Echoes of the Past: Detecting and Classifying Re-registered Domains in Enterprise DNS Traffic

Muhammad Muzammil[Γ]*, Zafir Ansari[γ]*, Nick Nikiforakis[Γ], and Darin Johnson[γ]
[Γ]Stony Brook University, New York, USA
[γ]Infoblox, Washington, USA
{mmuzammil, nick}@cs.stonybrook.edu, {zansari, djohnson}@infoblox.com

***Abstract*—The Domain Name System (DNS) is a critical component of the Internet, yet its foundational processes, such as domain registration and ownership changes, are generally opaque to end users. This lack of transparency enables adversaries to re-register expired domains and host malicious content that continues to receive traffic from users who trust and revisit the domain. In this paper, we introduce EchoLoc, a scalable system for detecting malicious re-registered domains across the entire TLD space that appear in live DNS resolution telemetry from Infoblox, a major DNS resolution and threat intelligence provider. We deploy EchoLoc for a one-month period, during which it analyzed 144.6M new domain registrations and identified 1.5M re-registrations, of which 66K were queried by customers. Using a machine learning-based website classification pipeline that combines structural features from web content with semantic signals derived from a large language model, we identify over 9K malicious re-registered domains. The classifier achieves 0.95 precision and recall for malicious domain detection, with an overall accuracy of 98.1%. Our analysis further shows that these domains exhibit user activity both prior to expiration and after re-registration.**

I. Introduction

DNS is often referred to as the "phonebook of the Internet," mapping human-readable domain names to resource records such as IP addresses. From banking and e-commerce to cloud platforms and email, virtually every digital service depends on DNS to remain discoverable and reachable. Yet, for the average user, the intricacies behind DNS remain invisible. Most users do not notice the domain registration process, such as when a domain was registered, when it expires, who owns the domain, or whether the domain is actually controlled by the entity they expect. This disconnect creates a significant blind spot, where users assume continuity and trust in the domains they visit, leaving them unaware that expired domains can be re-registered and exploited by malicious actors who have the ability to masquerade as the previous owner [1].

Lauringer *et al.* [2] demonstrated that deleted domain names are frequently re-registered almost immediately after becoming publicly available (a phenomenon known as "dropcatching"), sometimes with delays as short as under one second. This threat is further amplified by the absence of standards to alert users when a domain changes ownership. Consequently, if an adversary acquires an expired domain with residual trust, they can exploit the incoming traffic and leverage the repurposed domain for malicious purposes.

In the past, multiple research efforts have been made to understand the domain re-registration ecosystem and the security threats around it [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. Prior work on domain re-registrations has faced three main challenges, which we address in this paper. First, existing approaches have been constrained by limited data availability for capturing re-registrations across the full TLD space. Most prior studies rely on daily gTLD zone files, identifying re-registrations when a domain disappears from one zone file and reappears in a subsequent release. However, zone files are published for only a subset of TLDs, with most ccTLDs remaining opaque. Moreover, processing large TLD zone files at daily granularity is computationally expensive, which makes it a challenge to include a large amount of TLDs for analysis. As a result, prior work has typically restricted analysis to a small number of popular TLDs, often combined with rate-limited WHOIS queries needed to infer ownership changes.

Second, to assess the threats posed by re-registered websites, prior studies relied either on publicly available blocklists or on subsets of domains evaluated through services such as VirusTotal and Google Safe Browsing [1], [6]. However, these approaches lead to domain analysis constrained by the coverage of the chosen datasets or outsourced tools, making it difficult to rigorously evaluate the completeness of the presented results. Third, the security research community still lacks a clear understanding of how successful attackers are in attracting natural residual traffic to re-registered domains. In 2022, So *et al.* [4] re-registered 201 expired domains to measure the extent of residual traffic and assess attacker success. However, this leaves open the question of how often users interact with re-registered domains at a larger scale.

To bridge these gaps, we design EchoLoc, a system for automated, longitudinal detection of re-registered domains across all TLDs and their classification as benign or malicious. Rather than restricting analysis to a subset of TLDs, EchoLoc ingests a high-frequency WHOIS feed to identify re-registrations across the entire namespace and concentrates computational resources on domains that receive live user

*Equal Contribution. This work was done while Muhammad Muzammil was interning at Infoblox under supervision of full time researchers.

Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb) 2026
27 February 2026, San Diego, CA, USA
ISBN 978-1-970672-06-0
https://dx.doi.org/10.14722/madweb.2026.23002
www.ndss-symposium.org

traffic from customers of Infoblox, a major DNS resolution and threat intelligence provider.

For each such domain, we extract a variety of features from the website's DOM and a screenshot. Additionally, we generate semantic features using OpenAI's GPT-4o model to further understand the nature of web content hosted on the repurposed domain. These combined features are then used to train a random forest classifier that distinguishes between malicious and benign domains. Our trained model achieves an accuracy of 98.1%, with both precision and recall at 0.95 for malicious websites, and 0.99 for benign websites.

We deployed EchoLoc in the wild for 32 days, from July 11, 2025 to August 12, 2025, during which we observed 144.6M new DNS registrations of which 1.5M were re-registrations. Among these, 66K domains were queried by Infoblox customers after they were re-registered. Of this subset, we classify 9K domains as malicious, with gambling sites being the most prevalent threat category. We show that these domains attract significant user traffic both prior to expiration and after re-registration. Our contributions are summarized as follows:

- We present EchoLoc, a system for large-scale, longitudinal detection of domain re-registrations across the full TLD space.
- We develop a hybrid website classification approach that combines structural features from web content with semantic signals derived from large language models to identify malicious domains with high accuracy.
- We perform a large-scale analysis of DNS resolution telemetry for re-registered domains, showing that many are associated with substantial user traffic.

II. Background

DNS translates human-readable domain names (e.g., `example.com`) into machine-readable IP addresses. The last segment of a domain name is the top-level domain (TLD). TLDs fall into two major families: generic TLDs (gTLDs), such as `.com`, `.org`, and `.net`, which are overseen by ICANN; and country-code TLDs (ccTLDs), such as `.uk` or `.de`, which are administered by their respective national authorities. In addition, a smaller set of sponsored TLDs (sTLDs), such as `.edu`, operate under eligibility restrictions set by designated sponsoring organizations [11].

Domains are registered through accredited registrars, which act as intermediaries between registrants and the registries responsible for operating each TLD. Domain registrations must be periodically renewed to maintain ownership. Expiration and recovery policies differ significantly between ccTLDs and gTLDs as ccTLDs follow country-specific rules that vary widely in renewal windows, grace periods, and deletion procedures. In contrast, most gTLDs (and sTLDs) follow a broadly consistent lifecycle defined by ICANN policies and registry operational standards. Under these policies, gTLD registrars must provide expiration notifications and display an expiration notice page once a domain lapses, as required by the Expired Registration Recovery Policy [12]. Operationally, many gTLDs offer an auto-renew grace period that are commonly around 30–45 days, during which the registrant may renew the domain name without incurring a penalty. If the domain is not renewed, it typically enters a 30-day Redemption Grace Period and then a 5-day Pending Delete phase, after which it becomes available for public registration. While these intervals are widely adopted by gTLD registries, the exact durations remain registry-defined and may vary [13].

III. Data Sources

A. WHOIS

When a domain is registered, the registrar transmits the registration data to the registry, which then exposes this information through WHOIS records. These records typically include domain-related metadata such as creation and expiration dates. Historically, WHOIS also exposed registrant, administrative, and technical contact information (e.g., names, email addresses, and phone numbers), but many of these fields are now redacted or replaced with privacy-protected substitutes due to registry policies and privacy regulations [14].

B. Enterprise DNS Traffic

In this paper, we use enterprise DNS traffic logs from Infoblox, which consist of real-world DNS queries observed across their customer network deployments. Infoblox is a major DNS resolution provider that serves prominent enterprises and, by extension, millions of end users. These datasets include aggregated DNS resolution patterns as well as domains drawn from existing threat-intelligence feeds (i.e., domains the provider has already classified as malicious). Together, these datasets allow us to determine whether a re-registered domain is malicious and whether it is actively affecting real users in the wild as they continue to query them.

Importantly, a single *customer* refers to an enterprise rather than an individual user. A customer refers to an enterprise that delegates DNS resolution for its employees and devices to Infoblox. We note that all data used in this study is aggregated and anonymized to ensure that no personally identifiable information is accessed or retained.

IV. System Design

In this section, we describe the architecture of EchoLoc, which is illustrated in Figure 1.

A. Re-registration Detection

We require a comprehensive set of domains that were previously registered, expired, and subsequently re-registered by another entity.

To compile a daily list of dropped domains, past research work has relied on zone files for specific TLDs and on domain drop lists from dropcatch services [1], [6], [2]. A zone file contains every second-level domain in a given TLD that has at least one authoritative name server. When a domain enters

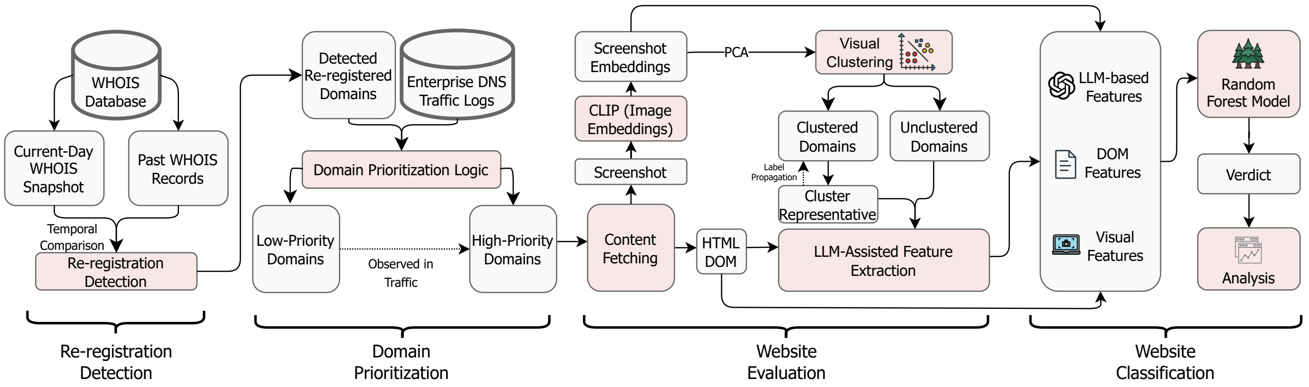


Fig. 1: *A schematic representation of EchoLoc's pipeline*

the pending-delete stage, it is removed from the zone file, and if it is re-registered, it appears again in subsequent zone files that are published on a daily-basis. A key limitation of this approach is that processing large zone files is computationally expensive, restricting prior studies to a small set of popular TLDs (e.g.,`.com`,`.net`,`.org`) [1], [6], [2]. Furthermore, many ccTLDs (such as `.uk`,`.fr`, and`.ca`) do not publish zone files, making them entirely invisible to such methods.

In contrast, we rely solely on WHOIS data for dropcatching detection. Our dataset is a high-frequency WHOIS feed with updates at regular intervals. Using this dataset, we retrieve registration records for millions or unique eTLD+1s each day. For each domain that appears as newly registered, we examine historical WHOIS entries to check whether it was previously registered and later allowed to expire. Concretely, if we find a prior record with a creation date predating the one in the current registration, we label the domain as "dropcaught" and include it in the next step of our analysis.

It is difficult to accurately verify whether the most recent registrant of an expired domain is the same entity as the prior registrant, since registrant identity fields in WHOIS records are typically redacted [14]. Nevertheless, treating a post-expiration registration as a new entity is a reasonable assumption in practice. Most gTLD registries offer generous renewal and redemption grace periods before a domain is permanently deleted, giving the original registrant multiple opportunities to renew the domain or recover it after expiration (as described in Section II) [12]. These safeguards make accidental loss of a domain unlikely, and thus re-registration by the same entity after deletion is uncommon. Even so, if benign re-registration by the original owner does occur, these cases will be identified as benign by downstream pipeline stages.

B. Domain Prioritization

Our prioritization step relies on active DNS telemetry and threat intelligence data from Infoblox. Using this dataset, we classify each re-registered domain as either low- or high-priority. A domain is considered high-priority if it is queried by one or more customers (i.e. enterprises) after re-registration and is not already listed as a malicious domain in Infoblox's existing threat intelligence data (Section III). Such domains are of particular interest because they are likely attracting residual traffic based on their prior reputation rather than their current content. Any low-priority domain becomes high-priority once it appears in customer traffic feeds. By focusing on these cases, we allocate our resources and cost budget toward domains that pose potential security risks and actively receive user traffic.

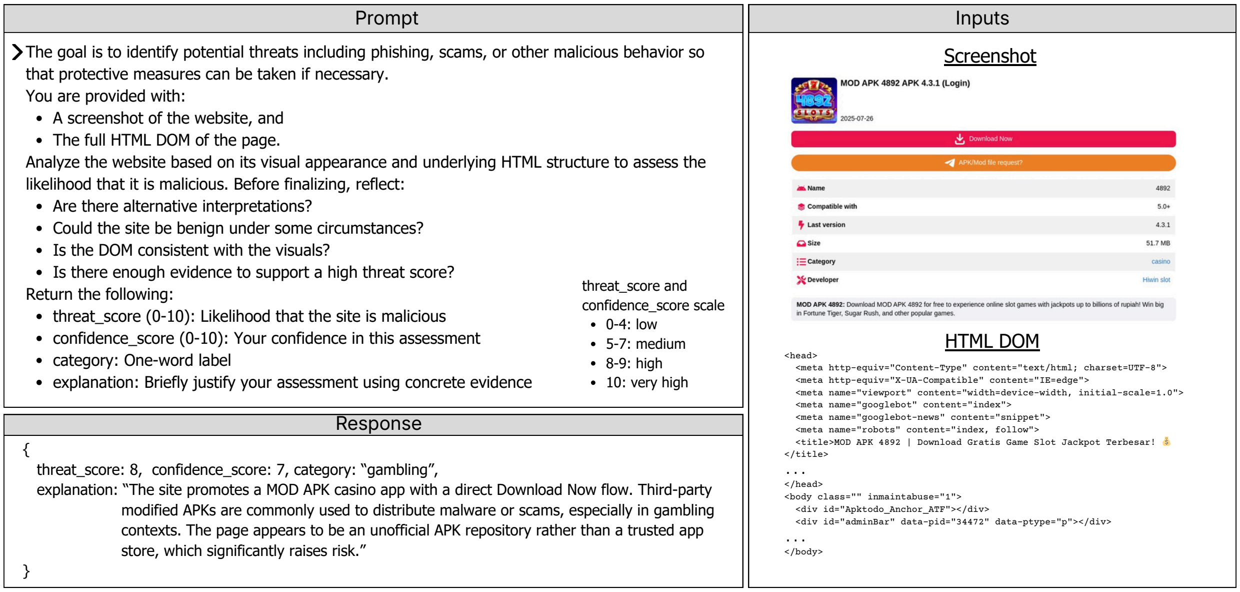
Over our 32-day data collection period, we observed traffic towards 66K re-registered domains. Out of these domains, 23K were not present in existing threat intelligence datasets, and are thus passed to the next stage of the pipeline.

C. Website Evaluation

Each high-priority domain undergoes a multi-stage evaluation to determine the nature of its hosted content.

1) Fetching: Using a webscraping tool, EchoLoc retrieves a screenshot of each website hosted on a re-registered domain, along with its complete HTML DOM. Domains that do not initially host active websites are reprocessed after one week. If no content is observed at that time, they are discarded from the pipeline. To mitigate the latency of webscraping requests, we employ a queue-based concurrent submission system that processes up to 75 domains in parallel. As domains complete processing, they exit the queue and are replaced by new submissions. Over the monitoring period, this approach enabled us to collect screenshots and DOM contents for 88% of all high-priority domains.

2) Clustering: Past work has shown that attackers frequently reuse common website templates to rapidly scale their operations, deploying nearly identical content across multiple domain names [15]. This strategy not only accelerates the setup but also increases resilience of the attack as if one domain is taken down, others remain active, ensuring continuity of their malicious campaign. In contrast, we observe that benign pages such as error screens or default WordPress templates are also common among newly re-registered domains. Processing these low-value pages increases resource consumption and time taken towards detecting malicious websites with little gain. To address this, we incorporate a three-stage clustering module that groups visually and contextually



Prompt

❯ The goal is to identify potential threats including phishing, scams, or other malicious behavior so that protective measures can be taken if necessary.

You are provided with:

- A screenshot of the website, and
- The full HTML DOM of the page.

Analyze the website based on its visual appearance and underlying HTML structure to assess the likelihood that it is malicious. Before finalizing, reflect:

- Are there alternative interpretations?
- Could the site be benign under some circumstances?
- Is the DOM consistent with the visuals?
- Is there enough evidence to support a high threat score?

Return the following:

- threat_score (0-10): Likelihood that the site is malicious
- confidence_score (0-10): Your confidence in this assessment
- category: One-word label
- explanation: Briefly justify your assessment using concrete evidence

threat_score and confidence_score scale

- 0-4: low
- 5-7: medium
- 8-9: high
- 10: very high

Response

```
{
  threat_score: 8,  confidence_score: 7, category: “gambling”,
  explanation: “The site promotes a MOD APK casino app with a direct Download Now flow. Third-party
                modified APKs are commonly used to distribute malware or scams, especially in gambling
                contexts. The page appears to be an unofficial APK repository rather than a trusted app
                store, which significantly raises risk.”
}
```

Inputs

Screenshot

MOD APK 4892 APK 4.3.1 (Login)

2025-07-26

Name	4892
Compatible with	5.0+
Last version	4.3.1
Size	51.7 MB
Category	casino
Developer	Hiwin slot

MOD APK 4892: Download MOD APK 4892 for free to experience online slot games with jackpots up to billions of rupiah! Win big in Fortune Tiger, Sugar Rush, and other popular games.

HTML DOM

```
<head>
  <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
  <meta http-equiv="X-UA-Compatible" content="IE=edge">
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
  <meta name="googlebot" content="index">
  <meta name="googlebot-news" content="snippet">
  <meta name="robots" content="index, follow">
  <title>MOD APK 4892 | Download Gratis Game Slot Jackpot Terbesar! 💰
</title>
...
</head>
<body class="" inmaintabuse="1">
  <div id="Apktodo_Anchor_ATF"></div>
  <div id="adminBar" data-pid="34472" data-ptype="p"></div>
...
</body>
```

Fig. 2: *Prompt and response used to assess website maliciousness from visual and DOM-based input. The model produces a structured risk assessment with threat and confidence scores, a category label, and a brief explanation of its reasoning.*

similar websites. By reducing redundant evaluation of near-duplicate content, this step optimizes the pipeline, lowering compute and API costs while enabling more efficient and scalable detection.

In the first stage, we extract OpenAI CLIP embeddings for each screenshot. CLIP is a vision-language model trained to align images and text in a shared high-dimensional space, producing semantic image representations that reflect both visual appearance and contextual cues [16]. We then apply HDBSCAN with a Euclidean distance metric to cluster these embeddings, grouping semantically similar images while marking outliers. In the second and third stage, we re-cluster the outliers identified in the first stage using the same CLIP embeddings but with more lenient HDBSCAN parameters. This step allows us to recover additional visually similar websites that were initially discarded due to parameter sensitivity, without compromising the semantic grouping enabled by CLIP embeddings. We choose to perform the clustering step three times based on manual analysis of cluster quality, prioritizing configurations that grouped visually similar websites together while minimizing the number of sites left in the outlier group that could have been clustered together.

Through this clustering approach, we effectively group parked domains, blank pages, and generic error pages (e.g., HTTP 404), ultimately reducing the downstream classification workload by 55%. The largest cluster consists of partially set-up websites, such as a largely white page with some text. This is followed by two clusters websites primarily comprising gambling websites and news sites. The remaining clusters include registrar parking pages, sites following similar templates and contexts (“Launching Soon” pages), and blogs.

All domains that are not assigned to any cluster are forwarded to the next stage for individual processing. For clustered domains, we randomly select a single representative domain from each cluster for analysis. The resulting classification of this representative is then propagated to all other domains within the same cluster. As new domains are added to the high-priority dataset, they are continuously assigned to existing or new clusters. If a domain is placed into a cluster that has already been evaluated, the previously determined classification is automatically propagated to the new domain to optimize cost and time.

3) LLM-assisted Evaluation: The next step is to evaluate the selected domains to determine whether or not they host malicious content. In the past, security researchers have relied on manual investigation to assign such verdicts [17], [18]. However, given the large number of domains processed daily by EchoLoc, manual categorization is infeasible. More recently, researchers have begun to use large-language models (LLMs) in various areas, such as to classify certain types malicious websites, such as cryptocurrency scams and phishing websites [19], [20], [15], [21]. Although the results are promising, relying solely on LLMs risks producing incorrect outputs due to hallucinations [22], [23]. This becomes even more pronounced when evaluating a diverse set of websites, as in our case where malicious domains vary widely in structure and behavior and cannot be captured by a single category. To address this, rather than soliciting a direct verdict from the LLM, we prompt the model to produce a calibrated assessment consisting of a `threat_score` and a `confidence_score` (each on a 0–10 scale), a one-word category label, and a brief evidence-based explanation. This structured output enables graded risk assessment while preserving interpretability across heterogeneous website types. The prompt used with OpenAI’s GPT-4o model is shown in

TABLE I: *Detection performance of three feature configurations: (1) DOM features with webpage screenshots, (2) LLM-derived features, and (3) a combined LLM + DOM + screenshot model. The combined feature set yields the highest overall accuracy.*

	DOM + Screenshot		**LLM**		**LLM + DOM + Screenshot**	
	Benign	**Malicious**	**Benign**	**Malicious**	**Benign**	**Malicious**
Precision	0.96	0.97	0.99	0.84	0.99	0.95
Recall	0.99	0.83	0.95	0.97	0.99	0.95
Accuracy	96.0%		95.5%		98.1%	

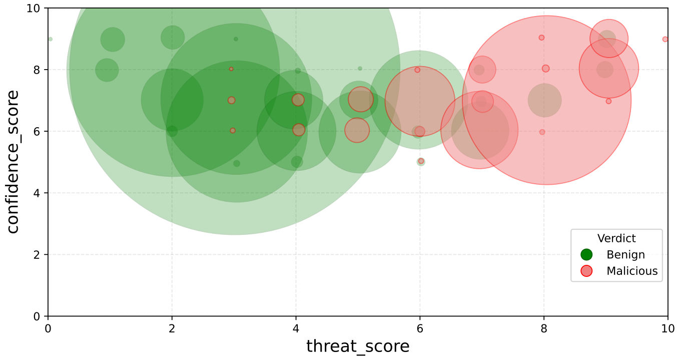


Fig. 3: *Threat and confidence scores assigned by the LLM during website evaluation for re-registered domains queried by users. Each bubble represents one or more domains sharing the same (threat_score, confidence_score) pair, with bubble size proportional to the number of domains assigned to that coordinate. Green bubbles indicate the final Random Forest classification as benign, while red bubbles indicate the final classification as malicious.*

Figure 2. These outputs are then incorporated as features into a random forest classifier, which produces the final verdict.

4) Random Forest Classifier: The final stage of EchoLoc's pipeline integrates outputs from the LLM with additional structural and content-based features to produce a reliable maliciousness verdict. We train a Random Forest classifier on a labeled dataset where a security professional manually classified 2,000 re-registered domains into benign or malicious categories, of which 26% of domains are labeled as malicious. To represent each website, we extract five groups of features:

- TF-IDF representations of the LLM's textual justifications, which often highlight behavioral cues not directly captured by DOM features.
- One-hot encodings of observed website categories, along with mapped risk levels. For instance, *phishing*, *gambling*, and *crypto* are treated as high-risk, while categories like *blogs*, *educational*, or *government* are considered low-risk.
- Dimensionality-reduced screenshot embeddings obtained via PCA over OpenAI CLIP vectors, which capture semantic similarity across visual templates.
- 15 indicators extracted from the HTML. These include the presence of excessive iframes or forms, obfuscated JavaScript, or suspicious keywords (e.g., *login*, *casino*, *bitcoin*). In total, we curated 50 suspicious keywords, split into categories such as *gambling*, *adult*, and *finance*, which frequently appear in malicious sites.
- `threat_score` and `confidence_score`

After feature extraction, all groups are combined, scaled, and used to train the Random Forest classifier. To determine the most effective feature combination, we conducted ablation experiments under three configurations:

- LLM-only features
- DOM and screenshot features
- A combined model with LLM, DOM, and screenshots

As shown in Table I, the combined configuration yielded the best overall performance on test data, with 98.1% accuracy and 0.95 precision and recall on malicious websites, demonstrating that the structural and visual features complement the LLM's reasoning and help reduce hallucination errors. Based on this result, we adopt the combined feature set as the final input to the Random Forest classifier. Any newly re-registered domain is automatically evaluated by EchoLoc, by applying the trained model to assign a final *benign* or *malicious* verdict. Figure 3 plots the threat scores against the confidence scores assigned by the LLM for websites hosted on re-registered domains, with bubble sizes proportional to the number of domains in the point on the center of the circle. Benign sites generally cluster towards the lower threat scores (0–4) while still showing high confidence, while malicious sites are concentrated at higher threat scores (6–10). The overlap between the two classes in the middle range (around threat scores of 4–6) highlights the challenge of relying solely on LLM-based ratings, since both benign and malicious domains may receive similar intermediate scores. In these ambiguous cases, the incorporation of DOM features provides discriminative power by capturing structural elements and content markers that are more consistently aligned with either benign or malicious websites, hence improving accuracy.

V. Analysis

We used EchoLoc to continuously collect and monitor re-registered domains as they appeared in the wild during the period from July 11, 2025 to August 12, 2025. In this section, we present a detailed analysis of our findings. Specifically, we provide an overview of the overall re-registration activity, examine the infrastructure and hosted content of malicious websites operating on re-registered domains, and analyze the customer traffic patterns that these domains attract.

The flow of domains through EchoLoc is shown in the Alluvial diagram in Figure 4. Across our measurement period, we observed 144.6M total WHOIS registrations, of which 1.5M

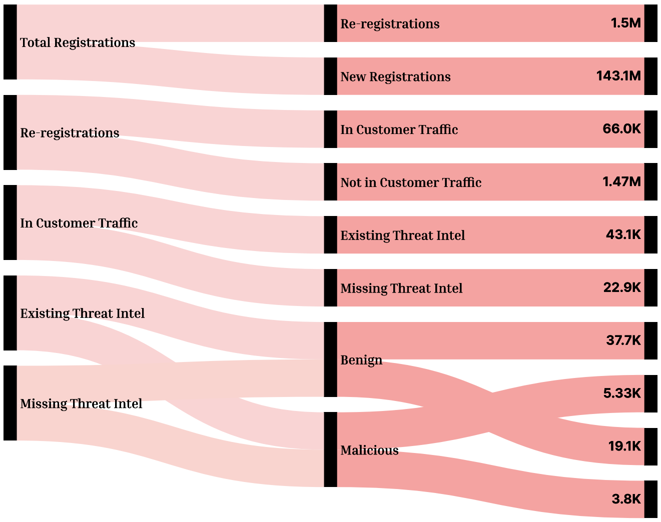


Fig. 4: Alluvial diagram illustrating the flow of domains through EchoLoc *from all WHOIS registrations to re-registrations, customer traffic, threat-intelligence coverage, and final labels.*

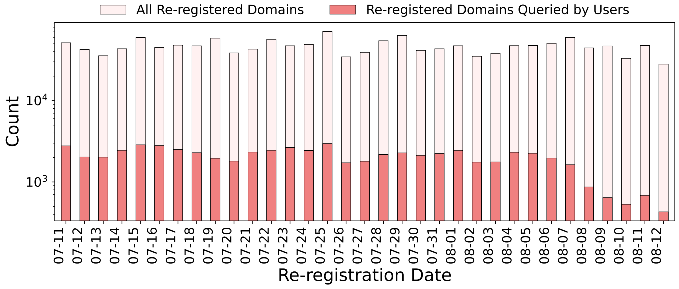


Fig. 5: Daily counts of all re-registered domains observed during our measurement period, compared with the subset of re-registered domains that were subsequently queried by Infoblox customers.

(1.04%) were re-registrations. Among these, 66K (4.28%) were queried by Infoblox customers; the remaining eTLD+1s exhibited no observed traffic. Of the queried set, 43K domains were already present in existing threat-intelligence feeds, while the remaining 23K had not been previously identified. Our classifier labeled 3,841 of these newly surfaced domains as malicious. Combining threat-intel coverage with our own detections, we find that 9,171 (13.9%) of all queried re-registered domains were malicious, while 56,795 (86.1%) were benign. The remainder of this section focuses exclusively on insights derived from the newly discovered malicious domains.

A. Re-registered Domain Characteristics

1) Volume: Figure 5 shows the daily re-registrations alongside the subset of domains queried by users. On average, about 46,600 domains were re-registered per day, with activity peaking at over 70,000 on July 25. Of these, an average of 2,000 domains were queried daily by users, highlighting that re-registered domains in fact do attract real user traffic, where each query carries potential exposure to malicious content driven by residual trust in the domain name.

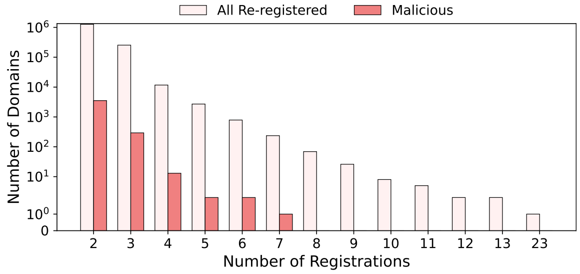


Fig. 6: Distribution of the total number of times domains detected during our measurement period have been re-registered over their lifetime.

TABLE II: Most frequent TLDs observed among re-registered domains and their corresponding counts of malicious re-registrations.

TLD	All Re-registered	Malicious
com	746,717	2,213
org	29,694	324
net	34,992	279
blog	4,246	179
ru	26,794	118
io	5,024	68
in	15,038	51
za	3,534	36
co	13,988	33
store	11,981	33

2) Number of Re-registrations: While the majority of domains underwent only a single re-registration cycle, more than 210,000 domains were re-registered at least twice, highlighting the sustained demand for certain expired domains. The maximum number of re-registrations observed for a single domain was 21. Figure 6 shows the distribution of the number of times domains in our dataset were re-registered. Notably, the most frequently re-registered domain that is currently classified as malicious is *buywithmurphy.com*, which has been registered a total of seven times since its creation. At the time of writing, the domain resolves to a gambling website, despite having hosted benign content in prior registration cycles.

3) TLDs: Table II presents the most common TLDs among the re-registered domains in our dataset. The *.com* TLD dominates with over 746,717 re-registrations. Several low-cost TLDs, such as *.xyz*, *.top*, *.online*, and *.site*, also appear prominently. Country-code TLDs such as *.cn*, *.cc*, *.pl*, and *.in* are also significant. Our results in Table II suggest that the demand for expired *.com* domains is significantly higher than other TLDs, although other TLDs with lower popularity also have significantly high amounts of re-registrations. Altogether, our results in Figure 6 and Table II indicate that expired domains remain in high demand for re-registration.

4) Domain Age: Figure 7 shows the CDF of the earliest registration dates for all domains that were re-registered within our data collection period along with the labeled malicious domains. The distribution reveals that more than half of re-registered domains were first created during or after 2023, suggesting that short-lived, disposable domains are more frequently re-registered soon after expiration. At the same time,

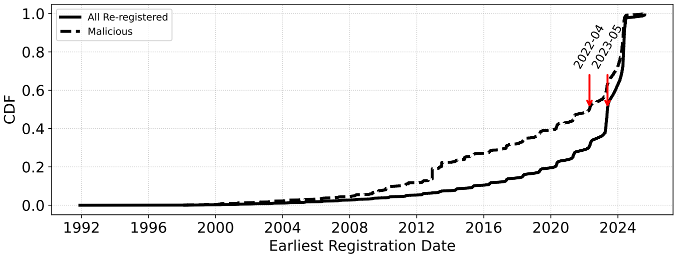


Fig. 7: *Cumulative distribution of original/first creation dates for all re-registered domains compared with the subset classified as malicious.*

the presence of older domains being re-registered indicates that abandoned names with historical usage or residual reputation also remain valuable targets.

For example, we find that *squirrel-run.com* was first registered in 1998. For nearly two decades, the domain legitimately hosted a website for a local golf club, with stable ownership and consistent benign content. However, after the domain expired and was re-registered, the website operates as a download portal for modified Android application packages. The landing page prominently advertises a slot-machine gambling app called *MOD APK 4892*, claiming unusually high payout rates and offering users the chance to win jackpots worth billions of Indonesian Rupiah. The screenshot in Figure 2 displays the current version of the site, which encourages visitors to download an unofficial APK and interact with a set of slot games. As modified applications bypass the vetting and security checks enforced by official marketplaces such as Google Play or the App Store, they carry significantly higher security risks. In many cases, the injected modifications introduce malicious code or harmful permission changes that can infect a user's device with malware [24], [25]. Similarly, *weavingroom.com*, which hosted a website about books on clothing since 1998, hosts a gambling portal that promotes "legal football betting sites in Vietnam" at the time of writing. The landing page lists multiple bookmakers with marketing claims such as high win rates, fast deposit and withdrawal, and large bonus payouts. The site is styled to mimic a review or ranking platform, but primarily functions as a directory redirecting users to betting services.

Gambling websites appear to be the most common repurposing of re-registered domains, and this trend is not limited to older names. Even more recently created websites are also victim to such exploitation. For instance, the official website of the IEEE ICDL conference, originally created in 2023, was re-registered and now similarly hosts gambling content.

5) Malicious Website Content: Using our random forest classifier, we were able to classify 3,841 unique malicious websites hosted on re-registered domains hosted on 2,290 unique IP addresses. Among the top categories assigned by the LLM in the evaluation phase of EchoLoc, the top three categories were *gambling* (3,039), *adult* (312), and *scam* (308).

Figure 8 presents a wordcloud generated from the page titles of websites classified as malicious. Prominent terms in the visualization reveal common themes across re-registered domains. Words such as *Casino*, *Slot*, *Togel*, and *Jackpot* indicate that online gambling and betting platforms dominate the malicious landscape. Beyond English, the wordcloud also contains terms from Vietnamese, Russian, Indonesian, and other languages. For example, a Vietnamese phrase (*Nha Cai*) meaning "bookmaker" frequently appears in repurposed gambling webpages, while *Uy Tin* translates to "prestige" or "trustworthy", often used to create a false sense of trust.

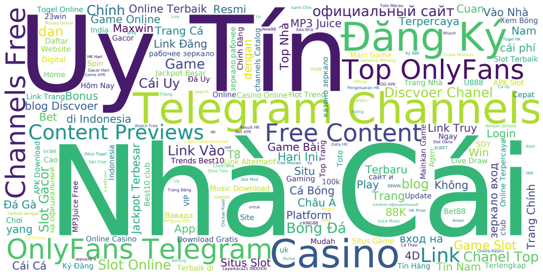


Fig. 8: *Wordcloud of page titles from malicious re-registered domains, showing gambling, adult content, and fraudulent campaigns across multiple languages.*

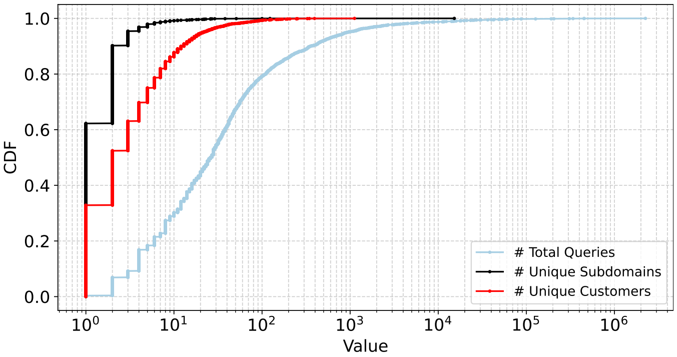


Fig. 9: *CDF of overall user traffic toward malicious re-registered domains, measured by total DNS queries received, the number of unique subdomains observed per domain, and the number of unique customers issuing those queries. Metrics include traffic to all associated subdomains.*

B. DNS Traffic

Beyond the classification and analysis of malicious re-registered domains, we further examine traffic patterns to assess their practical significance. Although each malicious domain identified in our study is queried at least once, traffic volume and persistence vary widely. Figure 9 presents CDFs of customer activity towards malicious re-registered domains including both before and after re-registration, capturing three dimensions of customer engagement. The red curve shows the distribution of unique customers per domain. Approximately 70% of malicious domains attracted traffic from more than 1 distinct user, highlighting that a majority of re-registered domains affect multiple customers. The black curve reflects the number of unique subdomains (QNames) observed per malicious re-registered domain, aggregated across all associated subdomains. Nearly half of the domains are associated with

TABLE III: *Top 10 malicious re-registered domains ranked by (a) total query count, (b) total unique QNames, and (c) total customers across all subdomains observed both before and after re-registration.*

#	eTLD+1	# Queries
1	netdna-ssl.com	2,251,878
2	l-tec.shop	454,092
3	score808.cfd	451,522
4	acs-education.com	338,999
5	nhansamkiv.com	308,455
6	bocongan113.com	206,087
7	hotelmuhabura.com	191,618
8	chillremote.tv	127,309
9	indymacbank.com	87,427
10	xunmzone.com	71,343

#	eTLD+1	# QNames
1	netdna-ssl.com	15,225
2	hpdex.org	123
3	sokol-shumen.com	100
4	centex-edu.net	51
5	onsen-trip.club	38
6	bwmar.com	29
7	bazzda.com	27
8	chinadanceusa.com	26
9	northug.net	26
10	jisinyosoku.com	26

#	eTLD+1	# Customers
1	netdna-ssl.com	1,121
2	loadinggif.com	394
3	faunatopsites.com	345
4	bwmar.com	327
5	zeda.blog	253
6	acs-education.com	246
7	dropload.tv	246
8	codinggear.blog	195
9	myimg.bid	194
10	antiquehomestyle.com	187

more than one distinct subdomain. Finally, the blue curve represents total DNS query volume per domain. Although around 80% of domains receive at most approximately one hundred queries, a small fraction accumulates tens of thousands to millions of queries, demonstrating that a limited number of malicious re-registrations account for a disproportionate share of overall traffic.

Table III presents the top 10 domains for each of the three categories. Although many of the highly ranked domains are associated with gambling content and were originally benign prior to re-registration, the case of *netdna-ssl.com* is particularly notable. Before its expiration, this domain functioned as a legitimate content delivery network (CDN) endpoint, widely embedded across websites to serve assets such as stylesheets and JavaScript. However, following its expiration and subsequent re-registration, it now redirects to *sstik.ai*, a site that facilitates unauthorized downloads from a social media platform. After expiration on 24th July 2025, thousands of unique subdomains of *netdna-ssl.com* have attracted customer traffic, which indicates that this domain is likely integrated into web applications that are fetching resources through the various subdomains (such as *1...1-wpengine.netdna-ssl.com* and *1...a-wpengine.netdna-ssl.com*). Moreover, *codinggear.blog* previously hosted a website focused on programming tutorials. Following its expiration and re-registration, it now routes users through intermediary tracking endpoints to a rotating set of adult-oriented domains. Repeated requests to the domain consistently return an HTTP 301 response with a dynamically changing `Location` header. Each response redirects through a Google ad click-tracking endpoint (*adservice.google.com/ddm/clk/*), where a different domain is embedded as the final destination parameter. This embedded domain rotates across successive requests.

VI. Related Work

In this section, we describe prior work that is most closely related to ours and highlight the novelty of our approach. Our study intersects two main areas of research in the security community: (i) investigating malicious activity on newly registered domains, and (ii) classification of malicious web pages using generative AI.

A. Domain Re-registrations

Several works in the security community have examined the malicious intent behind domain registrations, particularly for newly created domains using the DNS infrastructure and DNS lookup patterns from networks that perform initial lookups to the domain. Hao *et al.* [26] analyzed the early DNS activity of newly registered domains, showing how the observable behavior of malicious domains differs from that of legitimate ones. Halvorson *et al.* [27], [28], [29] later studied the registration patterns of more than 500 new TLDs introduced by ICANN, investigating the intent behind domains registered within each TLD. In contrast, our work focuses on domains that are not only newly registered but have also been used in the past. By studying re-registered domains, we capture the risks associated with residual trust and the substantial user traffic these domains continue to attract even after ownership changes.

In the past, the risks of re-registered domains and residual trust have attracted substantial research attention [9], [30], [7], [31], [3], [1], [32], [33], [34], [35], [36]. Nikiforakis *et al.* [9] demonstrated that expired domains linked through remote JavaScript inclusions enable adversaries to launch code-injection attacks upon re-registration. Moore and Clayton [10] examined domains used by U.S. banks following shutdowns or mergers, finding that 33% of expired domains were re-registered by malicious actors. Lever *et al.* [1] later formalized the concept of residual trust and showed that attackers increasingly exploit this property, identifying 27K re-registered domains present on public blocklists over six years. Miramirkhani *et al.* [6] conducted a longitudinal study of dropcatching in the *.com*, *.net*, and *.org* TLDs, tracking re-registrations over eight months. Lauringer *et al.* [2], [3] investigated domain expirations and re-registrations in terms of the time to re-registration and the competitive dynamics among registrars vying to capture high-value expired domains. More recently, So *et al.* [4] measured traffic to 201 expired domains and found that re-registered domains attract substantial residual traffic; in follow-up work [5], they showed that expired domains can persist in Android application logic, creating opportunities for abuse until developers update their apps. Finally, Muzammil *et al.* [7] explored the harms of expired domains in the Ethereum Name Service (ENS) [37], where expiration can lead to irreversible loss of cryptocurrency by end users.

In contrast to prior studies that relied on WHOIS records to confirm re-registrations [38], [6], [3], [39], we detect re-registrations at scale and use them as our primary vantage point, enabling coverage across the complete TLD space. By incorporating real-world DNS telemetry, we further prioritize computational resources toward re-registered domains that continue to receive traffic from active users. This design allows us to efficiently identify malicious domains under resource constraints and ensures that our analysis concentrates on domains that pose tangible risk in practice.

B. LLM-assisted Website Classification

Beyond re-registrations, a broad body of research has focused on the detection and classification of phishing, scam, and malware-serving web pages [40], [41], [42], [43], [44], [18], [45], [21], [6], [15], [19], [17], [46]. Early efforts in domain classification mainly relied on manual analysis. For instance, Li *et al.* [18] manually investigated newly created cryptocurrency scam websites on a daily basis, while Agten *et al.* [17] conducted manual inspections of websites hosted on typosquatting domains targeting popular sites. Although this approach results in accurate classifications, such approaches cannot scale to the volume and diversity of domains observed in the wild. To address scalability, subsequent work explored traditional machine learning models for phishing and scam detection [47], [48], [44]. However, as reported by Ji *et al.* [49], given the heterogeneity of the web content, training machine learning models that generalize across varied contexts remains challenging, and detections can be evaded by adversaries.

More recently, LLMs have been adopted for malicious website classification because of their ability to reason over contextual cues. Muzammil *et al.* [15] used LLMs to detect cryptocurrency investment scams, identifying over 43K websites with 88% accuracy against a ground truth dataset. Bitaab *et al.* applied fine-tuned *LLaMA-8B* to detect fraudulent shopping websites, while Nakano *et al.* [19] designed ScamFerret, which employs GPT-4 as an autonomous agent to crawl and classify scam sites.

In contrast, our approach avoids relying solely on the LLM's judgment, as LLMs remain susceptible to hallucination and misclassification [50], [51], [23], [22], [52]. To mitigate these issues while still leveraging the LLM's semantic reasoning capabilities, we incorporate safeguards through additional structural and visual features that counteract the effects of hallucinations. Accordingly, we adopt a hybrid methodology that combines structural features extracted from webpage screenshots and DOM content with the LLM's semantic assessments, and train a Random Forest classifier on top of these features. As demonstrated by our experiments, this design substantially reduces errors introduced by LLM hallucinations and achieves high precision and recall.

VII. Discussion

A. Limitations

While EchoLoc provides large-scale visibility into malicious re-registrations, some limitations remain. First, our identification of re-registrations assumes that a post-expiration registration corresponds to a change in ownership. Due to widespread redaction of registrant information in WHOIS records, we cannot definitively verify ownership transitions, and in rare cases, domains may be re-registered by the original owner. However, such cases are eventually evaluated in the later stages of our pipeline through content-based analysis and classification, which helps distinguish benign and malicious re-registrations, albeit at the cost of additional computational overhead. Second, our detections are limited to DNS traffic observed through Infoblox enterprise resolvers. Domains that primarily target users outside this ecosystem, including residential users or regions underrepresented in Infoblox's customer base, may not be captured. As a result, our findings reflect exposure within this specific measurement vantage point and may not fully represent the global prevalence or geographic distribution of re-registration abuse.

B. Ethics

All DNS traffic used in this study was aggregated and anonymized, ensuring that no personally identifiable user information was accessed or analyzed. Data was accesed only on Infoblox systems and only by Infoblox employees. Language models were used only to analyse publically accessible websites, which came from WHOIS data. No customer data went through a language model.

VIII. Conclusion

In this paper, we presented EchoLoc, a scalable system for the longitudinal detection of domain re-registrations across all TLDs. We further classify repurposed domains using a hybrid approach that combines LLM-derived semantic signals with structural website analysis. By leveraging real-time DNS traffic logs from Infoblox, we demonstrate that malicious re-registered domains not only persist in the ecosystem but also attract substantial user traffic. Our results highlight that domain re-registration remains an active attack surface, and highlight the need for stronger safeguards and greater transparency around domain expiration and ownership transitions to reduce such large-scale exploitation.

Acknowledgments

We thank the reviewers for their helpful comments. This work was completed by Muhammad Muzammil while interning at Infoblox under the support and supervision of Infoblox at their facilities. Muhammad Muzammil and Nick Nikiforakis are also supported by the National Science Foundation (NSF) under grants CNS-1941617, CNS-2126654, and CNS-2211575.

References

[1] C. Lever, R. Walls, Y. Nadji, D. Dagon, P. McDaniel, and M. Antonakakis, “Domain-z: 28 registrations later measuring the exploitation of residual trust in domains,” in *IEEE symposium on security and privacy*, 2016.

[2] T. Lauinger, K. Onarlioglu, A. Chaabane, W. Robertson, and E. Kirda, “Whois lost in translation: (mis) understanding domain name expiration and re-registration,” in *Proceedings of the Internet Measurement Conference*, 2016.

[3] T. Lauinger, A. Chaabane, A. S. Buyukkayhan, K. Onarlioglu, and W. Robertson, “Game of registrars: An empirical analysis of {Post-Expiration} domain name takeovers,” in *USENIX Security Symposium*, 2017.

[4] J. So, N. Miramirkhani, M. Ferdman, and N. Nikiforakis, “Domains do change their spots: Quantifying potential abuse of residual trust,” in *IEEE Symposium on Security and Privacy*, 2022.

[5] J. So, I. Sanchez-Rola, and N. Nikiforakis, “Lost in the mists of time: Expirations in dns footprints of mobile apps,” 2025.

[6] N. Miramirkhani, T. Barron, M. Ferdman, and N. Nikiforakis, “Panning for gold.com: Understanding the dynamics of domain dropcatching,” in *Proceedings of the Web Conference (WWW)*, 2018.

[7] M. Muzammil, Z. Wu, A. Balasubramanian, and N. Nikiforakis, “Panning for gold. eth: Understanding and analyzing ens domain dropcatching,” in *Proceedings of the ACM on Internet Measurement Conference*, 2024.

[8] A. Pitumpe, M. Muzammil, X. Li, N. Nikiforakis, and A. Rahmati, “Scams at scale: Unmasking crypto fraud and social engineering on the modern web,” *XRDS: Crossroads, The ACM Magazine for Students*, 2025.

[9] N. Nikiforakis, L. Invernizzi, A. Kapravelos, S. Van Acker, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, “You are what you include: large-scale evaluation of remote javascript inclusions,” in *Proceedings of the ACM conference on Computer and communications security*, 2012.

[10] T. Moore and R. Clayton, “The ghosts of banking past: Empirical analysis of closed bank websites,” in *International Conference on Financial Cryptography and Data Security*. Springer, 2014.

[11] Cloudflare, “What is a top-level domain (tld)?” Aug. 2025, https://www.cloudflare.com/learning/dns/top-level-domain/.

[12] ICANN, “Expired registration recovery policy,” Aug. 2025, https://www.icann.org/en/contracted-parties/consensus-policies/expired-registration-recovery-policy/expired-registration-recovery-policy-28-02-2013-en.

[13] N. W. Group, “Rfc 3915,” September 2004, https://datatracker.ietf.org/doc/html/rfc3915.

[14] Kevin Rollinson, “Gdpr and whois: Here’s what you need to know,” Aug. 2025, https://umbrella.cisco.com/blog/gdpr-and-whois.

[15] M. Muzammil, A. Pitumpe, X. Li, A. Rahmati, and N. Nikiforakis, “The poorest man in babylon: A longitudinal study of cryptocurrency investment scams,” in *Proceedings of the ACM on Web Conference*, 2025.

[16] OpenAI, “Clip: Connecting text and images,” August 2025, https://openai.com/index/clip/.

[17] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis, “Seven months’ worth of mistakes: A longitudinal study of typosquatting abuse,” in *Proceedings of the Network and Distributed System Security Symposium*, 2015.

[18] X. Li, A. Yepuri, and N. Nikiforakis, “Double and nothing: Understanding and detecting cryptocurrency giveaway scams,” in *Proceedings of the Network and Distributed System Security Symposium*, 2023.

[19] H. Nakano, T. Koide, and D. Chiba, “Scamferret: Detecting scam websites autonomously with large language models,” in *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2025.

[20] R. Haroon, K. Wigdor, K. Yang, N. Toumanios, E. T. Crehan, and F. Dogar, “Neurobridge: Using generative ai to bridge cross-neurotype communication differences through neurotypical perspective-taking,” in *ACM SIGACCESS Conference on Computers and Accessibility*, 2025.

[21] J. Lee, P. Lim, B. Hooi, and D. M. Divakaran, “Multimodal large language models for phishing webpage detection and identification,” in *APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, 2024.

[22] A. Ravichander, S. Ghela, D. Wadden, and Y. Choi, “Halogen: Fantastic llm hallucinations and where to find them,” *arXiv preprint arXiv:2501.08292*, 2025.

[23] G. Sriramanan, S. Bharti, V. S. Sadasivan, S. Saha, P. Kattakinda, and S. Feizi, “Llm-check: Investigating detection of hallucinations in large language models,” *Advances in Neural Information Processing Systems*, 2024.

[24] L. A. Saavedra, H. S. Dutta, A. R. Beresford, and A. Hutchings, “Modzoo: A large-scale study of modded android apps and their markets,” in *APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, 2024.

[25] aiplexantipiracy, “Is mod apk safe or a threat to your device; expert’s advice,” August 2025, https://www.aiplexantipiracy.com/blog/is-mod-apk-safe-or-a-threat-to-your-device-experts-advice.

[26] S. Hao, N. Feamster, and R. Pandrangi, “Monitoring the initial dns behavior of malicious domains,” in *Proceedings of the ACM SIGCOMM conference on Internet measurement conference*, 2011.

[27] T. Halvorson, J. Szurdi, G. Maier, M. Felegyhazi, C. Kreibich, N. Weaver, K. Levchenko, and V. Paxson, “The biz top-level domain: ten years later,” in *International Conference on Passive and Active Network Measurement*. Springer, 2012.

[28] T. Halvorson, K. Levchenko, S. Savage, and G. M. Voelker, “Xxxtortion? inferring registration intent in the. xxx tld,” in *Proceedings of the international conference on World wide web*, 2014.

[29] T. Halvorson, M. F. Der, I. Foster, S. Savage, L. K. Saul, and G. M. Voelker, “From. academy to. zone: An analysis of the new tld land rush,” in *Proceedings of the Internet Measurement Conference*, 2015.

[30] C. Tsoukaladelis, B. Kondracki, N. Balasubramanian, and N. Nikiforakis, “The times they are a-changin’: Characterizing post-publication changes to online news,” in *IEEE Symposium on Security and Privacy*. IEEE.

[31] R. Kirchner, C. Tsoukaladelis, M. Johns, and N. Nikiforakis, “The Power to Never Be Wrong: Evasions and Anachronistic Attacks Against Web Archives,” in *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2025.

[32] C. Tsoukaladelis, R. Perdisci, and N. Nikiforakis, “Uncontained Danger: Quantifying Remote Dependencies in Containerized Applications,” in *Proceedings of the International Symposium on Research in Attacks, Intrusions and Defenses*, 2025.

[33] D. Liu, S. Hao, and H. Wang, “All your dns records point to us: Understanding the security threats of dangling dns records,” in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[34] J. So, M. Ferdman, and N. Nikiforakis, “What gets measured gets managed: Mitigating supply chain attacks with a link integrity management system,” in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2025.

[35] C. Tsoukaladelis and N. Nikiforakis, “Manufactured Narratives: On the Potential of Manipulating Social Media to Politicize World Events,” in *Proceedings of the SecWeb Workshop (SecWeb)*, 2024.

[36] O. Starov, J. Dahse, S. S. Ahmad, T. Holz, and N. Nikiforakis, “No Honor Among Thieves: A Large-Scale Analysis of Malicious Web Shells,” in *Proceedings of the International World Wide Web Conference (WWW)*, 2016.

[37] ENS, “Ens,” Aug. 2025, https://ens.domains/.

[38] Z. Ma, A. Faulkenberry, T. Papastergiou, Z. Durumeric, M. D. Bailey, A. D. Keromytis, F. Monrose, and M. Antonakakis, “Stale tls certificates: Investigating precarious third-party access to valid tls keys,” in *Proceedings of the ACM on Internet Measurement Conference*, 2023.

[39] M. Korczynski, M. Wullink, S. Tajalizadehkhoob, G. C. Moura, A. Noroozian, D. Bagley, and C. Hesselman, “Cybercrime after the sunrise: A statistical analysis of dns abuse in new gtlds,” in *Asia Conference on Computer and Communications Security*, 2018.

[40] D. Canali, M. Cova, G. Vigna, and C. Kruegel, “Prophiler: a fast filter for the large-scale detection of malicious web pages,” in *Proceedings of the international conference on World wide web*, 2011.

[41] M. Felegyhazi, C. Kreibich, and V. Paxson, “On the potential of proactive domain blacklisting.” *LEET*, 2010.

[42] S. Hao, A. Kantchelian, B. Miller, V. Paxson, and N. Feamster, “Predator: proactive recognition and elimination of domain abuse at time-of-registration,” in *Proceedings of the ACM SIGSAC conference on computer and communications security*, 2016.

[43] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond blacklists: learning to detect malicious web sites from suspicious urls,” in *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009.

[44] K. Subramani, W. Melicher, O. Starov, P. Vadrevu, and R. Perdisci, "Phishinpatterns: measuring elicited user interactions at scale on phishing websites," in *Proceedings of the ACM Internet Measurement Conference*, 2022.

[45] P. Zhang, A. Oest, H. Cho, Z. Sun, R. Johnson, B. Wardman, S. Sarker, A. Kapravelos, T. Bao, R. Wang *et al.*, "Crawlphish: Large-scale analysis of client-side cloaking techniques in phishing," in *IEEE Symposium on Security and Privacy*, 2021.

[46] F. Trad and A. Chehab, "Large multimodal agents for accurate phishing detection with enhanced token optimization and cost reduction," in *International Conference on Foundation and Large Language Models*. IEEE, 2024.

[47] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "Cantina+ a feature-rich machine learning framework for detecting phishing web sites," *ACM Transactions on Information and System Security (TISSEC)*, 2011.

[48] Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, and J. S. Dong, "Phishpedia: A hybrid deep learning based approach to visually identify phishing webpages," in *USENIX Security Symposium*, 2021.

[49] F. Ji, K. Lee, H. Koo, W. You, E. Choo, H. Kim, and D. Kim, "Evaluating the effectiveness and robustness of visual similarity-based phishing detection models," in *USENIX Security Symposium*, 2025.

[50] N. Martin, A. B. Faisal, H. Eltigani, R. Haroon, S. Lamelas, and F. Dogar, "Llmproxy: Reducing cost to access large language models," *arXiv preprint arXiv:2410.11857*, 2024.

[51] H. Eltigani, R. Haroon, A. Kocak, A. B. Faisal, N. Martin, and F. Dogar, "Wallm–insights from an llm-powered chatbot deployment via whatsapp," *arXiv preprint arXiv:2505.08894*, 2025.

[52] J. Spracklen, R. Wijewickrama, A. N. Sakib, A. Maiti, and B. Viswanath, "We have a package for you! a comprehensive analysis of package hallucinations by code generating {LLMs}," in *USENIX Security Symposium*, 2025.