Understanding Privacy Norms through Web Forms

Hao Cui
University of California, Irvine
cuih7@uci.edu

Rahmadi Trimananda
University of California, Irvine
rtrimana@uci.edu

Athina Markopoulou
University of California, Irvine
athina@uci.edu

Abstract

Web forms are one of the primary ways to collect personal information online, yet they are relatively under-studied. Unlike web tracking, data collection through web forms is explicit and contextualized. Users (i) are asked to input specific personal information types, and (ii) know the specific context (*i.e.,* on which website and for what purpose). For web forms to be trusted by users, they must meet the common sense standards of appropriate data collection practices within a particular context (*i.e.,* privacy norms). In this paper, we extract the privacy norms embedded within web forms through a measurement study. First, we build a specialized crawler to discover web forms on websites. We run it on 11,500 popular websites, and we create a dataset of 293K web forms. Second, to process data of this scale, we develop a cost-efficient way to annotate web forms with *form types* and *personal information types*, using text classifiers trained with assistance of large language models (LLMs). Third, by analyzing the annotated dataset, we reveal common patterns of data collection practices. We find that (i) these patterns are explained by functional necessities and legal obligations, thus reflecting privacy norms, and that (ii) deviations from the observed norms often signal unnecessary data collection. In addition, we analyze the privacy policies that accompany web forms. We show that, despite their wide adoption and use, there is a disconnect between privacy policy disclosures and the observed privacy norms.

Keywords

Web form, privacy norm, privacy policy, measurement.

1 Introduction

Collection of personal information (PI) has been widely studied in the privacy community. On websites, PI collection is performed explicitly through web forms, and/or implicitly through web tracking. While web tracking has received much attention [3, 20, 31], web forms are rarely discussed from a privacy perspective despite being so ubiquitous and designed specifically to collect user inputs.

PI collection through web forms has unique characteristics that make it explicit and clear. First, it requires direct user involvement to input PI types that cannot be otherwise automatically collected. Second, web forms are set up to collect data in specific *contexts*, that is, on specific websites and for specific purposes. Figure 1 shows two web forms that collect different sets of PI in different contexts.

Privacy Norms. For web forms to be trusted by users, we argue that they must meet the standards of what is widely considered appropriate PI collection, which we broadly refer to as *privacy norms*. For example, account registration on a financial website may require personal tax ID; in contrast, subscribing to a news media site is unlikely to require such information. Privacy norms are often implicit and the research community has explored the topic primarily through user surveys [1, 7, 8, 29, 40, 68]. In this work, we provide a novel approach to understanding privacy norms, by observing common PI collection practices on web forms. Through a measurement study of web forms on popular websites, we connect PI types to contexts[1] in which they are collected.

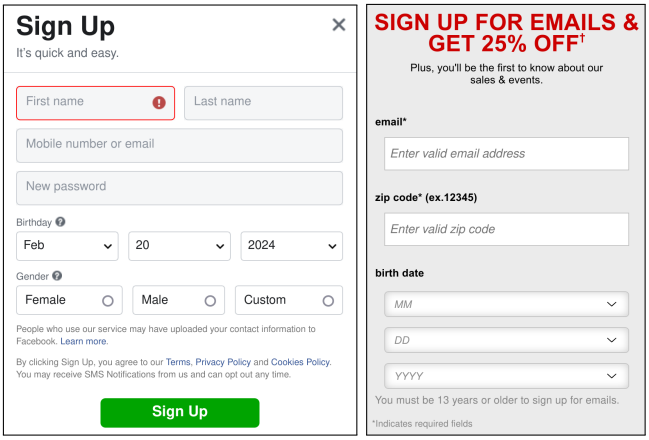


Figure 1: Web form examples. Left: the account registration form on *facebook.com* asks for name, phone number, email address, birth date, and gender. Right: the email list subscription form on *macys.com* asks for email address, zip code, and birth date. Users have to fill in the requested PI in the fields in order to use the functionality provided by the forms.

Web Forms Collection. To perform measurement at scale, we build a browser-based crawler that discovers and downloads web forms from websites. The crawler is able to simulate clicks on web elements to trigger web forms, including dynamic forms that are created at runtime. It also implements heuristics to detect links and buttons that likely point to web forms, thus improving efficiency. We run the crawler on websites from the Tranco list [56] and we collect nearly a million web forms from over 10K English websites.

Dataset Annotation. We process the HTML code of crawled web forms to infer the functionality (which we refer to as *form type*) and extract *PI types*. In order to efficiently annotate the huge amount of data, we build a machine learning system that distills the large language model (LLM) into a task-specific text classifier [79]. We use GPT-3.5 Turbo [52], the state-of-the-art LLM, for unsupervised data labeling, and transfer the knowledge to a smaller form type classifier. We adopt active learning, with LLM in loop, to improve model generalization on minority labels [60, 79]. Our methodology

[1] In this work, the *context* of PI collection on a web form refers to the *website category* (*e.g.,* financial) and the *functionality* of the web form (which reflects the purposes of PI collection, *e.g.,* account registration).

Proceedings on Privacy Enhancing Technologies 2025(1), 5–22

https://doi.org/10.56553/popets-2025-0002

is more efficient in terms of monetary and labor costs for labeling than both LLM-only zero-shot classification [34, 94] and traditional manual approach of training data labeling. After annotation and cleaning up, we create the first annotated web form dataset with 293K web forms from 11,500 websites.

Web Form Analysis. We analyze the annotated web forms and reveal common patterns of PI collection. By comparing the collection rates of each PI type across different website categories and form types, we reveal what PI types are collected often in what contexts. These patterns reflect privacy norms that can be explained by functionality, legal obligations and other reasons, as shown in the following examples from our findings. (1) The ubiquitous collection of email addresses reflects the perceived non-sensitivity of this PI type. (2) Other contact information, namely phone numbers and addresses, are used more often by websites that are directly related to real-world services, such as health and finance. (3) The collection of date of birth and age for account registration can be attributed to children's privacy regulations. (4) In the financial and health contexts, strict identity verification requirements are evident by the collection of extensive PI types. Conversely, uncommon PI collection practices that do not align with the privacy norms can indicate excessive PI collection. The privacy norms identified in our dataset can be used as a baseline to assess data minimization.

Privacy Policy Analysis. We also compare our observations on web forms to privacy policies – the legal documents that are supposed to disclose PI collection practices. We download the privacy policies that accompany web forms. We find that, while over 90% of websites provide privacy policies, less than half of the web forms include links within them, indicating that many policies may not be contextualized to explain the web forms. This is further confirmed by the differences between privacy policy disclosures and observed privacy norms. We use a state-of-the-art privacy policy analyzer, PoliGraph-er [19], to extract disclosures of PI collection. By comparing each website's actual PI collection practices with its privacy policy, we reveal the gap between the two. On the one hand, some websites do not disclose all the PI types collected in the web forms, indicating possible privacy violations. On the other hand, some websites simply use blanket disclosures, claiming to collect many PI types that we did not observe, and are unlikely appropriate in the corresponding contexts. These findings put in question whether privacy policies actually help in understanding websites' PI collection practices.

Contributions. This paper makes the following contributions.

- *Measurement of Web Forms:* We perform the first, to the best of our knowledge, large measurement study of PI collection through web forms. We create a large annotated dataset of 293K web forms on 11,500 popular English websites, which provides a comprehensive view of PI collection practices across different contexts of form types and website categories.
- *Understanding Privacy Norms:* We propose a novel approach to extracting privacy norms, by analyzing common PI collection practices on web forms. We show that these patterns can be attributed to reasons like functionality and legal obligation, which thus can be called privacy norms. We also extend our analysis to privacy policies, revealing their misalignment with the norms.
- *Methodological Contributions:* To facilitate the measurement study, we build a customized crawler to discover and download web forms. We also develop a cost-efficient machine learning system to annotate web forms, using a combination of LLM and active learning to help train task-specific classifiers.

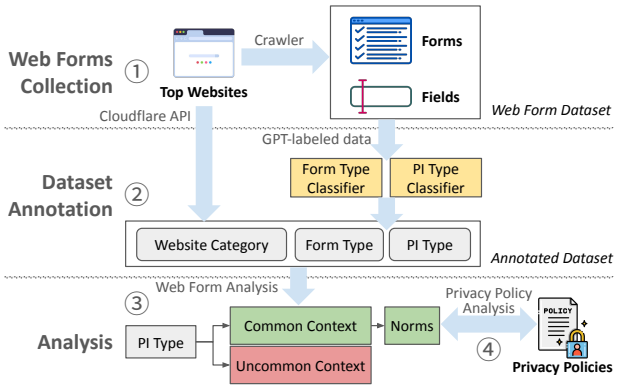


Figure 2: Overview. ① We collect web forms from top websites using a customized crawler (Section 3). ② We develop a machine learning system to annotate the web forms with form types and PI types (Section 4). ③ We analyze the web forms to reveal common patterns of PI collection that reflect privacy norms and uncommon cases (Section 5). ④ Finally, we also analyze privacy policies to compare the observed norms to disclosed PI collection practices (Section 6).

Outline. Figure 2 is an overview of our work. The rest of the paper is structured as follows. Section 2 presents related work. Section 3 describes the collection of the web form dataset. Section 4 describes the dataset annotation methodology. Section 5 presents the analysis of the web form dataset. Section 6 presents the analysis of privacy policies. Section 7 discusses limitations and future work.

2 Background and Related Work

Next, we position our work among related bodies of work on privacy norms, laws, privacy policies and privacy measurement.

2.1 Contextual Privacy Norms

Privacy norms refer to the common standards of acceptable privacy practices, and specifically in this paper, *acceptable PI collection practices*. Contextual integrity (CI) is a framework to analyze privacy norms [48]. CI states that privacy should be thought as appropriate information flows (rather than just hiding information) that conform to informational (privacy) norms specific to given social contexts. CI describes such contextual privacy norms using five parameters: (1) data *subject*; (2) *sender* of the data; (3) *recipient* of the data; (4) *information type*; and (5) *transmission principle*, such as the condition or the purposes of PI collection.[2]

Privacy norms are often implicitly defined by social and cultural norms [48]. A large group of prior work has applied CI to discover these implicit privacy norms. This is usually done through vignette surveys that present participants different contexts (*i.e.,* combinations of CI parameters) and ask them to score the acceptability [40]. Shvartzshnaider et al. [68] shows crowdsourcing can be used to

[2]For example, it is considered appropriate for a financial institution (*recipient*) to collect name, tax ID and date of birth (*information types*) from clients (*sender / subject*) during account registration for identity verification purposes (*transmission principle*).

discover contextual privacy norms. Other work surveyed public privacy expectations in specific contexts, such as IoT toy [8], smart home [1, 7], online photo sharing [29].

Our work aims to understand privacy norms through web forms. Because of their explicit and contextual nature, web forms are governed by implicit privacy norms. The parameters of web forms can be *roughly* mapped to CI parameters: the website (*recipient*) collects a set of PI (*information types*) from users (*sender & subject*) for the purposes implied by the form type (*transmission principle*). We propose a novel way to discover privacy norms by collecting and analyzing web forms from popular websites. Unlike vignette surveys that present hypothetical situations, the privacy norms we extract are based on actual PI collection practices in the wild.

2.2 Laws and Privacy Policy

Laws and regulations are other sources of privacy norms. For example, U.S. financial institutions are legally required to collect many PI types from customers for identity verification [23, 82]. In Section 5.1, we attribute some of the observed privacy norms to legal requirements. Privacy laws, such as the California Consumer Privacy Act (CCPA) [73, 74] in the U.S., and the General Data Protection Regulation (GDPR) [59] in the European Union (EU) and the European Economic Area (EEA), generally do not directly specify what PI types can or cannot be collected. Instead, they require businesses to provide transparent privacy notices, obtain informed consents (*i.e., notice and consent*) [32] and/or explain the legal basis for PI collection. Despite many criticisms about the effectiveness of *privacy policies* [9, 49, 54], they have been the main legally-binding mechanism for disclosing PI collection practices.

Privacy policy analysis, including by experts [67, 95] and via natural language processing (NLP) [5, 10, 19, 26, 80], has revealed compliance issues in privacy policies. Shvartzshnaider et al. [67], using CI as the analysis framework, report missing contextual details and vague policy language that make the interpretation ambiguous. Prior work of NLP analysis shows the prevalence of vague, missing or even self-contradicted [5, 49, 64] statements in the privacy policies of mobile apps [5, 6, 49, 97], virtual reality (VR) games [88] and smart home devices [38]. In Section 6, we apply PoliGraph-er [19], a state-of-the-art NLP privacy policy analyzer, to analyze privacy policies associated with web forms. We show the misalignment between privacy policies and observed PI collection, which questions the very role of privacy policies in understanding privacy norms.

Another relevant legal concept is *data minimization*. For example, the CCPA [73, 74] requires PI collection to be limited to what is "reasonably necessary and proportionate to achieve the purposes for which the personal information was collected or processed..." While the law does not define what PI types are necessary under which circumstances, our research, along with prior work on privacy norms [1, 7, 8, 29, 40, 68], provides a baseline for data minimization in different contexts. In Section 5.2, we discuss how uncommon cases that do not align with privacy norms may indicate *over-collection*, which violates the data minimization principle.

2.3 Data Collection Measurement

Extensive research has been conducted to reveal privacy issues of software systems. In the web ecosystem, web tracking (*i.e.,* the automatic data collection happening in the background as users interact with websites) has received much attention [3, 20]. Iqbal et al. [31] report that stateless tracking was used by over a quarter of the top-10K websites as of 2019. Similar tracking technologies are also used in other non-web platforms, such as smartphones [6, 97], smart TVs [93], VR headsets [88], and voice assistants [30].

Unlike the opaqueness and invisibility of web tracking, PI collection through web forms appears transparent,[3] thus better represent privacy norms. We also note that web forms involve many sensitive PI types that cannot be otherwise automatically collected, such as contact information, government IDs, gender and ethnicity, *etc.*

As for measurements related to web forms, Preibusch et al. [58] find that web users often disclose optional information in survey forms despite the fact that they have a choice. Acar et al. [2], Lin et al. [36] report that malicious web forms can exploit browsers' autofill features to steal credentials. Prior work also shows that trackers can leak sensitive PI in the web forms to third parties, even before submission [13, 66, 72]. In contrast to these studies that investigate "bad" actors, our work focuses on the ordinary and transparent usage of web forms by first parties to collect PI, and aims to extract privacy norms, instead of anomalies, from them. To the best of our knowledge, this work is the first paper that analyzes PI collection through web forms in general.

3 Web Forms Collection

In this section, we describe the collection of web form dataset that supports the measurement study. We develop a browser-based crawler to discover and download web forms. We run the crawler on over 10K top websites to build the web form dataset.

3.1 Web Form Crawler

The goal of the web form crawler is to discover and save web forms that potentially collect PI. As it is impossible to download the entire website, the main technical challenge is to efficiently discover web pages that contain forms.

We base the crawler on Playwright [44], a browser automation library. Playwright allows the crawler to programmatically access websites using a headless but real Google Chrome browser, and simulate actions (*e.g.,* clicks on elements) in the browser like a real user. It enables the crawler to interact with dynamic web pages that create web forms using JavaScript code instead of static HTML.

The web form crawler starts by visiting the homepages of websites (*i.e.,* `http://<domain>/` or `http://www.<domain>/`). As the crawler visits web pages, it saves any encountered web forms. It also determines potential next steps (*i.e.,* what pages to visit next) and adds them to the list of crawler tasks. The task list, which we refer to as *crawl frontier* [27], is implemented as a priority queue. The crawler assigns each task a priority according to the strategy that we will explain later. Tasks with higher priority values are tried first. As many web forms need to be triggered dynamically by clicking elements, the crawl frontier does not simply record URLs to visit. Instead, each crawl task is described as a starting URL plus a sequence of click actions on clickable elements (including HTML

[3] While it is possible that websites with forms also use web tracking, we consider this out of scope, and we focus on measuring the data explicitly collected by web forms.

<button>, <a> elements, and other elements that are bound with JavaScript onclick event handlers).

The main crawler loop is as follows. For each website, the crawl frontier is initialized with a task to access the homepage (and no click actions). In each run, the crawler gets the task with the highest priority from the crawl frontier and runs the following steps:

- *Navigation*: The crawler navigates to the starting URL and completes the sequence of click actions specified in the task to restore the full-page state of this crawl task.
- *Downloading web forms*: The crawler identifies any web forms (*i.e.,* HTML <form> elements) on the web page and stores information about each form locally on disk.
- *Defining next steps*: For each clickable element on the page, the crawler generates a new task as a potential next step. It assigns each task a priority based on the text on it. If the element is a hyperlink, the new task will simply set the linked URL as the starting URL. Otherwise, if the element is button-like, the new task will inherit the starting URL and click actions from the current task, and append a new click action to it.

The crawler repeats the steps above to try more and more links and buttons in the crawl frontier, potentially being redirected to new pages in the process, and save any discovered HTML forms. The crawler limits crawling to each website by skipping tasks that redirect to a different apex domain.

Priority Assignment. As it is impossible to visit every web page and try every clickable element, we set the crawler to stop after finishing 100 tasks for each website. The crawler thus needs to prioritize steps that likely lead to web forms. To find such steps, the crawler checks how similar the text on each clickable element is to a list of seed phrases. The seed phrases are 100 phrases that are manually curated and indicative of web forms, such as "Sign Up", "Contact Us", and "Subscribe". The crawler uses *all-MiniLM-L6-v2*, a lightweight sentence transformer model [62] to compute embedding vectors for both the seed phrases ($t_s \in Seeds$) and the text on elements (t). Each element is scored as the max cosine similarity from any seed phrases: $Score(t) = \max_{t_s \in Seeds} \text{CosSim}(\text{Embed}(t_s), \text{Embed}(t))$ where $\text{CosSim}(\mathbf{v_1}, \mathbf{v_2}) = \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{\|\mathbf{v_1}\| \|\mathbf{v_2}\|}$.

To avoid crawls of repetitive high-score texts that appear on different pages, the crawler: (1) adds a random number ε to the score to shuffle steps with similar scores, and (2) discounts the score with a factor that is exponential with the crawl depth (d), assuming that most useful web forms are not nested deeply. The final priority score is calculated as:[4] $Priority(t) = 0.9^d(Score(t) + 0.05\varepsilon)$, $\varepsilon \sim U(0, 1)$ Elements with higher priority values are tried first.

Web Form Discovery and Processing. The crawler recognizes any HTML <form> elements as web forms. This includes invisible <form> elements which are not yet triggered by the crawler. The crawler stores the entire HTML code of each web form locally. It also takes advantage of the ability to call Web APIs in the browser to parse the forms. It extracts form fields (<input> elements) and their labels (<label> elements). The label text is a useful feature for classifying PI types (see Section 4.2).

The crawler also handles some irregularity of HTML code. First, many web pages contain irregular web forms that are made of form fields (<input>, <select> and <textarea>) not enclosed by <form> elements. The crawler identifies these isolated field elements and heuristically groups elements that are in close proximity as additional web forms. Second, some web forms use other text containers that are not <label> elements to show field labels. If the crawler cannot detect the label element of a field, it extracts visible text appearing before the field as the label.

Ethical Considerations. (1) To avoid interfering with websites' operation, the crawler follows the good practice of limiting the rate per website (at most 10 tasks / minute) and stopping on errors (*e.g.,* permission denied due to crawler protection). (2) Although technically possible, the crawler *does not* try to fill in any forms or bypass authentication to discover subsequent content. Due to these (self-)restrictions, the crawler will not discover multi-page forms that appear only after the first page is filled in, or forms that require logging in (*e.g.,* account settings).

Table 1: Website category statistics.

Website Category	# Websites	
Technology	3,206	27.9%
- Technology	1,983	17.2%
- Information Technology	904	7.9%
Entertainment	3,041	26.4%
- News & Media *	992	8.6%
- Video Streaming *	548	4.8%
- Gaming *	407	3.5%
Business & Economy	2,211	19.2%
- Business	1,662	14.5%
- Economy & Finance *	509	4.4%
Education	1,519	13.2%
- Education	916	8.0%
- Educational Institutions *	538	4.7%
Society & Lifestyle *	1,098	9.5%
- Fashion	188	1.6%
- Clothing	179	1.6%
- Food & Drink	164	1.4%
Shopping & Auctions *	972	8.5%
- Ecommerce	733	6.4%
- Shopping	130	1.1%
Government & Politics *	921	8.0%
- Politics, Advocacy, and Government-Related	767	6.7%
- Government	154	1.3%
Internet Communication *	818	7.1%
- Personal Blogs	418	3.6%
- Information Security	176	1.5%
- Forums	130	1.1%
Health *	448	3.9%
Travel *	328	2.9%

Note: Due to a lack of space, we only show the most popular website categories.
* The marked categories, which make up 53.2% of the dataset, are selected for discussion in Section 5.1.

3.2 Web Form Dataset

We choose top websites from the Tranco list [56], a research-oriented domain ranking, as the subjects of our study.[5] The Tranco list ranks apex domains mainly based on the amount of traffic they receive. Some domains in the list are not meant to be accessed directly (*e.g.,* content delivery services). We write a Python script to probe HTTP services on each domain. We filter out domains that (1) do not serve HTTP, (2) serve HTTP but redirect to a different domain, or (3) serve non-English content on the homepage. We also skip websites identified as malware, adult theme or other questionable content by Cloudflare [17] to prevent our researchers from being exposed to disturbing content when analyzing the data. We crawled websites

[4] We tested different parameters on a small number of websites. The parameters are chosen empirically to improve the chance of finding more web forms.

[5] In this paper, a *website* refers to all the content served under an *apex domain*.

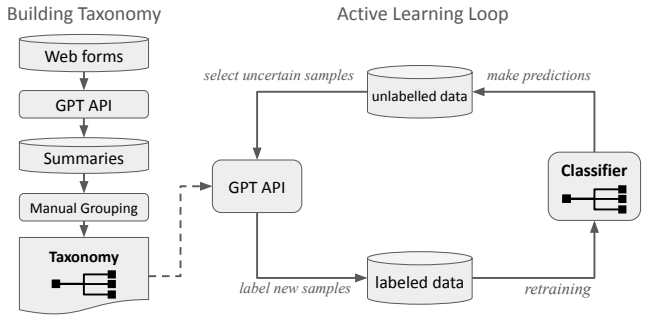


Figure 3: Overview of form type classification.

in the order listed in the Tranco list, on a server located in California, U.S., from Dec. 2023 through Jan. 2024. Excluding websites that the crawler did not finish exploring due to errors, we stopped the crawler after collecting approximately 10K valid websites. In total, the raw web form dataset comprises 938,324 HTML forms from 11,500 websites.[6] In Section 4.3, we describe dataset cleaning that leads to an annotated dataset of 293K forms to be further analyzed.

Website Categorization. We further categorize websites based on their topics. The website category is part of the context information that we need to describe the privacy norms (see Sections 1 and 5). We use the domain categorization from Cloudflare domain intelligence API [15], which has also been used in prior web measurement studies [21, 63]. Cloudflare categorizes websites into a two-level taxonomy, for example, Entertainment (level 1), News & Media and Gaming (both from level 2). Table 1 lists the major website categories in our dataset. Cloudflare may classify each website into zero to multiple categories. In our dataset, 11,263 (97.9%) websites are in at least one category.

4 Dataset Annotation

In this section, we describe how we process the raw web form dataset to (1) classify the form type (*i.e.,* the functionality of the web form), and (2) extract PI types requested in form fields. To that end, we build machine learning classifiers to efficiently process nearly a million web forms (and many more form fields).

LLM Assisted Classification. We considered using large language models (LLMs) to annotate the dataset, since modern LLMs provide a general solution for many NLP tasks, including text classification, without the need to retrain task-specific models [34, 94]. However, prompting LLMs to process almost 1 million HTML code snippets turns out to be unacceptably slow and costly despite the great accuracy. Instead, we use LLMs (specifically OpenAI's GPT-3.5 Turbo and GPT-4 Turbo models [4, 52]) to assist classification (for both form type and PI type classification) and to generate training samples (for form type classification) in a cost-efficient way.

4.1 Form Type Classification

We classify web forms into types according to their functionality. The input for classification is the HTML code of each form saved by the crawler. The functionality is inferred from information in the HTML code, including visible text and HTML attributes. Figure 3 shows an overview of our form type classification methodology.

Web Form Taxonomy. The first step is to build the taxonomy of output labels to use in classification. To our knowledge, there is no widely accepted taxonomy of web forms. Table 2 shows the taxonomy that we create for form type classification.

We derive this taxonomy in a data-driven way. We randomly sample 2,000 web forms[7] and prompt the GPT-4 Turbo model to summarize the functionality of each form in a short phrase (see Appendix B.1 for the prompt). We do not give any specific examples in the prompt to avoid biases. Example responses from GPT-4 are "newsletter signup", "gift purchase", "student loan application", *etc.* Then, we manually read through the responses and group similar functionality together as one label. We do not validate the correctness of the GPT-4 responses at this point, but we manually validate the trained classifier later. In this process, we tried different ways of grouping. Our objectives are that the labels: (1) cover the functionality of a majority of forms in our dataset; (2) have clear definitions; and (3) have proper granularity so that all labels are sufficiently represented in the dataset. For the granularity, we use regular expressions to match strings in the web forms to get a rough estimation of the numbers. If a label is too specific to have enough samples for training, we merge it into another label.

The taxonomy in Table 2 is designed to cover web forms in our dataset. As such, it does not include all possible web forms on the Internet. First, due to the limitation of our crawler (see Section 3.1), we cannot observe forms that require account login, *e.g.,* account preferences. Second, it does not include web form types that are not for collecting PI, *e.g.,* searching forms. We also note that, to strike a balance between granularity, clarity and coverage, these labels are not fine-grained and by no means accurately describe the diverse use cases of all existing web forms. The coverage of the taxonomy is illustrated in Table 4 in Section 4.3.

Table 2: Web form taxonomy for form type classification.

Label	Description
Account Registration	For creating new online user accounts.
Account Login	For users to log into existing accounts using their credentials.
Account Recovery	For retrieving or resetting forgotten account credentials.
Payment	For financial transactions, such as bill payments, online purchases or donations.
Financial Application	For applying to financial services like credit cards, loans, financial aid, insurance, investment accounts.
Role Application	For applying to positions such as employment, school admissions, or volunteer opportunities.
Subscription	For users to sign up for newsletters, mailing lists, or similar channels of periodic updates.
Reservation	For users to book services, schedule appointments, register for events, or similar.
Contact	For users to send private messages, inquiries, or feedback to the website owner.
Content Submission	For submitting user-generated content like comments, reviews, or ratings, intended to be published on the website.

Note: Also see Appendix B.2 for representative examples of each form type.

[6] Please see Appendix A for statistics on discarded and failed websites. Note that the crawler simply stores all the HTML forms without cleaning. There are many duplicated and irrelevant (*e.g.,* not collecting PI) samples in the raw dataset.

[7] The taxonomy is intended to cover the functionality of the majority of web forms in our dataset. In general, the more web forms sampled, the better coverage of the functionality of web forms. However, there are bottlenecks: human workload and budget (the GPT API cost). Also note that returns are diminishing: more samples cover increasingly rare form types that have few samples for training. For example, Table 2 shows that, with the 2,000 forms sampled, the rarest form type (Financial Application) was only found on 1.5% of websites.

Knowledge Distillation from LLM. The traditional way of text classification is to train a machine learning classifier from labeled training data. Without the dataset, a modern alternative is zero-shot classification using LLMs [34, 94]. We only need to input the task description, along with the HTML code to label, and the LLM will do the tedious reading task and give the result. However, labeling millions of HTML snippets would be costly and slow. To control the cost, we use OpenAI's GPT 3.5 Turbo model[8] to bootstrap a training dataset, and train a BERT-based text classifier from the GPT-labeled dataset. This technique is generally known as *knowledge distillation*, which distills the knowledge of a powerful but costly model (*teacher model*) into a smaller task-specific model (*student model*) [79].

Specifically, each time we provide the GPT with the HTML code of a web form, along with a prompt that asks it to choose a label in Table 2 that best describes the form. We also provide a few more options that are interpreted as a special Unknown label (*i.e.,* none of the above). First, we include two common form types that most likely do not collect PI in the prompt: (1) Search forms, for searching contents on websites, and (2) Setting forms, such as cookie consent dialogs, language settings, *etc.* Second, we prompt the GPT to answer "Unknown" if there is not enough information (*e.g.,* due to incomplete HTML). See Appendix B.1 for the full prompt.

In some cases, the functionality of a web form can be ambiguous and more than one category may apply. To characterize the ambiguity, we take advantage of the GPT's non-determinism to generate *soft labels*. We set the GPT's temperature parameter to 0.8 (more deterministic than the default 1.0) and get 10 outputs for each sample. The 10 outputs are then converted into soft labels that represent the probabilities of each label in Table 2. For example, if a form is used for both account registration and login, the GPT may output the former label 7 times and the latter 3 times, which is converted to soft labels $[0.7, 0.3, 0.0, ...]$. The Unknown label does not contribute to any of the label probabilities.

Form Type Classifier. As for the student model, we fine-tuned a MarkupLM model [35], a BERT variant, to learn from the GPT-labeled data. We favor MarkupLM because it is pretrained on HTML code and has a specialized text processor to encode HTML efficiently. Other general-purpose BERT models would have to truncate inputs more often because HTML markups waste a lot of tokens. A limitation of MarkupLM is that it does not process HTML attributes. In web forms, much information can be found in the attributes. To preserve the information, we improve MarkupLM's processor to extract selected HTML attributes (*e.g.,* `placeholder` of `<input>` elements, and text on buttons). The classifier is fine-tuned with the GPT-labeled probabilities as the learning objective. To accommodate soft labels, sigmoid activation is used after the classification head, meaning that each label is predicted independently.

Active Learning. Next, we also need a strategy to choose which samples the GPT should label. Random sampling is not efficient because our dataset is extremely unbalanced – in terms of not only label distribution (*e.g.,* Account Login forms are far more common than Role Application forms) but also other features (*e.g.,* some websites have more web forms than others). Imbalanced training data would harm classifier performance on less-represented labels and samples. This is a real problem because we are interested in each form type equally, not just most common ones. We use *active learning* to interactively expand the training data based on the current classifier performance [60, 79]. The right part in Figure 3 shows the active learning loop. At each round, we run the current classifier to label the dataset. We pick samples for which the classifier is less certain (according to a query strategy that we will explain later), and use the GPT to label them. Then, we retrain the classifier with the expanded training data and go to the next round.

We use a combination of uncertainty-based and random sampling as the active learning query strategy. First, uncertainty sampling means selecting samples for which the model is least certain. In our case, uncertainty means the predicted probability is close to 0.5 (for any label). Second, instead of directly picking the probability closest to 0.5, we add some randomness to the target probability and pick the probability closest to $0.5 + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 0.15^2)$. This prevents similar samples from being picked repeatedly.[9]

The above query strategy helps dataset balance because it picks up uncertain samples of all labels, and ideally less-trained labels tend to have more uncertain samples. However, we still find that the majority labels significantly outnumber the minority ones. This is a side effect of the soft label – the classifier is trained to output uncertain probabilities, which voided the assumption that less-trained labels have higher uncertainty. This is especially an issue in the initial rounds – minority labels are not predicted at all because of insufficient training. To workaround the issue, we enforce label balancing by giving the minority labels (*i.e.,* those that are not predicted) a preference.[10] That is, we only consider the probability of those labels when picking up uncertain samples.

Training & Validation. We train the form type classifier as shown in Figure 3. Specifically, we start by using the GPT to label about 500 samples. To increase the chance of covering the minority labels, we sample web forms that collect different sets of PI types (labeled by the PI type classifier; see Section 4.2).[11] In each round, the GPT-labeled data is split into 80% for training and 20% for per-epoch validation, and we train the new classifier for 10 epochs. After model training, we expanded the sample size by about 1,000 using the active learning strategy. We stopped after 5 rounds, at which point we determined that the classifier performance became stable. The GPT labeled about 5,400 samples in total.

After we finalize the classifier, we run it to label the entire dataset. We validate the precision by manually labeling a subset of the dataset. An author and a volunteer independently labeled 50 random samples for each predicted form type (500 samples in total, no overlap with training data), with a final discussion to resolve any disagreements. The classifier shows 85.6% macro average precision. The full validation results are provided in Table 3 (left columns). The

[8] We use *the GPT* to refer to OpenAI's GPT 3.5 Turbo model for the rest of the paper. The other LLM used in this work is GPT 4 Turbo, which is more powerful than GPT 3.5 Turbo, for taxonomy creation. We use GPT 3.5 Turbo to label web forms for its lower cost (10 times cheaper than GPT 4 Turbo).

[9] In our dataset, many web forms have the same or similar predicted probabilities of form type labels because (1) the use of bfloat16 format limits the numeric precision of classifier predictions; (2) there are many web forms that are different in HTML code but have the same embedding after MarkupLM's processor cleaned the code. The standard deviation of ε is chosen empirically to sufficiently avoid repetitive samples and not to significantly distort uncertainty-based sampling.

[10] Our objective is to make sure all labels have *enough* training samples. However, we do not need to enforce perfect balance. Once the classifier starts to predict all the labels, we stop using the workaround.

[11] Technically, we make a weighted random selection over all the web forms, with each form weighted by one over the number of forms that collect the same set of PI types.

Table 3: Manual validation of form type and PI classification.

Form Type	Precision	PI Type	Precision
Role Application	0.70	Email Address	0.98
Financial Application	0.72	Phone Number	1.00
Payment	0.90	Person Name	1.00
Reservation	0.84	Address	0.92
Contact	0.78	Coarse Location	0.98
Content Submission	0.92	Postal Code	1.00
Subscription	0.86	Age	0.90
Account Registration	0.94	Date of Birth	0.90
Account Login	0.98	Bank Account Num.	0.82
Account Recovery	0.92	Government ID	0.76
macro average	0.856	Tax ID	0.96
		Online Alias	0.88
		Ethnicity	0.98
		Gender	1.00
		Immigration Status	0.96
		Military Status	0.92
		macro average	0.935

precision varies with form types. Account Registration, Account Login, Account Recovery and Payment forms have high precision, presumably due to their clearly distinguishable functionality and textual features. For Role Application and Financial Application forms, we find that the classifier can confuse miscellaneous information gathering forms (*e.g.,* price quote forms for services) with them, resulting in lower precision. In addition, the functionality of some forms can be ambiguous. For example, we generally recognize lead generation forms on commercial websites as Contact forms, but similar forms are also used by some websites for collecting user feedback. In Appendix B.2, we provide representative examples of correctly classified forms of each type. In Appendix B.3, we show that input sequence length does not significantly affect precision.

We purposely avoid reporting recall, which measures how many samples of the entire dataset (*i.e.,* the real distribution) are correctly predicted. Our validation data is sampled from *positive* samples, on which calculating recalls would be misleading, because they do not represent the real distribution. Due to the extremely unbalanced label distribution (as discussed above and presented in Table 4 in Section 4.3), if we sampled the entire dataset, minority labels would have small support for effective validation. Furthermore, our dataset is already limited in coverage (due to restrictions explained in Section 3.1). Therefore, we conservatively avoid reporting recalls.

Cost Estimation. Knowledge distillation and active learning reduce both monetary and labor costs of the classification task. Here we give a rough cost comparison of our method with pure classifier and pure GPT-based methods. The traditional way of classifier training with manual dataset creation is labor intensive, namely manual labor is the main cost. Suppose we were to hire 10 people to do the GPT's task, *i.e.,* each person labeled 5,400 samples and that labeling each sample took 30 seconds. This would cost, at minimum, 450 human hours, which translates to about 3,260 USD if they were paid based on the current U.S. federal minimum wage [92]. For the fully GPT-based annotation, LLM inference is the main cost. We estimate that an average web form's HTML code, plus the prompt, is converted to about 2,740 tokens (after we clean up many useless text features). Our raw dataset contains 938K forms. Even after cleaning up (see Section 4.3), there are still about 293K web forms to label. With the current price of GPT-3.5 API (0.001 USD / 1K tokens), this would cost about 800 USD to classify the entire dataset.[12] In our case, we use the GPT to label only 5,400 training samples in total, which costs about 15 USD.

4.2 PI Type Classification

We train another classifier to identify the PI type requested in each web form field. We choose to manually label a small training dataset, for two reasons: (1) most PI types are sparse (*e.g.,* tax IDs are collected on few websites) – it would be more efficient to search for relevant samples to label; (2) PI types are usually clearly indicated in the text or HTML attributes, so manual labeling is easy if we display the right features (instead of HTML code) to the annotators.

List of PI Types. Similar to form type classification, we start by prompting the GPT-4 Turbo model to list all the PI types collected in about 4,000 randomly sampled web forms (see Appendix B.1 for the prompt), and manually build the list of PI types. To maximize the chance of capturing diverse PI types, the forms are sampled across different website categories. We consider PI types that are generally recognized as personal. We refer to CCPA Section 1798.140(v) (definition of "personal information") and 1798.140(ae) (definition of "sensitive personal information") for example PI types [74]. We also use regular expressions to estimate if a PI type has a significant presence to support training.

We examine the list of PI types and consider the following 16 labels for PI type classification: Address, Date of Birth, Email Address, Ethnicity, Gender, Tax ID (*i.e.,* social security number or equivalents in other countries), Government ID (*i.e.,* passport, driver's license number, voter ID or other national IDs), Coarse Location (city-level or coarser), Postal Code, Bank Account Number (including credit card number), Person Name, Phone Number, Online Alias (username or other online IDs used in specific websites), Age (including age groups), Immigration Status (including citizenship, nationality, and residency status) and Military Status (including veteran status).

This list is by no means exhaustive as we exclude many PI types that are rare or limited to few contexts (*e.g.,* meal preference, size of one's household). For model training purposes, we used three more labels for other common field types: Password, Business Info (*e.g.,* contacts of companies, which we do not consider as PI), and Fingerprints (hidden fields that contain machine-readable information beyond the scope of this study). These additional labels are equivalent to Unknown labels as in form type classification.

Feature Extraction and Labeling. To facilitate manual labeling and the use of pre-trained NLP models, we do not directly use HTML code as input features. The PI types are usually indicated by text and a few HTML attributes (*e.g.,* `name`, `placeholder`, `type` and `aria-` attributes) of the field elements and label elements, so we only extract these values and convert them to YAML format which ensures human readability. For example, the following is the featurized string of a field that requests date of birth information:

```
tagName: INPUT
label: DATE OF BIRTH
attributes:
- placeholder: MM/DD/YYYY
- id: dateOfBirth
```

This format is used both for training data labeling and as the classifier's input format. One of the authors manually labeled about

[12]We run the GPT to get 10 predictions for each form. However, the cost does not increase with the number of outputs requested, presumably due to internal caching.

2,200 web form fields as the training data, including about 80~100 positive samples for each of 16 PI types and additional samples of Unknown labels. We use Label Studio [87], an open-source dataset annotation platform, to facilitate manual labeling.

Training & Validation. We use SetFit to fine-tune a classifier based on *bge-small-en-v1.5* [96], a pre-trained sentence transformer model. SetFit is a few-shot learning framework that uses contrastive learning to learn the differences between labels effectively from a relatively small dataset like ours [89]. To evaluate classifier performance, we randomly select 50 positive samples for each PI type and manually validate the precision, following the same procedure as in Section 4.1. We obtain 93.5% macro average precision. The result is shown in Table 3 (right columns). For most PI types, the field labels are clear so the classifier is expected to learn the textual features well, resulting in over 90% precision. We find that false positives are often caused by ambiguous label text. For example, the label *ID number* is likely recognized as Government ID, but in some contexts the same term refers to other account identifiers.

4.3 Annotated Dataset

Dataset Cleaning. We run the form type and PI type classifiers on the 938K web forms. With the annotations ready, the dataset is cleaned and processed as follows. First, we discard about 27K web forms on non-English web pages (as determined by HTML `lang` attributes and lingua-py library [71]) because our NLP models only understand English. Second, about 587K forms do not collect any recognizable PI types. These are mostly search forms, cookie consent dialogs and hidden forms used only for programming purposes. Lastly, as we are only interested in web forms that collect PI, we discard 31K forms that do not require any personal identifiers (*i.e.,* Address, Email Address, Government ID, Bank Account Number, Person Name, Phone Number, Online Alias, and Tax ID). As per the CCPA [74], *personal information* is information that relates to an identified or identifiable individual. It is unclear whether an age verification switch, or gender selection dropdown in a product search form, can count. Note that the information collected in these discarded forms may still constitute PI when combined with PI collected elsewhere (notably, web tracking [13, 66, 72]). We conservatively choose precision over guesswork. After cleaning, there are 292,655 web forms from 11,500 websites in the final annotated dataset for analysis.

Dealing with Duplication. Websites often reuse components, resulting in duplicated web forms in our dataset.[13] However, two forms can be visually the same but different in HTML (*e.g.,* due to randomized tokens), or visually different but serving the same purpose (*e.g.,* two versions of login forms). Because of such ambiguity, and to be able to trace forms back to where they are located, we do not remove duplicates from the dataset. Duplication will not affect any analysis in the following sections, where we report the number of unique websites with specific contexts.

Annotated Dataset. The annotated dataset has 293K web forms, labeled with information about website categories (Section 3.2), form types (Section 4.1), and PI types (Section 4.2). Table 4 shows the occurrences of form type and PI type labels, in terms of *number of unique websites* where these occur.

[13]For reference, if we simply deduplicate the HTML code in the raw dataset, there will be 173,606 unique web forms.

Table 4: Form type and PI type statistics.

Form Type	#Websites		PI Type	#Websites	
Role App.	959	8.3%	Email Address	9,805	85.3%
Financial App.	175	1.5%	Phone Number	4,704	40.9%
Payment	1,096	9.5%	Person Name	7,804	67.9%
Reservation	413	3.6%	Address	2,244	19.5%
Contact	6,304	54.8%	Coarse Location	3,792	33.0%
Content Submission	831	7.2%	Postal Code	2,493	21.7%
Subscription	4,613	40.1%	Age	249	2.2%
Account Reg.	5,138	44.7%	Date of Birth	1,324	11.5%
Account Login	7,159	62.3%	Bank Account Num.	530	4.6%
Account Recovery	5,054	43.9%	Government ID	155	1.3%
Unknown	1,748	15.2%	Tax ID	170	1.5%
			Online Alias	4,244	36.9%
			Ethnicity	137	1.2%
			Gender	466	4.1%
			Immigration Status	280	2.4%
			Military Status	69	0.6%

Note: Please note that a website can have more than one web form and collect more than one PI type. Therefore, the percentages do not sum to 100%.

More formally, let $\mathbf{W}$ be the set of possible website categories (see Table 1), $\mathbf{F}$ be the set of possible form types (see Table 2, also including Unknown label), $\mathbf{T}$ be the set of possible PI types (see Section 4.2), x_i be the i-th web form in the dataset. The annotated dataset can be described as $\langle d(x_i), \mathbf{w}(x_i), f(x_i), \mathbf{t}(x_i)\rangle$, where $d(x_i)$ is (the apex domain of) the website, $\mathbf{w}(x_i) \subseteq \mathbf{W}$ is the set of website categories to which the website belong[14], $f(x_i) \in \mathbf{F}$ is the form type label, $\mathbf{t}(x_i) \subseteq \mathbf{T}$ is the set of PI types collected by x_i.

5 Web Form Analysis

In this section, we analyze patterns of PI collection revealed in the web form dataset. For each PI type, we seek to understand in what *contexts* (which we define as website category and form type) this PI type is typically collected. As we argue in Section 2.1, the common patterns[15] should reflect privacy norms, the common standards on appropriate PI collection.

We mainly compare the *collection rates* of a PI type across contexts. Formally, we define $N[t|w,f] = \|\{d(x_i)|w\in\mathbf{w}(x_i)\wedge f(x_i)=f\wedge t\in\mathbf{t}(x_i)\}\|$, *i.e.,* the number of *unique websites* in the website category w that use the form type f to collect the PI type t. And specially, we use p=$*$, w=$*$ or f=$*$ to match all labels in that field[16]. The collection rate of p in the context (w, f) is $P[t|w,f] = \frac{N[t|w,f]}{N[*|w,f]}$.

In this analysis, we compare the collection rate in a specific context, or website category (*i.e.,* $P[t|w,*]$), or form type (*i.e.,* $P[t|*,f]$) to the average collection rate over the entire dataset (*i.e.,* $P[t|*,*]$).

Results. We report our web form analysis results in the following figures. First, Figure 4 shows the collection rates of 4 common contact information types: Email Address, Phone Number, Person Name and Address. Second, other PI types appear less frequently (P<10%) in the dataset. To zoom in into contexts where these PI types are collected frequently, Figure 5 heatmaps show the collection rates by combinations of website categories and form types. Note that some contexts do not have enough samples to be statistically significant. We use Welch's t-test, with p threshold 0.05, to

[14]Recall that a website can be classified into multiple categories (see Section 3.2).

[15]We use the term *common pattern* to refer to contexts where a PI type is relatively frequently collected, *i.e.,* significantly more frequently than average.

[16]That is, $N[*|w,f]=\|\{d(x_i)|w\in\mathbf{w}(x_i)\wedge f(x_i)=f\}\|$, $N[t|*,f]=\|\{d(x_i)|f(x_i)=f\wedge t\in\mathbf{t}(x_i)\}\|$, and $N[t|w,*]=\|\{d(x_i)|w\in\mathbf{w}(x_i)\wedge t\in\mathbf{t}(x_i)\}\|$.

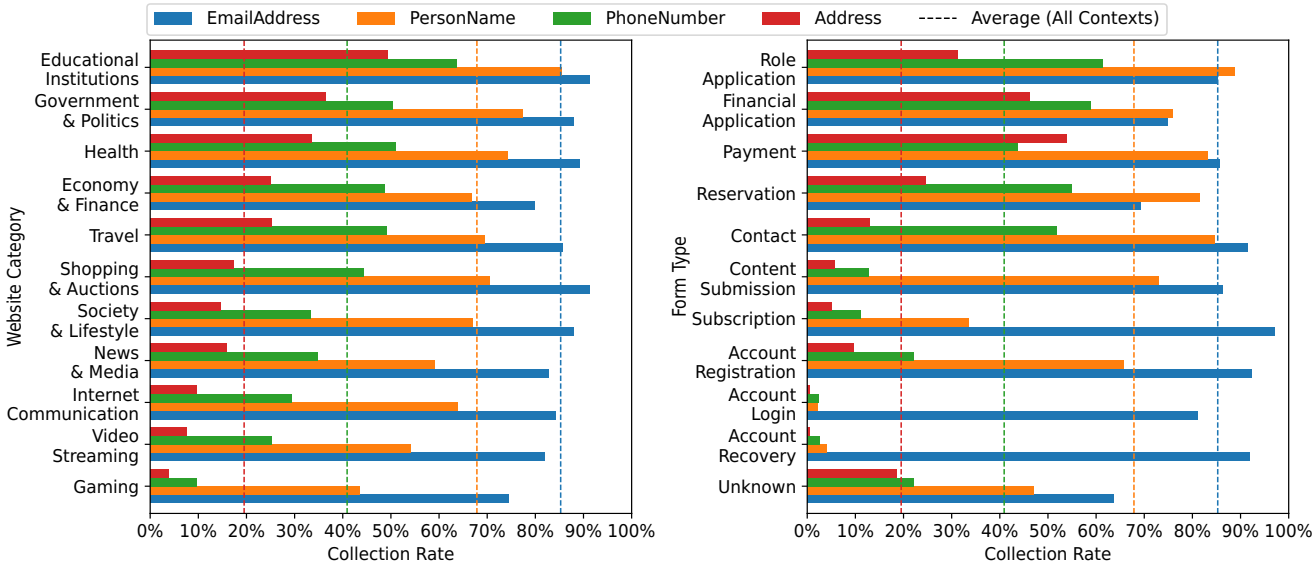


Figure 4: Collection rates of major PI types (contact information) by website category (left) and form type (right).

filter out insignificant results.[17] Due to a lack of space, and to facilitate readability, we only show columns and rows with significantly higher or lower collection rates than average. We also only choose 11 website categories among others to report (marked in Table 1).

5.1 Common Patterns and Privacy Norms

Next, we discuss common patterns of web form PI collection, and we show that they can be attributed to functional necessities, legal obligations or other norms, thus reflecting privacy norms. As we collect the dataset in the U.S. and consider only English websites, our discussion focuses on U.S. laws and regulations.

Norm 1. *Ubiquity of Email Addresses.* The most common PI type is Email Address, collected by over 85% of websites and ubiquitously used across all website categories and form types. While email address serves as a common contact method and identifier, some studies have shown that users perceive it as less sensitive than other identifiers [39, 53, 65]. The easiness of creating new email accounts, through either traditional email providers or emerging email masking services [22, 45], makes it possible, even for average users, to hide their real identities behind email addresses.

Norm 2. *Phone Numbers and Addresses in Specific Contexts.* In contrast, Figure 4 shows that the collection rates of Phone Number and Address vary significantly across website categories and form types. Both contact methods have a closer connection to users' real-world identity. Accordingly, websites with a direct link to real-world services (*e.g.,* education, government, health, finance, and travel) are more inclined to collect both PI types, reflecting the functional necessities, than online services (*e.g.,* video streaming, gaming).

Norm 3. *Card Number in Combination with Name and Address for Payments.* In Payment forms, card numbers (Figure 5c) are collected along with card holder names and billing addresses (Figure 4). The norm is tied to standard anti-fraud protocols of the payment card industry, including the PCI DSS standard [55] and address verification services [77] implemented by major card issuers.

Norm 4. *Age Verification*[18] *during Account Registration.* Date of Birth is collected by many website categories for account registration (Figure 5a). The norm is explained by COPPA requirements. For all *general audience websites*, the COPPA rules require websites to verify the age of users if they collect children's personal information, with some exceptions like one-time contact [81, 86]. This requirement is likely agnostic to website categories. Account registration by definition collects and stores users' PI and does not qualify for the one-time contact exception.

Norm 5. *Identity Verification in Government, Healthcare and Financial Contexts.* These services often require extensive PI types, including contact information (Figure 4), government-issued IDs (Figure 5d and 5e) and date of birth (Figure 5a), for identity verification. Specifically, healthcare providers are required to verify the identity of a person requesting health information under HIPAA rules [83]. Financial institutions are required to verify customer's identity using a combination of name, date of birth, address and government-issued ID numbers, which is known as the "Know Your Customer" procedure [23, 82].

Norm 6. *Personal Attributes in Education and Job Applications.* Gender, Immigration Status, Ethnicity, and Military Status (Figure 5f, 5g, 5h and 5i) are more likely collected in the Role Application forms and Educational Institutions contexts. This can be explained by legal requirements. Specifically, most employers in the U.S. are required to report employment data by ethnicity, gender and national origin [85]. And federal contractors are further required to report veteran status data [84]. The Higher Education Act applies similar requirements to educational institutions [46].

[17] The collection rate can be viewed as the mean of Bernoulli trials (collect, or not). We compare the collection rate in each context to *other* contexts (*i.e.,* dataset samples that do not match the current context). We also acknowledge that the use of t-test may lack mathematical rigour because our dataset is subject to labeling errors (see Table 3).

[18] Note that our study does not cover all *age-screening mechanisms* in COPPA. Many websites use standalone age verification forms that only ask for age but no other identifiers. These web forms are not included in the dataset (see Section 4.3).

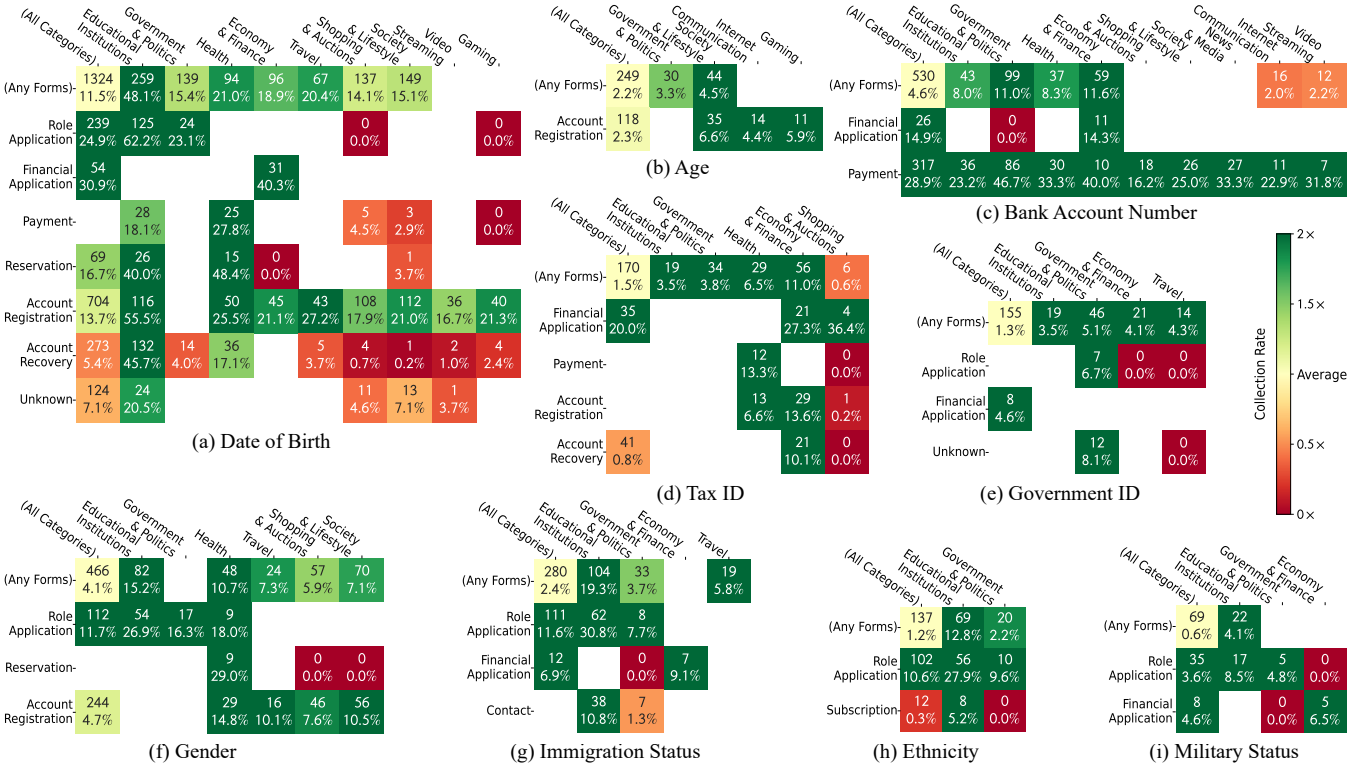


Figure 5: Collection rates of less frequent PI types. For each PI type, we show the average collection rate (top-left cell) and contexts where the collection rates are significantly different from the average. Each cell shows the collection rate (*i.e.,* $P[t|w, f]$, the percentage at the bottom) and the number of websites that collect the PI type in the corresponding context (*i.e.,* $N[t|w, f]$).

Norm 7. *Personal Attributes for Personalization and Marketing.* Some PI collection patterns are, presumably, linked with personalization and marketing purposes. Gender and Date of Birth are frequently collected by Society & Lifestyle, Shopping & Auctions and Travel categories, especially for account registration (Figure 5f). These website categories often provide personalized contents (including ads) based on personal attributes. We can confirm such usage from the privacy policies of some websites. For example, *facebook.com*'s privacy policy explicitly states that age and gender information can be used for "providing ads" [41]. *casio.com*, the shopping website, mentions that gender can be used for "marketing and promoting our products" and "conducting post-purchase surveys" [12].

Despite the frequent use, personalization and marketing are not purposes that users find generally worthy or acceptable for giving up their PI [7, 14]. This indicates that privacy norms found in the web forms do not always align with users' expectations. We hypothesize that web forms represent the websites' view, and not necessarily the users' view, of acceptable PI collection practices.

5.2 Uncommon Cases

Our analysis also provides a way to identify rare practices that deviate from the privacy norms. We scrutinize them to understand their implications, and whether they are appropriate or not. More specifically, we analyze the data underlying Figures 4 and 5 to look for contexts with relatively low collection rates. The first step is to determine "uncommon" cases. In this paper, we simply choose a collection rate threshold p_o to find outliers. Specifically, for web form x_i that collects PI type t, we consider such PI collection uncommon if $P[t|w, f_i] < p_o, \forall w \in \mathbf{w}(x_i)$. For example, with $p_o = 2.5\%$ (chosen empirically), there are 855 (7.4%) websites with uncommon PI collection under this criterion.[19] Once uncommon cases are identified, the second step is to interpret them, by manually looking into the corresponding forms and websites. Because our dataset is subject to errors, and our context definition may not fully describe each web form's functionality, we are cautious in interpreting these uncommon cases (see Section 7.2 for further discussion). We find that uncommon cases can indicate unnecessary PI collection or dark patterns, as shown in the representative examples below.

First, *macys.com*, a shopping website, asks for date of birth in its email list subscription form (also shown in Figure 1) [37]. Our dataset shows that $P[\text{Date of Birth}|*, \text{Subscription}] = 2.3\%$ ($\ll 11.5\%$ overall). The form states that "You must be 13 years or older", seemingly indicating the information is for age verification. However, when we actually test the form, we found that the field is optional. The most relevant explanation in its privacy policy indicates the information is used for "loyalty program" (marketing). The misleading language in the form suggests a potential dark pattern to trick users into giving unnecessary information.

Second, *metopera.org*, the website of an opera house, optionally asks for ethnicity for account registration [42]. Our dataset shows

[19]The parameter p_o, can be tuned by the user of our methodology, and depending on the dataset. Furthermore, there are other principled methodologies, beyond thresholds, for identifying outliers that can be incorporated.

that $P[\text{Ethnicity}|*, \text{Account Registration}] = 0.2\%$ ($\ll 1.2\%$ overall). Ironically, its privacy policy states "We ask that you not send us... information related to racial or ethnic origin..." [43], leaving the purpose of PI collection unclear.

Third, some shopping websites ask for users' gender and/or age, linking with the name, in the product review form (*i.e.,* Content Submission), such as *colgate.com* and *sleepnumber.com* [70]. Our dataset shows that $P[\text{Gender}|*, \text{Content Submission}] = 1.0\%$ ($\ll 4.1\%$ overall) and $P[\text{Age}|...] = 1.8\%$ ($< 2.2\%$ overall). Both privacy policies briefly mention public reviews and merely say that the submitted information may become public [18, 69].

Excessive collection violates the data minimization principle in the CCPA [73, 74] and other privacy laws (see Section 2.2). While the CCPA does not define what constitutes the minimal set of PI types under which circumstances, our investigation suggests that the observed privacy norms can be used as a baseline to quantify excessive collection. Also note that, in two of the above cases, the PI types in question are all optional. In our privacy norm analysis, we do not take optionality into account due to technical challenges in recognizing it. Separating optional from mandatory PI types when evaluating privacy norms is a useful direction for future work.

6 Privacy Policy Analysis

In this section, we turn our attention to the privacy policies that accompany web forms. In accordance with the notice requirements of privacy laws (see Section 2.2), one may expect that the privacy policy disclosures cover and, may even conservatively go beyond, what we can observe through web forms, thus helping the understanding of privacy norms. But is this the case? We extend our analysis by comparing privacy policy disclosures to observed privacy norms.

6.1 Privacy Policy Availability

We first process the HTML code of web forms and the pages where they locate to search for links (`<a>` elements) to privacy policies. We use the same text scoring method as our web form crawler (see Section 3.1) to fuzzily match link texts and target URLs against seed phrases (*e.g.,* "privacy policy", "privacy notice"). Note that we only consider websites that have at least one web form that collects PI[20] (see Section 4.3), leaving 10,143 websites for this analysis. Please see Appendix C for details on the detection of privacy policy links, including a discussion on the selection of seed phrases.

At the website level, we find privacy policy links on 94.2% (9,559) of websites, showing the wide adoption of privacy policies. Table 5 shows the number of domains without privacy policies by website category. Notably, Educational Institutions and Government & Politics, which are more inclined to collect more PI types, have the worst availability. While missing privacy policies can be a legal compliance issue for businesses, some privacy laws, such as the CCPA, only apply to for-profit entities, which may explain the low availability in these categories. The same reason may also apply to Internet Communication, which includes subcategories, such as Personal Blogs and Forums (see Table 1), likely non-profit.

[20]We do not exclude them in Section 5, in order to keep a consistent number of websites (11,500) and not to overstate the collection rates. However, if a website collect no PI (*e.g.,* personal blogs), it seems appropriate for it to omit privacy policies.

Table 5: Number of websites without privacy policies by website category.

Website Category	# w/o privacy policies	
Educational Institutions	53	(10.4% of 511)
Government & Politics	82	(10.0% of 820)
Internet Communication	56	(8.00% of 700)
Economy & Finance	23	(5.07% of 454)
Video Streaming	20	(4.37% of 458)
Gaming	12	(3.81% of 315)
News & Media	29	(3.48% of 834)
Travel	11	(3.75% of 293)
Health	13	(3.17% of 410)
Society & Lifestyle	23	(2.62% of 878)
Shopping & Auctions	18	(1.99% of 903)
All Categories	584	(5.76% of 10,143)

Table 6: Number of websites that include privacy policy links inside the form by form type.

Form Type	# w/ privacy policy link in the form	
Account Registration	2,309	(45.0% of 5,138)
Payment	401	(36.6% of 1,096)
Financial Application	59	(33.7% of 175)
Subscription	1,491	(32.3% of 4,613)
Role Application	299	(31.2% of 959)
Contact	1,783	(28.3% of 6,304)
Reservation	82	(19.9% of 413)
Content Submission	124	(14.9% of 831)
Account Login	772	(10.8% of 7,159)
Account Recovery	210	(4.16% of 5,054)

At the web form level, we further check if privacy policy links are provided *in context* with the forms. Technically, we check if the links are inside `<form>` elements (see Appendix C for details). This usually indicates a good practice that allows users to see the notice before submitting the form. Table 6 shows the statistics. We can see that Account Registration forms are more likely to provide the links close by. Account Login and Account Recovery forms, possibly assuming that the user has read the privacy policies, tend not to provide the links. Overall, fewer than half of web forms provide the privacy policy links in context. Many websites put the link in the footer or even on a different web page (usually the homepage). This raises questions on whether a website-level privacy policy is contextualized to explain the specific PI collection practices happening in the web forms.

6.2 Content Analysis

We look further into the text of privacy policies for disclosures of collected PI types. We use PoliGraph-er [19], a state-of-the-art NLP privacy policy analyzer, to extract what PI types are disclosed to be collected. As PoliGraph-er's PI types do not fully align with ours, we only consider 9 PI types that can be clearly mapped: Email Address, Phone Number, Person Name, Address, Date of Birth, Age, Tax ID, Gender, and Ethnicity for this analysis. We are able to download privacy policies for 9,013 websites and run PoliGraph-er on them.[21] To avoid false results due to failed HTML parsing, we remove privacy policies that do not disclose any of the 9 PI types,

[21]Privacy policy links detected on 546 websites are not accessible. Also note that, as we extract the privacy policy link from each web form and page, there can be multiple privacy policy links on one website. In this case, we combine all the versions by taking the disclosed PI types from all of them into account.

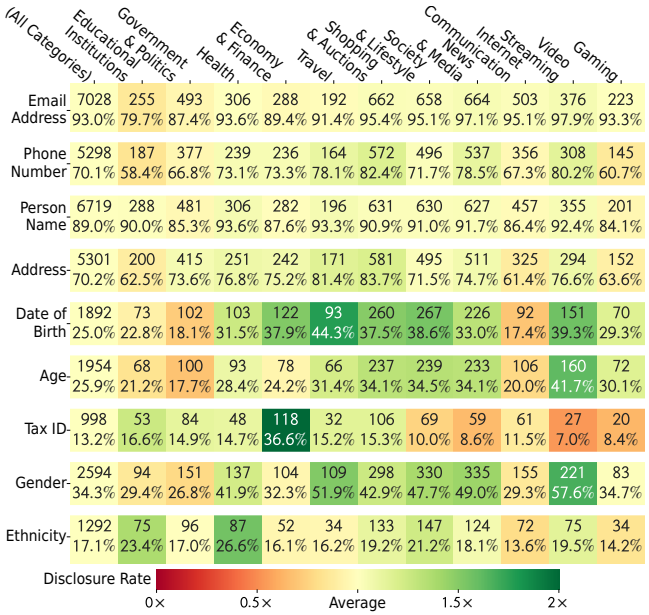


Figure 6: Disclosed PI collection by website category.

limiting our analysis to 7,553 websites. This likely underestimates the number of bad privacy policies (*e.g.,* due to vague language). However, as PoliGraph-er is optimized for precision rather than recall, we conservatively choose to focus on reliable results.

Patterns of PI Collection Disclosures. We first check the overall trend of disclosed PI collection, and compare it with privacy norms observed in Section 5.1. In Figure 6, we show, for each PI type, the percentage of websites that disclose the collection of it (*i.e., disclosure rate*) by website category. It can be seen as collection rates inferred from privacy policies. We note that the pattern of disclosures is different from our observations in Section 5.1.

First, Educational Institutions and Government & Policies, the categories observed to collect extensive PI types in web forms, appear to less often disclose many PI types (*e.g.,* date of birth, gender, ethnicity) than average. The contradiction agrees with the previous finding that these categories are doing worse in privacy policies. For example, *vermont.gov*, the state government's website, provides a privacy policy but only details automatic data collection [76]. Some web forms on this website require other PI types, including addresses and government-issued ID numbers. In this case, the collection of these non-disclosed PI types aligns with the observed privacy norms. Indeed, few would be concerned about providing governments with personal identifiers for registering a license [75].

Then, some other categories, notably Video Streaming and News & Media, disclose the collection of personal attributes (*e.g.,* Gender, Ethnicity) and most PI types more frequently than average, contradicting our observations that these online-based services are less PI-intensive. For example, *netflix.com* claims to collect ethnicity and gender information for "research surveys" [47]. *ft.com*, a news website, claims to collect this information for "diversity and inclusion goals" [24]. Both websites provide rather comprehensive privacy policies. While it is possible our crawler does not capture the corresponding web forms, the takeaway here is that privacy policy disclosures do not correlate well with observed privacy norms.

Table 7: Privacy policy statements *vs.* PI collection observed.

Data Type	C&P	C\P	$P_{Omitted}$	P\C	$P_{NotCollected}$	ϕ
Email Address	6,859	504	6.80%	169	2.40%	0.026
Phone Number	2,759	822	23.0%	2,539	47.9%	0.142
Person Name	5,287	609	10.3%	1,432	21.3%	0.043
Address	1,282	370	22.4%	4,019	75.8%	0.085
Date of Birth	392	580	59.7%	1,500	79.3%	0.134
Age	71	125	63.8%	1,883	96.4%	0.039
Tax ID	59	58	49.6%	939	94.1%	0.137
Gender	155	176	53.2%	2,439	94.0%	0.056
Ethnicity	29	71	71.0%	1,263	97.8%	0.037

Note: C = the set of websites that collect the PI type; P = the set of websites that disclose the collection of the PI type in their privacy policies; $P_{Omitted} = \|C\backslash P\|/\|C\|$; $P_{NotCollected} = \|P\backslash C\|/\|P\|$.

Practices vs. Disclosures. To further investigate the gap, we compare the disclosures to collected PI types on a per-website basis. For each PI type, we consider two sets of websites: (1) C are websites that collect the PI type, as observed in our web form dataset; (2) P are websites that disclose to collect the PI type, as PoliGraph-er determines based on privacy policies. We compare the two and Table 7 summarizes the results.

We distinguish three cases *w.r.t.* consistency between the two. (1) $C\&P$ is the ideal case, in which the PI type is both collected and disclosed. (2) $C\backslash P$ means that the collected PI type is not disclosed, which may indicate privacy violations. Most prior work focuses on the previous two cases and shows the prevalence of omitted (undisclosed) PI collection [6, 19, 88, 97]. In our case, the high rate of omissions ($P_{Omitted}$) for many PI types confirms that. (3) However, we are also interested in a third case: $P\backslash C$ means that the PI type is disclosed to be collected, but we do not observe it in our dataset. The high numbers of both $C\backslash P$ and $P\backslash C$ for many PI types raise the question: *Are privacy policy disclosures really associated with actual PI collection practices?* In Table 7, we use ϕ-coefficients as a statistic to measure the association strength between the *observed* PI collection and disclosures. Interestingly, the association appears to be weak (< 0.20) for all PI types.

The disconnect is evident in many privacy policies that claim to collect a lot of PI types. For example, *swarovski.com*, a jewelry brand and a website under Society & Lifestyle category, literally claims to have collected "Identifiers such as... social security number; driver's license number, passport number..." and the information may even have been disclosed to third parties [78]. In fact, this privacy policy simply copies the definition of identifiers from the CCPA. *redbox.com*, a video rental company, provides a seemingly detailed privacy policy that explains the purposes for collection and third-party sharing for each PI category [61]. A closer look reveals another case of blanket disclosures – supposedly collecting all PI types, each of which can be used for "other purposes as disclosed" and shared with "other service providers". On a positive note, it indicates that some PI types are not shared for targeted advertising. Note that, due to the limited coverage of our web form dataset (see Section 3.1) and other possible sources of PI collection than web forms, we cannot conclusively verify if these additional PI types ($P\backslash C$ cases) are never collected in the wild. Nevertheless, "overly inclusive and broad" privacy policies have been a common practice for companies to avoid litigation [33].

To conclude, our privacy policy analysis hints the gap between (i) the disclosures made in privacy policies and (ii) the observed privacy norms. This gap is due to both omissions and potentially irrelevant (blanket) information disclosed. This challenges the very notion that privacy policies help the understanding of websites' PI collection practices.

7 Discussion

7.1 Summary

In this paper, we propose a novel approach to understanding privacy norms through web forms. We conduct a large-scale measurement study on nearly 293K web forms from 11,500 websites, using a combination of web crawling and NLP classifiers. Our analysis of the dataset reveals common PI collection patterns that reflect privacy norms, which are influenced by functionality, legal obligations, and other reasons. We also show that deviations from these norms may indicate excessive data collection. In addition, our analysis of privacy policies shows a disconnect between the observed norms and the disclosed PI collection practices, thus questioning the role of privacy policies in understanding privacy norms.

7.2 Limitations

We acknowledge the following limitations of our study.

Limited and Skewed Coverage. It is impossible to access every web form on a website. As discussed in Section 3.1, the crawler does not submit forms or bypass authentication to discover new forms. Thus, our web form dataset only represents a lower bound, and our analysis may underestimate the collection rates, which likely affects different PI types unevenly. For example, in multi-page account registration forms, detailed personal attributes are more likely to be requested after basic contact information is verified, and our crawler cannot discover PI types beyond contact information in this case. To mitigate the issue, we focused on comparing the collection rates of *each PI type* across contexts in Section 5.

Moreover, as discussed in Section 3.2, some websites are excluded for various reasons. Particularly, we collected data in the U.S. and excluded non-English websites, and our discussion of privacy norms is U.S.-centric. As explained in Section 2.2, laws and regulations impact privacy norms, so do their regional differences. For example, in contrast to the CCPA [73, 74], which allows opt-out consent for non-sensitive PI, the GDPR [59] has stricter consent requirements, mandating opt-in consent, when the consent is the legal basis for processing [32]. Studies have shown regional differences in privacy policy writing [28] and users' privacy decisions [11], influenced by legislation and cultural factors. This being said, our methodology for extracting privacy norms from measurements can be applied beyond English and U.S. websites.

Context Definition. Recall that we defined the context of PI collection in this study as the combination of website category and form type (functionality). Due to limitations in categorization and label granularity, the labels may not align with intuition. For website categories, Cloudflare's categorization is based on content topics [15, 16], which do not always indicate the types of services. For example, many financial institutions fall in the Economy & Finance category, but the category also includes financial news websites (which are also in the News & Media category). For form types, our taxonomy is coarse and oversimplifies the variety of actual form usages. It also does not directly map to the purposes of PI collection as outlined in privacy policies and laws. For example, Contact forms cover many lead generation forms (*i.e.,* marketing purposes) and general customer service forms (*i.e.,* functional purposes).

Statistical Analyses *vs.* Automated Auditing. We have focused on statistical analyses to reveal common trends. In Section 5.2, we investigate whether PI collection that is uncommon compared to the norms can indicate excessive collection. However, due to the aforementioned limitations, our statistical analysis of privacy norms falls short in the granularity of context definitions. Our methodology is not meant to be an *automated auditing tool* that reports privacy issues of individual websites. Therefore, we have been purposely cautious in interpreting uncommon cases and we opted for manual inspection of example cases. A similar limitation applies to the privacy policy analysis in Section 6.2 – we show the statistical evidence that the observed PI collection misaligns with privacy policies, but we cannot verify if the (seemingly) overly disclosed PI types are never collected by individual websites.

NLP Methodology. Our study extensively relies on machine learning models for NLP tasks, which are subject to errors. For example, despite our effort to improve the form type classifier, it still has low precision for some labels (see Table 3). We also use PoliGrapher [19] for privacy policy analysis. The tool has a limited recall and may miss some privacy policy disclosures.

7.3 Implications & Future Work

We envision the following applications and future work.

Privacy Risk Assessment. Our analysis of privacy norms can be used as a baseline to quantify the risk of PI collection. Section 5.2 shows that uncommon cases, which do not align with privacy norms, can indicate excessive PI collection. A user-facing tool (*e.g.,* a web extension) can be implemented based on our infrastructure and analysis to warn users about unusual PI collection, based on form types and website categories.

User Perceptions. Web forms provide an alternative way to discover privacy norms. It does not replace vignette surveys that directly measure user perceptions. An interesting research question would be how much the privacy norms observed through web forms align or misalign with users' expectations. Our high-level intuition is that web forms represent more of websites', rather than user surveys', standpoint *w.r.t.* the norms. We cannot refer to prior studies [40, 68] to answer the question yet because their surveyed contexts do not fit well with the web form contexts.

Deep Dive into Specific Contexts. As a reflection on the limitations of our context definition, it would be interesting to perform a deeper analysis of specific contexts. For example, COPPA imposes strong restrictions on collecting PI from children, thus we expect that users' age would impact how web forms collect PI. For children, some functions might be removed (*e.g.,* payment forms) or require a restricted set of information (*e.g.,* in user surveys). This study would require extensions to our infrastructure, including signing in with different user profiles and collecting web forms on one website for an extended period of time to observe the differences.

To facilitate future extensions and applications, we have released the source code [90] and the datasets [91] associated with this study.

Acknowledgments

This work was supported in part by the National Science Foundation under award numbers 1956393 and 1900654, and a gift from the Noyce Initiative. We would like to thank @NyaMisty for the help with the collection of domain categorization data, and Jingning Zhang for her help with the manual validation of classification results. We appreciate the anonymous PoPETs reviewers for their insightful feedback that helped improve the paper.

The authors used AI-based tools, including OpenAI ChatGPT [50] and Grammarly [25], to correct typos, grammatical errors, and awkward phrasing throughout the paper.

References

[1] Noura Abdi, Xiao Zhan, Kopo M. Ramokapane, and Jose Such. 2021. Privacy Norms for Smart Home Personal Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery.

[2] Gunes Acar, Steven Englehardt, and Arvind Narayanan. 2020. No boundaries: data exfiltration by third parties embedded on web pages. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, Vol. 2020. De Gruyter. Issue 4.

[3] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, Frank Piessens, and Bart Preneel. 2013. FPDetective: dusting the web for fingerprinters. In *Proceedings of the 2013 ACM SIGSAC conference on Computer and communications security (CCS '13)*. Association for Computing Machinery.

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[5] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association.

[6] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions Speak Louder than Words: Entity-Sensitive Privacy Policy and Data Flow Analysis with PoliCheck. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association.

[7] Noah Apthorpe, Yan Shvartzshnaider, Arunesh Mathur, Dillon Reisman, and Nick Feamster. 2018. Discovering smart home internet of things privacy norms using contextual integrity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–23.

[8] Noah Apthorpe, Sarah Varghese, and Nick Feamster. 2019. Evaluating the Contextual Integrity of Privacy Regulation: Parents' IoT Toy Privacy Norms Versus COPPA. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association.

[9] Solon Barocas and Helen Nissenbaum. 2009. On Notice: The Trouble with Notice and Consent. In *Proceedings of the Engaging Data Forum: The First International Forum on the Application and Management of Personal Electronic Information*.

[10] Duc Bui, Yuan Yao, Kang G. Shin, Jong-Min Choi, and Junbum Shin. 2021. Consistency Analysis of Data-Usage Purposes in Mobile Apps. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21)*. Association for Computing Machinery.

[11] Weicheng Cao, Chunqiu Xia, Sai Teja Peddinti, David Lie, Nina Taft, and Lisa M. Austin. 2021. A Large Scale Study of User Behavior, Expectations and Engagement with Android Permissions. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association.

[12] CASIO AMERICA, INC. 2023. Privacy Policy (Internet Archive snapshot)). https://web.archive.org/web/20231031111205/https://www.casio.com/us/privacy/.

[13] Manolis Chatzimpyrros, Konstantinos Solomos, and Sotiris Ioannidis. 2019. You Shall Not Register! Detecting Privacy Leaks Across Registration Forms. In *International Workshop on Information and Operational Technology Security Systems*. Springer.

[14] CI&T. 2023. The Art and Science of Retail Personalization. https://ciandt.com/us/en-us/the-art-and-science-retail-personalization-ciandt.

[15] Cloudflare, Inc. 2024. Cloudflare API Documentation – Get Domain Details. https://developers.cloudflare.com/api/operations/domain-intelligence-get-domain-details.

[16] Cloudflare, Inc. 2024. Cloudflare Radar docs – Glossary > Content Categories. https://developers.cloudflare.com/radar/glossary/#content-categories.

[17] Cloudflare, Inc. 2024. Introducing 1.1.1.1 for Families. https://blog.cloudflare.com/introducing-1-1-1-1-for-families.

[18] Colgate-Palmolive Company. 2024. Colgate-Palmolive Company Privacy Policy (Internet Archive snapshot). https://web.archive.org/web/20240610050345/https://www.colgatepalmolive.com/en-us/legal-privacy-policy.

[19] Hao Cui, Rahmadi Trimananda, Athina Markopoulou, and Scott Jordan. 2023. PoliGraph: Automated Privacy Policy Analysis using Knowledge Graphs. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association.

[20] 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 (CCS '16)*. Association for Computing Machinery.

[21] Aurore Fass, Caitlin Sadowski, Emma Thomas, Jon Azose, Kimberly Ruth, Mark Pearson, and Zakir Durumeric. 2022. A World Wide View of Browsing the World Wide Web. In *Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22)*. Association for Computing Machinery.

[22] Fastmail Pty Ltd. 2024. Masked Email - Fastmail. https://www.fastmail.help/hc/en-us/articles/4406536368911-Masked-Email.

[23] Financial Industry Regulatory Authority (FINRA). 2024. 2090. Know Your Customer. https://www.finra.org/rules-guidance/rulebooks/finra-rules/2090.

[24] Financial Times Ltd. 2024. Privacy policy (Internet Archive snapshot). https://web.archive.org/web/20240222234107/https://help.ft.com/legal-privacy/privacy-policy/.

[25] Grammarly, Inc. 2024. Grammarly. https://www.grammarly.com/.

[26] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *27th USENIX Security Symposium (USENIX Security 18)*. USENIX Association.

[27] Allan Heydon and Marc Najork. 1999. Mercator: A scalable, extensible web crawler. *World Wide Web* 2, 4 (1999), 219–229.

[28] Henry Hosseini, Christine Utz, Martin Degeling, and Thomas Hupperich. 2024. A Bilingual Longitudinal Analysis of Privacy Policies Measuring the Impacts of the GDPR and the CCPA/CPRA. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, Vol. 2024. Issue 2.

[29] Roberto Hoyle, Luke Stark, Qatrunnada Ismail, David Crandall, Apu Kapadia, and Denise Anthony. 2020. Privacy norms and preferences for photos posted online. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 4 (2020), 1–27.

[30] Umar Iqbal, Pouneh Nikkhah Bahrami, Rahmadi Trimananda, Hao Cui, Alexander Gamero-Garrido, Daniel J. Dubois, David Choffnes, Athina Markopoulou, Franziska Roesner, and Zubair Shafiq. 2023. Tracking, Profiling, and Ad Targeting in the Alexa Echo Smart Speaker Ecosystem. In *Proceedings of the 2023 ACM on Internet Measurement Conference (IMC '23)*. Association for Computing Machinery.

[31] Umar Iqbal, Steven Englehardt, and Zubair Shafiq. 2021. Fingerprinting the fingerprinters: Learning to detect browser fingerprinting behaviors. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE.

[32] Scott Jordan. 2022. Strengths and Weaknesses of Notice and Consent Requirements under the GDPR, the CCPA/CPRA, and the FCC Broadband Privacy Order. *Cardozo Arts & Entertainment Law Journal* 40, 1 (2022), 113.

[33] Thorin Klosowski. 2023. Here's What You're Actually Agreeing To When You Accept a Privacy Policy. https://www.nytimes.com/wirecutter/blog/what-are-privacy-policies/.

[34] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, Vol. 35. 22199–22213.

[35] Junlong Li, Yiheng Xu, Lei Cui, and Furu Wei. 2021. Markuplm: Pre-training of text and markup language for visually-rich document understanding. *arXiv preprint arXiv:2110.08518* (2021).

[36] Xu Lin, Panagiotis Ilia, and Jason Polakis. 2020. Fill in the Blanks: Empirical Analysis of the Privacy Threats of Browser Form Autofill. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS '20)*. Association for Computing Machinery.

[37] Macy's. 2024. Macy's – Email list subscription (Internet Archive snapshot). https://web.archive.org/web/20240120213854/https://emails.macys.com/pub/sf/FormLink?_ri_=X0Gzc2X=AQjkPkSTUQG3bKdtB7O7SqopgMzdy4Ce1kN97mv2qzazc4Nh5hDhqmOVXyjLNpLOfhKLX=VXMtX=AQjkPkSTUQGzcsJ9i2wJSyIEqXDBauoOaPzcP7XvHfn7eMB7zc6mC8ePb&_ei_=ENDNv86fYm3RMqM7NyyF_i2y4PhAxC229igXVjQVRiJ6q1OMfKcgDM.&_di_=q86jqhufo0l1kgdgdbioo1hcqok66ljgcqn1s0bcgkb7takm7ncg.

[38] Sunil Manandhar, Kaushal Kafle, Benjamin Andow, Kapil Singh, and Adwait Nadkarni. 2022. Smart Home Privacy Policies Demystified: A Study of Availability, Content, and Coverage. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association.

[39] Ereni Markos, George R Milne, and James W Peltier. 2017. Information sensitivity and willingness to provide continua: a comparative privacy study of the United States and Brazil. *Journal of Public Policy & Marketing* 36, 1 (2017), 79–96.

[40] Kirsten Martin and Helen Nissenbaum. 2016. Measuring privacy: An empirical test using context to expose confounding variables. *Columbia Science and Technology Law Review, Forthcoming* 18 (2016), 176.

[41] Meta Platforms, Inc. 2023. Meta Privacy Policy (archive.today snapshot). http://archive.today/2023.12.07-112023/https://www.facebook.com/privacy/policy/.

[42] Metropolitan Opera. 2024. Metropolitan Opera – Register (Internet Archive snapshot). https://web.archive.org/web/20240111011311/https://www.metopera.org/account/register/.

[43] Metropolitan Opera. 2024. Privacy Policy (Internet Archive snapshot). https://web.archive.org/web/20240117111106/https://www.metopera.org/user-information/privacy-policy/.

[44] Microsoft. 2022. Fast and reliable end-to-end testing for modern web apps – Playwright. https://playwright.dev/.

[45] Mozilla. 2024. Firefox Relay. https://relay.firefox.com/.

[46] National Center for Education Statistics. 2024. Statutory Requirements for Reporting IPEDS Data. https://surveys.nces.ed.gov/ipeds/public/statutory-requirement.

[47] Netflix. 2024. Netflix Privacy Statement (Internet Archive snapshot). https://web.archive.org/web/20240225172032/https://help.netflix.com/legal/privacy.

[48] Helen Nissenbaum. 2009. *Privacy in Context - Technology, Policy, and the Integrity of Social Life*. Stanford University Press, Redwood City, CA, USA.

[49] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, Serge Egelman, et al. 2019. On the ridiculousness of notice and consent: Contradictions in app privacy policies. In *Workshop on Technology and Consumer Protection (ConPro 2019), in conjunction with the 39th IEEE Symposium on Security and Privacy*. IEEE.

[50] OpenAI. 2024. ChatGPT. https://openai.com/chatgpt/.

[51] OpenAI. 2024. JSON mode. https://platform.openai.com/docs/guides/text-generation/json-mode.

[52] OpenAI. 2024. Models - OpenAI API. https://platform.openai.com/docs/models.

[53] Martin Ortlieb and Ryan Garner. 2016. Sensitivity of personal data items in different online contexts. *it - Information Technology* 58, 5 (2016), 217–228.

[54] Claire Park. 2020. How "Notice and Consent" Fails to Protect Our Privacy. https://www.newamerica.org/oti/blog/how-notice-and-consent-fails-to-protect-our-privacy/.

[55] PCI Security Standards Council, LLC. 2018. Payment Card Industry (PCI) Data Security Standard – Requirements and Security Assessment Procedures (Version 3.2.1). https://www.pcisecuritystandards.org/documents/PCI_DSS_v3-2-1.pdf.

[56] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Network and Distributed Systems Security (NDSS) Symposium 2019*.

[57] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2023. Information on the Tranco list with ID 82NJV. https://tranco-list.eu/list/82NJV/full

[58] Sören Preibusch, Kat Krol, and Alastair R Beresford. 2013. The privacy economics of voluntary over-disclosure in web forms. *The Economics of Information Security and Privacy* (2013), 183–209.

[59] Publications Office of the European Union. 2016. General Data Protection Regulation (GDPR). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504.

[60] Wenjun Qiu, David Lie, and Lisa Austin. 2023. Calpric: Inclusive and Fine-grain Labeling of Privacy Policies with Crowdsourcing and Active Learning. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association.

[61] Redbox Entertainment Inc. 2024. Redbox Privacy Policy (archive.today snapshot)). http://archive.today/2024.02.27-074600/https://www.redbox.com/privacy.

[62] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[63] Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric. 2022. Toppling top lists: evaluating the accuracy of popular website lists. In *Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22)*. Association for Computing Machinery.

[64] Nikita Samarin, Shayna Kothari, Zaina Siyed, Oscar Bjorkman, Reena Yuan, Primal Wijesekera, Noura Alomar, Jordan Fischer, Chris Hoofnagle, and Serge Egelman. 2023. Lessons in VCR Repair: Compliance of Android App Developers with the California Consumer Privacy Act (CCPA). In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, Vol. 2023. Issue 3.

[65] Eva-Maria Schomakers, Chantal Lidynia, Dirk Müllmann, and Martina Ziefle. 2019. Internet users' perceptions of information sensitivity – insights from Germany. *International Journal of Information Management* 46 (2019), 142–150.

[66] Asuman Senol, Gunes Acar, Mathias Humbert, and Frederik Zuiderveen Borgesius. 2022. Leaky Forms: A Study of Email and Password Exfiltration Before Form Submission. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association.

[67] Yan Shvartzshnaider, Noah Apthorpe, Nick Feamster, and Helen Nissenbaum. 2019. Going against the (appropriate) flow: A contextual integrity approach to privacy policy analysis. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7.

[68] Yan Shvartzshnaider, Schrasing Tong, Thomas Wies, Paula Kift, Helen Nissenbaum, Lakshminarayanan Subramanian, and Prateek Mittal. 2016. Learning privacy expectations by crowdsourcing contextual informational norms. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4.

[69] Sleep Number Corporation. 2024. Sleep Number – Legal Privacy Policy (Internet Archive snapshot). https://web.archive.org/web/20240224060445/https://www.sleepnumber.com/pages/legal-privacy-policy.

[70] Sleep Number Corporation. 2024. Sleep Number – Write Your Review (Internet Archive snapshot). https://web.archive.org/web/20240226221915/https://www.sleepnumber.com/reviews/pse-special-edition/write.

[71] Peter M. Stahl. 2023. pemistahl/lingua-py: The most accurate natural language detection library for Python, suitable for short text and mixed-language text. https://github.com/pemistahl/lingua-py.

[72] Oleksii Starov, Phillipa Gill, and Nick Nikiforakis. 2016. Are you sure you want to contact us? quantifying the leakage of pii via website contact forms. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, Vol. 2016. De Gruyter. Issue 1.

[73] State of California Department of Justice. 2018. CCPA Regulations. https://oag.ca.gov/privacy/ccpa/regs.

[74] State of California Department of Justice. 2020. California Consumer Privacy Act of 2018 (amended by the California Privacy Rights Act of 2020). https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?lawCode=CIV&division=3.&title=1.81.5.&part=4..

[75] State of Vermont. 2023. Educator Registration (Internet Archive snapshot). https://web.archive.org/web/20230921114221/https://alis.edlicensing.vermont.gov/Registration.aspx.

[76] State of Vermont. 2024. Privacy & Security Policy (Internet Archive snapshot). https://web.archive.org/web/20231209224635/https://www.vermont.gov/policies/privacy.

[77] Stripe, Inc. 2023. What is address verification service (AVS)? What businesses need to know. https://stripe.com/resources/more/what-is-address-verification-service.

[78] Swarovski. 2024. SWAROVSKI DATA PRIVACY POLICY (Internet Archive snapshot). https://web.archive.org/web/20240125023818/https://www.swarovski.com/en-US/s-dataprotection/Privacy-Policy/.

[79] Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large Language Models for Data Annotation: A Survey. *arXiv preprint arXiv:2402.13446* (2024).

[80] Welderufael B. Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. PrivacyGuide: Towards an Implementation of the EU GDPR on Internet Privacy Policy Evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics (IWSPA '18)*. Association for Computing Machinery.

[81] The Office of the Federal Register. 2017. 16 CFR Part 312 – Children's Online Privacy Protection Rule. https://www.ecfr.gov/current/title-16/chapter-I/subchapter-C/part-312.

[82] The Office of the Federal Register. 2017. 31 CFR 1020.220 – Customer identification program requirements for banks. https://www.ecfr.gov/current/title-31/section-1020.220.

[83] The Office of the Federal Register. 2017. 45 CFR 164.514 – Other requirements relating to uses and disclosures of protected health information. https://www.ecfr.gov/current/title-45/section-164.514.

[84] The Office of the Federal Register. 2023. 41 CFR Part 60-300 – Affirmative Action and Nondiscrimination Obligations of Federal Contractors and Subcontractors Regarding Disabled Veterans, Recently Separated Veterans, Active Duty Wartime or Campaign Badge Veterans, and Armed Forces Service Medal Veterans. https://www.ecfr.gov/current/title-29/part-1602.

[85] The Office of the Federal Register. 2024. 29 CFR Part 1602 – Recordkeeping and Reporting Requirements Under Title VII, the Ada, Gina, and the Pwfa. https://www.ecfr.gov/current/title-29/part-1602.

[86] The Office of the Law Revision Counsel of the United States House of Representatives. 2024. 15 USC Ch. 91: Children's Online Privacy Protection. https://uscode.house.gov/view.xhtml?path=/prelim@title15/chapter91&edition=prelim.

[87] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. https://github.com/heartexlabs/label-studio

[88] Rahmadi Trimananda, Hieu Le, Hao Cui, Janice Tran Ho, Anastasia Shuba, and Athina Markopoulou. 2022. OVRseen: Auditing Network Traffic and Privacy Policies in Oculus VR. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association.

[89] Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient Few-Shot Learning Without Prompts. *arXiv preprint arXiv:2209.11055* (2022).

[90] UCI Networking Group. 2024. GitHub: UCI-Networking-Group/webform (source code of this paper). https://github.com/UCI-Networking-Group/webform.

[91] UCI Networking Group. 2024. Web Form Artifacts (datasets and other artifacts of this paper). https://athinagroup.eng.uci.edu/projects/auditing-and-policy-analysis/webform-artifacts/.

[92] U.S. Department of Labor. 2024. State Minimum Wage Laws (Internet Archive snapshot). https://web.archive.org/web/20240229030752/https://www.dol.gov/agencies/whd/minimum-wage/state.
[93] Janus Varmarken, Jad Al Aaraj, Rahmadi Trimananda, and Athina Markopoulou. 2022. FingerprinTV: Fingerprinting Smart TV Apps. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, Vol. 2022. Sciendo. Issue 3.
[94] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).
[95] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
[96] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed Resources For General Chinese Embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery.
[97] Yue Xiao, Zhengyi Li, Yue Qin, Xiaolong Bai, Jiale Guan, Xiaojing Liao, and Luyi Xing. 2023. Lalaine: Measuring and Characterizing Non-Compliance of Apple Privacy Labels. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association.

A Website Selection Details

In Section 3.2, we briefly explain website selection in our dataset, We provide more details in this appendix.

We choose top websites from the Tranco list version 82NJV [56, 57]. The successfully crawled 11,500 websites (apex domains) rank from 1 (`google.com`) to 33,515 (`ebayinc.com`) in the list, *i.e.,* 22,015 websites in between were not crawled due to various reasons as explained in Section 3.2 and detailed below.

The following websites are removed according to the results from the Python HTTP probing script:

- 7,924 domains do not resolve, do not serve HTTP(s) on standard ports (TCP 80 and 443), or return HTTP errors (4XX or 5XX) when accessing homepages. These include many CDN domains (*e.g.,* `googleapis.com` and `tiktokcdn.com`). These also include domains that completely block our script (*e.g.,* `whatsapp.com`).
- 8,044 websites host non-English contents on their homepages, according to the HTML `lang` tags (*e.g.,* `mail.ru`).
- 2,513 websites redirect to different apex domains when accessing their homepages, usually indicating they are not meant to be accessed directly (*e.g.,* `amazonaws.com`).

Then, the following websites are filtered out based on the Cloudflare domain intelligence [15] results:

- 32 are classified as "public suffixes". These include some dynamic DNS services that host many websites (*e.g.,* `duckdns.org`).
- 1,148 websites are classified as unsafe contents (*i.e.,* any of CIPA, Adult Themes, Questionable Content and Blocked labels).
- 255 are classified as "applications" that belong to a parent organization – we only crawl the first among all applications under each organization (*e.g.,* `google.co.jp` is removed because `google.com` supersedes it).

Finally, the crawler failed to finish on 2,099 websites due to errors, including network errors and errors caused by crawler protection. To rule out sporadic network issues, we retried crawling failed websites at least three times on different days before giving them up. Top-ranking examples in this category include `cloudflare.com`, `icloud.com` and `msn.com`.

B Dataset Annotation Details

B.1 GPT Prompts for Dataset Annotation

Section 4 details our dataset annotation methodology, including using the GPT to help build the taxonomy and label training samples that are used to train the form type and PI type classifiers. We provide the prompts that we used to query the GPT in this appendix.

We use the following prompt for creating the form type taxonomy (*i.e.,* summarizing web forms) in Section 4.1:

> Analyze the provided HTML code of a web form, along with the URL and title of the web page to determine the type of the form based on its usage.
>
> URL: {...}
> Page Title: {...}
> HTML Code of the Web Form:
> ```
> {...}
> ```
>
> Please use a simple phrase to describe the usage of the form.
>
> If insufficient information is available to determine the usage, output "unknown".
>
> The response should be in JSON format with a single key "Classification".

We use the following prompt for labeling training samples of form type classification in Section 4.1:

> Analyze the provided HTML code of a web form, along with the URL and title of the web page to determine the type of the form based on its usage.
>
> URL: {...}
> Page Title: {...}
> HTML Code of the Web Form:
> ```
> {...}
> ```
>
> Classify the web form based on its intended usage.
>
> Here are the possible categories for classification:
> - "Account Registration Form": For creating new online user accounts.
> - {*... see Table 2 for the full list of label definitions*}
> - "Search Form": Used to search or filter website content, typically featuring a search query field and/or filter options.
> - "Configuration Form": For customizing the user experience on the website, like setting preferences for cookies, language, or display settings.
>
> Please choose the category that best describes the form.
> If none of the above categories accurately describe the form, suggest a new category.
> If the information is insufficient to make a confident classification, label it as "Unknown".
>
> Format the response in JSON with one key "Classification".

We use the following prompt for bootstrapping the list of PI types in Section 4.2:

> I will provide the HTML code of a web form. Please analyze the form and identify the types of personal data that are being requested in the form fields.
>
> "Personal data" (or "personal information") should be understood according to the following definitions in privacy laws:
>
> 1. **California Consumer Privacy Act (CCPA)**: {*... quoting CCPA Section 1798.140(v)(1)*}
> 2. **General Data Protection Regulation (GDPR)**: {*... quoting GDPR Art. 4 – Definition of "personal data"*}

> Please analyze the given HTML code of a web form and identify fields that may collect personal data as per these definitions.
>
> The output should be in JSON format and include concise and easily interpretable noun phrases that clearly indicate each type of personal data being requested, for example: `` `{"personal_data_types": ["Name", "Email Address", "Phone Number"]}` ``
>
> Remember, the focus is on identifying personal data that are being requested in the form. Exclude any data that does not fit above-mentioned definitions. If no personal data is being collected, simply output an empty list:
> `` `{"personal_data_types": []}` ``
>
> Here is the HTML code of the form:
> ```
> {...}
> ```

The placeholders (*i.e.,* {...}) are filled in programmatically with corresponding data to be processed. We also turn on the JSON mode [51] of the OpenAI GPT API to be able to process the GPT's responses programmatically.

B.2 Form Type Classification Examples

In Table 2 in Section 4.1, we define 10 labels for classifying web forms according to their functionality. In this appendix, we show representative examples of each form type in Figure 7, and summary their visual characteristics below:

- *Account Login*: These forms typically ask for, at least, an account identifier (*e.g.,* email address or username) and a password (Figure 7a). The submit button is usually labeled with "Log In" or similar phrases.
- *Account Recovery*: These forms typically ask for a piece of contact information that is also used as the account identifier (Figure 7b).
- *Account Login*: These forms, depending on the services, may ask for various contact information, along with a password to be set (Figure 7c1), or a minimal set of information similar to Account Login forms (Figure 7c2). The submit button is usually labeled with "Sign Up", "Register" or similar.
- *Contact*: This form type is usually indicated by a free-form text box and requires some contact information. The purposes of "contact" are broad. Examples include: feedback (Figure 7d1), service inquiry (Figure 7d2), and lead generation (Figure 7d3).
- *Content Submission*: This form type looks similar to Contact forms. The difference is the submitted content is intended to be published, which is usually indicated by words "comment", "review" or similar in the forms (Figure 7e).
- *Financial Application*: This form type can require extensive user information, including name, address and tax IDs. The functionality is typically indicated by phrases like "open an (banking / credit card) account" in the forms (Figure 7f).
- *Payment*: This form type typically asks for payment methods, *i.e.,* card number, name and address (Figure 7g).
- *Reservation*: This form type typically asks for contact information, as well as time preferences for the reservation (Figure 7h).
- *Role Application*: This form type can require extensive user information. The functionality is typically indicated by phrases like "(school / job) application" in the forms (Figure 7i).
- *Subscription*: This form type typically asks for very minimum contact information, usually the email address (Figure 7j).
- *Unknown*: These forms are not classified into any of the previous types, because (1) their functionality does not fit into any form types (*e.g.,* Figure 7k1, the form is used for checking service coverage), or (2) there is insufficient information for classification (*e.g.,* Figure 7k2 is the first step of a login form, however, the HTML code and the page title do not clearly indicate the usage). The second type can also be counted as false results. However, the classifier would have to read additional textual features on the webpages, beyond the HTML code of the web forms themselves, to be able to work.

B.3 Impact of Input Sequence Length on Form Type Classification

In Section 4.1 and Table 4, we briefly discuss factors that impact the precision of form type classification. In response to a question by the reviewers, we additionally analyze the impact of the input sequence lengths on the classifier performance.

In the form type classification, the base MarkupLM model [35] supports a max input length of 512 tokens, and we truncate any longer HTML inputs. In this evaluation, we split the 500 validation samples into equal-size quantiles of input sequence lengths (after truncation): $(0, 37]$, $(37, 84]$, $(84, 205]$, $(205, 512]$. The classifier gets 85.9%, 84.6%, 87.1% and 84.8% precision, respectively, in the four quantiles. In addition, on the 55 truncated inputs, the classifier gets 85.5% precision. Therefore, we conclude that the input sequence length and input truncation do not significantly affect the classificaiton performance.

C Detection of Privacy Policies

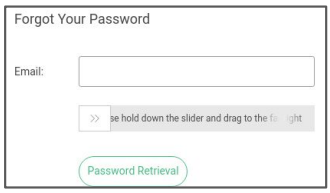
In Section 6.1, we briefly explain how we extract the privacy policy link associated with each website or web form. We provide more details in this appendix.

The web form crawler (see Section 3) saves the HTML of all the visited web pages, besides the web forms. We use a Python script to parse the HTML code to detect the privacy policy links. It starts by extracting all the links (*i.e.,* <a> elements) in the HTML code. Then, similar to the priority assignment strategy explained in Section 3.1, the text and URL of each link are compared with a list of seed phrases using cosine similarity between the text embeddings. If the link with the highest cosine similarity score has a score higher than 0.75, it is recognized as the privacy policy link. Additionally, it also tries to match any link text and URLs that have seed phrases in it (*e.g.,* `http://.../privacy-policy-1`).

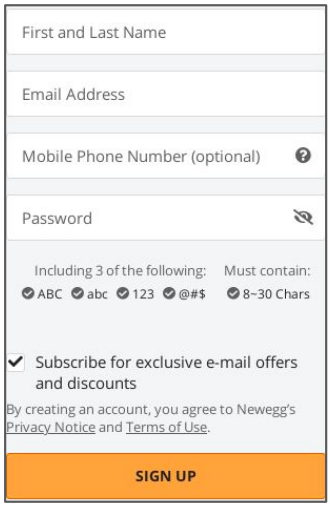
We use six seed phrases: "privacy policy", "privacy notice", "privacy statement", "privacy center", "privacy & terms", and "privacy & cookies notice". The list was obtained as follows. We started with three commonly used phrases: "privacy policy", "privacy notice", and "privacy statement". Then, we added three more phrases which were later found on websites but have low similarity scores to the initial three, so we could cover more privacy policy links. Our manual validation later shows that the six selected phrases sufficiently cover privacy policy links in our dataset.

To determine if privacy policy links are provided within the forms (*i.e.,* Table 6), we simply run the detection logic above on the HTML code of individual forms, which effectively limits the detection within the <form> elements. For each website, we consider

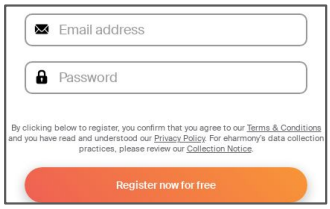
(a) Account Login (archive-it.org)



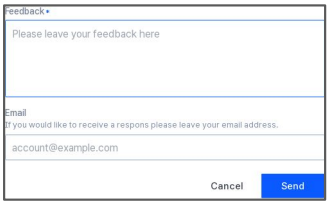
(b) Account Recovery (yupoo.com)



(c1) Account Registration #1 (newegg.com)

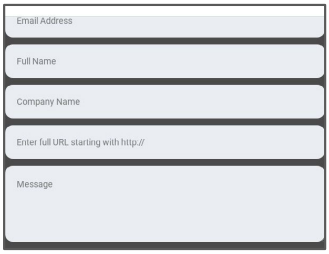


(c2) Account Registration #2 (eharmony.com)



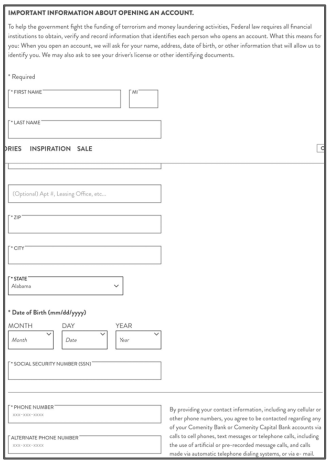
(d1) Contact #1 (airbridge.io)

(d2) Contact #2 (play-asia.com)

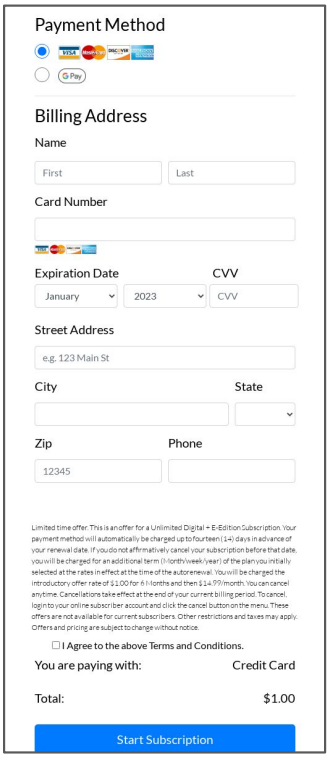


(d3) Contact #3 (avantlink.com)

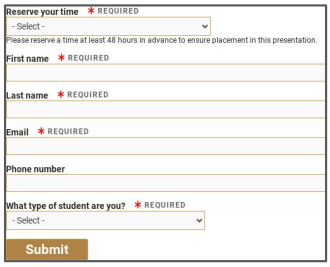
(e) Content Submission (cookieandkate.com)



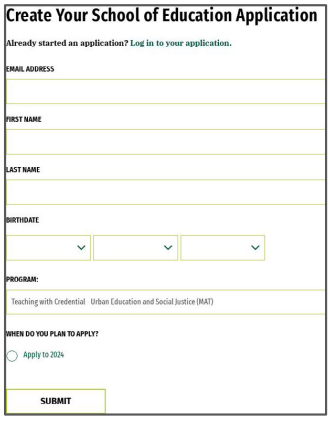
(f) Financial Application (jjill.com)



(g) Payment (greensboro.com)

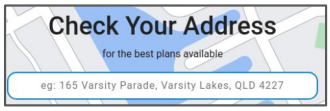


(h) Reservation (wright.edu)



(i) Role Application (usfca.edu)

(j) Subscription (spring.io)



(k1) Unknown #1 (onthenet.com.au)

(k2) Unknown #2 (zedge.net)

Figure 7: Typical examples of each form type. For all form types except Unknown, we pick the examples from our validation data. For Unknown forms, we pick the examples from the annotated dataset.

all the privacy policy links found on its web pages as its privacy policies (for the availability analysis in Table 5, and the content analysis in Section 6.2).

Validation. We randomly selected 200 websites to manually validate the accuracy of privacy policy link detection, including 17 privacy policy links found inside the web forms, 164 found on the web pages, and 19 without any privacy policy. One author opened each containing web page, visually checked the form (if the link is detected inside a form), the page footer and also the homepage (in the case of no privacy policy) to search for the privacy policy links. Overall, the detection results are accurate for 188 (94.0%) websites. Of the 12 incorrect results, 4 were due to incorrect third-party privacy policy links being recognized, 6 were due to unexpected link text (*e.g.*, a website uses "learn more" in the cookie consent dialog), and 2 were due to HTML parsing issues.