-of-the-art proprietary models such as GPT4?
- **RQ3:** How does the length of the HTML code influence both analysis runtime and detection performance?

The contributions of this article are:

- We benchmark 15 **SLMs**, regarding various numeric and text-based criteria on the task of website phishing detection to develop a detection method that can compete with state-of-the-art larger proprietary **LLMs**.
- The benchmarking methodology, dataset, and source code are publicly available on Github⁹, so other researchers and practitioners can reproduce the results and test new models in the future.
- We discuss benefits and challenges of small local models compared to commercially available options and provide recommendations for application of **SLMs** for website phishing detection.

This work is structured as follows: Section 2 presents related work focusing on using **LLMs** for website phishing detection. In Section 3, the experimental design setup is presented, with Section 4 containing the results of the experiments. Section 5 discusses the advantages and costs of local **LLMs** compared to commercial applications, providing a foundation for application recommendations. Section 6 concludes the work and outlines future research directions.

2. Related Work

Although the field of **LLM**-based website phishing detection has only recently gained significant traction, several studies have reported promising results using different models and analysis strategies. Table 1 provides an overview of these studies, highlighting input data modalities, models, and their reported classification performance.

Table 1: Summary of related work classified into data modes, **LLMs** used, dataset size, and reported classification performance (F1-score).

Publication	Input Data	LLM	Dataset Size	F1-score
Lee et al. (2024)	Text + Vision	Gemini Pro-Vision 1, GPT-4-turbo, Claude 3 Opus	~3,000 benign ~1,500 phish	0.81 - 0.92
Koide et al. (2024)	Text + Vision	GPT-4V4;40, GPT-3.5-turbo, Gemini Pro 1.0, Command R+, Llama 2 70b, Llama 3 70b, Gemma 2 9b	1,000 benign 1,000 phish	0.72 - 0.99
Li et al. (2025)	Text	Qwen2.5-v1.72b-instruct, Gemini-2.0-Flash, GPT-4o, GPT-4o-mini	500 benign 500 phish	0.85-0.97
Trad & Chehab (2024)	Text	GPT-3.5-turbo Claude 2	500 benign 500 phish	0.78-0.93
Kulkarni et al. (2025)	Text + Vision	Gemini-1.5-Flash	17,794 benign 9105 phish	0.75-0.98
Liu et al. (2024)	Text + Vision	GPT-3.5-turbo-16k	6,000 benign 6,000 phish	0.86

Regarding input data, two main approaches are commonly used in the literature: text-only, which consists of website HTML code and URLs, and text combined with screenshot images. The choice of data modality typically depends on data availability and the capabilities of the selected models. Since most **LLMs** are designed to process and generate text, relying solely on HTML code is frequently chosen as a starting point for a phishing detection framework. Combining text and images is a common next step and a method often used to detect brand impersonations.

The most frequently used models are versions of GPT-3.5 and GPT-4, followed by iterations of the Gemini models. Open-source alternatives, such as Llama and Qwen, are less commonly chosen. The dataset sizes used in the different studies vary but generally range of a few thousand websites. Typically, researchers use balanced datasets, with an equal number of phishing and benign websites. As a comparative performance measure in phishing detection setups, the F1-score is commonly reported. The reported F1-scores range from 0.72 to 0.99. Detailed descriptions of the cited detection architectures can be found below.

¹<https://ollama.com/library/gemma3>

²<https://ollama.com/library/deepseek-r1>

³<https://ollama.com/library/qwen3>

⁴<https://ollama.com/library/llama3.1>, <https://ollama.com/library/llama3.2>, <https://ollama.com/library/llama3.3>

⁵<https://ollama.com/library/mistral-nemo>

⁶<https://ollama.com/library/gpt-oss>

⁷<https://ollama.com/library/phi3>

⁸<https://ollama.com/library/dolphin3>

⁹<https://github.com/sbaresearch/benchmarking-SLMs>

[Lee et al. \(2024\)](#) employ a multimodal approach that involves both brand identification and domain verification. In the first step, **LLMs** must identify the brand based on the website's HTML code and a screenshot of the website. The recognised brand, combined with the input domain name, then serves as input for the second **LLM**, which is tasked with classification. Regarding the prompt design, there are three different versions depending on the input data mode (Screenshot, HTML, Screenshot + HTML), which instruct the **LLM** to analyse the given input and return a list of indicators, like `brand`, `has credentials` or a `confidence score`. The dataset used in their work is freely available in a GitHub repository¹⁰

[Koide et al. \(2024\)](#) use a different approach for handling the website screenshot, utilising an OCR model to extract text elements from the image. The extracted elements, along with the HTML code, are simplified and combined with the full screenshot and the URL, forming a data foundation for the analysis. The prompt instructs the **LLM** to analyse the input data and gives a detailed step-by-step description of how to proceed. The output should then be structured as a JSON-style document containing certain keys, such as 'phishing score', 'brands', or 'suspicious domain'. Although the method for deriving the dataset is provided, the final dataset is not freely available.

Compared to other approaches that rely on a single agent (or multiple non-collaborating agents), [Li et al. \(2025\)](#) present a multi-agent system where each model represents a different expert to analyse website HTML code, content, URL, and screenshot. A moderator agent tracks if a common answer has been found, and if yes, provides the result to a judge agent who makes the final decision. The prompt template is minimalist compared to other approaches, but also instructs the model to give a structured list of items as a response. The dataset is derived from various sources that can be accessed for free, but the final version is not freely available.

Unlike other related work, [Trad & Chehab \(2024\)](#) do not build an involved framework to detect malicious websites, where each website goes through in-depth analysis steps involving URL, HTML code, brand identification, or screenshot. Instead, in their approach, 50 URLs are batched together and are simultaneously analysed by an **LLM**. To enhance model performance, a classification head is added to the model architecture, whose weights are trained using a sample of the utilised URL dataset. The base dataset can be found in an online repository [Hannousse & Yahiouche \(2021\)](#), and while the exact sample used in their work is not directly available, the sampling was performed using a random seed, allowing for reproducibility. Three versions of the prompt were tested: a zero-shot version, a role-playing version, and a chain-of-thought version, to assess the output quality for different prompt styles. All prompts instruct the model to output only a Boolean decision, indicating whether a URL corresponds to a phishing website or not.

[Kulkarni et al. \(2025\)](#) have a different focus, testing the resilience of phishing detection systems to evasion strategies.

Based on the original HTML code of legitimate websites, deviating versions are created, which incorporate phishing techniques. The dataset containing these changed websites is available in a GitHub repository¹¹. The detection performance of the models is assessed based on all the available website versions. All the tested approaches are prone to evasion techniques, thus losing classification performance; however, **LLMs** show the strongest resilience to HTML perturbations and are deemed the most resilient of the tested techniques. However, the performance loss of the **LLM** still shows the need to further improve upon the existing detection techniques.

A different direction is taken by [Liu et al. \(2024\)](#), who emphasise brand recognition as a mechanism to collect potential URLs a brand might use and also account for changes such as logo or web presence that a brand might have gone through. This is done to build a dependable database for each brand, with the goal of reducing detection errors and minimising reliance on a reference database. In contrast to other works, which only use website screenshots, a system is built to iteratively navigate to credential-taking pages when starting on a non-credential-taking page. In their pipeline, **LLMs** are used to detect the brand and predict whether a website is credential taking or not. For both tasks, the prompting technique relies on structured prompts including few-shot examples. The datasets used to answer their research questions can be freely downloaded from a repository¹².

While this work focuses solely on detecting suspicious websites, adversarial phishing attack vectors also include emails and SMS messages. While **LLMs** get used for both website analysis and email/SMS analysis, the detection goals differ [Koide et al. \(2025\)](#); [Zhang et al. \(2025\)](#); [Nair et al. \(2025\)](#). With emails and SMS, the focus is on the used language, potential URLs in the text, and the sender. In contrast, website phishing detection goes beyond just human language analysis, as HTML code and structure hold crucial information for detecting malicious websites.

3. Methodology

This work examines the costs and benefits of using small local **LLMs** for phishing detection compared to commonly used proprietary models. To gain detailed insights into the advantages and disadvantages of using such models and comparing them to proprietary models, two experiments are conducted, which cover different input data and data processing steps. In the Model Selection experiment (Experiment 1), the behaviour of a local **LLMs** is examined on a small subset of the full dataset used in the Scaled Evaluation experiment (Experiment 2). The goals of the first experiment are to uncover random behaviour, model runtime and model output quality in terms of completeness and formatting. In experiment 2, models that performed satisfactorily in experiment 1 are used to classify a larger set of

¹⁰https://github.com/JehLeeKR/Multimodal_LLM_Phishing_Detection/tree/main/data/MMLLM_Benign/1

¹¹<https://github.com/LetsBeSecure/PhishOracle-Project?tab=readme-ov-file>
¹²<https://sites.google.com/view/phishllm/experimental-setup-datasets>

websites. In this experiment, the focus lies on the actual phishing detection performance, output coherence and the incurred analysis cost. Figure 1 summarises the experiments, dataset and models that are subsequently described in detail.

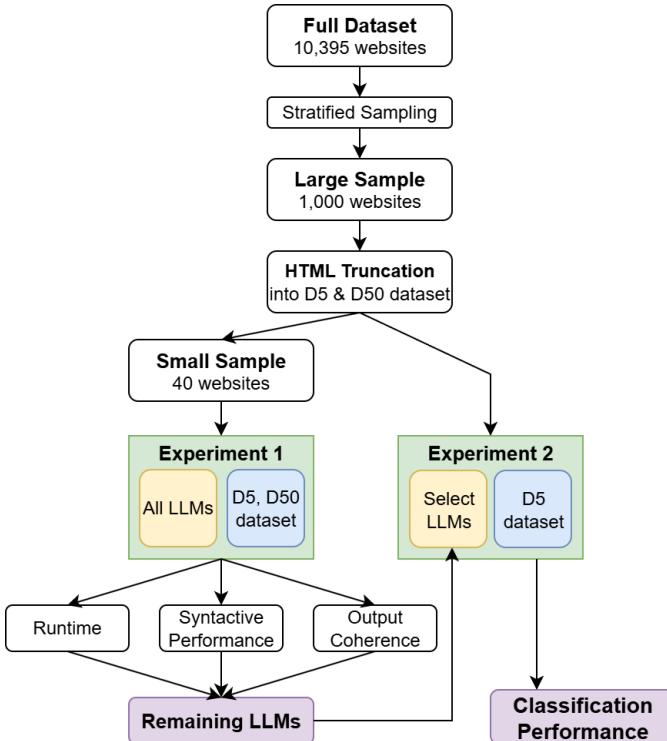


Figure 1: Methodological approach of the [LLM](#) benchmarking pipeline.

3.1. Dataset

The dataset used as a basis for this work can be accessed freely online and is provided to improve phishing detection techniques [Putra \(2023\)](#). The full dataset contains detailed information on 10,395 websites, comprising 5,244 legitimate and 5,151 phishing websites. This dataset includes HTML files, URLs, screenshots, CSS files, and information about JavaScript components and their execution. In this work, only the index HTML file, which displays the page with all JavaScript code executed, was selected, while the rest of the data was discarded. Focusing only on the HTML file ensures that our approach is comparable to publications mentioned in the related work and reproducible, as scraping HTML code is easier than accessing all the files in the background of the website.

The data repository was chosen because, with approximately 7,000 downloads, it is frequently used and contains an extensive data collection.

Based on this initial dataset, a stratified sample of 1,000 websites is selected, which is then further processed for use in the two experiments. The processing steps are described below. Data reduction is necessary to ensure a feasible workload in terms of available computational resources and the resulting analysis runtime.

- **Experiment 1:** For the initial Model Selection experiment, we sampled 20 benign and 20 phishing websites of

the 1,000 website large sample using a stratified random sampling approach. This ensured that the dataset covers a broad range of HTML document lengths, thereby reducing the risk of bias in the evaluation. For the sampled websites, we created two different truncated versions, referred to as the D5 and D50 datasets. Both datasets contain only the HTML code elements most relevant for phishing detection, as these provide key information about the contents and structure of the websites [Koide et al. \(2024\)](#). The two datasets differ in the amount of HTML retained: the D5 dataset includes only a maximum of 5% of the original websites' HTML code, while the D50 dataset includes a maximum of 50%. This truncation serves two purposes: 1) to avoid information that is unnecessary for phishing detection, and 2) to reduce the computational costs when processing the data with [LLMs](#).

To identify the most relevant HTML code elements, we built on the work of [Koide et al. \(2024\)](#), who selected the following tags for a similar phishing detection task: *head*, *title*, *meta*, *body*, *h1*, *h2*, *h3*, *h4*, *h5*, *h6*, *p*, *strong*, *a*, *img*, *hr*, *table*, *tbody*, *tr*, *th*, *td*, *ol*, *ul*, *li*, *ruby*, and *label*. Our truncation approach extends this by ranking the tags according to their importance for phishing detection [Kulkarni et al. \(2025\)](#). Specifically, the elements: *a*, *img*, *meta* and *link* are deemed the most important, followed by *HTML* and *head*, while all remaining elements were assigned the lowest importance. Typically, these elements convey the most expressive information about a website, including brand cues, indications of potential phishing links, and high semantic meaning. The truncation process works as follows: first, it unwraps all elements, and then removes empty elements. In an iterative reduction, the least important remaining elements are removed until the HTML token count falls below the target threshold, either 5% or 50% of the total token count. If the removal of a component would eliminate an entire class of elements, elements from other classes with multiple occurrences are removed first. The only difference between the D5 and D50 dataset truncation processes is that the allowed HTML token threshold is higher for the D50 dataset.

- **Experiment 2:** In a setup similar to Experiment 1, a larger dataset consisting of 500 benign and 500 phishing websites was selected using a stratified random sampling strategy to ensure a balanced and representative distribution across different website characteristics. Based on the findings from Experiment 1, this dataset was truncated following the D5 dataset, retaining a maximum of 5% of the original HTML content per website.

It is noteworthy that balanced datasets were used in both experiments, even though in real-world scenarios, the vast majority of websites encountered are expected to be benign. Unlike supervised learning approaches, [LLMs](#) analyse each website independently and do not rely on aggregated training data. Consequently, when websites are analysed in isolation rather than

Table 2: **SLM** selection criteria

SLM Selection Criteria
In the top 30 “Popular” models on Ollama in July 2025
Model has approximately 70 billion parameters or fewer
Model has 128,000 context window length
If there are multiple model variants available, choose the smallest model, a model in the centre of the parameter range and the model closest to 70b
Exclusion if tagged as “embedding”
If the exact popular model does not have the required context window length, choose a model of the same family that falls within the scope (e.g., mistral:7b’s context window is too short, therefore, it is replaced by mistral-nemo)

within a continuous conversation context, there is no need to reflect the real-world class distribution. Using a balanced dataset for analysis in this setting offers two key advantages: 1) it simplifies the interpretation of classification metrics, as results are not affected by class skew, and 2) it ensures a sufficient number of phishing samples, enabling more robust analysis of the **LLMs**’ phishing detection capabilities.

3.2. Model Selection

Given the large number of local **LLMs**—Ollama alone lists over 1,000 models, versions, and specialisations—it is infeasible to evaluate all options exhaustively ¹³. Therefore, a subset of models was selected for this study, based on clearly-defined inclusion and exclusion criteria, summarised in Table 2.

Two key constraints informed our selection process. First, only models up to 70 billion parameters were included to focus on **LLMs** that can run locally on limited hardware. This allows organisations to retain sensitive data on-premise, thereby addressing privacy and compliance concerns, while also reducing operational costs. Furthermore, studies evaluating commercial **LLMs**, such as GPT models, for phishing detection already exist (see Section 2). Hence, our work focuses on local, open-source alternatives to provide a complementary perspective. Second, a minimum context window of 128,000 tokens was chosen to ensure that the HTML input could be fully processed by the **LLMs**, avoiding automatic truncation by the model itself. Despite the truncation process, the HTML code often remains very long. Allowing models with smaller windows to internally truncate prompts would compromise comparability. Based on the above model selection criteria, 15 models were selected, spanning a diverse range of architectures and parameter sizes: deepseek-r1: 1.5b, 14b, 70b; gemma3: 4b, 12b, 27b; qwen3: 4b, 30b; llama: 3.2:1b, 3.1:8b, 3.3:70b; mistral-nemo:12b; phi3-medium: 14b; dolphin3: 8b and gpt-oss: 20b.

Considerations:

- Since no singular llama3.x model on Ollama comes with all required parameter sizes; multiple versions were included to represent a range of models.

- The chosen Qwen3 models feature an exceptionally large context window length of 256,000 tokens, which allows them to process more input, compared to the other selected models. To maintain comparability, the maximum prompt length was capped at 128,000 tokens, ensuring that no model receives additional information.
- Since **LLM**-based phishing detection research often relies on GPT-family models, we decided to include gpt-oss:20b as an open-source GPT alternative with a comparable model size to the other models tested in our experiments.

3.3. Prompt Design

Designing an effective prompt is a critical step in **LLM**-based phishing website detection. The primary objective of the prompt is to instruct the model to analyse a given HTML document and corresponding URL, identify potential phishing indicators, and produce a structured, interpretable output that can be used for quantitative evaluation. Using HTML and URL is common practice for LLM-based phishing detection but also other approaches Aljofey et al. (2025). The final prompt, shown in Prompt Template 1, is based on the template proposed by Koide et al. (2024) and was iteratively refined through multiple design cycles. The prompt consists of three core elements: 1) the task description, which instructs the model to classify a website as benign or phishing based on the provided HTML code and URL, 2) guidance examples (few-shot prompting) that outline common phishing indicators to guide the model’s reasoning, and 3) the output specification, which defines the required JSON output format including fields and score ranges. While the models are clearly instructed on how to deliver their output, any JSON-parsable response is accepted during the analysis phase. We anticipate inaccuracies when working with smaller models, and therefore do not discard results solely due to minor formatting deviations.

The following considerations shaped the final design:

- **Flat Prompt Design:** We use the generate mode (includes no conversational history), since each website must be analysed independently. As a result, the prompt had to be written in a flattened format, i.e., without explicit role labels like system or user.
- **Few-Shot Prompting:** To guide the model in detecting phishing characteristics, the prompt includes a limited number of examples of common phishing indicators (e.g., suspicious URLs, hidden forms, misleading redirects). These examples provide context and decision-making cues.
- **Output Formatting:** A strict JSON-parsable output format was defined to standardise model responses. The output contains three fields: `phishing_score` (integer from 0 to 10, segmented into predefined risk brackets to reduce ambiguity and facilitate consistent interpretation across different models). `is_phishing` (boolean classification), `reasoning` (brief textual explanation citing key indicators).

¹³<https://ollama.com>

Furthermore, the output length is limited to 1,000 tokens in an attempt to encourage models to retain the required output formatting and obtain concise answers.

Prompt Template 1

You are a cybersecurity expert analysing websites for phishing attempts. Your task is to examine the provided HTML code and the original website URL to determine if the website is likely a phishing site.

****Important:**** The HTML may be truncated to reduce costs, so CSS styles and JavaScript code may be missing. Focus on the HTML structure, text content, URLs, and the original URL.

Original URL: '{original_url}'

HTML: '{html_text}'

Original HTML character count: '{original_character_count}'

When assessing the likelihood of phishing, consider both your general knowledge of phishing techniques and the evidence you see in the data - focus on HTML structure, content, and domain/URL consistency.

" Below are common examples of phishing indicators you may find useful, but you are not limited to them:

1. **Suspicious URLs/domains**
2. **Form analysis**
3. **Content and language**
4. **HTML structure issues**
5. **Link analysis**
6. **Brand impersonation**
7. **Missing legitimacy markers**

Indicator details have been omitted for brevity. See Appendix A for a full version.

****Note:**** Since CSS/JS may be truncated, focus on HTML tags, text content, and URL analysis rather than visual styling or dynamic behaviour.

****Scoring guide:****

- 0-2: Very unlikely phishing (legitimate site)
- 3-4: Low risk (minor suspicious elements)
- 5-6: Medium risk (several concerning indicators)
- 7-8: High risk (multiple clear phishing signs)
- 9-10: Very high risk (obvious phishing attempt)

****Required output format (JSON only):****

```
{  
phishing_score: int [0-10],  
is_phishing: boolean [true/false],  
reasoning: string [Brief explanation of your decision based  
on specific indicators found]  
}
```

****Output Constraints:****

Do only output the JSON-formatted output and nothing

else.

3.4. Used Hardware

For all computations, initially, the Nvidia A100 80GB PCIe GPU and an AMD EPYC 9554 CPU were tested. For the 70b models, a switch to the Nvidia H100 80GB PCIe GPU was necessary. All experiments were conducted on cloud infrastructure from runpod.io¹⁴.

3.5. Experimental Design Setup

The goal of this work is to evaluate the performance of LLMs for phishing detection by assessing their output quality and correctness, runtime, and variability. To ensure a clear structure for this work, the evaluation process is divided into two experiments. The first experiment focuses on randomness, runtime, and output completeness, establishing a baseline for model stability and feasibility. The second experiment then evaluates the models' phishing classification performance and computational costs.

3.5.1. Experiment 1 - Syntactic Performance and Runtime

When using LLMs, randomness is critical factor, especially when aiming to build a reliable and reproducible. Output variability arises because models predict the next output token based on a probability distribution learned during training. Even when presented with identical prompts, the model may select different tokens, leading to variations in output Google Developers (2025); Vaswani et al. (2017).

To mitigate this behaviour, most LLMs have specific parameters, such as `temperature`, `top_p`, and `top_k`, that can be adjusted to reduce output variance.

Temperature: Values near 0 make outputs deterministic by always selecting the most probable token, while values near 1 increase variability by allowing lower-probability tokens to be chosen.

Top_p: Sets a cumulative probability threshold. The model samples only from the smallest set of tokens whose combined probability reaches this threshold.

Top_k: Limits selection to the k most probable tokens. A value of 5 means that only the top five tokens are considered.

For our experiments, the parameters for each model are set to `temperature = 0`, `top_p = 0`, `top_k = 1`, ensuring outputs are as deterministic as possible.

Using these settings, each model uses the prompt from Section Section 3.3 and analyses both the D5 and D50 phishing websites datasets, truncated to different levels (5% and 50% of the original HTML length). Each model analyses the same website five times to measure model output variability and to estimate the total runtime. This results in 3,000 total runs per

¹⁴<https://www.runpod.io/>

dataset (40 websites \times 15 models \times 5 repetitions). For testing the variability and stability of the model results, using the small sample dataset already provides valuable insights that can be applied to larger samples without requiring days of waiting for the results to be computed. To illustrate the point: If the full 1,000 website sample were chosen, instead of 3,000 analysis runs, this number would drastically increase to 75,000.

The goals for this experiment are to answer the following questions:

- Is the output deterministic or do variations occur?
- How long are the model runtimes, and how does the prompt length influence it?
- To what extent do models adhere to the specified JSON output format?

At the end of Experiment 1, models that fail to produce consistent and correctly formatted output or cannot run efficiently on the available hardware are excluded from Experiment 2.

3.5.2. Experiment 2 - Phishing Detection Performance

The findings of Experiment 1 directly inform the design of Experiment 2. Only the models that passed Experiment 1 are included in the analysis, ensuring the evaluation focuses on practical and reliable candidates. At this stage, the full data sample of 1,000 websites is utilised, comprising 500 benign and 500 phishing sites. Due to computational resource constraints, only the D5 version (truncated to 5% of the original HTML length) is utilised. The prompt remains identical to the one from Experiment 1 (see Section 3.3).

The primary goal of this experiment is to evaluate the classification performance of the models using standard metrics such as accuracy, precision, recall, and F1 score. Furthermore, the output coherence is examined by comparing the output variables to ensure that they lead to consistent decisions. For instance, if the boolean phishing variable is `False`, a corresponding low phishing score should be obtained, and the reasoning should reflect this decision. Finally, the experiment assesses the computational costs of running these models, allowing a comparison to commercial approaches. Ultimately, the experiment provides insights into the practical trade-offs between small open-source LLMs in phishing detection and commercial alternatives.

4. Results

This section presents the results of the two experiments. The evaluation experiment (Experiment 1) serves as the setup and feasibility study, focusing on runtime performance, output correctness in terms of the required formatting, and output coherence relating to the required output parameters, which means all parameters should indicate the same decision regarding the website classification. A discussion of the implications of the results for the setup of Experiment 2 follows this. Experiment 2 evaluates the classification performance of the selected LLMs on the phishing detection task.

4.1. Experiment 1

The goal of the first experiment is to assess the practicality of different LLMs for phishing detection scenarios. To this end, the focus lies on three key aspects: runtime performance, output variability, and output correctness. These factors offer initial insights into the stability, reliability, and suitability of each model for potential deployment in productive environments.

4.1.1. LLM Runtime

A first consideration in evaluating practicality is runtime, i.e., the time it takes a model to analyse a single website. Runtime performance is especially critical in operational settings, where large volumes of websites must be processed, and rapid decision-making is required to prevent harm caused by phishing attacks.

Figure 2 shows the total runtime per model for both datasets. While running the experiment using the A100 GPU with a 131,072k context window length, the two 70b models (llama3.3:70B and the deepseek-r1:70B) exceeded the VRAM capacity and therefore produced a runtime error. To still be able to test these two models, a switch to the more powerful H100 GPU was made. Since the 131,072k token context window still exceeded the VRAM, we decided to reduce the window length to 65,536k tokens.

According to the Ollama model list, all models have a context window of 128k tokens. However, we set the value to 131,072k tokens, which potentially allows for longer context than the models can handle, as we found it common practice to use powers of two as a context window limit. Since no prompt used in this work exceeds 128k tokens, this has no impact on the model's performance. If a prompt were to exceed a model's context limit, Ollama automatically truncates the prompt to fit the model's abilities.

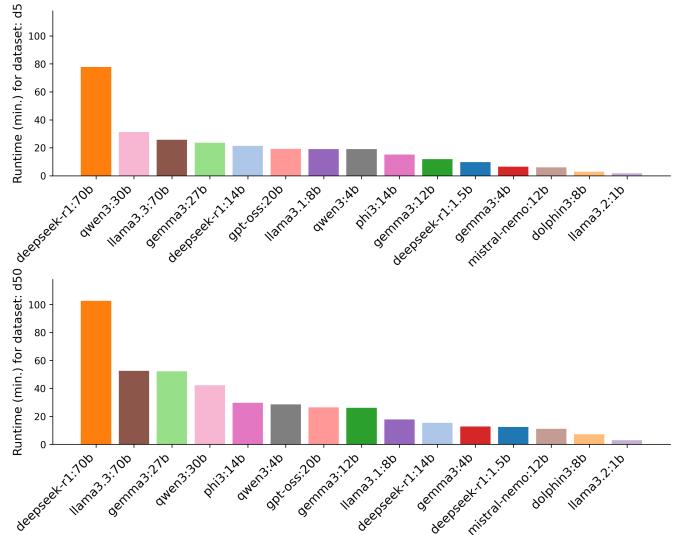


Figure 2: Runtime for each model with the D5 and D50 dataset.

Analysing the 40 websites in the D5 dataset required a total of 291 minutes (~ 5 hours), while analysing the same websites in the D50 dataset took 440.5 minutes (~ 7 hours and 25

minutes), which is about 51.4% longer. This clearly indicates that the prompt length has a significant influence on total runtime, which is a key consideration for Experiment 2, where the dataset size increases to 1,000 websites.

As expected, larger models, such as deepseek-r1:70b and llama3.3:70b, consistently required the most time for each of the two dataset configurations, ranging from 25.8 to 102.5 minutes, whereas the smallest models, like llama3.2:1b and dolphin3:8b, completed the task within just 1.8 to 7 minutes. In terms of runtime per analysis run, this means that deepseek-r1:70b took on average 23.3 and 30.8 seconds per analysis run for the D5 and D50 datasets, respectively. The fastest model, llama3.2:1b, managed an analysis run on average in 0.5 and 0.9 seconds for each of the two datasets. Table A.8 in the Appendix A contains the mean runtime per analysis run for all models.

Interestingly, runtime was influenced not only by model size but also by model family. For instance, the Gemma and Llama models processed data significantly more efficiently compared to others, such as DeepSeek or Qwen models. Interestingly, these two model families (as well as the gpt-oss model) are the only models tested in this work, which are classified as “thinking” models on ollama. This observation suggests that reasoning models take more time to return a result, highlighting that the model architecture plays a major role in runtime performance.

An interesting observation was the significantly longer overall runtime, which indicates that the prompt length plays a crucial role in that. Potential sources of this occurrence might be the result of several factors, including the tokenisation process, as converting longer text into tokens takes more time. The self-attention mechanism scales quadratically with respect to sequence length Wang et al. (2020); Duman Keles et al. (2023) and the key-value (KV) caches, which need to be built and get larger for longer prompts Fu (2024). Figure 3 illustrates the relationship between HTML token count and analysis runtime. The data includes all five analysis runs for each website across the D5 and D50 datasets for each model, totalling 400 data points. The large discrepancy of the HTML token count can be explained by the dataset sampling scheme, which exactly aimed at including websites of all lengths. Most models show a positive correlation, with correlation coefficients ranging from 0.27 to 0.95. This suggests that, on average, all models experience an increase in runtime with longer prompts; however, the strength of this relationship varies between model families rather than being directly tied to model size. For example, Gemma3 or the DeepSeek models exhibit a similar correlation coefficient within their respective families, underscoring the importance of architectural design in runtime efficiency.

Another way to compare the runtime is to compute the relative runtime per model between the two datasets. To do this, the runtime for each model and website in the D5 dataset was divided by the runtime for the same model and website in the D50 dataset. The results were then logarithmised to handle outliers and improve interpretability.

Figure 4 shows the resulting boxplots.

For most models, the boxes fall below 10^0 , indicating that

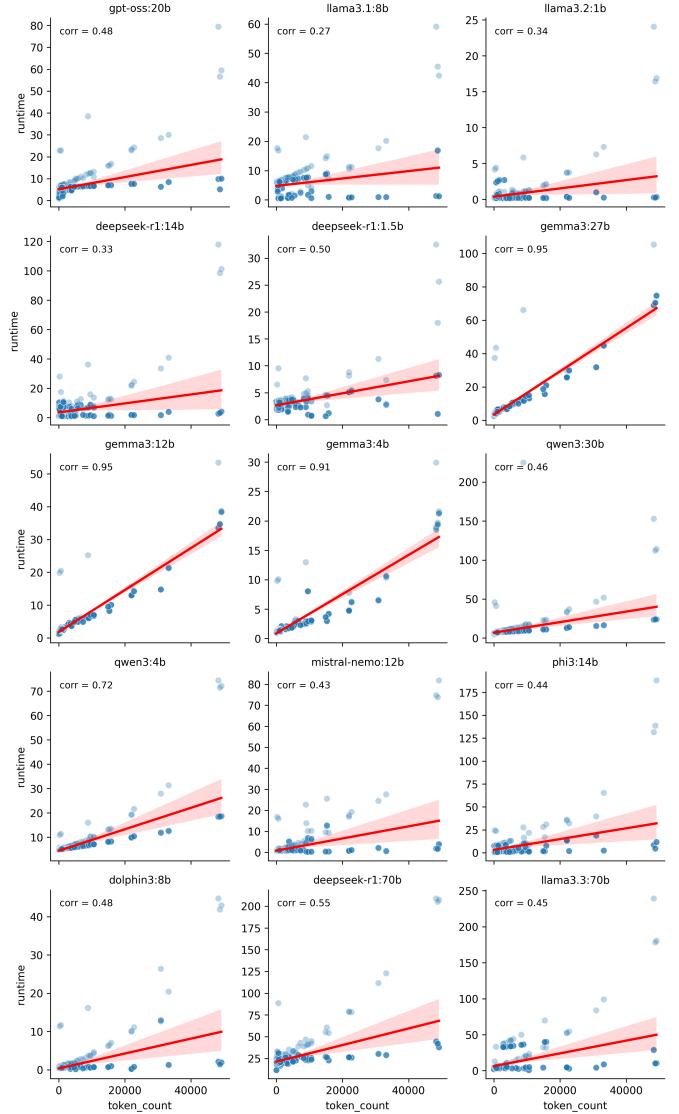


Figure 3: Correlation between model analysis runtime and prompt length. Each dot represents the token count of one website’s HTML code and the corresponding analysis time. Due to websites being analysed multiple times, there is significant overlap between the 200 dots. As a consequence of the sampling scheme, there are a lot more websites towards the lower end of the token count, as they are more prevalent in the dataset.

shorter website token counts (D5) lead to lower runtimes, as expected. However, there are notable exceptions, including deepseek-r1:14b and llama3.1:8b models, which show reduced runtimes for longer prompts. Especially the smallest models, like the llama3.2:1b or llama3.1:8b only take a very short time (around 1 second) to analyse a website. Therefore, even small changes in analysis time can skew the results, making it take longer to process the D5 websites. The systematic behaviour of the deepseek-r1:14b and the numerous outliers of the mistral-nemo models cannot be explained as easily. Potential sources for these results are the internal model structure and optimisations, or the type of model quantisation. Especially the quantisation might be the reason why the other two deepseek models perform as expected. The boxplots also highlight runtime vari-

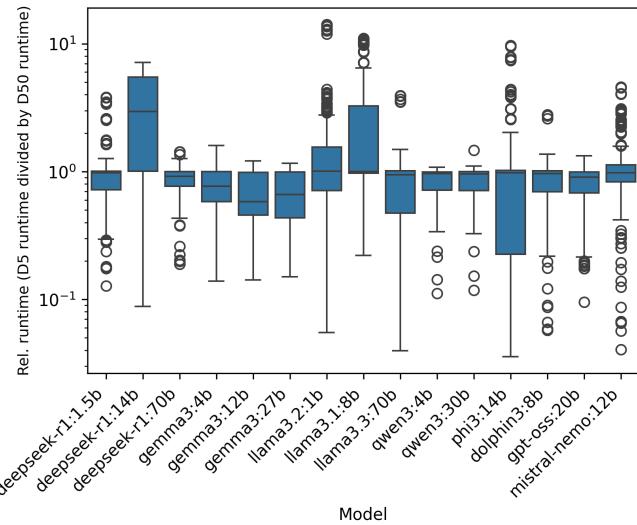


Figure 4: Distribution of the logarithm of the relative analysis time differences between the D5 and D50 datasets for each model. The distributions are based on 200 analysis runs per model, meaning that each website’s time difference is included five times (once per run). Values below 0 indicate that analysing a website with fewer HTML tokens (D5) was faster than analysing the same website with more tokens (D50).

ability. Some models, like the Gemma3, exhibit low runtime variability, indicated by compact boxes and a lack of outliers. In contrast, models like mistral-nemo and deepseek-r1:1.5b exhibit high variance, with outliers in both directions.

In summary, Experiment 1 examined the runtime of the tested LLMs. The key findings are:

- Prompt length strongly affects runtime, with longer inputs leading to higher computational costs, for almost all models.
- Model family, rather than model size alone, is a key determinant of runtime efficiency.
- Gemma and Llama models are generally faster and more consistent than DeepSeek and Qwen models.
- Runtime variability plays a crucial role in determining model practicality, with low-variance models being more predictable.

4.1.2. Syntactic Performance

Adherence to the specified output format is another important quality of a model’s reliability and usability. Consistent formatting allows the results to be analysed automatically and compared systematically. In this study, models were instructed to produce their output strictly in a JSON format (see Prompt Template Section 3.3), containing exactly three items: `is_phishing`, `phishing_score`, and `reasoning`.

Table 3 shows for each model the number of analysis runs (per website) in which the required JSON output was successfully included. In practice, some models, particularly the DeepSeek family, produced the requested format but also added additional information, such as their thought process, alongside the JSON

object. An example of this behaviour is shown in Model Output 1 in Appendix A. As long as the correct JSON format was present within the result, the run was counted as successful.

Table 3: This table shows, for each model and website, how many of the five analysis runs produced valid JSON output. For example, the deepseek-r1:14b model generated valid JSON output in all five analysis runs for 78 websites. For one website, it never produced valid JSON output (column 0), and for another website, valid output was produced in only one run (column 1). In total, this covers all 80 websites (40 from the D5 dataset and 40 from the D50 dataset). The table is sorted in descending order according to column “5”.

Model Name	# Runs per website with correct JSON output					
	0	1	2	3	4	5
deepseek-r1:70b	0	0	0	0	0	80
gemma3:12b	0	0	0	0	0	80
gemma3:27b	0	0	0	0	0	80
deepseek-r1:14b	1	1	0	0	0	78
gemma3:4b	2	0	0	0	0	78
llama3:3:70b	2	1	0	0	1	76
dolphin3:8b	8	0	0	0	1	71
phi3:14b	1	5	0	0	3	71
mistral-nemo:12b	8	2	0	0	0	70
qwen3:4b	23	0	0	0	0	57
llama3:1:8b	23	6	0	0	8	43
deepseek-r1:1.5b	33	10	0	0	7	30
gpt-oss:20b	45	6	1	0	5	23
llama3:2:1b	77	0	0	0	1	2
qwen3:30b	77	1	0	0	0	2

Since all model parameters were configured to produce deterministic results (see Section 3.5.1), the observed behaviour is somewhat surprising. Under the assumption of determinism, we would expect that when analysing the same website with a given model, the output format would be consistent in all five runs—either always correct or always incorrect, which is represented by the categories 0 and 5 in Table 3.

However, some models exhibit inconsistent behaviour, producing correct JSON output only in some of the five runs (once, twice, or four times). This indicates that their output was not entirely stable. A closer inspection of the results revealed a common pattern: in some cases, the first analysis run differed from all subsequent runs. An example of this behaviour for the gemma3:4b model is shown in Model Output 2 in the Appendix A. The most plausible reason for this phenomenon is the model loading and warm-up behaviour, which includes factors such as kernel caching and background telemetry, combined with the inference framework in use, that can introduce unwanted variability Ratul et al. (2025). Since the problem cannot be easily eliminated, we decided to use a practical mitigation strategy: discard the first run and treat only the subsequent runs as true deterministic output. An exception was observed with gpt-oss, which produced the correct JSON format in only two out of five runs for exactly one website. This behaviour can be attributed to the token output limit we applied in our setup. Some models ignored the instruction and included additional

explanatory text before the actual JSON output. When the maximum token limit was reached, the output was truncated by the model, sometimes cutting off parts or all of the JSON structure, thereby rendering the result unparsable. We decided that models that failed to produce the required output format at least four times per website across more than half of all analysed websites will be excluded from Experiment 2.

4.1.3. Output Stability

In this section, we focus on output stability, specifically whether models consistently produce the same decisions across all runs. Each model is required to generate a JSON format output containing three items (see Section 3.3): `is_phishing`, `phishing_score`, and `reasoning`. Therefore, in cases where the model delivered the required format, the variables were extracted and compared to determine if the information was consistent. Ensuring coherent output is a key criterion for evaluating a model’s reliability and is essential for building trust in model decisions. There are two desirable scenarios:

- `is_phishing = True, phishing_score ≥ 5, reasoning says it is phishing,`
- `is_phishing = False, phishing_score < 5, reasoning says it is not phishing.`

Since the phishing score can take on eleven possible values (0-10), it is unclear whether a score of five should be considered phishing detection or not. In cybersecurity-related tasks, it is often unclear which type of error (missing a phishing attempt or falsely flagging a benign website) is worse, as the former can potentially be costly when an intrusion occurs. The latter disrupts the user experience [Dalvi et al. \(2019\)](#); [Kavya & Sumathi \(2024\)](#); [Bold et al. \(2022\)](#).

We prioritise detecting phishing websites, as we deem the cost of missing an attack, especially in critical infrastructure, to be higher than incorrectly blocking a benign site. Therefore, a **LLM** score of five is considered as `True` when translating it into a binary classification.

In our analysis, we analyse coherence from two complementary perspectives: i) a quantitative consistency analysis, examining the alignment between the binary decision (`is_phishing`) and the certainty score (`phishing_score`), and ii) qualitative coherence, examining the reasoning texts, with respect to the occurrence of certain key words that are often connected to the classification into phishing and benign websites. The quantitative part provides a direct measure of consistency within the result of one website analysis, ensuring alignment between the numeric outputs. The qualitative part focuses on the frequency of phishing-related and legitimacy-related words and phrases, as well as their negations, which captures how models articulate their decision-making and provides insight into explanation strategies and their plausibility.

Quantitative Analysis

The results of the quantitative consistency analysis can be found in Table 4, which aims to highlight the relation between the two

quantitative phishing measures - the boolean decision and the phishing score.

The model output shows coherence in almost all cases between the boolean decision and the assigned score. Almost all inconsistencies can be attributed to cases where the boolean decision was `False` while the corresponding score was 5. An exception was observed with the Mistral model, which produced a different type of inconsistency by assigning a score of 6 to the decision `False`. The `Nan` rows and columns indicate instances where the model failed to produce a valid JSON output.

Notably, no model ever assigned the highest phishing score of 10, while the lowest score of 0 occurred frequently across models. In four cases, the llama3.1:8b model determined that twelve is an appropriate score, which, of course, is considered an error, as the score should be ten or lower. The tendency of the models to assign the lowest available scores and never the highest possible score indicates that models might find it easier to conclude when a website is benign based on the absence of phishing indicators. This behaviour could also be influenced by the prompt design chosen, where models check all the criteria of the provided examples, and if none are met, assign a low score. Another reason this result might appear is that phishing websites often contain elements that can also be found in benign websites. This can lead to the model being less certain when deciding on phishing, as indicators from both benign and malicious websites are present simultaneously.

Overall, the models showed a very similar behaviour when giving scores for the respective boolean decision, as well as the amount of missing JSON outputs (NaNs). This indicates that using longer HTML code does not guarantee a different (or better) classification result.

The few exceptions are the gemma3:4b model, which assigned significantly more scores `<5` for the D5 dataset compared to the D50 dataset, and the gpt-oss:20b, llama3.2:1b and the qwen3:4b models, where the number on NaN outputs was noticeably smaller for the D50 dataset. Overall, the results suggest that the D5 dataset provides sufficient data for a reasonable analysis.

Diving deeper into the dataset differences reveals that most models retain their assessment for the same website, as shown in Table 5. Eight models score 30 or above (out of 40 possible websites) in column “0”, which means that for 75% or more of the websites, the models produced the same boolean result in all 5 analysis runs, regardless of the dataset. Mainly, the larger tested models stay consistent between the two datasets, with the notable exception being the llama3.2:1b model. However, this model produced almost exclusively NaN values in both cases.

The second most common case is that models changed all five of their assessments, shown in column “5”. According to Table 3, the number of NaN classifications for the concerning models changes notably between the two datasets. This indicates that models did not change their opinion but rather that the changes are mostly related to the output being JSON or not.

The remaining observed differences, denoted in columns “1” to “4”, are related to the fact that the output of some models

¹³The model assigned a score of 12.

Table 4: This table shows the phishing scores and the classification decision (Phishing Cat) assigned by the models across the five analysis runs per website (200 runs per dataset). Each model and dataset combination has three possible classification decisions (True, False, NaN). Each decision is displayed as a separate row. In cases where models did not include any of the decisions, the corresponding row is omitted. The three rightmost columns display the number of times no JSON output was returned (NaN) and summarise the assigned scores in two groups: <5 and ≥ 5 . In red we highlight where the score does not match the decision.

Dataset	Model	Phishing Cat	Phishing Score										<5	≥ 5		
			0	1	2	3	4	5	6	7	8	9	10			
d5	deepseek-r1:1.5b	False	31	0	0	45	0	8	0	0	0	0	0	0	76	8
d5	deepseek-r1:1.5b	True	0	0	0	0	0	0	0	5	30	0	0	0	35	
d5	deepseek-r1:1.5b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	81	0
d5	deepseek-r1:14b	False	56	9	10	73	1	0	0	0	0	0	0	0	149	0
d5	deepseek-r1:14b	True	0	0	0	0	0	23	0	19	5	0	0	0	0	47
d5	deepseek-r1:14b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	4	0
d5	deepseek-r1:70b	False	4	64	43	28	6	0	0	0	0	0	0	0	145	0
d5	deepseek-r1:70b	True	0	0	0	0	0	14	5	17	17	2	0	0	0	55
d5	dolphin3:8b	False	10	5	76	34	0	0	0	0	0	0	0	0	125	0
d5	dolphin3:8b	True	0	0	0	0	0	10	25	30	0	0	0	0	0	65
d5	dolphin3:8b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	10	0
d5	gemma3:12b	False	0	20	20	60	5	0	0	0	0	0	0	0	105	0
d5	gemma3:12b	True	0	0	0	0	0	65	25	5	0	0	0	0	0	95
d5	gemma3:27b	False	0	0	0	35	35	30	0	0	0	0	0	0	70	30
d5	gemma3:27b	True	0	0	0	0	0	5	40	45	10	0	0	0	0	100
d5	gemma3:4b	False	0	0	0	25	0	0	0	0	0	0	0	0	25	0
d5	gemma3:4b	True	0	0	0	0	0	5	40	90	35	0	0	0	0	170
d5	gemma3:4b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	5	0
d5	gpt-oss:20b	False	15	48	8	8	0	0	0	0	0	0	0	0	79	0
d5	gpt-oss:20b	True	0	0	0	0	0	4	0	0	7	2	0	0	0	13
d5	gpt-oss:20b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	108	0
d5	llama3:1:8b	False	6	0	14	1	0	0	0	0	0	0	0	0	21	0
d5	llama3:1:8b	True	0	0	0	0	0	9	0	0	34	44	0	0	0	87
d5	llama3:1:8b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	92	0
d5	llama3:2:1b	False	0	0	0	0	0	0	5	0	0	0	0	0	0	5
d5	llama3:2:1b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	195	0
d5	llama3:3:70b	False	5	14	101	0	0	0	0	0	0	0	0	0	120	0
d5	llama3:3:70b	True	0	0	0	0	0	0	5	66	9	0	0	0	0	80
d5	llama3:3:70b	NaN	0	0	0	33	51	1	4	6	0	0	0	0	85	10
d5	mistral-nemo:12b	False	0	0	0	0	0	5	8	32	22	33	0	0	0	100
d5	mistral-nemo:12b	True	0	0	0	0	0	0	0	0	0	0	0	0	0	5
d5	mistral-nemo:12b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d5	phi3:14b	False	84	6	21	5	0	0	0	0	0	0	0	0	116	0
d5	phi3:14b	True	0	0	0	0	0	5	4	37	15	10	0	0	0	71
d5	phi3:14b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d5	qwen3:30b	False	5	4	0	0	0	0	0	0	0	0	0	0	0	9
d5	qwen3:30b	True	0	0	0	0	0	1	0	0	1	0	0	0	0	2
d5	qwen3:30b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	189	0
d5	qwen3:4b	False	95	0	0	0	0	0	0	0	0	0	0	0	0	95
d5	qwen3:4b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	105	0
d50	deepseek-r1:1.5b	False	1	0	0	4	0	12	0	0	0	0	0	0	5	12
d50	deepseek-r1:1.5b	True	4	0	0	0	1	0	11	0	36	0	0	0	4	48
d50	deepseek-r1:1.5b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	131	0
d50	deepseek-r1:14b	False	50	0	0	79	0	0	0	0	0	0	0	0	129	0
d50	deepseek-r1:14b	True	0	0	0	0	0	31	0	21	14	0	0	0	66	
d50	deepseek-r1:14b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	5	0
d50	deepseek-r1:70b	False	11	78	31	0	14	0	0	0	0	0	0	0	134	0
d50	deepseek-r1:70b	True	0	0	0	0	0	12	4	30	19	1	0	0	0	66
d50	dolphin3:8b	False	0	5	29	20	0	0	0	0	0	0	0	0	54	0
d50	dolphin3:8b	True	0	0	0	0	0	26	48	41	0	0	0	0	0	115
d50	dolphin3:8b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	31	0
d50	gemma3:12b	False	0	10	5	40	15	0	0	0	0	0	0	0	70	0
d50	gemma3:12b	True	0	0	0	0	0	65	45	20	0	0	0	0	0	130
d50	gemma3:27b	False	0	0	0	15	45	10	15	0	0	0	0	0	60	25
d50	gemma3:27b	True	0	0	0	0	0	35	50	30	0	0	0	0	0	115
d50	gemma3:4b	True	0	0	0	0	0	40	70	85	0	0	0	0	0	195
d50	gemma3:4b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	5	0
d50	gpt-oss:20b	False	5	27	5	4	0	0	0	0	0	0	0	0	41	0
d50	gpt-oss:20b	True	0	0	0	0	0	0	0	0	4	6	0	0	0	10
d50	gpt-oss:20b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	149	0
d50	llama3:1:8b	False	10	0	4	0	1	0	0	0	0	0	0	0	0	15
d50	llama3:1:8b	True	0	0	0	0	0	5	0	1	81	39	4 ¹⁵	0	0	130
d50	llama3:1:8b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	55	0
d50	llama3:2:1b	False	0	0	0	0	0	0	0	0	0	0	0	0	0	5
d50	llama3:2:1b	True	0	0	0	0	0	0	0	0	4	0	0	0	0	4
d50	llama3:2:1b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	191	0
d50	llama3:3:70b	False	5	5	59	0	0	1	5	0	0	0	0	0	0	69
d50	llama3:3:70b	True	0	0	0	0	0	0	0	0	101	9	0	0	0	110
d50	llama3:3:70b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	15	0
d50	mistral-nemo:12b	False	5	0	15	16	0	4	0	0	0	0	0	0	36	4
d50	mistral-nemo:12b	True	0	0	0	0	0	0	1	42	39	35	0	0	0	117
d50	mistral-nemo:12b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	43	0
d50	phi3:14b	False	19	5	5	14	0	0	0	0	0	0	0	0	0	43
d50	phi3:14b	True	0	0	0	0	0	4	0	116	12	10	0	0	0	142
d50	phi3:14b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	15	0
d50	qwen3:30b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	200	0
d50	qwen3:4b	False	190	0	0	0	0	0	0	0	0	0	0	0	0	190
d50	qwen3:4b	NaN	0	0	0	0	0	0	0	0	0	0	0	0	10	0

Table 5: Overview of the classification consistency across models for the D5 and D50 analyses. For each website, the number of differing classifications across five runs is counted. A value of 0 indicates identical results in all runs, while a value of 5 means that each run produced different results. The counts are then aggregated over all 40 websites for each model. The models are listed in descending order according to column “0”.

Classification result differences between the D5 and D50 dataset per website runs

Qualitative Analysis of Reasoning Outputs

To complement the coherence checks, we conducted a qualitative analysis of the models’ reasoning outputs, i.e., the explanatory texts contained in the `reasoning` key of the JSON output. The analysis focuses on the language used by the models to justify their decisions. Specifically, we scanned the reasoning texts for the occurrence of predefined keywords: (i) phishing indicators, terms associated with malicious or suspicious behaviour (e.g. “phishing”, “suspicious”, “fraudulent”, “deceptive”), and (ii) legitimacy indicators, terms associated with safe or trusted websites (e.g. “legitimate”, “safe”, “no phishing indicators”, “trusted”). These keywords were derived by manually analysing model outputs during the experimentation phase. For each model–dataset pair, we counted the total mentions of each category and reported them as `phishing_score_tendency` and `legit_score_tendency`, respectively. While this keyword-based approach is inherently high-level and depends on the defined keyword set, it provides an indication of the style and focus of the models’ explanation.

The results, summarised in Table A.9, which can be found in Appendix A, reveal differences in how models frame their reasoning. Some models, such as gemma3:4b, use many terms that we, based on our list, would associate with a phishing website. This result is also in line with the observation that the gemma3:4b model assigned almost all websites to be phishing. Others, such as mistral-nemo, return a more balanced amount of phrases and keywords associated with phishing or benign web-

sites, which is again aligned with a more balanced boolean decision profile. We also observe variability across datasets: for instance, dolphin3:8b uses more language associated with phishing websites on the d5 dataset compared to the d50 dataset.

4.1.4. Consequences for the setup of Experiment 2

The results from Experiment 1 were instrumental in determining an efficient and reliable setup for Experiment 2. We excluded the following models, due to their failures to generate the required JSON result format: llama3.1:8b, llama3.2:1b, gpt-oss:20b, qwen3:4b, qwen3:30b and deepseek.r1:1.5b.

Furthermore, we observed that models are consistent after the first analysis run Section 4.1.2. To account for this, only two runs per website will be conducted. The first run serves as initiation and will be discarded, while the second run will be used as final decision output.

To maintain reasonable runtimes analysing the larger dataset, all websites will be truncated using the D5 truncation scheme, retaining at most 5% of the original HTML code for analysis. This choice is supported by the quantitative coherence check, which showed minimal performance differences between the D5 and the D50 datasets.

Finally, while the context window size could be reduced due to the smaller token number in the D5 scheme, we decided against modifying model parameters, as doing so could negatively impact comparability between the experiments.

4.2. Experiment 2

In this experiment, the objective is to evaluate the classification performance of the models on a larger dataset comprising 1,000 websites. Performance is measured using well-established evaluation metrics, namely accuracy, precision, recall, and F1-score.

$$\text{Accuracy} = \frac{TP+TN}{P+N}$$

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

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

$$\text{F1-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

where: P = Positive, N = Negative, TP = True Positive, TN = True Negative,

FP = False Positive, FN = False Negatives

Table 6 and Table 7 show the performance results.

For Table 6, the boolean decision was taken as the predicted label, whereas Table 7 shows the metrics based on the numeric decision of the models. Numeric values smaller than five correspond to a decision of “not phishing” and all values of five or larger correspond to the decision “phishing”.

The evaluation in Table 6, based on the boolean decision output, highlights the relative strengths of the different models in phishing website classification. When prioritising overall effectiveness, the F1-score and accuracy are particularly informative, as both false positives and false negatives can have serious consequences in this domain.

Table 6: Summary of performance metrics per model based on the boolean classification decision, sorted in descending order by F1 score.

Model	F1_score	Accuracy	Precision	Recall	Nan Proportion
llama3.3:70b	0.893	0.887	0.845	0.948	0.001
deepseek-r1:70b	0.873	0.865	0.824	0.929	0.073
mistral-nemo:12b	0.858	0.849	0.811	0.909	0.009
deepseek-r1:14b	0.842	0.842	0.810	0.877	0.086
gemma3:27b	0.835	0.809	0.737	0.964	0.004
gemma3:12b	0.828	0.804	0.734	0.951	0.008
gemma3:4b	0.690	0.562	0.533	0.980	0.007
dolphin3:8b	0.651	0.729	0.913	0.506	0.000
phi3:14b	0.462	0.601	0.701	0.345	0.014

Table 7: Summary of performance metrics per model based on the numeric classification decision, sorted in descending order by F1 score.

Model	F1_score	Accuracy	Precision	Recall	Nan Proportion
llama3.3:70b	0.892	0.885	0.842	0.948	0.001
deepseek-r1:70b	0.870	0.862	0.821	0.927	0.073
mistral-nemo:12b	0.843	0.823	0.761	0.946	0.009
deepseek-r1:14b	0.842	0.839	0.795	0.895	0.086
gemma3:12b	0.828	0.803	0.732	0.951	0.008
gemma3:27b	0.764	0.690	0.618	1.000	0.004
dolphin3:8b	0.651	0.729	0.913	0.506	0.000
phi3:14b	0.462	0.601	0.701	0.345	0.014

Llama3.3:70b emerges as the strongest overall model, achieving the highest F1-score (0.893) and accuracy (0.887). It maintains a good balance between precision (0.845) and recall (0.948), with an almost negligible proportion of NaNs (0.001), making it highly reliable for deployment. Similarly, deepseek-r1:70b achieves a strong F1-score (0.873) and accuracy (0.865), balancing precision (0.824) and recall (0.929), while maintaining consistent output formatting (NaN proportion: 0.073).

Mistral-nemo:12b also performs strongly, with an F1-score of 0.858 and accuracy of 0.849. It emphasises a balanced trade-off between precision (0.811) and recall (0.909), and a low rate of missing predictions (0.009). The best-performing models all exhibit higher recall than precision, indicating greater success in identifying actual phishing sites than in correctly identifying benign websites.

Mistral-nemo:12b, deepseek-r1:14b, gemma3:27b and gemma3:12b show solid all-round performance (F1-score 0.828-0.858, accuracy 0.804-0.865), and similar error patterns compared to the top models. While the overall performance is slightly worse than for the 70b models, this tier of models is considerably smaller, which affects deployability and analysis runtime.

At the lower end, gemma3:4b, dolphin3:8b, and phi3:14b show the weakest results, with F1-Scores (0.462-0.690) and accuracies (0.562-0.729).

Overall, models such as llama3.3:70b and deepseek-r1:70b provide the most balanced and reliable performance, excelling in both F1-score and accuracy. Mistral-nemo:12b and gemma3:27b also perform strongly when maximising recall is essential, while deepseek-r1:14b offers dependable balanced detection. Overall, model size impacts the result performance, with the largest models performing the best. However, considerably smaller models only perform marginally worse.

The results in Table 7 present a similar picture, where instead of the boolean decision, the numeric categorisation was

taken as the decision criterion. Consistent with Experiment 1, we observe that models only rarely assign conflicting Boolean decisions and numeric assessments. Overall, however, measuring the model’s performance based on the phishing score results in a slightly worse performance across most performance metrics in almost all cases. A notable outlier is gemma3:27b, which now achieves only a 0.764 F1-score, down from 0.835 for the Boolean decision. In conclusion, the models’ binary decisions yield better results across almost all tested models, indicating that relying solely on a binary decision can be sufficient.

In Section 2, the results of comparable approaches are summarised. Most prior studies report results using large, proprietary models, such as Gemini Pro 1.0 or GPT-4, with F1-scores reaching up to 0.99. However, due to differences in the analysis architecture and datasets, these results should be interpreted with caution, as direct comparisons are not always meaningful. Overall, the evidence suggests that large models tend to outperform smaller ones, which is consistent with expectations. Nevertheless, it is encouraging to see that the tested 70b models come close to 90% F1-score and accuracy, while more compact models sit around 85% for both metrics. This shows the improvement in small open-source models, which previously achieved only around 74% accuracy or were worse, for state-of-the-art models one or two years ago Koide et al. (2024). This suggests that smaller, more efficient models are closing the performance gap, making them increasingly viable for practical phishing detection scenarios.

5. Discussion

The goal of this study was to evaluate how small, non-fine-tuned LLMs perform on a phishing detection task. While the phishing detection performance of SLMs is promising, these models still lag behind proprietary large-scale models. In the following, we discuss the advantages and disadvantages of both approaches, highlighting the trade-offs related to performance, feasibility, costs, and data privacy.

5.1. Costs

One of the primary advantages of small local LLMs is their low operational cost. Running models locally eliminates recurring API fees that are typical of proprietary services, making them particularly attractive for organisations with limited budgets that aim for large-scale deployments, where models are queried almost continuously. The following hosting options are typically available:

- **Proprietary API:** In related literature, GPT4o or GPT4-turbo are frequently used for phishing detection. At the time of writing, GPT4o is priced at \$2.5 per million input tokens and \$10 per million output tokens, while GPT4-turbo costs \$10 per million input tokens and \$30 per million output tokens. The newly released GPT5 models are priced at a similar level OpenAI (2025).

Across both experiments in this study, each model processed approximately 8 million input tokens. While output token counts varied between models, using gpt-oss as

an example, which belongs to the same family as GPT4, generated roughly 1.6 million output tokens in total. Using these figures, the total cost of running the experiments would be approximately \$36 (31€) for GPT4o, or \$128 (109€) for GPT4o-turbo.

- **Renting GPUs:** In this study, models were executed on rented GPUs via runpod, with the following rates: Nvidia A100, 1.4€ per GPU hour, and Nvidia H100, 2.03€ per GPU hour. The total runtime for Experiment 1 and 2 was slightly over 28 hours, resulting in a total cost of approximately 41€ for the A100 setup, including disk storage fees. For the two largest 70b models, which required H100 GPU, the total runtime was 22 hours and 20 minutes, amounting to roughly 46€, including disk storage.
- **Fully local Setup:** As the Nvidia A100 is no longer widely available, the latest GPU generation (H100 series) is an option, which costs between 27,000€ to 39,000€, depending on the provider and specifications ^{16, 17}. In addition to the initial investment, electricity, cooling, and maintenance costs must be factored in, which can vary significantly by region and infrastructure.

Taking all these factors into account, the break-even point where investing in local hardware becomes more economical than relying on proprietary APIs ranges between 625 million and 3.6 billion tokens (input and output combined). This range depends on both the GPU price and the selected GPT model. With a high GPU price and a cheaper GPT version, the hardware costs only balance out after a substantial number of tokens are processed. For perspective, in this study, gpt-oss processed about 9.6 million tokens to analyse 2,400 websites. At this rate, the break-even point would be reached after roughly 160,000 to 900,000 analysis runs. Considering that the APWG reported over 1 million phishing websites in Q2 2025 alone APWG (2025), running a GPU with local LLMs becomes economically viable compared to using commercial models within just a few months. However, the results in Section 4.2 show that models take several seconds to analyse a website. Therefore, (repeated) scans of hundreds of thousands of websites call for more efficient models or, more likely, parallelised analysis setups. Considering the costs for the three approaches, none of them remains economically feasible. A potential solution is to reduce the number of websites to analyse by combining classical ML approaches as preselectors and LLMs as final decision-makers.

5.2. Benefits of SLMs

The following advantages make (locally-hosted) SLMs attractive for phishing detection and similar security-critical applications.

¹⁶<https://www.newegg.com/p/pl?d=h100> (11 September 2025)

¹⁷<https://geizhals.at/nvidia-h100-nvl-900-21010-0020-000-a3356480.html> (11 September 2025)

- **Data Privacy, Control and Security:** Running all inference on local infrastructure ensures that sensitive information such as URLs, HTML content, and user metadata remains internal and is not transmitted to external providers. This is particularly important for organisations subject to strict data protection regulations or operating in sensitive domains. In the context of phishing detection, local models allow organisations to maintain full control over their data, reducing exposure to external systems. Furthermore, keeping the analysis methods and intellectual property (IP) in-house provides a competitive advantage, preventing sensitive prompts and detection strategies from being shared with third parties. It also lowers the risk of manipulation or infiltration by adversaries.
- **Customisability:** Although models in this study were evaluated in their out-of-the-box state, local models can be fine-tuned for phishing detection. Organisations with relevant expertise can leverage their proprietary data to partially retrain model weights or implement a retrieval-augmented generation (RAG) system to improve performance. The current open-source **LLM** ecosystem offers a wide variety of models for domain-specific adaptations. Furthermore, fine-tuned models for related tasks are often shared on platforms such as Hugging Face. However, to the best of our knowledge, no fine-tuned models specifically for phishing detection are publicly available at this time.
- **Independence and Availability:** Local model deployment eliminates vendor lock-in, ensuring that operations are not dependent on the availability, pricing policies, or strategic decisions of external providers. Local models can also offer greater reliability, as they are not affected by cloud service outages or external network issues. In addition, they can deliver lower latency and faster response times, which is particularly beneficial for time-critical applications such as real-time phishing detection.

5.3. Challenges of SLMs

Despite these benefits, small local **LLMs** come with a set of limitations that must be acknowledged.

- **Performance:** The most evident drawback is the lower performance of small models compared to larger proprietary ones. The results achieved in this study are promising; the tested 70b models, for instance, demonstrated solid overall performance, while other models excelled in specific areas such as precision or recall. However, the best proprietary large models used in related work outperformed local models across all performance metrics. Although differences in datasets and detection pipelines make direct comparisons difficult, it is reasonable to assume that a similar performance gap would persist in this experimental setup. This gap has practical implications, leading to higher rates of false positives or false negatives. These errors can have severe consequences in

security-critical applications such as phishing detection, including missed threats or unnecessary disruptions.

- **Customisability:** Customisability is one of the key strengths of local models, but realising it is non-trivial. First, a suitable base model must be selected, which requires in-depth knowledge of available models and careful evaluation of their strengths and weaknesses. Second, high-quality, domain-specific data is needed to achieve a meaningful fine-tuning result. Finally, depending on the specific approach, fine-tuning itself can be a computationally intensive process, requiring specialised hardware and expertise.
- **Hardware cost:** Running even relatively small models locally requires specialised hardware, such as GPUs or high-performance CPUs, and the technical expertise to set up, optimise, and maintain the infrastructure. For organisations without existing machine learning infrastructure, the initial hardware investment can be substantial. In short-term or exploratory projects, this upfront investment may be too high, making proprietary models or rental GPU services more cost-effective alternatives. Ultimately, the exact cost must be carefully calculated based on the organisation's specific strategy.
- **Scalability:** Proprietary cloud-based solutions offer elastic scaling with demand. In contrast, local deployments are constrained by on-site hardware capacity. Handling sudden spikes in phishing detection workloads may require over-provisioning hardware or accepting degraded performance.

5.4. Application Recommendation

Based on the highlighted advantages and disadvantages in Section 5, determining a single optimal setup is challenging, as the decision depends heavily on organisational priorities and operational constraints. With the results of the current study, two primary paths emerge:

Proprietary models for high performance. If high performance and accuracy are the primary objectives and economic considerations are less critical, proprietary models are the preferred choice. This is particularly true when the dataset size of websites to analyse is small to moderate, or if the analysis is only performed infrequently. Ease of deployment, scalability, and superior performance outweigh the recurring costs.

*Local **LLMs** for cost efficiency and privacy.* The experiments conducted in this work demonstrate that there is potential in running smaller local **LLMs** for phishing detection, particularly when cost, data control, and privacy are key priorities. Using local models in combination with customisation techniques could become sustainable long-term solutions. While the initial investment for local model deployment may be substantial, these upfront costs can be offset over time. Furthermore, having full ownership of the model enables continuous improvements through iterative fine-tuning.

Among the tested local models, llama3.3:70b, deepseek-r1:70b and mistral-nemo:12b stand out as the most promising options out of the box. The two 70b models demonstrated particularly strong performance when given raw HTML from websites and their corresponding URLs. While all three models are small enough to be fine-tuned at reasonable computational cost, especially the Mistral model, which only consists of 12 billion parameters, is a promising candidate for testing fine-tuned setups. Regarding analysis runtime, the Llama and Gemma models are very similar, while the Mistral model, among the fastest three tested models, delivers its analysis rapidly.

Another consideration that needs to be accounted for is the type of response variable that the models should use. In Experiment 2, we observed that relying on a Boolean classification decision yielded better results than relying on the numeric score. A potential reason may originate from the chosen few-shot prompt design. Models may base their score just on the provided examples by matching them to the techniques found in the HTML code. Due to the nature of few-shot prompting, the list of provided examples is not exhaustive, which means models may miss phishing techniques not included and, therefore, receive a worse score. The llama3.2:1b model at some point responded with a Python script, which basically counted the number of occurrences of the phishing techniques mentioned in the prompt.

Generally, models with high runtime variability are harder to manage in production environments, as their runtime behaviour is less consistent and may lead to uncertainty in planning.

6. Limitations

Although we carefully conducted this work, we acknowledge some limitations.

A more detailed investigation into the linguistic patterns of the reasoning outputs could provide additional insights. Potentially, more advanced [Natural Language Processing \(NLP\)](#) methods or even [LLMs](#) could be used to analyse the text more thoroughly. However, such analysis is beyond the scope of this work. In this study, we focused exclusively on HTML code as input for phishing detection models. This choice enabled the inclusion of a broader range of [SLMs](#) and allowed for clearer, more controlled comparisons between models. While we expect performance improvements from incorporating multimodal data (e.g., text and screenshots), a careful analysis of such approaches is left for future research.

7. Conclusion

In this work, we evaluated and benchmarked 15 small, local (non-fine-tuned) [LLMs](#) for the task of website phishing detection. The tested models represent a diverse selection of frequently used architectures, ranging from 1b to 70b parameters. Over the course of two experiments, we systematically assessed analysis runtime, output coherence and classification performance. Of the 15 models, nine were deemed practical

and provided accurate results on a small sample dataset. These models were then further evaluated on a larger dataset of 1,000 websites.

The results show that overall classification accuracy for most models ranges from 56% to 89%, with most models achieving 80% or more. However, more substantial differences were observed in precision. For example, the gpt:oss model achieved an impressive precision of 98% in test runs, but numerous missing JSON outputs and a low recall score hinder its applicability; it was therefore excluded. Models with a more balanced F1-score, such as llama3.3:70, deepseek-r1:70b or mistral-nemo:12b, are more practical in most use cases, as they strike a balance between relatively high accuracy and moderate error rates.

A key observation is that even some mid-sized models in the 10b-20b parameter range achieve F1 Scores comparable to those of previous generations of 70b models. While the current-generation 70b models can outperform proprietary models such as GPT3.5. This reflects a promising trend toward smaller, more efficient models becoming increasingly viable for phishing detection. Tracking future developments of these models will be particularly interesting.

While the out-of-the-box local models tested in this work do not fully match the performance of state-of-the-art large proprietary [LLMs](#), they offer several advantages, including greater data privacy, customisation potential, and operational independence. Regarding deployment cost, [SLMs](#) become a viable option when a heavy workload over an extended period of time is expected, as then the upfront cost of the hardware level out with the API cost of proprietary alternatives. These benefits underscore the importance of conducting further research to enhance classification performance.

8. Future Work

In this work, we provide runtime and performance benchmarks of out-of-the-box (local) [SLMs](#) for phishing detection. While the results offer valuable insights, several opportunities for future research remain. First, customisation is a key advantage of open-source models. Future work could focus on fine-tuning these models or integrating RAG techniques to improve performance. Second, the dataset used in this work includes additional information, such as website screenshots and other website components, which could be utilised for (multimodal) phishing detection. Finally, integrating [LLMs](#) into a larger framework for phishing detection, combining multiple tools and analysis stages, could further improve the results. Studying efficient architectures and workflows is subject to future work.

Acknowledgments

This work was supported by the Austrian Research Promotion Agency (FFG) under the “Digitale Technologien 2023” programme (FFG Project No. FO999915293 – LLM4CTI). SBA Research (SBA-K1 NGC) is a COMET Center within the COMET – Competence Centers for Excellent Technologies Programme and funded by BMIMI, BMWET, and the federal state

of Vienna. The COMET Programme is managed by FFG. We also thank Johannes Kosche for working on the infrastructure and assisting with early experiments.

Appendix A.

Prompt Template 1 shows the complete, non-truncated prompt template used in this work, which includes the detailed list of the few-shot examples provided to the model.

Prompt Template 1

You are a cybersecurity expert analysing websites for phishing attempts. Your task is to examine the provided HTML code and the original website URL to determine if the website is likely a phishing site.

****Important:**** The HTML may be truncated to reduce costs, so CSS styles and JavaScript code may be missing. Focus on the HTML structure, text content, URLs, and the original URL.

Original URL: '{original_url}'

HTML: '{html_text}'

Original HTML character count: '{original_character_count}'

When assessing the likelihood of phishing, consider both your general knowledge of phishing techniques and the evidence you see in the data - focus on HTML structure, content, and domain/URL consistency.

" Below are common examples of phishing indicators you may find useful, but you are not limited to them:

1. ****Suspicious URLs/domains**** - Check href attributes, form actions, image sources for:

- Misspelled brand names, unusual domains, suspicious subdomains
- IP addresses instead of domains, excessive hyphens, unusual TLDs

2. ****Form analysis**** - Login/input forms with:

- Action URLs pointing to wrong domains
- Password/sensitive data collection for mismatched brands
- Excessive personal information requests (SSN, full address, etc.)

3. ****Content and language**** - Text containing:

- Urgent threats: 'Account suspended', 'Verify immediately', 'Limited time'
- Fear tactics: 'Security breach', 'Unauthorised access detected'
- Reward baits: 'You have won', 'Free gift', 'Exclusive offer'

4. ****HTML structure issues****:

- Spelling/grammar errors in text content
- Inconsistent or poor HTML structure
- Missing or suspicious meta tags (title, description)

5. ****Link analysis**** - Check all href attributes for:

- Links to different domains than expected
- Shortened URLs (bit.ly, tinyurl, etc.)
- Misleading anchor text vs actual URL

6. ****Brand impersonation**** - Look for:

- Company names in text that don't match domain
- References to legitimate services (PayPal, Amazon, banks, ...) on wrong domains
- Official-sounding but incorrect terminology

7. ****Missing legitimacy markers****:

- No contact information or privacy policy links
- Missing proper company details in footer
- No legitimate copyright notices

****Note:**** Since CSS/JS may be truncated, focus on HTML tags, text content, and URL analysis rather than visual styling or dynamic behaviour.

****Scoring guide:****

- 0-2: Very unlikely phishing (legitimate site)

- 3-4: Low risk (minor suspicious elements)
- 5-6: Medium risk (several concerning indicators)
- 7-8: High risk (multiple clear phishing signs)
- 9-10: Very high risk (obvious phishing attempt)

Required output format (JSON only):

```
{
phishing_score: int [0-10],
is_phishing: boolean [true/false],
reasoning: string [Brief explanation of your decision based on specific indicators found]
}
```

Output Constraints:

Do only output the JSON-formatted output and nothing else.

Table A.8 is referenced in Section 4.1.1, where only the fastest and slowest models are discussed.

Table A.8: Average analysis runtime (in seconds) per model and website run for the D5 and the D50 dataset. The table is sorted by dataset and in ascending order according to runtime.

Dataset	Model	Runtime
d5	llama3.2:1b	0.550
d5	dolphin3:8b	0.892
d5	mistral-nemo:12b	1.793
d5	gemma3:4b	1.966
d5	deepseek-r1:1.5b	2.916
d5	gemma3:12b	3.545
d5	phi3:14b	4.515
d5	qwen3:4b	5.725
d5	llama3.1:8b	5.729
d5	gpt-oss:20b	5.784
d5	deepseek-r1:14b	6.386
d5	gemma3:27b	7.077
d5	llama3.3:70b	7.731
d5	qwen3:30b	9.375
d5	deepseek-r1:70b	23.318
d50	llama3.2:1b	0.876
d50	dolphin3:8b	2.162
d50	mistral-nemo:12b	3.358
d50	deepseek-r1:1.5b	3.752
d50	gemma3:4b	3.840
d50	deepseek-r1:14b	4.618
d50	llama3.1:8b	5.353
d50	gemma3:12b	7.846
d50	gpt-oss:20b	7.946
d50	qwen3:4b	8.601
d50	phi3:14b	8.913
d50	qwen3:30b	12.709
d50	gemma3:27b	15.658
d50	llama3.3:70b	15.748
d50	deepseek-r1:70b	30.775

The qualitative reasoning output analysis in Table A.9 is part of Section 4.1.3 and contains the complete information of the phishing and benign wording tendencies per model and dataset.

Table A.9: Overview of the qualitative output analysis per model and dataset, summarised in the Phishing Score and Benign Score tendencies.

Model	Dataset	Phishing Score Tendency	Benign Score Tendency
deepseek-r1:1.5b	d5	336	86
deepseek-r1:14b	d5	322	109
deepseek-r1:70b	d5	333	127
dolphin3:8b	d5	257	125
gemma3:12b	d5	485	212
gemma3:27b	d5	483	186
gemma3:4b	d5	518	123
gpt-oss:20b	d5	194	95
llama3.1:8b	d5	116	19
llama3.2:1b	d5	10	0
llama3.3:70b	d5	306	151
mistral-nemo:latest	d5	219	166
phi3:medium	d5	304	117
qwen3:30b	d5	14	2
qwen3:4b	d5	101	106
deepseek-r1:1.5b	d50	207	31
deepseek-r1:14b	d50	301	166
deepseek-r1:70b	d50	318	176
dolphin3:8b	d50	197	83
gemma3:12b	d50	475	190
gemma3:27b	d50	425	195
gemma3:4b	d50	470	115
gpt-oss:20b	d50	107	63
llama3.1:8b	d50	170	19
llama3.2:1b	d50	18	0
llama3.3:70b	d50	246	113
mistral-nemo:latest	d50	156	112
phi3:medium	d50	235	57
qwen3:30b	d50	0	0
qwen3:4b	d50	164	239

Model Output 1 shows the typical result of the DeepSeek models, highlighting the combination of the thinking process and the required JSON output. Although other models used in this work are also reasoning models, only the DeepSeek models provide a thinking process in this way.

Model Output 1

Okay, so I need to analyze this website's HTML code for phishing attempts. The user provided an HTML snippet that seems very minimal—just a title link. But since CSS and JS might be missing, I have to focus on the structure and content.
First, let me look at the HTML structure. It starts with <html>, then <head> which includes meta tags, charset, and a title. The body is empty except for some empty <body></body></html>. So, no actual content in the page beyond the title link.
Now, looking at the phishing indicators:
1. **Suspicious URLs/Domain**: The HTML doesn't have any links or href attributes because it's just an empty body. But wait, maybe there are other elements? No, the structure is very minimal—probably no form fields or images either since they're not present.
2. **Form Analysis**: There are no login inputs or forms in this snippet. So, that's probably not a concern here.
3. **Content and Language**: The only text is the title: "Amazon". That doesn't seem like it contains any phishing indicators. It's just a link to an account sign-in page for Amazon. I don't see any urgency, threats, or reward baits here.
4. **HTML Structure Issues**: There are no errors in the HTML structure because there are no tags beyond what's already present. So, that doesn't add anything.
5. **Link Analysis**: The only link is to "Amazon", which is a legitimate account sign-in page for Amazon. It doesn't point to any other domain or have any suspicious links.
6. **Brand impersonation**: The text has "Amazon" in the title, but it's just a regular brand name. There are no references to other companies or services on this page. So, that's not a phishing indicator either.
7. **Missing Legitimacy Markers**: The HTML doesn't have any meta tags like title or description. It does have a link tag, which is fine because it's pointing to a legitimate domain. There are no proper company details in the footer or any privacy policy links. So, that's not

an issue here.
Putting this all together: The website has very minimal HTML structure and content. It doesn't contain any phishing indicators like suspicious URLs, form fields, incorrect domains, or references to other companies. The only link is to a legitimate account sign-in page for Amazon, which isn't phishing material.
Therefore, based on the analysis, there are no clear signs of phishing attempts here. So, the phishing score should be 0-2, indicating very unlikely.
`</think>`
`““json\n{\n “phishing_score”: 0,\n “is_phishing”: false,\n “reasoning”: “The website’s HTML code is minimal and lacks any links or content that would indicate phishing attempts. The only link points to a legitimate Amazon account sign-in page, which does not contain phishing indicators.”\n}\n“”`

Model Output 2 illustrates a common occurrence, where the model's first output differs from all subsequent runs, which are then identical.

Model Output 2

Run 1: `““json\n{\n “phishing_score”: 3,\n “is_phishing”: false,\n “reasoning”: “The HTML code is minimal and lacks significant content. It contains basic HTML tags for a webpage, but there are no suspicious URLs, forms, or text content that would immediately indicate a phishing attempt. The lack of content makes it difficult to assess the legitimacy of the site.”\n}\n“”`

Run 2: `““json\n{\n “phishing_score”: 3,\n “is_phishing”: false,\n “reasoning”: “The HTML code is minimal and lacks significant content. It contains basic HTML tags for a webpage, but no suspicious URLs, forms, or text content are present. It’s difficult to assess the legitimacy of the site without more information. The score reflects the lack of indicators, but it’s not a definitive determination of legitimacy.”\n}\n“”`

Run 3: `““json\n{\n “phishing_score”: 3,\n “is_phishing”: false,\n “reasoning”: “The HTML code is minimal and lacks significant content. It contains basic HTML tags for a webpage, but no suspicious URLs, forms, or text content are present. It’s difficult to assess the legitimacy of the site without more information. The score reflects the lack of indicators, but it’s not a definitive determination of legitimacy.”\n}\n“”`

Run 4: `““json\n{\n “phishing_score”: 3,\n “is_phishing”: false,\n “reasoning”: “The HTML code is minimal and lacks significant content. It contains basic HTML tags for a webpage, but no suspicious URLs, forms, or text content are present. It’s difficult to assess the legitimacy of the site without more information. The score reflects the lack of indicators, but it’s not a definitive determination of legitimacy.”\n}\n“”`

Run 5: `““json\n{\n “phishing_score”: 3,\n “is_phishing”: false,\n “reasoning”: “The HTML code is minimal and lacks significant content. It contains basic HTML tags for a webpage, but no suspicious URLs, forms, or text content are present. It’s difficult to assess the legitimacy of the site without more information. The score reflects the lack of indicators, but it’s not a definitive determination of legitimacy.”\n}\n“”`

References

- Abdelnabi, S., Krombholz, K., & Fritz, M. (2020). VisualPhishNet: Zero-Day Phishing Website Detection by Visual Similarity. *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS '20)* (pp. 1681–1698). Association for Computing Machinery. URL <https://doi.org/10.1145/3372297.3417233>; Accessed: 2025-09-25.
- Abuadbba, A., Wang, S., Almashor, M., Ahmed, M. E., Gaire, R., Camtepe, S., & Nepal, S. (2022). Towards Web Phishing Detection Limitations and Mitigation. *arXiv preprint arXiv:2204.00985*. URL <https://doi.org/10.48550/arXiv.2204.00985>; Accessed: 2025-09-25.
- Aljofey, A., Bello, S. A., Lu, J. & Xu, C. (2025). Comprehensive phishing detection: A multi-channel approach with variants TCN fusion leveraging URL and HTML features. *Journal of Network and Computer Applications*, **238**, 104170. URL <https://doi.org/10.1109/ACCESS.2025.3540075>; Accessed: 2025-11-03.
- Anti-Phishing Working Group (APWG). (2025). *Phishing Activity Trends Report: 2nd Quarter 2025*. APWG. URL https://docs.apwg.org/reports/apwg_trends_report_q2_2025.pdf; Accessed: 2025-09-25.
- Bahaghight, M., Ghasemi, M., & Ozen, F. (2023). A high-accuracy phishing website detection method based on machine learning. *Journal of Information Security and Applications*, **77**, 103553. URL <https://doi.org/10.1016/j.jisa.2023.103553>; Accessed: 2025-09-25.
- Barberá, I. (2025). *AI Privacy Risks & Mitigations – Large Language Models (LLMs)*. European Data Protection Board, Brussels, Belgium. URL <https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-langs.pdf>; Accessed: 2025-09-25.
- Bold, R., Al-Khateeb, H., & Ersotelos, N. (2022). Reducing False Negatives in Ransomware Detection: A Critical Evaluation of Machine Learning Algorithms. *Applied Sciences*, **12**(24), 12941. URL <https://doi.org/10.3390/app122412941>; Accessed: 2025-09-25.
- Bountaka, P., & Xenakis, C. (2023). HELPHED: Hybrid Ensemble Learning PHishing Email Detection. *Journal of Network and Computer Applications*, **210**, 103545. URL <https://doi.org/10.1109/ACCESS.2025.3540075>; Accessed: 2025-11-03.
- Chiew, K. L., Yong, K. S. C., & Tan, C. L. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Systems with Applications*, **106**, 1–20. URL <https://doi.org/10.1016/j.eswa.2018.03.050>; Accessed: 2025-09-25.
- Dalvi, S., Gressel, G., & Achuthan, K. (2019). Tuning the False Positive Rate / False Negative Rate with Phishing Detection Models. *International Journal of Engineering and Advanced Technology*, **9**(1S5), 7–13. URL <https://doi.org/10.35940/ijeat.A1002.1291S519>; Accessed: 2025-09-25.
- Duman Keles, F., Wijewardena, P. M., & Hegde, C. (2023). On The Computational Complexity of Self-Attention. *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, pp. 597–619. PMLR. URL <https://proceedings.mlr.press/v201/duman-keles23a.html>; Accessed: 2025-09-25.
- Europol. (2024). *Internet Organised Crime Threat Assessment (IOCTA) 2024*. European Union Agency for Law Enforcement Cooperation, Luxembourg. URL <https://doi.org/10.2813/442713>; Accessed: 2025-09-25.
- Federal Bureau of Investigation, Internet Crime Complaint Center (IC3). (2025). *Internet Crime Report 2024*. Federal Bureau of Investigation, Washington, D.C. URL https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf; Accessed: 2025-09-25.
- Fu, Y. (2024). Challenges in Deploying Long-Context Transformers: A Theoretical Peak Performance Analysis. *arXiv preprint arXiv:2405.08944*. URL <https://doi.org/10.48550/arXiv.2405.08944>; Accessed: 2025-09-25.
- Google Developers. (2025, August). Introduction to Large Language Models. URL <https://developers.google.com/machine-learning/resources/intro-llms>; Accessed: 2025-09-25.
- Hannousse, A., & Yahiouche, S. (2021). *Web page phishing detection* (Version v3). Mendeley Data. URL <https://doi.org/10.17632/c2gw7fy2j4.3>; Accessed: 2025-09-25.
- Irugalbandara, C., Mahendra, A., Daynauth, R., Arachchige, T. K., Dantanarayana, J. L., Flautner, K., Tang, L., Kang, Y., & Mars, J. (2024). Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's LLM with Open Source SLMs in Production. *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* (pp. 280–291). IEEE. URL <https://doi.org/10.1109/ISPASS61541.2024.00034>; Accessed: 2025-09-25.

- Kavya, S., & Sumathi, D. (2024). Staying ahead of phishers: a review of recent advances and emerging methodologies in phishing detection. *Artificial Intelligence Review*, **58**(2), 50. URL <https://doi.org/10.1007/s10462-024-11055-z>; Accessed: 2025-09-25.
- Koide, T., Fukushi, N., Nakano, H., & Chiba, D. (2025). ChatSpamDetector: Leveraging Large Language Models for Effective Phishing Email Detection. In *Security and Privacy in Communication Networks: 20th EAI International Conference, SecureComm 2024, Dubai, UAE, Proceedings, Part III* (pp. 297–319). Springer. URL https://doi.org/10.1007/978-3-031-94455-0_14; Accessed: 2025-09-25.
- Koide, T., Nakano, H., & Chiba, D. (2024, October). ChatPhishDetector: Detecting Phishing Sites Using Large Language Models. *IEEE Access*, **12**, 154381–154400. IEEE, Piscataway, NJ, USA. URL <https://doi.org/10.1109/ACCESS.2024.3483905>; Accessed: 2025-09-25.
- Kulkarni, A., Balachandran, V., Divakaran, D. M., & Das, T. (2025). From ML to LLM: Evaluating the Robustness of Phishing Web Page Detection Models against Adversarial Attacks. *Digital Threats: Research and Practice*, **6**(2), Article 10. URL <https://doi.org/10.1145/3737295>; Accessed: 2025-09-25.
- Lee, J., Lim, P., Hooi, B., & Divakaran, D. M. (2024, September). Multimodal Large Language Models for Phishing Webpage Detection and Identification. *2024 APWG Symposium on Electronic Crime Research (eCrime)* (pp. 1–13). IEEE, Boston, MA, USA. URL <https://ieeexplore.ieee.org/document/10896256>; Accessed: 2025-09-25.
- Li, W., Manickam, S., Chong, Y.-W., & Karuppayah, S. (2025). PhishDebate: An LLM-Based Multi-Agent Framework for Phishing Website Detection. *arXiv preprint arXiv:2506.15656*. URL <https://doi.org/10.48550/arXiv.2506.15656>; Accessed: 2025-09-25.
- Lin, Y., Liu, R., Divakaran, D. M., Ng, J. Y., Chan, Q. Z., Lu, Y., Si, Y., Zhang, F., & Dong, J. S. (2021). Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages. *30th USENIX Security Symposium (USENIX Security 21)* (pp. 3793–3810). USENIX Association. URL <https://www.usenix.org/conference/usenixsecurity21/presentation/lin>; Accessed: 2025-09-25.
- Liu, R., Lin, Y., Teoh, X., Liu, G., Huang, Z., & Dong, J. S. (2024). Less Defined Knowledge and More True Alarms: Reference-based Phishing Detection without a Pre-defined Reference List. *33rd USENIX Security Symposium (USENIX Security 24)* (pp. 523–540). USENIX Association. URL <https://www.usenix.org/conference/usenixsecurity24/presentation/liu-ruofan>; Accessed: 2025-09-25.
- Mahboubi, A., Luong, K., Abutorab, H., Bui, H. T., Jarrad, G., Bahutair, M., Camtepe, S., Pogrebna, G., Ahmed, E., Barry, B., & Gately, H. (2024). Evolving techniques in cyber threat hunting: A systematic review. *Journal of Network and Computer Applications*, **232**, 104004. URL <https://doi.org/10.1109/ACCESS.2025.3540075>; Accessed: 2025-11-03.
- Nahapetyan, A., Khare, K., Schwarz, K., Reaves, B., & Kapravelos, A. (2025). Characterizing Phishing Pages by JavaScript Capabilities. *arXiv preprint arXiv:2509.13186*. URL <https://doi.org/10.48550/arXiv.2509.13186>; Accessed: 2025-09-25.
- Nair, R., Abbasi, F., & Pervez, S. (2025). PhishEmailLLM: A Meta Model Approach to Detect Phishing Emails by Leveraging LLMs and Machine Learning Models. *Proceedings of the 2025 Australasian Computer Science Week (ACSW '25)* (pp. 19–29). Association for Computing Machinery. URL <https://doi.org/10.1145/3727166.3727169>; Accessed: 2025-09-25.
- OpenAI. (2025). OpenAI API Pricing. URL <https://platform.openai.com/docs/pricing>; Accessed: 2025-09-25.
- Panum, T. K., Hageman, K., Hansen, R. R., & Pedersen, J. M. (2020). Towards Adversarial Phishing Detection. *Proceedings of the 13th USENIX Workshop on Cyber Security Experimentation and Test (CSET '20)*. USENIX Association. URL <https://www.usenix.org/conference/cset20/presentation/panum>; Accessed: 2025-09-25.
- Putra, I. K. A. A. (2023). *Phishing Website Dataset* (Version v1). Zenodo. URL <https://doi.org/10.5281/zenodo.8041387>; Accessed: 2025-09-25.
- Ratul, I. J., Zhou, Y., & Yang, K. (2025). Accelerating Deep Learning Inference: A Comparative Analysis of Modern Acceleration Frameworks. *Electronics*, **14**(15), 2977. MDPI. URL <https://www.mdpi.com/2079-9292/14/15/2977>; Accessed: 2025-09-25.
- RunPod, Inc. (2025). RunPod: The cloud built for AI. URL <https://www.runpod.io/>; Accessed: 2025-09-25.

- Trad, F., & Chehab, A. (2024). Prompt Engineering or Fine-Tuning? A Case Study on Phishing Detection with Large Language Models. *Machine Learning and Knowledge Extraction*, **6**(1), 367–384. URL <https://doi.org/10.3390/make6010018>; Accessed: 2025-09-25.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017, December). Attention Is All You Need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 5998–6008). Curran Associates, Inc., Long Beach, CA, USA. URL <https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf>; Accessed: 2025-09-25.
- Wang, S., Li, B. Z., Khabsa, M., Fang, H., & Ma, H. (2020). Linformer: Self-Attention with Linear Complexity. *arXiv preprint arXiv:2006.04768*. URL <https://doi.org/10.48550/arXiv.2006.04768>; Accessed: 2025-09-25.
- Yang, R., Zheng, K., Wu, B., Wu, C., & Wang, X. (2021). Phishing Website Detection Based on Deep Convolutional Neural Network and Random Forest Ensemble Learning. *Sensors*, **21**(24), 8281. URL <https://doi.org/10.3390/s21248281>; Accessed: 2025-09-25.
- Yerima, S. Y., & Alzaylaee, M. K. (2020). High Accuracy Phishing Detection Based on Convolutional Neural Networks. In *Proceedings of the 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)* (pp. 1–6). IEEE. URL <https://doi.org/10.1109/ICCAIS48893.2020.9096869>; Accessed: 2025-09-25.
- Zhang, J., Wu, P., London, J., & Tenney, D. (2025). Benchmarking and Evaluating Large Language Models in Phishing Detection for Small and Midsize Enterprises: A Comprehensive Analysis. *IEEE Access*, **13**, 28335–28352. URL <https://doi.org/10.1109/ACCESS.2025.3540075>; Accessed: 2025-09-25.