

LLM-Assisted Web Measurements

Simone Bozzolan
Università Ca' Foscari Venezia
simone.bozzolan@unive.it

Stefano Calzavara
Università Ca' Foscari Venezia
stefano.calzavara@unive.it

Lorenzo Cazzaro
Università Ca' Foscari Venezia
lorenzo.cazzaro@unive.it

Abstract

Web measurements are a well-established methodology for assessing the security and privacy landscape of the Internet. However, existing top lists of popular websites commonly used as measurement targets are unlabeled and lack semantic information about the nature of the sites they include. This limitation makes *targeted* measurements challenging, as researchers often need to rely on ad-hoc techniques to bias their datasets toward specific categories of interest. In this paper, we investigate the use of Large Language Models (LLMs) as a means to enable targeted web measurement studies through their semantic understanding capabilities. Building on prior literature, we identify key website classification tasks relevant to web measurements and construct datasets to systematically evaluate the performance of different LLMs on these tasks. Our results demonstrate that LLMs may achieve strong performance across multiple classification scenarios. We then conduct LLM-assisted web measurement studies inspired by prior work and rigorously assess the validity of the resulting research inferences. Our results demonstrate that LLMs can serve as a practical tool for analyzing security and privacy trends on the Web.

CCS Concepts

• Information systems → Web mining; • Security and privacy → Web application security.

Keywords

web measurements, large language models

ACM Reference Format:

Simone Bozzolan, Stefano Calzavara, and Lorenzo Cazzaro. 2025. LLM-Assisted Web Measurements. In . ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Web measurements are a popular tool to establish the current state of security and privacy on the Internet. Starting from a dataset of websites to analyze, web measurements leverage web crawling and automated analysis techniques to determine whether existing websites comply with security best practices [11, 12], suffer from known vulnerabilities [30, 47], or are aligned with current privacy regulations [23, 37]. Naturally, the representativeness of web measurements and the quality of the conclusions they draw are only as

good as the quality of the datasets they rely on. Legacy work on web measurements largely relied on public lists of popular websites (*top lists*) created by private companies, e.g., the now discontinued Alexa ranking [48]. Unfortunately, these lists turned out to be brittle, unstable and ultimately unreliable to draw meaningful conclusions, which motivated the creation of the Tranco ranking as a more robust alternative for security and privacy research [40]. Tranco aggregates multiple top lists to mitigate their bias and reduce popularity fluctuations over time, hence it is now considered the reference dataset for modern web measurements.

Unfortunately, many web measurements cannot be meaningfully performed over Tranco as is, because Tranco is an *unlabeled* dataset, i.e., it is just a list of popular websites with no additional information about them. This makes *targeted* web measurement studies particularly challenging to carry out or significantly limited in practice. For example, prior work analyzed the privacy guarantees of the governmental websites ecosystem [25, 42] or studied website compliance with respect to country-level privacy regulations [17, 38]. These studies require the classification of existing websites as governmental or not, or even call for multiclass classification to associate websites to different countries, respectively.

In general, the process of labeling a dataset of websites to carry out targeted measurements may be complex, costly and error-prone. High-quality labels can be collected through human evaluators with sufficient domain expertise. Unfortunately, manual labeling has a significant cost, does not scale and makes it difficult to expand, or even replicate, existing studies. Automated labeling, in turn, is cheap and easy to scale, however it is normally based on heuristics that can introduce bias or lead to inaccuracies. For example, the country of a website can be determined by extracting its top-level domain, like .br, .de, or .it [38]. This approach is useful, yet inferior to manual labeling, because it does not allow labeling websites with a generic top-level domain like .com or .net.

Motivated by the explosive growth of generative AI, and the many success stories of Large Language Models (LLMs) in particular, we here explore the use of LLMs for creating labeled datasets of websites that enable the execution of representative, targeted web measurements. The key intuition of this proposal is that LLMs can perform automated website classification by leveraging contextual information, natural language understanding, and the extensive knowledge they have gained from training on massive datasets constructed by scraping diverse web sources. This makes them significantly more advanced than custom ad-hoc heuristics, e.g., based on selected keywords, and better equipped to rival the performance of human experts.

Contributions

In this paper, we make the following contributions:

- (1) We present a curated benchmark for key website classification tasks previously considered in web measurement papers

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, Washington, DC, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

published at major academic conferences. Starting from existing work, we go through the process of creating high-quality, labeled website datasets that are amenable for a systematic evaluation of automated classifiers.

- (2) Using our curated datasets, we investigate the effectiveness of three LLMs on different website classification tasks. Our analysis includes both a cloud-based LLM and two self-hosted LLMs of different size, thus covering different representative scenarios appealing to diverse use cases.
- (3) We demonstrate that LLMs can effectively support targeted web measurements by comparing the results of our LLM-assisted experiments against independent ground truths derived using methodologies from previous studies.

In the end, our work shows that LLMs are a versatile tool to support web measurements and can enable empirical studies that are accurate, require limited manual work, and are compliant with widely accepted practices, such as the use of the Tranco list. To support reproducible science, we release our datasets and prompts [8].

2 Methodology

We here present our research questions and motivate the design of our experiments starting from them. We then explain and motivate our experimental setup.

2.1 Research Questions

Our study is centered around three main research questions:

- RQ1: To what extent is website classification a common and useful practice in the field of web measurements?
- RQ2: Can LLMs perform website classification correctly and at scale?
- RQ3: Can LLMs be leveraged to support representative, targeted web measurements?

To answer RQ1, we perform a systematic literature review to identify research papers relying on website classification to carry out targeted web measurements (see Section 3). We leverage this analysis both to motivate the importance of the problem at hand and to identify publicly available datasets of labeled websites.

Then, to answer RQ2, we select representative website classification tasks to assess how LLMs perform on them (see Section 4). After validating the quality of the available datasets, we create curated variants suitable for a principled experimental evaluation. We then design corresponding prompts to guide the LLMs and assess their performance using standard measures.

Finally, to answer RQ3, we perform targeted web measurements based on our best-performing LLM for website classification (see Section 5). We then carefully validate the results of the measurements to determine whether the LLM was effective enough to ensure the correctness of the drawn research inferences.

2.2 Experimental Setup

We here clarify the most important details of our experimental setup, so that the next sections can focus on the results.

2.2.1 Choice of the LLMs. Nowadays, there are a plethora of LLMs available and their performance remains a subject of debate. We consider one popular cloud-based, closed-sourced LLM developed,

Gemini 2.5 Flash [15] by Google, which we access through its Developer API [24]. Moreover, we consider two self-hosted and open-source LLMs: Llama 4:109B [4], developed by Meta, and Gemma 3:27B [16], developed by Google, which we run through Ollama [26].

We test these two setups because they are both popular and appeal to different use cases. Cloud-based, closed-source LLMs like Gemini are particularly easy to use and do not require significant computational resources, however their use may introduce ethical concerns because sensitive research data may not be shared with third parties. Self-hosted, open-source LLMs, in turn, require a careful technical setup and sufficient computational power, however they enable local computations over sensitive data. Moreover, although the number of parameters of Gemini 2.5 Flash is not publicly available, Llama4:109B and Gemma3:27B are good representatives of large and medium-sized language models, respectively. The use of models of different sizes also allows us to appreciate the impact of the number of parameters on classification performance. We thus believe that our experimental setup captures representative models that are widely available, cover different use cases, and are expected to differ in performance, providing a meaningful basis for our investigation. Of course, our study could be readily extended to more LLMs with additional engineering effort.

2.2.2 Prompt Design. The design of our prompts for LLMs follows established prompt engineering practices. In particular, we employ *persona assignment* to enhance task adherence and focus [39], as well as *one-shot prompting* [10] to improve response accuracy by providing a single, well-defined example of the desired output. We consider two different prompt types: one where the LLMs are just provided with the website URLs and one where the LLMs are also fed the screenshots of the website homepages. We rely on screenshots rather than raw HTML documents, since state-of-the-art LLMs are multimodal and can readily process images. By contrast, providing HTML as input to LLMs requires additional care: HTML pages are often too long to fit within prompt limits, and deciding which HTML elements to prioritize is far from trivial and may depend on the specific website classification task.

2.2.3 Measurement Setup. When performing web measurement tasks, we visit the landing page of each domain from within an academic network using a lightweight crawler built on Playwright [36], that relies on the Chromium browser running in headful mode. Each domain is accessed through a fresh browser instance with a clean profile, and we allow 10 seconds for the page to load before performing any required interaction. Finally, we ground our investigation on the Tranco list generated on 06 April 2025 [1].

3 Website Classification

Here we review existing work on website classification, in particular in the field of web measurements, and we identify three case studies that we investigate in our research.

3.1 Literature Review

Website classification is a broad research area, appealing to different audiences and communities. Since the goal of our work is exploring the use of LLMs to support web measurements, we restrict the focus of our literature review to measurement papers and we

Table 1: Website classification tasks in web measurements.

Classification	Example Applications
By category	Analysis of governmental websites [25, 42, 43]
	Privacy analysis of pornographic websites [49]
	Social studies and disinformation [44, 55]
	Breakdown results by website category [35, 45]
By country	Compliance with privacy regulations [17, 38]
By functionality	Identify websites with a private area [5, 6, 22]
	Identify websites with SSO access [7, 18]

identify those requiring (or making use of) some form of website classification. Our methodology consists of the following steps:

- (1) We extract from DBLP all the papers published from 2018 to 2024 at the major academic conferences in the following fields: computer security (IEEE S&P, NDSS, ACM CCS, USENIX Security), privacy (PETS), Internet measurements (IMC) and the Web (WWW and WebSci).
- (2) We filter papers so as to only keep those that are most likely to present a web measurement. To do this, we first identify as potential candidates all the papers including a case insensitive match for the sub-strings “web” or “measur” in their title. When DBLP lists sessions or tracks for conferences, we also consider papers falling in the web security and web privacy categories as potential candidates. Finally, we read the abstract of the candidate papers to identify those actually performing a web measurement.
- (3) We inspect the matching papers to determine whether their web measurement involves any website classification step.

Our methodology identified 89 measurement papers published at the surveyed top venues, including 38 papers performing some form of website classification in at least one experiment (43%). At a high level, we observe that several papers rely on website classification to create new datasets, which are essential for drawing the primary conclusions of their study. For example, they are only interested in specific website categories [25, 42, 49] or they classify websites by country to check compliance against local privacy regulations [17, 38]. The other papers instead perform website classification as a complementary part of a broader analysis, e.g., they check whether specific website categories are correlated with the security or privacy aspects under study [18, 22, 45]. The main website classification tasks identified in the web measurement literature are reported in Table 1, along with a few representative papers for the different tasks.

3.2 Case Study Selection

Based on Table 1, we select key classification tasks to evaluate the capabilities of LLMs in supporting targeted web measurements.

3.2.1 Governmental Websites. Prior research analyzed relevant privacy risks for citizens associated with e-government practices [25, 42, 43]. We consider the automated detection of governmental websites as a first task to test the classification power of LLMs for multiple reasons. First, the importance of the topic: e-government is

becoming more widespread nowadays, thus drawing attention from the research community in the last few years. Moreover, previous studies crucially relied on the creation of datasets of governmental websites, which can be effectively used as a starting point for a systematic evaluation of the classification performance of LLMs. It is worth noticing that constructing these datasets is challenging, because many governmental websites are not hosted under dedicated top-level domains like .gov [25], meaning that the semantic understanding of LLMs can be helpful for their identification. Finally, the detection of governmental websites can be interpreted as a binary classification task (governmental vs. non-governmental), which is typically regarded as a baseline task in automated classification.

3.2.2 Website Country. Prior privacy studies analyzed website compliance with local privacy regulations [17, 38]. This requires associating websites with the country of their primary target audience, e.g., the GDPR imposes specific regulations on all websites offering services in the European Union. Unfortunately, classifying websites based on the country of their primary target audience is far from a simple task and prior work leveraged ad-hoc heuristics, e.g., by inferring the country from the top-level domain of the analyzed websites [38]. For instance, ebay.co.uk is labeled as a British website according to this methodology. LLMs can serve as an effective tool for the automated identification of the country where a website primarily operates, even when the top-level domain provides no useful information. This naturally leads to a multiclass classification task of particular interest to the web privacy community.

3.2.3 Website Category. Multiple studies rely on website categorization out of necessity (because they carry out targeted measurements, e.g., [49, 55]) or just to provide complementary insights (they break down analysis results by website category, e.g., [35, 45]). Traditional approaches to website categorization broadly fall into two categories. On the one hand, we have annotated website datasets like DMOZ / Open Directory Project [20], Yahoo Directory [53], and the already mentioned Alexa ranking. Most of these datasets have been discontinued and are no longer maintained, hence they cannot be used to meaningfully categorize today’s websites. On the other hand, we have online website classification services such as McAfee SiteAdvisor [34], Virus Total [50], and Cloudflare Radar [14]. These commercial services normally require premium access or put restrictions in their terms of service that complicate their adoption at scale. Additionally, they rely on fixed taxonomies that may provide insufficient granularity for specific web measurement studies. This state of affairs supports the case for LLMs as a convenient and widely available website categorization service, offering a great deal of flexibility with respect to the categories of interest.

3.2.4 Exclusions. In this work, we do not focus on the detection of private areas and SSO access in existing websites, as this can be automated using web crawlers designed to detect registration and login pages at scale [5, 6, 22, 27]. We believe that enhancing such crawlers with LLMs could improve accuracy, however this falls out of the scope of our investigation. Recent work on the use of LLMs for web crawling [46] may be inspiring for this line of research.

Table 2: Dataset statistics.

Dataset	#Instances	#Classes
Governmental	8,140	2
Countries	2,466	22
Categories	7,000	14

4 LLMs for Website Classification

A systematic assessment of the performance of classifiers requires the availability of high-quality datasets that are correctly labeled and represent all classes of the problem at hand. In this section, we create benchmark datasets for different classification tasks starting from existing work. We then report on the performance of the tested LLMs on our benchmark, complementing quantitative measures with a qualitative analysis of the results.

4.1 Benchmark Construction

We start by rigorously assessing the quality of existing datasets and understanding the details of the underlying classification tasks, which is important both to construct benchmarks supporting a principled experimental evaluation and to design prompts for LLMs.

Since our primary goal is assessing whether LLMs can support web measurements, which are normally performed on live websites, our dataset contains just websites that are correctly accessible using a standard web browser at the time of our experiments. This allows us to perform a careful validation of the actual website classes by accessing them when needed. We do not further stress this technical detail in the subsequent description and we often leave it implicit so as to maintain readability. Table 2 reports statistics about the constructed datasets, which we present in the following.

4.1.1 Governmental Websites. To test LLMs on the task of detecting governmental websites, we started from the dataset by Gotze et al. [25]. Their dataset was created by visiting the official webpage of the government of different countries and collecting all the links to ministries and agencies that were listed therein. In particular, their study focuses on those websites that are “associated with a domain that is registered and used by a national government” [25]. Unfortunately, this simple definition does not seem to fully reflect the actual nature of the dataset, which complicates its validation. For example, the dataset includes <https://laeggs.com/>, which is the website of the Louisiana Egg Commission in the United States. Although not directly managed by the U.S. federal government, it is operated by the Louisiana Department of Agriculture and Forestry and serves an educational service by informing consumers about the nutritional value of eggs and egg products.

To better understand the quality of the dataset in [25], we reviewed the original paper and proposed the following revised definition of governmental website for label validation: “a governmental website is an official online platform created and maintained by a government entity, or an organization significantly controlled or owned by a government. A primary goal of a governmental website must be to deliver government services, such as announcements, communication, exchange of information, and point of service to its citizens”. This definition extends the notion of governmental

website to those websites that are managed by any entity with strong connections with a national government, while it enforces the additional restriction that the website must deliver some services to the citizens. These choices are in line with the goals of the original study, which aims at understanding the privacy risks of e-governance as a point of interaction with useful services available to a wide audience of citizens [25].

To confirm the correctness of our revised definition, we sampled a random subset of 200 websites from the original dataset [25] and manually confirmed that our definition correctly captures 183 of them (92%), while the original definition just covers 138 websites (69%). As for the 17 cases that were not yet captured by our revised definition, we observed that 16 are false positives of the original dataset (these websites do not appear to be governmental in any substantive sense). This shows that the original dataset largely reflects a meaningful definition of governmental website, with a true positive rate of 92%, hence it can serve as a reliable source of governmental websites in our benchmark construction.

To properly assess whether LLMs can actually distinguish governmental from non-governmental websites, we constructed a balanced dataset that assigns equal weight to both classes. Our benchmark includes the 4,070 governmental websites from the original dataset [25] that are still accessible nowadays and a random subset of 4,070 websites from the Tranco list, which are not hosted under a known governmental top-level domain, e.g., .gov. To confirm the correctness of this random sampling, we accessed a subset of 200 websites from the sampled set and confirmed that 197 of them (99%) were indeed non-governmental as expected.

4.1.2 Website Country. To test LLMs on the task of detecting website countries, we constructed a new dataset of websites associated with the country of their primary target audience. In particular, we started from the methodology by Ogut et al. [38], who performed this association just by using the website top-level domain. Similarly, we randomly sampled 100 websites from Tranco for each of the 20 country-specific top-level domains (the 13 of the original study and 7 others), leading to an initial dataset of 2,000 websites from 20 classes. To confirm the quality of this labeling process, we randomly sampled 200 websites from the dataset and confirmed that 191 cases (96%) were labeled with the correct country.

Of course, websites hosted under generic top-level domains like .com and .net cannot be labeled using this simple methodology, which significantly complicates country attribution in large-scale measurements. Websites under a generic top-level domain would benefit from the semantic understanding of LLMs for an effective classification. Hence, to better capture the complexity of the classification task, we extended the initial dataset by sampling 500 random websites from Tranco hosted under four generic top-level domains (.com, .net, .org, and .io). This is five times the number of websites sampled for each of the chosen countries, thus reflecting that websites with a generic top-level domain are more common than others. We manually assigned labels to these websites, using the “international” label to denote those targeting a global audience rather than users from a single country. Out of the 500 classified websites, 221 (44%) serve an international audience, 120 (24%) target an American audience, while the other 159 (32%) are fragmented across 38 countries. We then filtered out websites associated with

extremely underrepresented countries with fewer than 10 instances each, where it would be impossible to compute any representative performance measure. The final dataset thus includes 2,466 websites from 22 classes, with the “international” class being roughly twice as large as the others. The subset of 466 websites with a generic top-level domain is particularly challenging to classify and subject to careful scrutiny in our experimental evaluation.

4.1.3 Website Category. To create our benchmark dataset, we leveraged the curated snapshot of the Curlie dataset by Lugeon et al. [33] and its 14 website categories. Curlie is a community-driven successor of DMOZ, inheriting its taxonomy and URLs while providing ongoing updates. Lugeon et al. performed a careful curation of a snapshot of Curlie to ensure appropriate data cleaning and label consistency within the community-contributed dataset, hence we take advantage of their effort in our research. The original snapshot includes around 90,000 websites, each labeled with a set of categories. Since we do not want to prioritize any category over the others and we consider all of them as equally important in practice, we created a balanced dataset of 7,000 websites by considering just the 86,680 websites with a single category (98%) and picking 500 websites at random for each of the 14 categories.

4.1.4 Label Noise. Although we carefully assessed the quality of the created datasets, primarily by manual validation, we acknowledge that some websites may have been labeled incorrectly. For example, we mentioned that the governmental websites dataset we start from [25] may contain some false positives (we estimated its true positive rate at 92%). Also, the country of a website may not always coincide with the country of its top-level domain (we estimated the accuracy of this heuristic at 96%). Finally, Lugeon et al. [33] observed that the Curlie dataset for website categorization is largely correct, but not exhaustively labeled, because contributors often select only one among all the relevant categories, e.g., the site of a football magazine may be labeled just as Sport or News.

Our adoption of existing datasets and methodologies, complemented by manual labeling and validation, gives strong assurance that our performance evaluation is representative, because the amount of label noise is estimated to be relatively small. To further reduce the impact of label noise, after evaluating the performance of LLMs on our datasets, we investigate the prediction errors of the best-performing model to identify cases where the labels of our datasets may have been wrong, and we perform a qualitative analysis of the findings to better understand performance.

4.2 Classification Performance

Now that we have curated datasets for the different classification tasks, we are ready to assess whether LLMs are effective on them.

4.2.1 Preliminaries. The general structure and an example of the prompts provided to the LLMs are reported in Appendix A. We note that we ask LLMs to provide two categories in output for the Categories dataset and we consider the output to be correct if and only if any of the two predicted categories matches the category reported in our dataset. The reason is that the Curlie dataset we start from is not exhaustively labeled, hence even a perfect classifier that always predicts a relevant category may be unduly penalized by missing website categories [33].

We use two standard performance measures: *accuracy* and *macro F1 score*. Accuracy, defined as the ratio of correct predictions to the total number of predictions, provides an overall measure of classifier performance on balanced datasets. For unbalanced datasets, the F1 score is commonly used; it is defined as the harmonic mean of precision and recall, thereby combining both aspects into a single metric. The macro F1 score extends this concept to the multiclass classification setting by computing a per-class F1 score and averaging them, thus giving each class the same weight. This way, we can assess whether performance degrades substantially on specific classes. Results of our experiments at a glance are shown in Table 3.

4.2.2 Governmental Websites. All the tested LLMs show good to excellent performance on the Governmental dataset, even when using URLs alone. The most accurate is Gemini, reaching an accuracy and a macro F1 score of 0.93 when having access to the screenshots, a small improvement over the use of URLs alone (+0.02 for both measures). All the tested models correctly classify at least 85% of the websites, with no significant performance differences between classes or between the two types of errors, as shown by the macro F1 score being very close to the accuracy. The inclusion of screenshots also benefits Llama and Gemma, both of which achieve an accuracy and a macro F1 score of 0.88 and 0.90, respectively, thus showing that self-hosted LLMs can also be highly effective for the classification task at hand.

We manually investigated a random subset of 20 websites apparently misclassified by Gemini to better understand our results from a qualitative perspective. As it turns out:

- 8 classification errors are clear-cut, because there is clear evidence that Gemini was wrong. Five of these errors come from university websites that Gemini incorrectly flagged as governmental, meaning they can likely be fixed through prompt engineering, i.e., by instructing Gemini that most universities are not significantly run by governments.
- 6 websites (3 governmental and 3 non-governmental) were incorrectly labeled in our dataset. This shows that Gemini can be even more precise than manual labeling and can be an effective tool for label validation. In particular, the amount of label errors in the sampled subset (30%) is significantly higher than the percentage of label errors previously estimated in the Governmental dataset (less than 8%), meaning Gemini can help target label validation efforts.
- 6 websites are in a gray area that is difficult to assess even for human evaluators. Most of these cases are associated with organizations where a national government has (or had) some share, but it is difficult to judge its actual involvement.

4.2.3 Website Country. All the evaluated LLMs achieve good to excellent performance also on the Countries dataset, even when relying solely on URLs. The best-performing LLM is again Gemini, with an accuracy of 0.95 and a macro F1 score of 0.96 when using URLs alone. In other words, 95% of the websites are correctly assigned to the country of their primary target audience and no class appears to be excessively penalized with respect to the others.

Perhaps surprisingly, Gemini performs worse when provided with screenshots alongside URLs: accuracy drops to 0.90 (-0.05) and macro F1 score lowers to 0.92 (-0.04). The reason might be

Table 3: Classification performance of different LLMs.

Dataset	URL alone						URL + Screenshot					
	Accuracy			Macro F1			Accuracy			Macro F1		
	Gemini	Llama	Gemma	Gemini	Llama	Gemma	Gemini	Llama	Gemma	Gemini	Llama	Gemma
Governmental	0.91	0.87	0.86	0.91	0.87	0.85	0.93	0.88	0.90	0.93	0.88	0.90
Countries	0.95	0.86	0.89	0.96	0.90	0.92	0.90	0.90	0.90	0.92	0.93	0.92
Generic TLD	0.72	0.50	0.44	0.61	0.50	0.54	0.77	0.59	0.49	0.78	0.63	0.73
Categories	0.77	0.70	0.74	0.77	0.70	0.73	0.76	0.60	0.64	0.76	0.60	0.65

that the URL itself is generally a strong predictor of a website’s country, thanks to the presence of the top-level domain, hence the screenshot might actually mislead the LLM. In contrast, by including screenshots Llama’s and Gemma’s performance improves as expected, e.g., Llama gains 0.04 in accuracy and 0.02 in macro F1 score.

Screenshots prove particularly useful when the classification task becomes more challenging, i.e., if we focus on websites hosted under a generic top-level domain. If we recompute the performance measures over this subset, Gemini achieves an accuracy of 0.77 and a macro F1 score of 0.78 using screenshots, significantly better than using URLs alone (+0.05 in accuracy, +0.17 in macro F1 score). Screenshots are thus fundamental for classification when the top-level domain does not provide any useful information about the website’s country, meaning the domain name is likely a weak predictor by itself. Llama and Gemma also benefit from screenshots, but their performance on the most challenging task remains insufficient, with accuracies of 0.59 and 0.49. In the end, our analysis indicates that Gemini can reliably identify the country of a website’s primary target audience even for generic domains, whereas self-hosted LLMs struggle with this more challenging task.

To better understand the performance of Gemini, we sampled a random subset of 20 websites hosted under generic top-level domains that have been classified to an apparently incorrect country:

- 11 websites are international (resp. US-based), but Gemini classified them as US-based (resp. international). Many are hard to judge even for human experts, as company sites are often in English, making it difficult to determine if they serve a global audience, even after extensive crawling.
- 5 websites are borderline cases, challenging even for human experts. For example, two websites host Hindi movies but have English interfaces, making the primary target audience unclear (human label: IN, Gemini label: international). In two cases, Gemini labeled websites with the country of the company owning them. While the primary markets are national, some services are also available internationally (as reflected by the human label).
- 4 websites have been classified by Gemini to a factually wrong class. In three of these cases, Gemini incorrectly assigned the “international” label.

Generalizing on the patterns identified in our sample, we observe that 29 of the total prediction errors occurred because Gemini classified international websites as US-based, while 31 of the total prediction errors happened because Gemini incorrectly flagged a local website as international. These two categories represent 50%

of the total number of prediction errors. Notably, these misclassifications often highlight cases that are inherently challenging to discriminate even for human evaluators (see the discussion above). This suggests that, in practical terms, Gemini’s performance is likely better than raw error counts might indicate, as the model successfully handles the majority of clearly distinguishable instances.

4.2.4 Website Category. The best-performing LLM on the Categories dataset is again Gemini, with an accuracy and a macro F1 score of 0.77 when using URLs alone. This means that in over three quarters of the websites, the correct category appears within the model’s two best predictions, with no category overly penalized compared to the others. If Gemini is asked to return a single category per website, its accuracy and macro F1 score drop to 0.64 and 0.63, suggesting that asking for just one category is relatively unlikely to return the label available in our dataset. This is expected because the Curly dataset is under-labeled and misses relevant categories [33]. Llama and Gemma also show reasonable performance, with accuracies of 0.70 and 0.74 and macro F1 scores of 0.70 and 0.73, respectively, when using URLs alone. This shows that self-hosted LLMs can also be effective for website categorization.

Including screenshots in the prompt does not improve the classification performance of the tested LLMs and may even penalize them. For example, Gemma loses 0.10 in accuracy and 0.09 in macro F1 score when given access to screenshots, suggesting that the URL may provide a more reliable signal than visual content for website categorization. This is likely because URLs activate knowledge about specific domains acquired during pre-training, while screenshots may introduce noisy or irrelevant visual information.

Finally, to better appreciate the actual performance of our best-performing model (Gemini), we sampled a random subset of 20 websites that have been assigned to apparently wrong categories. We observed that Gemini was factually wrong just in 6 cases, while in the other 14 cases the categories predicted by the model were just not reflected in our dataset. In most cases, both the categories predicted by Gemini were relevant in practice.

5 LLM-Assisted Web Measurements

In the previous section, we evaluated the effectiveness of LLMs for website classification on our benchmark datasets. Still, this does not yet show to what extent LLMs can provide effective support for web measurements in practice. Here we close this gap by leveraging our best-performing LLM (Gemini 2.5 Flash) as the backbone of selected web measurements, whose results we carefully review to assess the correctness of the drawn research inferences.

5.1 Privacy Risks of Single Sign-On

We start by exploring the privacy risks of SSO. Dimova et al. [18] presented a tool to assess such risks in existing SSO deployments and used it to measure their prevalence in the wild. The tool detects the presence of SSO on websites, identifies the available identity providers, and extracts the *scopes* requested during authentication, i.e., the type and amount of user information a website requests. A scope is considered *minimal* when it includes only the necessary information required for a given identity provider, and a website mitigates privacy risks when it consistently requests minimal scopes. As a part of their study, Dimova et al. [18] further reported a breakdown of the websites minimizing privacy risks (in terms of requested scopes) across different website categories. This work is a good example of how website categorization can complement web measurements by uncovering correlations between website categories and privacy/security best practices. For website classification, the study relies on the commercial McAfee SiteAdvisor tool [34], which is freely available for occasional individual use but not intended for large-scale analyses under its terms of service.¹

We replicate the study performed by Dimova et al. about the privacy risks per website category using Gemini, to understand if LLMs can help in drawing similar inferences. In particular, we downloaded the original dataset by Dimova et al. along with its website categories from SiteAdvisor and reassigned the categories using Gemini. For simplicity and readability, our analysis covers just the 10 most popular categories in the dataset, accounting for 2,912 websites from the original study (47%). Note that the categories assigned by SiteAdvisor are different from those available in the Curlie dataset considered in Section 4, hence our prompt for Gemini was adapted accordingly. Moreover, we instructed Gemini to assign only one label per website. Recall that in Section 4.2.4 we requested two labels from Gemini to compensate for the underlabeling of Curlie, while, in this case, requiring a single label allows us to be coherent with the categorization provided by SiteAdvisor.

Figure 1 compares the results of the privacy analysis using the original SiteAdvisor classification and the Gemini classification. As it turns out, the privacy inferences are nearly identical for the two classification tools across the vast majority of the categories. For example, the difference observed in the Forum/Bulletin category amounts to 1.4%, where websites requesting minimal scope amount to 72.0% and 70.6% according to Gemini’s and SiteAdvisor’s classification, respectively. The largest observed discrepancy between the two categorizations, and arguably the only notable one, is 8.2% for the Blogs/Wiki category. However, this difference does not significantly affect the overall conclusions of the analysis, because Blogs/Wiki still ranks among the last two categories in terms of the percentage of websites requesting minimal scopes. This result confirms that a commercial website classification service can be effectively replaced by a general-purpose LLM, provided it is properly instructed about the semantics of the target categories.

To further assess the consistency of the inferences that can be drawn using LLMs for website categorization, we conduct an additional experiment focusing on the two most popular identity providers in the dataset, i.e., Google and Facebook. Precisely, we compare the percentage of websites requesting minimal scopes

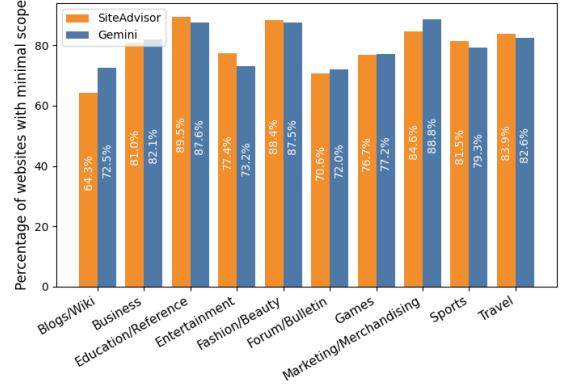


Figure 1: Percentage of websites with minimal scope by category, comparing SiteAdvisor and Gemini classification.

across categories for these providers, using both Gemini’s and SiteAdvisor’s classifications. This way, we can evaluate if the LLM-based categorization leads to similar privacy-related conclusions when the analysis is centered on identity providers rather than on websites alone. The results again confirm that the conclusions derived from both classifications are consistent. Detailed results are reported in Appendix B.

5.2 Privacy Analysis of Governmental Websites

Previous work explored the privacy risks of e-government practices by analyzing the cookies set by governmental websites [25, 42]. A significant challenge of these studies is the construction of an ad-hoc dataset of governmental websites for the measurement task. This process lacks standardized grounds and prior work relied on seeds of known governmental websites to identify more websites to include in the dataset by means of web crawling. We here explore to what extent the adoption of LLM-filtered versions of the Tranco list may allow performing representative web measurements, leading to similar conclusions to those observed on a carefully constructed, ad-hoc dataset of governmental websites. This task is much more ambitious than what we explored in the previous case study, where we used Gemini just to label an existing dataset rather than to construct a new dataset from scratch.

Concretely, we select the top and bottom 100k websites from the Tranco list to build an initial website dataset spanning a wide popularity range, and then apply Gemini to filter out non-governmental sites. To determine whether the resulting dataset is representative and allows one to draw meaningful research inferences, we use it to perform a privacy analysis by measuring the prevalence of third-party cookies set by known trackers on governmental websites, based on the Disconnect list [19]. We then compare our results against those obtained on two independent, ad-hoc datasets of governmental websites [25, 43] using the same analysis methodology.²

Results are shown in Figure 2, where we report the percentage of governmental websites with at least one third-party cookie set by a known tracker for different countries. We note that the figure

¹Dimova et al. likely relied on SiteAdvisor under a large-scale use agreement [34].

²We cannot compare also with the dataset of [42] because it is not publicly available.

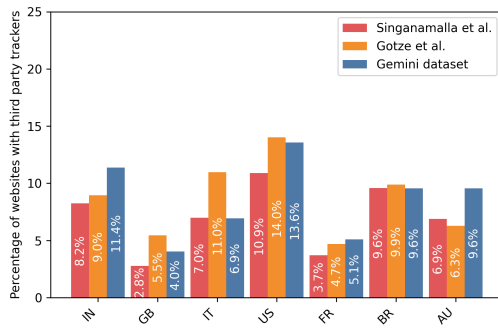


Figure 2: Percentage of websites with third-party trackers by country in the three datasets.

includes just the seven countries where we found at least 100 governmental websites in all three datasets, otherwise numbers would be too small to observe any meaningful trends. The experiment shows that the conclusions drawn using our LLM-filtered version of Tranco are accurate, because they are close to those observed on two independent, curated datasets of governmental websites. The largest observed difference between any two datasets is for Italy, with a percentage gap of around 4%. In particular, the lowest percentage of governmental websites including third-party cookies from known trackers is 6.9% in our dataset, which is aligned with the 7.0% observed in the dataset of [43], but differs from the 11.0% estimated on the dataset of [25]. We consider this result quite positive: specific differences across datasets are expected, because they include different sets of websites - indeed, also the two datasets [25, 43] that have not been constructed using LLMs are not perfectly aligned and particularly differ on Italy. Nevertheless, the gap between 6.9% and 11.0% is sufficiently small to support the conclusion that roughly one in ten governmental websites in Italy includes third-party tracking cookies. Observed differences across datasets for other countries are even smaller than this, meaning that privacy trends can be reliably measured for all countries where we have enough governmental websites to analyze.

Since the tracking ecosystem has been changing over the last years, with trackers relying less on third-party cookies given the improved privacy protections available in web browsers [37], we performed a second experiment to better appreciate other privacy trends measurable in our dataset. Concretely, we extracted all the script tags available in the body of the HTML documents accessed in our measurement and used the Disconnect list to identify the most prevalent (first-party) trackers for each country. Our analysis shows that all three datasets support similar conclusions, with the top five trackers for each country being roughly the same, sometimes even in the same relative order of popularity. The highest similarity is observed for the US, where the four most popular trackers have identical relative rankings across datasets, while the lowest occurs for France and Italy, where the sets of top trackers computed on the different datasets share three common elements, but their relative order is not preserved. For all the other countries, the datasets agree on at least three top trackers and partially preserve their relative order. Detailed results are reported in Appendix C.

6 Related Work

Web measurements are popular in the security and privacy communities to understand the current state of the Web. Security measurements have been performed to assess the adoption and configuration of important HTTP headers, like Content Security Policy [11, 52], HTTP Strict Transport Security [28], X-Frame-Options [12] and Cross Origin Resource Sharing [13]. Other work instead measured the prevalence and impact of significant web vulnerabilities, such as cross-site scripting [30], cross-site request forgery [47] and web cache poisoning [31]. In the privacy field, web measurements largely focused on cookies [2, 23], browser fingerprinting [29] and compliance with privacy regulations [17, 38]. As noted in Table 1, several measurement studies involved some form of website classification either as a fundamental component of the study or to provide a complementary perspective on specific website categories. We are not aware of any methodological study on how LLMs can serve web measurements at the time of writing.

Website classification is also an important task [41], appealing to different communities. While some research focuses on general classification tasks [33, 54], most of the works in the web security field focus on the effective detection of malicious activity, such as fraudulent e-commerce pages [9], phishing websites [32] and other types of malicious web pages. The main goal of this line of work is to improve the performance of existing detection approaches, rather than measuring the prevalence of malicious websites in the wild. We expect that LLMs can be successfully applied to this field as well, however the focus of our paper is investigating the use of LLMs as a useful support for web measurement studies.

Finally, LLMs have been recently applied to web crawling [46]. In this field, LLMs are used to process web pages after extracting semantic information from them, to improve crawling coverage and trigger complex interactions. This ability of LLMs can certainly be useful in web measurements, given the importance of the crawler on research inferences [3]. However, this line of research is orthogonal to our work, which instead focuses on the ability of LLMs to effectively perform website classification.

7 Conclusion

In this paper, we addressed the challenge of creating labeled datasets for targeted web measurements by proposing the use of LLMs for automated website classification. Traditional manual or heuristic-based labeling is often not scalable and can be inaccurate, while we showed that LLMs provide a robust alternative. We first reviewed existing web measurement studies to identify common website classification tasks. Then, we introduced a curated benchmark for key classification tasks and evaluated three LLMs on it, observing that they can effectively classify websites. Finally, we applied LLMs to support two targeted web measurements, leveraging community standards such as the Tranco list and obtaining results consistent with prior studies. Our findings demonstrate that LLMs are a versatile and powerful tool for web measurement research, enabling accurate and scalable analyses with minimal manual effort.

In future work, we would like to apply LLMs also to identify *relevant* websites and web pages to analyze for the specific security or privacy measurement at hand. Moreover, we plan to enhance LLM performance on website classification by exploring how to

employ fine-tuning [51] and in-context learning [21], since our results already indicate that LLMs are accurate for this task but there is margin of improvement.

References

- [1] 2025. *Tranco list 06 April 2025*. <https://tranco-list.eu/list/4QY5X> [Accessed 23-September-2025].
- [2] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. 2014. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, Scottsdale, AZ, USA, November 3-7, 2014, Gail-Joon Ahn, Moti Yung, and Ninghui Li (Eds.). ACM, 674–689. doi:10.1145/2660267.2660347
- [3] Syed Suleman Ahmad, Muhammad Daniyal Dar, Muhammad Fareed Zaffar, Narseo Vallina-Rodriguez, and Rishab Nithyanand. 2020. Apophanies or Epiphanies? How Crawlers Impact Our Understanding of the Web. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 271–280. doi:10.1145/3366423.3380113
- [4] AI@Meta. 2025. *Llama4*. https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md [Accessed 23-September-2025].
- [5] Suood Abdulaziz Al-Roomi and Frank Li. 2023. A Large-Scale Measurement of Website Login Policies. In *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, Joseph A. Calandrino and Carmela Troncoso (Eds.). USENIX Association, 2061–2078. <https://www.usenix.org/conference/usenixsecurity23/presentation/al-roomi>
- [6] Suood Alroomi and Frank Li. 2023. Measuring Website Password Creation Policies At Scale. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, Weizhi Meng, Christian Damsgaard Jensen, Cas Cremers, and Engin Kirda (Eds.). ACM, 3108–3122. doi:10.1145/3576915.3623156
- [7] Calvin Ardi and Matt Calder. 2023. The Prevalence of Single Sign-On on the Web: Towards the Next Generation of Web Content Measurement. In *Proceedings of the 2023 ACM on Internet Measurement Conference, IMC 2023, Montreal, QC, Canada, October 24-26, 2023*, Marie-José Montpetit, Aris Leivadreas, Steve Uhlig, and Mobin Javed (Eds.). ACM, 124–130. doi:10.1145/3618257.3624841
- [8] Artifacts 2025. Artifacts. <https://anonymous.4open.science/r/LLM-Assisted-Web-Measurements-Artifacts-2B0F/>. Repository containing all of the artifacts related to this paper.
- [9] Marzieh Bitaab, Haehyun Cho, Adam Oest, Zhuoer Lyu, Wei Wang, Jorij Abraham, Ruoyu Wang, Tiffany Bao, Yan Shoshitaishvili, and Adam Doupe. 2023. Beyond Phish: Toward Detecting Fraudulent e-Commerce Websites at Scale. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, IEEE, 2566–2583. doi:10.1109/SP46215.2023.10179461
- [10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). <https://proceedings.neurips.cc/paper/2020/hash/1457c0db6fcb4967418bfb8ac142f64a-Abstract.html>
- [11] Stefano Calzavara, Alvise Rabitti, and Michele Bugliesi. 2018. Semantics-Based Analysis of Content Security Policy Deployment. *ACM Trans. Web* 12, 2 (2018), 10:1–10:36. doi:10.1145/3149408
- [12] Stefano Calzavara, Sebastian Roth, Alvise Rabitti, Michael Backes, and Ben Stock. 2020. A Tale of Two Headers: A Formal Analysis of Inconsistent Click-Jacking Protection on the Web. In *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020, Srdjan Capkun and Franziska Roesner (Eds.)*. USENIX Association, 683–697. <https://www.usenix.org/conference/usenixsecurity20/presentation/calzavara>
- [13] Jianjun Chen, Jian Jiang, Hai-Xin Duan, Tao Wan, Shuo Chen, Vern Paxson, and Min Yang. 2018. We Still Don't Have Secure Cross-Domain Requests: an Empirical Study of CORS. In *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*, William Enck and Adrienne Porter Felt (Eds.). USENIX Association, 1079–1093. <https://www.usenix.org/conference/usenixsecurity18/presentation/chen-jianjun>
- [14] Cloudflare, Inc. 2020–. *Cloudflare Radar Domain Categorization*. <https://radar.cloudflare.com/domains>
- [15] Google Deepmind. 2025. *Gemini 2.5 Flash*. <https://deepmind.google/models/gemini/flash/> [Accessed 23-September-2025].
- [16] Google Deepmind. 2025. *Gemma3*. <https://ai.google.dev/gemma/docs/core?hl=it> [Accessed 23-September-2025].
- [17] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*, The Internet Society. <https://www.ndss-symposium.org/ndss-paper/we-value-your-privacy-now-take-some-cookies-measuring-the-gdprs-impact-on-web-privacy/>
- [18] Yana Dimova, Tom van Goethem, and Wouter Joosen. 2023. Everybody's Looking for SSOmething: A large-scale evaluation on the privacy of OAuth authentication on the web. *Proc. Priv. Enhancing Technol.* 2023, 4 (2023), 452–467. doi:10.56553/POPETS-2023-0119
- [19] Disconnect, Inc. 2025. *Disconnect – Tracker Protection Services List*. <https://disconnect.me/trackerprotection> Open-source list of domains used to identify trackers; used in many privacy and web measurement research projects..
- [20] DMOZ Contributors. 1998–2017. *DMOZ – The Open Directory Project*. <https://dmoz-odp.org/>
- [21] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 1107–1128. doi:10.18653/V1/2024.EMNLP-MAIN.64
- [22] Kostas Drakonakis, Sotiris Ioannidis, and Jason Polakis. 2020. The Cookie Hunter: Automated Black-box Auditing for Web Authentication and Authorization Flaws. In *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, Jay Ligatti, Xinming Ou, Jonathan Katz, and Giovanni Vigna (Eds.). ACM, 1953–1970. doi:10.1145/3372297.3417869
- [23] Steven Englehardt and Arvind Narayanan. 2016. Online Tracking: A 1-million-site Measurement and Analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi (Eds.). ACM, 1388–1401. doi:10.1145/2976749.2978313
- [24] Google. 2025. *Google Developer API*. <https://ai.google.dev/> [Accessed 23-September-2025].
- [25] Matthias Gotze, Srdjan Matic, Costas Iordanou, Georgios Smaragdakis, and Nikolaos Laoutaris. 2022. Measuring Web Cookies in Governmental Websites. In *WebSci '22: 14th ACM Web Science Conference 2022, Barcelona, Spain, June 26 - 29, 2022*, ACM, 44–54. doi:10.1145/3501247.3531545
- [26] Ollama Inc. 2025. *Ollama: Run, create, and share large language models locally*. <https://ollama.com> Accessed: 2025-08-26.
- [27] Louis Jannett, Christian Mainka, Maximilian Westers, Andreas Mayer, Tobias Wich, and Vladislav Mladenov. 2024. SoK: SSO-MONITOR - The Current State and Future Research Directions in Single Sign-on Security Measurements. In *9th IEEE European Symposium on Security and Privacy, EuroSecP 2024, Vienna, Austria, July 8-12, 2024*, IEEE, 173–192. doi:10.1109/EUROSP60621.2024.00018
- [28] Michael J. Kranch and Joseph Bonneau. 2015. Upgrading HTTPS in mid-air: An empirical study of strict transport security and key pinning. In *22nd Annual Network and Distributed System Security Symposium, NDSS 2015, San Diego, California, USA, February 8-11, 2015*, The Internet Society. <https://www.ndss-symposium.org/ndss2015/upgrading-https-mid-air-empirical-study-strict-transport-security-and-key-pinning>
- [29] Pierre Laperdrix, Natalia Bielova, Benoit Baudry, and Gildas Avoine. 2020. Browser Fingerprinting: A Survey. *ACM Trans. Web* 14, 2 (2020), 8:1–8:33. doi:10.1145/3386040
- [30] Sebastian Lekies, Ben Stock, and Martin Johns. 2013. 25 million flows later: large-scale detection of DOM-based XSS. In *2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4-8, 2013*, Ahmad-Reza Sadeghi, Virgil D. Gligor, and Moti Yung (Eds.). ACM, 1193–1204. doi:10.1145/2508859.2516703
- [31] Yuejia Liang, Jianjun Chen, Run Guo, Kaiwen Shen, Hui Jiang, Man Hou, Yue Yu, and Haixin Duan. 2024. Internet's Invisible Enemy: Detecting and Measuring Web Cache Poisoning in the Wild. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security, CCS 2024, Salt Lake City, UT, USA, October 14-18, 2024*, Bo Luo, Xiaoqing Liao, Jun Xu, Engin Kirda, and David Lie (Eds.). ACM, 452–466. doi:10.1145/3658644.3690361
- [32] Yun Lin, Ruofan Liu, Dinil Mon Divakaran, Jun Yang Ng, Qing Zhou Chan, Yiwen Lu, Yuxuan Si, Fan Zhang, and Jin Song Dong. 2021. Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, Michael D. Bailey and Rachel Greenstadt (Eds.). USENIX Association, 3793–3810. <https://www.usenix.org/conference/usenixsecurity21/presentation/lin>
- [33] Sylvain Lugeon, Tiziano Piccardi, and Robert West. 2022. Homepage2Vec: Language-Agnostic Website Embedding and Classification. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, Ceren Budak, Meeyoung Cha, and

- Daniele Quercia (Eds.). AAAI Press, 1285–1291. <https://ojs.aaai.org/index.php/ICWSM/article/view/19380>
- [34] McAfee Corp. 2006–. *McAfee SiteAdvisor*. <https://sitelookup.mcafee.com/>
- [35] Abner Mendoza, Phakpoom Chinpruthiwong, and Guofei Gu. 2018. Uncovering HTTP Header Inconsistencies and the Impact on Desktop/Mobile Websites. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 247–256. doi:10.1145/3178876.3186091
- [36] Microsoft. 2025. *Playwright*. <https://playwright.dev> [Accessed 23-September-2025].
- [37] Shaor Munir, Sandra Siby, Umar Iqbal, Steven Englehardt, Zubair Shafiq, and Carmela Troncoso. 2023. CookieGraph: Understanding and Detecting First-Party Tracking Cookies. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, Weizhi Meng, Christian Damsgaard Jensen, Cas Cremers, and Engin Kirda (Eds.). ACM, 3490–3504. doi:10.1145/3576915.3616586
- [38] Aysun Ogut, Berke Turanlioglu, Doruk Can Metiner, Albert Levi, Cemal Yilmaz, Orçun Çetin, and A. Selcuk Uluagac. 2024. Dissecting Privacy Perspectives of Websites Around the World: "Acceptar Todo, Alle Akzeptieren, Accept All...". In *33rd USENIX Security Symposium, USENIX Security 2024, Philadelphia, PA, USA, August 14-16, 2024*, Davide Balzarotti and Wenyuan Xu (Eds.). USENIX Association. <https://www.usenix.org/conference/usenixsecurity24/presentation/ogut>
- [39] Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche (Eds.). ACM, 2:1–2:22. doi:10.1145/3586183.3606763
- [40] Victor Le Pochat, Tom van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczynski, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society. <https://www.ndss-symposium.org/ndss-paper/tranco-a-research-oriented-top-sites-ranking-hardened-against-manipulation/>
- [41] Xiaoguang Qi and Brian D. Davison. 2009. Web page classification: Features and algorithms. *ACM Comput. Surv.* 41, 2 (2009), 12:1–12:31. doi:10.1145/1459352.1459357
- [42] Nayanamana Samarasinghe, Aashish Adhikari, Mohammad Mannan, and Amr M. Youssef. 2022. Et tu, Brute? Privacy Analysis of Government Websites and Mobile Apps. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini (Eds.). ACM, 564–575. doi:10.1145/3485447.3512223
- [43] Sudheesh Singanamalla, Esther Han Beol Jang, Richard Anderson, Tadayoshi Kohno, and Kurtis Heimerl. 2020. Accept the Risk and Continue: Measuring the Long Tail of Government https Adoption. In *IMC '20: ACM Internet Measurement Conference, Virtual Event, USA, October 27-29, 2020*. ACM, 577–597. doi:10.1145/3419394.3423645
- [44] Alexander Spangher, Gireeja Ranade, Besmira Nushi, Adam Fourney, and Eric Horvitz. 2020. Characterizing Search-Engine Traffic to Internet Research Agency Web Properties. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 2253–2263. doi:10.1145/3366423.3380290
- [45] Marco Squarcina, Mauro Tempesta, Lorenzo Veronese, Stefano Calzavara, and Matteo Maffei. 2021. Can I Take Your Subdomain? Exploring Same-Site Attacks in the Modern Web. In *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, Michael D. Bailey and Rachel Greenstadt (Eds.). USENIX Association, 2917–2934. <https://www.usenix.org/conference/usenixsecurity21/presentation/squarcina>
- [46] Aleksei Stafeyev, Tim Recktenwald, Gianluca De Stefano, Soheil Khodayari, and Giancarlo Pellegrino. 2025. YuraScanner: Leveraging LLMs for Task-driven Web App Scanning. In *32nd Annual Network and Distributed System Security Symposium, NDSS 2025, San Diego, California, USA, February 24-28, 2025*. The Internet Society. <https://www.ndss-symposium.org/ndss-paper/yurascanner-leveraging-llms-for-task-driven-web-app-scanning/>
- [47] Avinash Sudhodanan, Roberto Carbone, Luca Compagna, Nicolas Dolgin, Alessandro Armando, and Umberto Morelli. 2017. Large-Scale Analysis & Detection of Authentication Cross-Site Request Forgeries. In *2017 IEEE European Symposium on Security and Privacy, EuroS&P 2017, Paris, France, April 26-28, 2017*. IEEE, 350–365. doi:10.1109/EUROSP.2017.45
- [48] The Daily Star. 2021. *Amazon closing down Alexa, the popular web traffic ranking site*. <https://www.thedailystar.net/tech-startup/news/amazon-closing-down-alexa-the-popular-web-traffic-ranking-site-2913401> Accessed on October 10, 2025.
- [49] Pelayo Vallina, Álvaro Feal, Julien Gamba, Narseo Vallina-Rodriguez, and Antonio Fernández Anta. 2019. Tales from the Poro: A Comprehensive Privacy Analysis of the Web Porn Ecosystem. In *Proceedings of the Internet Measurement Conference, IMC 2019, Amsterdam, The Netherlands, October 21-23, 2019*. ACM, 245–258. doi:10.1145/3355369.3355583
- [50] VirusTotal Team. 2025. *VirusTotal*. <https://www.virustotal.com/> Accessed: 2025-09-25.
- [51] Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. 2025. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artif. Intell. Rev.* 58, 8 (2025), 227. doi:10.1007/S10462-025-11236-4
- [52] Lukas Weichselbaum, Michele Spagnuolo, Sebastian Lekies, and Artur Janc. 2016. CSP Is Dead, Long Live CSP! On the Insecurity of Whitelists and the Future of Content Security Policy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*, Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi (Eds.). ACM, 1376–1387. doi:10.1145/2976749.2978363
- [53] Yahoo! Inc. 1994–2014. *Yahoo Directory*. <https://dir.yahoo.com/>
- [54] Eric Ye, Xiao Bai, Neil O'Hare, Eliyar Asgarieh, Kapil Thadani, Francisco Perez-Sorrosal, and Sujyothi Adiga. 2024. Multilingual Taxonomic Web Page Categorization Through Ensemble Knowledge Distillation. *IEEE Trans. Knowl. Data Eng.* 36, 11 (2024), 6614–6627. doi:10.1109/TKDE.2024.3406368
- [55] Eric Zeng, Miranda Wei, Theo Gregersen, Tadayoshi Kohno, and Franziska Roesner. 2021. Polls, clickbait, and commemorative \$2 bills: problematic political advertising on news and media websites around the 2020 U.S. elections. In *IMC '21: ACM Internet Measurement Conference, Virtual Event, USA, November 2-4, 2021*, Dave Levin, Alan Mislove, Johanna Amann, and Matthew Luckie (Eds.). ACM, 507–525. doi:10.1145/3487552.3487850

A LLM Prompts

In the following, we report the overall structure used in our prompts, along with an example of the prompt employed to categorize governmental websites that also leverage screenshots.

You are a classifier used to categorize websites into governmental and non-governmental websites. A governmental website is an official online platform created and maintained by a government entity, or an organization significantly controlled or owned by a government. A primary goal of a governmental website must be to deliver government services, such as announcements, communication, exchange of information, and point of service to their citizens. You are used by a research team conducting web measurements. You will be given a list of websites. For each website:

- Identify whether the website is a governmental website or not based on our definition.
- If it is a governmental website also tell us the country of the government.
- Do not modify the provided URLs.
- Use both the URL and its corresponding screenshot to decide whether each website is governmental or not.
- Do not excessively rely on the .gov TLD: although this is likely a strong signal of governmental websites, some websites are operated by governments, but do not offer any services to citizens.

For example, <https://www.pagopa.gov.it/> would be categorized as a governmental website, since it is run by the Italian government and allows citizens to perform online payments for governmental services.

Please analyze all websites and give an answer for all of them. Reply with the following json format: '`<url1>: gov_site: <true/false>, country: <country (provide only if it is a governmental website)>, <url2>: gov_site: <true/false>, country: <country (provide only if it is a governmental website)>, ...`' and nothing else. Here is the list of websites together with their screenshot:

Above is the prompt used to categorize governmental websites, which follows the techniques described in Section 2.2.2 and uses the same structure that was used for all of our prompts. The prompt starts with defining the LLM's role (*persona assignment* [39]), followed by the definition of what the LLM has to categorize. It then provides a list of instructions, an example for reference (*one-shot prompting* [10]), the desired output format, and the list of websites (with screenshots if required) to be categorized. The full list of prompts used can be found in our online repository [8].

B SSO IDPs

Figure 3 reports the percentage of websites requesting minimal scopes for the two most popular identity providers in the dataset by Dimova et al. [18], i.e., Facebook and Google, grouped according to the categories assigned by Gemini and SiteAdvisor. For simplicity

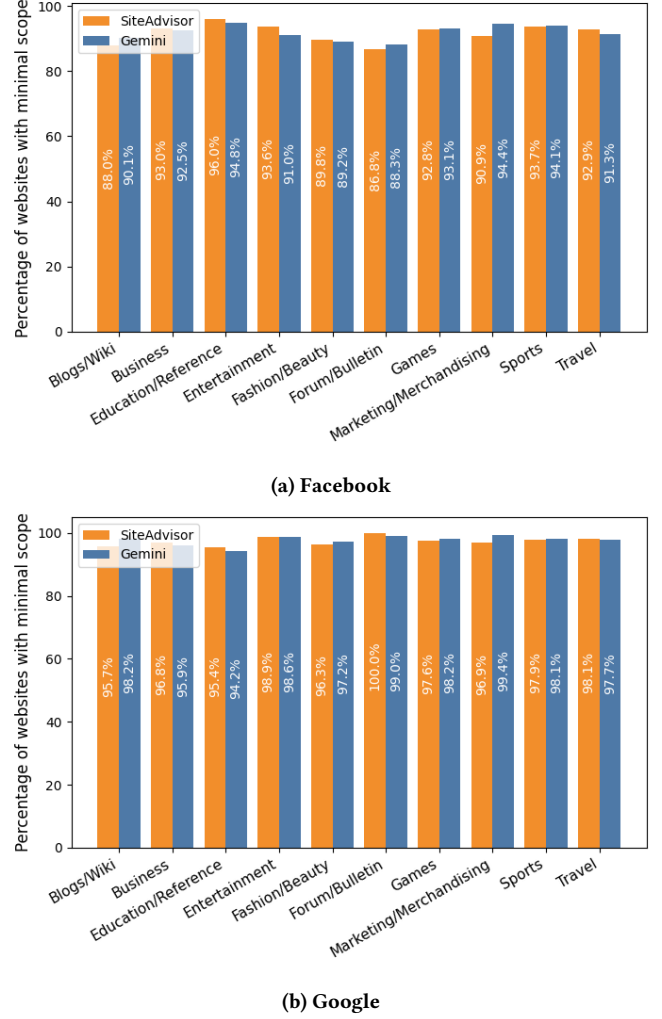


Figure 3: Percentage of websites with minimal scope by category for the two most popular IDPs.

and readability, we focus on the 10 most common categories in the dataset. We select these two identity providers because they include a sufficient number of websites per category to observe meaningful trends. The results indicate that the percentage of websites requesting minimal scopes across identity providers and categories is closely aligned when adopting the website classification operated by the two different tools, revealing consistent privacy trends. In particular, at least 95% of websites (for Facebook) and 86% (for Google) request minimal scopes across all categories, regardless of the tool used for website categorization.

C First-Party Trackers

Table 4 reports the most popular first-party trackers observed for different countries in three different datasets of governmental websites. Entries in bold mark cases where a tracker occurs in the same position in all datasets, while entries in italics mark cases where

Table 4: Popular script trackers by country.

Country	Rank	Gemini dataset	Gotze et al. [25]	Singanamalla et al. [43]
AU	1	.google.com	.google.com	.google.com
	2	.cdn.jsdelivr.net	.cdn.jsdelivr.net	.cdn.jsdelivr.net
	3	<i>.recaptcha.net</i>	<i>.unpkg.com</i>	<i>.unpkg.com</i>
	4	<i>.siteimproveanalytics.com</i>	<i>.recaptcha.net</i>	<i>.siteimproveanalytics.com</i>
	5	<i>.unpkg.com</i>	<i>.siteimproveanalytics.com</i>	<i>.recaptcha.net</i>
BR	1	<i>.cdn.jsdelivr.net</i>	<i>.unpkg.com</i>	<i>.cdn.jsdelivr.net</i>
	2	.google.com	.google.com	.google.com
	3	<i>.unpkg.com</i>	<i>.cdn.jsdelivr.net</i>	<i>.unpkg.com</i>
	4	.youtube.com	.youtube.com	.youtube.com
	5	<i>.gstatic.com</i>	<i>.bing.com</i>	<i>.gstatic.com</i>
FR	1	<i>.cdn.jsdelivr.net</i>	<i>.cdn.jsdelivr.net</i>	<i>.google.com</i>
	2	<i>.google.com</i>	<i>.unpkg.com</i>	<i>.cdn.jsdelivr.net</i>
	3	<i>.unpkg.com</i>	<i>.google.com</i>	<i>.unpkg.com</i>
	4	<i>.hcaptcha.com</i>	<i>.siteimproveanalytics.com</i>	<i>.youtube.com</i>
	5	<i>.youtube.com</i>		<i>.siteimproveanalytics.com</i>
GB	1	.cdn.jsdelivr.net	.cdn.jsdelivr.net	.cdn.jsdelivr.net
	2	.google.com	.google.com	.google.com
	3	<i>.unpkg.com</i>	<i>.unpkg.com</i>	<i>.siteimproveanalytics.com</i>
	4	<i>.gstatic.com</i>	<i>.youtube.com</i>	<i>.unpkg.com</i>
	5	<i>.googleadservices.com</i>	<i>.recaptcha.net</i>	<i>.gstatic.com</i>
IN	1	.cdn.jsdelivr.net	.cdn.jsdelivr.net	.cdn.jsdelivr.net
	2	.google.com	.google.com	.google.com
	3	<i>.jsc.mgid.com</i>	<i>.unpkg.com</i>	<i>.gstatic.com</i>
	4	<i>.instagram.com</i>	<i>.gstatic.com</i>	<i>.unpkg.com</i>
	5	<i>.unpkg.com</i>	<i>.statcounter.com</i>	<i>.instagram.com</i>
IT	1	<i>.google.com</i>	<i>.cdn.jsdelivr.net</i>	<i>.cdn.jsdelivr.net</i>
	2	<i>.cdn.jsdelivr.net</i>	<i>.google.com</i>	<i>.unpkg.com</i>
	3	<i>.unpkg.com</i>	<i>.hcaptcha.com</i>	<i>.google.com</i>
	4	<i>.siteimproveanalytics.com</i>	<i>.unpkg.com</i>	<i>.youtube.com</i>
	5	<i>.youtube.com</i>	<i>.siteimproveanalytics.com</i>	<i>.clarity.ms</i>
US	1	.cdn.jsdelivr.net	.cdn.jsdelivr.net	.cdn.jsdelivr.net
	2	.google.com	.google.com	.google.com
	3	.siteimproveanalytics.com	.siteimproveanalytics.com	.siteimproveanalytics.com
	4	.unpkg.com	.unpkg.com	.unpkg.com
	5	<i>.youtube.com</i>	<i>.youtube.com</i>	<i>.static.cloud.coveo.com</i>

a tracker occurs in the top five positions of all datasets, but the relative position is not preserved.