

An Empirical Assessment of Security and Privacy Risks of Web-based Chatbots

Nazar Waheed¹, Muhammad Ikram², Saad Sajid Hashmi³, Xiangjian He¹, and Priyadarsi Nanda¹

¹ University of Technology Sydney, NSW 2007, Australia

² Macquarie University, NSW, Australia

³ University of Wollongong, Wollongong, Australia

Abstract. Web-based chatbots provide website owners with the benefits of increased sales, immediate response to their customers, and insight into customer behaviour. While Web-based chatbots are getting popular, they have not received much scrutiny from security researchers. The benefits to owners come at the cost of users' privacy and security. Vulnerabilities, such as tracking cookies and third-party domains, can be hidden in the chatbot's iFrame script. This paper presents a large-scale analysis of five Web-based chatbots among the top 1-million Alexa websites. Through our crawler tool, we identify the presence of chatbots in these 1-million websites. We discover that 13,515 out of the top 1-million Alexa websites (1.59%) use one of the five analysed chatbots. Our analysis reveals that the top 300k Alexa ranking websites are dominated by *Intercom* chatbots that embed the least number of third-party domains. *LiveChat* chatbots dominate the remaining websites and embed the highest samples of third-party domains. We also find that 850 (6.29%) of the chatbots use insecure protocols to transfer users' chats in plain text. Furthermore, some chatbots heavily rely on cookies for tracking and advertisement purposes. More than two-thirds (68.92%) of the identified cookies in chatbot iFrames are used for ads and tracking users. Our results show that, despite the promises for privacy, security, and anonymity given by the majority of the websites, millions of users may unknowingly be subject to poor security guarantees by chatbot service providers.

1 Introduction

A Web-based chatbot (or bot) is a computer program interacting with users via a conversational user interface that simulates a conversation with a human user via textual methods [33]. Web-based chatbots offer improved customer services and efficiently manage human resources [28, 32]. For example, a website owner performs customer acquisition tasks (such as new customer query, or after-sales services) through customer service (or sales and marketing) personnel. As the business gets bigger and busier, the traditional way of interacting with the online customers gets choked up resulting in increased waiting queue. Besides, the customer service representative may not be available around the clock. Web-based

chatbot provides a website owner with the benefits of increased sales, immediate response to their customers' queries and insights into customers' behaviours. While Web-based chatbots are getting popular, they have not received much scrutiny from security researchers. The benefits of chatbots can come at the cost of privacy and security threats. These threats are inherited by third-party domains and cookies, which might be built-in to the script. These domains and cookies can be used for the purpose of tracking users and providing personalised advertisements. There has been plethora of work done based on the security and privacy issues of a complete website [17, 21, 31]. However, as per our knowledge, there is no research study that focuses explicitly on the privacy and security issues of Web-based chatbots.

While Web-based chatbots are getting popular and they come with several above-mentioned benefits, their advantages are inherited with several disadvantages. *Firstly*, consumers are concerned about their privacy and security [32]. Despite the remarkable improvements in Web-based chatbots being able to mimic a human conversation, they are vulnerable to the Reconnaissance and Man-in-the-Middle attacks [12]. *Secondly*, since the chatbot is a computer program, it does not have its own identity or emotions like a real human. Customers often tend to make connections during conversations, which is lacking when engaging with chatbots. The lack of personality in chatbots and their inability to make an emotional connection is a concern for some customers. *Finally*, a Web-based chatbot is still in its infancy since natural language processing is not the core competency in chatbot applications and is still in the development phase [32]. Web-based chatbots are prone to common communication errors, therefore, companies and organisations are very careful in using them to avoid any brand damage.

Although several studies have taken place to study chatbots in general, none of them covers their security and privacy comprehensively. There has been extensive research on the security and privacy issues of websites, however, to the best of our knowledge, we did not find any study that focuses on the iFrame component of the Web-based chatbot for the same issues.

An overview of our methodology is presented in Figure 1. In this paper, we present our methodology (depicted in Figure 1) to analyse Web-based chatbots at scale. We begin by inspecting how to filter chatbot websites by manually analysing the Alexa top 100 websites. We develop a Selenium-based web crawler to automatically detect these websites based on our analysis and assure the accuracy is 100%. We also search for popular chatbots on the internet and select them based on their prominence. We find a total of 13,515 chatbot websites for our five selected chatbots as our dataset.

We then inspect 10 different categories of websites in our dataset. We observe that Web-based chatbots present predominantly in the non-IT business category (21.78%), IT category (16.16%), and shopping category (5.89%). The complete list can be seen in the Figure 2. We confirm that at least 4.2% of the Alexa top 500k Web-based chatbots, 14.88% of the second half of the Alexa top 1-million, and at least 6.29% in the top 1-million Alexa Web-based chatbots are still using insecure HTTP. Although seemingly small, the fact that these are

the most popular websites is a big security concern. We then proceed to inspect the Web-based chatbot iFrame in particular instead of entire website DOM to find the vulnerabilities in our selected chatbots. We find that among the three chatbots that agree to write cookies on a customer’s visit to their website, **Drift** chatbot writes nine different types of cookies 5,396 times out of which at least 94.62% are tracking cookies. **Hubspot** chatbot writes twelve different types of cookies 15,829 times out of which 79.35% are tracking cookies, and **Intercom** chatbot writes fourteen different types of cookies 18,995 times out of which 34.85% are tracking and analytics cookies while 18.07% are Ads and marketing cookies. We cannot find any cookies for **Tidio** and **LiveChat** chatbots, and their support team confirms this as well. Note that we do not take the cookies of entire websites into account, rather we focus on the cookies that a chatbot is used for essential and tracking/advertisement purposes. Our focus is on the chatbot and its iFrame only, which, to the best of our knowledge, has not been discussed in any study thus far.

Despite the assurances for privacy, security, and anonymity given by the websites and privacy policies, users are victims of personally identifiable information (PII) leakages [19]. Similarly, by using chatbot services, users may inadvertently be exposed to the privacy and security risks [31].

In summary, the contributions of this paper are as follows.

1. We present the first large-scale study of security and privacy issues in chatbots on Alexa top 1-million popular websites [4]. We detect 13,515 (1.59%) websites leveraging web chatbots for customers’ interaction. We release our data and scripts for future research.
2. We analyse the 13,515 (1.59%) websites for the type of chatbots and analyse the coverage of the detected chatbots. We find that 21.78% of the chatbot websites belong to the non-IT business category, while the percentage of Information Technology (IT) chatbot websites is 16.16%, and shopping with 5.89% is the third most dominant category. We also analyse the security and privacy issues of our dataset chatbot websites. We explore the chatbot websites and find that 6.29% of them are still using the insecure HTTP protocol, where an alarming 14.88% of the websites ranking >500k still transfer their visitors’ data in plain-text. This shows that among the most popular websites, non-IT business, IT and shopping websites are more vulnerable than any other website categories.
3. Our analysis illuminates that chatbots have a disproportionate use of cookies for tracking and *essential* or *useful* functionalities. We discover 5,396 cookies in 2,110 websites leveraging **Drift** chatbot. 5,113 (94.62%) and 283 (5.24%) of the cookies are used for Tracking and essential functionalities, respectively. On the other hand, 2,185 websites rely on **Hubspot** for the provision of chat services via a total number of 15,829 unique cookies with 79.35% (12561) for tracking while the rest are essential cookies.
4. We identify the top 10 third-party domains embedded in the iFrames of each web-based chatbot. The most common third-parties are well-known operators, for example, googleapis, cloudflare, w3, and facebook. These operators

are imported by 39.67% (5361), 15.43% (2085), 6.1% (822), and 3.35% (453) web-based chatbots, respectively.

The rest of this paper is organized as follows: In Section 2, we present concepts and terms related to web tracking and services. We present our methodology for web-based chatbot detection and data collection in Section 3. In Section 4, we analyse our chatbots in the top 1-million Alexa websites and present our findings such as the presence of chatbots on websites, tracking cookies, and third-party domains. Section 5 presents the related work while we conclude our work by presenting the gaps with some future directions in Section 6.

2 Concepts and Terms

We begin by introducing the general concepts and terms used in the paper.

Advertising and tracking domain: The *advertising and tracking* domain (or tracker) is the URL of an entity embedded in a web page. The purpose of a tracker is to re-identify a user’s visit on the web page again for loading custom themes or analytics (*first-party* tracking) or to re-identify a given user across different websites for building the user’s browsing profile or providing personalised advertisements (*third-party* tracking).

iFrame: An iFrame or inline Frame is an HTML document embedded within an HTML web page. The purpose of an iFrame is to display embedded HTML contents from a different web page into the current web page. The contents of iFrames can be videos, maps, advertisements, chatbot services, as well as tracking components like cookies and JavaScript codes. Hence, besides providing utilities and services, iFrames can also be used for third-party tracking.

Cookie: A cookie (or HTTP cookie) is a text file that is stored on the user’s device by the web browser. The content of a cookie is in plain text format. A cookie is generated by the web server (of a web page) and is sent back from the user’s device to the web server at each subsequent visit by the user. A cookie can store shopping carts, theme preferences of the user, or the user’s authentication status. Cookies generated by third-parties via iFrames can be used for third-party tracking. Different types of cookies are discussed in detail in Section 4.2.

3 Chatbot Detection Methodology and Dataset

We begin by presenting our methodology, over-viewed in Figure 1, for detecting chatbots employed in the top 1-million Alexa websites. We then characterise our dataset.

3.1 Discovering Chatbots

Using Selenium Web Driver, we develop an automated web crawler to automate the visiting and rendering process of analysed websites. To increase our chatbot coverage and maximise the number of detected chatbots, we implement a crawler

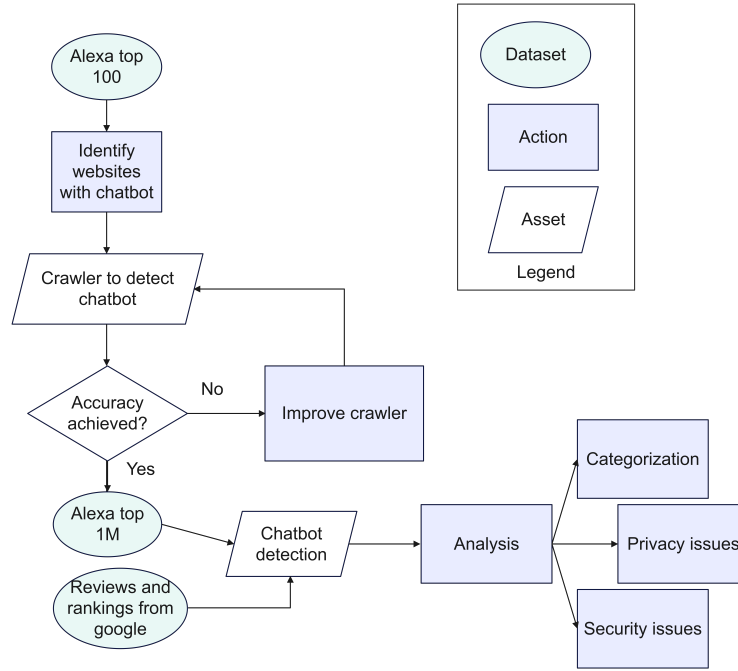


Fig. 1: Overview of our crawling and analysis methodology: We manually inspect the top 100 Alexa websites for chatbots to identify chatbot services and to construct keywords list for automatic detection of chatbots in the top 1-million Alexa websites. We then perform an analysis to categorise websites and to analyse security and privacy issues.

framework. We begin by discovering web-based chatbot services on the Alexa top 1-million websites. To this end, we find the difference between a normal website and the website with a chatbot service. We manually inspect the first Alexa top 100 websites for *potential* web chatbot services. Typically, websites implement chatbot services in iFrames, therefore, we explicitly focused on the iFrame of the chatbot on these 100 websites. The keywords include: ‘*chat widget*’, ‘*let’s chat*’, ‘*drift-widget*’, ‘*chat now*’, and ‘*chatbot*’. While we acknowledge that our keywords list is not exhaustive to include chatbots on non-English language websites, we do consider our method for chatbots as a *lower-limit* on the number of chatbots on the top 1-million Alexa websites.

Next, we crawl through the chatbot websites and extract their chatbot iFrame cookies only instead of the whole website, since we are specifically interested in the security and privacy issues of the web-based chatbots. We then analyse the embedded third-party domains in each of those chatbots. To extract the third-party domains, we only check the contents of the iFrame of a chatbot, instead of the complete website’s DOM. Overall, we find 13,515 (1.6%) chatbot websites,

out of which 566 (4.2%) websites do not render, either due to the website being closed or moved permanently to a new domain name.

Issues and Limitations. For chatbot websites, once a website is completely rendered, the chatbot icon is found at the bottom right corner of the screen. Sometimes, the chatbot is not visible on the respective website. This is mostly due to one of the following reasons (*i*) the chatbot is only available during certain office hours, and (*ii*) the chatbot is offline/hidden as the developers may be working on it.

3.2 Data Augmentation

Next, to analyse the coverage of chatbots in various categories of websites, we aim to categorise the Alexa top websites. There are several databases and tools available and website categories stored. However, we use crawling techniques on *Fortiguard* website classification tool [1] to gather this information. The websites that return errors while rendering in the first phase are manually labelled. We find the categories of each chatbot website in our dataset (13,515 websites). The top 10 categories, depicted in Figure 2, are selected based on their frequencies of occurrence. These ten categories comprise 75% of our dataset, while the remaining 25% are categorised as *Others*. It is found that most of the chatbot websites are used by *non-IT Business* and *IT* category websites. However, chatbots are not a popular choice among *Games* and *Government and Legal Organizations* related website owners.

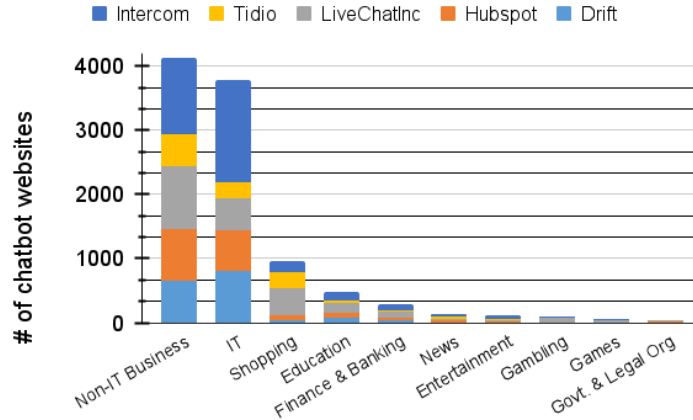


Fig. 2: Categories of chatbot websites in the top 1-million Alexa websites.

	Chatbots					# Total (%)
Alexa Rank	Drift	LiveChat	Hubspot	Tidio	Intercom	
1-100K	619 (0.62%)	436 (0.44%)	210 (0.21%)	100 (0.10%)	940 (0.94%)	2,305 (2.31%)
100K-200K	453 (0.45%)	591 (0.59%)	386 (0.39%)	185 (0.19%)	757 (0.76%)	2,372 (2.37%)
200K-300K	407 (0.41%)	586 (0.59%)	394 (0.39%)	343 (0.34%)	681 (0.68%)	2,411 (2.41%)
300K-400K	295 (0.30%)	573 (0.57%)	413 (0.41%)	365 (0.37%)	531 (0.53%)	2,177 (2.18%)
400K-500K	150 (0.15%)	433 (0.43%)	305 (0.31%)	286 (0.29%)	429 (0.43%)	1,603 (1.60%)
500K-600K	65 (0.07%)	319 (0.32%)	160 (0.16%)	218 (0.29%)	273 (0.27%)	1,035 (1.04%)
600K-700K	51 (0.05%)	341 (0.34%)	143 (0.14%)	125 (0.13%)	182 (0.19%)	843 (0.84%)
700K-800K	36 (0.04%)	131 (0.13%)	64 (0.06%)	2 (0.002%)	136 (0.14%)	369 (0.37%)
≥ 800K	26 (0.05%)	86 (0.18%)	104 (0.22%)	50 (0.11%)	93 (0.20%)	359 (0.75%)
Overall (1-million)	2,110 (0.25%)	3,507 (0.41%)	2,185 (0.26%)	1,676 (0.20%)	4,037 (0.48%)	13,515 (1.59%)

Table 1: Frequency (and percentage) of chatbot services amongst the Alexa top 1-million websites. Highlighted trends show **Intercom** chatbot is the preferred choice for the most popular set of websites followed by **LiveChat** which is also the preferred choice for the next tier of popular websites.

3.3 Dataset

Our comprehensive analysis is done by breaking the dataset into parts with each part having 10,000 websites to get an in-depth measurement of our study. Based upon the keywords (cf. § 3), we run our crawler that detect chatbots on 3.5% of the analysed websites. To check the accuracy of our crawler, we manually label the first hundred Alexa ranking websites and perform manual testing on them. It is learnt that our model is 61% accurate. The reason is that there are several possible ways to write a website script, and using the keywords alone is not an optimum solution.

Finding a common script, or tag among all of them is not possible. However, we find some unique keywords/tags/elements. Figure 3 shows the iFrame of a chatbot website www.synology.com. It has a tag `id='chat-widget-container'`, which can be used to filter the **LiveChat** chatbot websites. Similarly, we select five chatbots: **LiveChat**, **Drift**, **Intercom**, **Tidio** and **Hubspot** based on their frequencies of occurrence in the top 10k Alexa ranking websites. Overall, our crawler identifies 13,515 chatbot websites from Alexa top 1-million websites.

Table 1 summarises our findings. We observe that the **Intercom** chatbot is the preferred choice for the most popular set of websites (top 300k) followed by **LiveChat** for the next tier of Alexa ranking websites. Overall, **Intercom** chatbot is found on 29.87% of them, **LiveChat** on 25.95%, **Drift** on 15.61%, **Hubspot** on 16.17%, and **Tidio** on 12.40% only.

Based on the above findings, in the first round, we crawl the top 10k websites and render their DOMs. After optimising our crawler, we can filter all chatbots with 100% accuracy.

We also search for the top Web-based chatbots by using different keywords over the google search. We find chatbot rankings and reviews on the websites in [3, 7, 8, 16, 29, 30, 36] (accessed in Feb 2022). We choose the top three chatbots. After selecting *MobileMonkey*, *Aivo*, and *Pandorabots* from the blogs and

```

▶<div style="display: none; visibility: hidden;">...</div>
▼<div id="chat-widget-container" style="opacity: 1; visibility: visible; z-index: 0px; width: 84px; height: 84px; max-width: 100%; max-height: calc(100% - 0px); background-color: transparent; border: 0px; overflow: hidden; right: 0px; top: 0px;">
  ▼<iframe allow="autoplay;" src="https://secure-fra.livechatinc.com/customer-service/widget?group=28&embedded=1&widget_version=3&unique_groups=0" allowtransparency="true" title="LiveChat chat widget" scrolling="no" style="width: 100%; height: 100%; margin: 0px; padding: 0px; background-image: none; background-position: 0% 0%; background-attachment: scroll; background-origin: initial; background-clip: initial; border-width: 0px; float: none; position: absolute; inset: 0px; transition: none;">
    ▼#document
      <!DOCTYPE html>
      ▼<html lang="en">
        ▶<head>...</head>
        ▼<body>
          <noscript>You need to enable JavaScript to run this app.</noscript>
          ▼<div id="root" style="width: 100%; height: 100%; position: static;">
            ▼<div id="widget-global-r0abkde8drh">
              <span hidden></span>

```

Fig. 3: An example of an iFrame enabling a typical chatbot service on a given website.

reviews, we run our automated scraper for the top 200k websites and find only two chatbots belonging to *MobileMonkey*, four chatbots to *Botsify*, and zero for both *Aivo* and *Pandorabots*. Therefore, due to their insignificant presence, we do not consider them in our analysis further.

As a second attempt, we manually re-analyse the top 100 websites and find two relevant chatbots (*SF-chat* and *SnatchBot*) and search for them over the top 10k websites using an automated script. For *Salesforce* chatbot, we only find it to be on their own websites, for example, *cloudforce* and *exactforce*. On all other top 10k websites, we do not find any other websites having either of these chatbots. 729 websites do not render in the first phase, and they are analysed again in the second attempt (we learn that rendering chatbot websites take longer than our previous timeout). We also manually analyse the 100 chatbots from 100,000 to 100,100 range and find three chatbots only, i.e., (i) *Drift*, (ii) *Intercom*, and (iii) *eLum*⁴. *Drift* is already included in our study, *Intercom* is found on numerous websites (after initial automated crawler verification), and *eLum* is not found anywhere else since it is a private custom chatbot. Moreover, please note that social-media related chatbots like Facebook messenger are not valid chatbots since they require human interaction and are not automated. Therefore, we do not include them in our analysis. For the rest of the study, we use only five chatbots, which are *Drift*, *Hubspot*, *LiveChat*, *Tidio*, and *Intercom*.

⁴ <https://eluminoustechnologies.com>

4 Exploring Web Chatbots

4.1 Analysis of HTTP Chatbot Websites

To check whether a website uses HTTP, our crawler defaults to communicating with the site over HTTP by simply concatenating the 'http://' or 'http://www.' string with the hostname provided in the Alexa data. Once the crawler receives a final response and it does not redirect the client requests from HTTP to HTTPS, it is marked as HTTP. We also check the websites that have errors by manually inspecting each one of them and discover that such websites are very few and the main reason for the errors is that they do not exist anymore (something that Alexa should take care of as it is not updated). The trend in the Figure 4b shows that less popular websites are less secure. We find that 850 (6.29%) out of 13,515 chatbot websites are still using the insecure HTTP version.

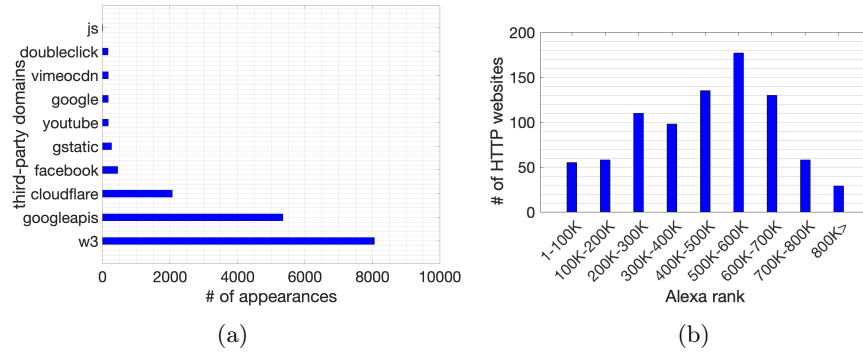


Fig. 4: (a) Breakdown of third-parties found in web-based chatbots. (b) Number of web-based chatbots using insecure HTTP websites in top Alexa websites.

4.2 Analysis of Cookies

The online ecosystem is composed of a large number of organizations engaging in tracking user behaviour across the web [13]. This is accomplished by a variety of techniques including tracking cookies, pixel tags, beacons, and other sophisticated mechanisms. Below, we provide an overview of the most common cookies.

Identification Cookies These cookies can track visitor's conversations and interactions with a website. The customer service representative uses such information to offer better service. It is challenging to learn about any old chat with the customer without these cookies.

Tracking Cookies These are the most common cookies used now to track user behaviour, user information and visits to a website.

Performance and Functionality Cookies: These cookies are used to enhance the performance and functionality of a website but are non-essential to their use. However, certain functionalities like videos may become unavailable or the login details are required every time a user visits the website.

Conditional Cookies These cookies may be written onto a website since they depend on using a specific feature of a website.

Marketing Cookies: These are account-based marketing cookies used to identify prospects and personalise sales and marketing interactions.

Analytics and Customization Cookies: These cookies are used to determine the effectiveness of marketing campaigns. Website owners use them to collect limited data from end-user browsers to enable them to understand the use of their websites.

Advertising Cookies: These cookies collect information over time about users' online activity on the websites and other online services to customise online advertisements.

The details about all cookies used on every chatbot can be read on their website [2, 20, 25–27].

Drift: According to **Drift**, the primary reason it uses cookies is to track user interactions with the visited website. It also uses cookies to customise products to the need of a customer. **Drift** claims that the data is never sold or sent to third-parties. Instead, it is used in their platform to allow for more personalised and specific messaging [25].

Hubspot: According to **Hubspot**, it uses cookies to track users who visit a **Hubspot** chatbot website. The purpose of these cookies is to keep track of visit counts and information about the sessions (such as session start timestamp). When the **Hubspot** software is run on a website, it leaves behind these cookies to help **Hubpost** identify the users on future visits [20].

LiveChat: We search for **LiveChat** cookies manually by inspecting several websites. We do not find any tracking cookie in our manual search. To confirm, we inquire from the **LiveChat** support team to ensure that none of the cookies is used for tracking purposes. The support team confirmed the same. The **LiveChat** chatbots automatically save and store two essential cookies on the user's device when a user visits a website with **LiveChat** widget [26]. The two essential cookies are as follows:

`__lc_cid` (*customerID*) This is a functional cookie that **LiveChat** account service uses. The purpose of this cookie is to verify the identity of a customer created.

Chatbots	Categories of Cookies			Total
	Essential	Tracking & Analytics	Ads & Marketing	
Drift	283 (5.38%)	5,113 (94.62%)	-	5,396
Hubspot	3,268 (20.65%)	12,561 (79.35%)	-	15,829
Intercom	8,942 (47.08%)	6,620 (34.85%)	3,433 (18.07%)	18,995
Total	12,493 (31.06%)	24,294 (60.4%)	3,433 (8.54%)	40,220

Table 2: Distribution of cookies across Web-based chatbots.

`__lc_cst` (*customerSecureToken*) This is also a functional cookie that **LiveChat** account service uses to identify a user, for example, name, IP address, geolocation etc.

Tidio: According to **Tidio**, it uses cookies to maintain, improve and customise the user experience. Additionally, the cookies are used to remember the visitor’s choice, such as language preference. **Tidio** claims to collect information, including PII and assures that it will be used by them only. We cannot find any evidence of **Tidio** cookies on any of the websites using their chatbots, nor can we find any information about what cookies are used on their website [27].

Intercom: According to **Intercom**, its chatbot writes “first-party” cookies only and assures that its cookies are strictly private and confidential. The purpose of these cookies is to identify users and keep track of sessions. Intercom states that it uses two cookies only [2]; however, this claim is contradictory to our findings discussed below

Findings/Discussion: To distinguish between a first-party and a third-party cookie, we consider any cookie with the same name as the respected chatbot as a first-party. We also consider the cookies that chatbot service providers have mentioned on their websites as first-party. We declare any other cookie as a third-party. We find a total number of 2,110 websites using **Drift**. From these, a total of 5,396 cookies are discovered. 5,113 (94.62%) of them are used for Tracking, and 283 (5.24%) are essential cookies. Hubspot is used on 2,185 websites, which have 15,829 cookies. 12,561 (79.35%) are tracking, while the rest are *essential* cookies. **Intercom** chatbot websites are 4,037, generating 18,995 cookies, out of which 52.92% are either tracking, advertisement or marketing cookies, while 47.08% are essential cookies for functionality. No cookies are found on either **LiveChat** or **Tidio** chatbots. Overall, more than two thirds of the discovered cookies are used for tracking or advertisement purposes.

Third-party Domain	Drift	LiveChat	Intercom	Tidio	Hubspot	Total
w3.org	742	5	0	6,501	813	8,061
googleapis.com	28	3,502	0	1,537	282	5,349
cloudflare.com	2,063	1	0	0	10	2,074
facebook.com	0	0	0	0	453	453
gstatic.com	0	0	0	0	268	268
youtube.com	0	0	0	0	174	174
google.com	0	0	0	0	172	172
vimeocdn.com	0	0	0	0	171	171
doubleclick.net	0	0	0	0	166	166
rlts.com	0	0	5	0	0	5
<i>other domains</i>	0	19	0	7	3,872	3,989
Total	2,833	3,527	5	8,045	2,053	20,791

Table 3: Distribution of top ten third-parties embedded in the iFrames of Web-based chatbots.

4.3 Analysis of Third-party Domains

We parse the URLs from the chatbot iFrames, extract the second-level domains using *tldextract*⁵, and compare them with the respective website. If they match, it is declared a first-party domain; otherwise, it is stated as a third-party domain. For instance, we extract `googleapis.com` and `drift.com` domains from the iFrame of **Drift** chatbot embedded in the landing page of `https://www.drift.com`. Given that `googleapis.com` does not match with `drift.com`, our method labelled `googleapis.com` as third-party whilst `drift.com` as first-party. For instance, we extract `googleapis.com` and `drift.com` domains from the iFrame of **Drift** chatbot embedded in the landing page of `https://www.drift.com`. Given that `googleapis.com` does not match with `drift.com`, our method labelled `googleapis.com` as third-party whilst `drift.com` as first-party.

Figure 4a depicts, and Table 3 lists the top 10 third-party domains embedded in the iFrames of chatbots. We observe that all chatbots rely on third-party services such as W3, Google APIs, and CloudFlare for iFrame template, fonts, and hosting as well as storing content, respectively. We observe that only one third-party domain (`rlts.com`) is found on the **Intercom** websites. Since **Intercom** dominates the top 300k Alexa websites (52% of total web-based chatbot websites) suggesting that the top websites do not rely much on advertising and analytical services revenues funneled from chatbots. On the other hand, less popular websites generate 99.9% of the top ten third-party domains. **Hubspot** based websites have the most variety⁶ of third-party domains, making it the most vulnerable. One hundred forty five different third-party domains are present in **Hubspot** websites.

⁵ <https://pypi.org/project/tldextract/>

⁶ Drift=3, Livechat=10, Hubspot=145, Tidio=4, Intercom=1

5 Related Work

To the best of our knowledge, no prior work has been done to address the privacy and security risks of cookies or third-party scripts embedded in web-based chatbots. Previous work has analysed PII leaks via advertisements and third-party scripts on various domains such as Facebook [5, 6, 14, 35], mobile eco-system [18, 22, 24], and web forms [34].

There are security and privacy risks associated with chatbots [15]. In *financial* chatbots, Bhuiyan et al. proposed a chatbot leveraging a private blockchain platform to conduct secure and confidential financial transactions [9]. Chatbots have also been developed to remove sensitive information from the conversation before passing it to its NLP engine [10]. Meanwhile, threats on the chatbot’s client-side (such as unintended activation attacks and access control attacks) and network-side (such as MITM attacks and DDoS attacks) have been studied in the literature [37]. Bozic et al. conducted a preliminary security study on an open-source chatbot to identify XSS and SQLi vulnerabilities [11]. Their work did not find any XSS and SQLi vulnerabilities and was limited to analysing only one chatbot. No prior work has been done to study the iFrames of Web-based chatbots and to determine the types of cookies embedded. In this paper, to fill the gap, we study the prevalence of five chatbots in Alexa top 1-million websites and analyse the chatbot cookies and third-party domains embedded in the iFrames of chatbots.

6 Conclusion and Future Work

In this paper, *firstly*, we have presented the difference between websites with and without chatbots. We have found the keywords to detect chatbots on the analysed websites. We have also manually inspected the top 1,000 websites to validate chatbot detection. *Secondly*, we have designed and implemented a crawler tool that systematically explores and collects DOMs from the top 1-million Alexa websites. We have discovered that a subset of 13,515 (1.59%) of these websites use our five selected chatbots. We have found the frequencies of these chatbots in ten different categories and discovered that non-IT business websites had used 21.78% of them. Our analysis has revealed that the top 300k Alexa ranking websites are dominated by **Intercom**, while **LiveChat** dominates the remaining chatbot websites. We have also found that 6.29% of the chatbot use insecure protocols to transfer users’ chats in plain-text. Our results show that, despite the promises for privacy, security, and anonymity given by the majority of the websites, millions of users may be unawarely subject to poor security guarantees by chatbot service providers on the same websites.

In the future, we want to extend our findings to the distribution of third-party domains and trackers in categories of web-based chatbots websites. This will help analyse and identify the dependence of chatbot websites on advertising and analytical services. Another area to explore is whether any chatbot websites render content that it does not directly load. Informed by the study by Ikram

et al. [23], this work can be extended to analyse the dependency web-resources chains of the chatbots.

References

1. Web filter lookup (2021), <https://www.fortiguard.com/webfilter>
2. Intercom cookie policy (2022), <https://www.intercom.com/legal/cookie-policy>
3. Aaron Brooks: 10 best chatbot builders in 2021 (2020), <https://www.ventureharbour.com/best-chatbot-builders/>
4. Amazon: Alexa top websites (2021), <https://www.alexa.com/topsites>
5. Andreou, A., Silva, M., Benevenuto, F., Goga, O., Loiseau, P., Mislove, A.: Measuring the Facebook Advertising Ecosystem. In: Network and Distribution Systems Security Symposium. San Diego (2019)
6. Andreou, A., Venkatadri, G., Goga, O., Gummadi, K.P., Loiseau, P., Mislove, A.: Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. In: NDSS. San Diego (2018)
7. Balkhi, S.: 14 best ai chatbots software for your website (compared) (2021), <https://www.wpbeginner.com/showcase/best-chatbots-software-ai/>
8. Barker, S.: 15 best ai chatbot platforms to boost your conversations in 2022 (2021), <https://shanebarker.com/blog/best-ai-chatbot/>
9. Bhuiyan, M.S.I., Razzak, A., Ferdous, M.S., Chowdhury, M.J.M., Hoque, M.A., Tarkoma, S.: BONIK: A blockchain empowered chatbot for financial transactions. In: TrustCom (2020)
10. Biswas, D.: Privacy Preserving Chatbot Conversations. Proceedings - 2020 IEEE 3rd International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2020 pp. 179–182 (2020). <https://doi.org/10.1109/AIKE48582.2020.00035>
11. Bozic, J., Wotawa, F.: Security testing for chatbots. In: IFIP International Conference on Testing Software and Systems. pp. 33–38. Springer (2018)
12. Carter, E., Knol, C.: Chatbot - an organisation's friend or foe? Research in Hospitality Management **9**(2), 113–115 (2019)
13. Cook, J., Nithyanand, R., Shafiq, Z.: Inferring tracker-advertiser relationships in the online advertising ecosystem using header bidding. arXiv preprint arXiv:1907.07275 (2019)
14. Ghosh, A., Venkatadri, G., Mislove, A., Kharagpur, I.: Analyzing Political Advertisers' Use of Facebook's Targeting Features. Workshop on Technology and Consumer Protection (ConPro '19) (2019), <https://facebook-targeting.ccs.neu.edu>
15. Gondaliya, K., Butakov, S., Zavarsky, P.: SLA as a mechanism to manage risks related to chatbot services. IEEE Intl Conference on Intelligent Data and Security (2020). <https://doi.org/10.1109/BigDataSecurity-HPSC-IDS49724.2020.00050>
16. Group, Z.: Top 10 best ai chatbots (2020), <https://medium.datadriveninvestor.com/top-10-best-ai-chatbots-f68705a8f559>
17. Hashmi, S.S., Ikram, M., Kaafar, M.A.: A longitudinal analysis of online ad-blocking blacklists. In: 2019 IEEE 44th LCN Symposium on Emerging Topics in Networking (LCN Symposium). pp. 158–165. IEEE (2019)
18. Hashmi, S.S., Ikram, M., Smith, S.: On optimization of ad-blocking lists for mobile devices. In: Proceedings of the 16th EAI International Conference on Mobile and

- Ubiquitous Systems: Computing, Networking and Services. p. 220–227. MobiQuitous '19 (2019). <https://doi.org/10.1145/3360774.3360830>
19. Hashmi, S.S., Waheed, N., Tangari, G., Ikram, M., Smith, S.: Longitudinal compliance analysis of android applications with privacy policies. In: Mobile and Ubiquitous Systems: Computing, Networking and Services. pp. 280–305. Springer International Publishing, Cham (2022)
 20. Hubspot: Cookies set on Hubspot’s websites (2021), <https://knowledge.hubspot.com/account/hubspot-cookie-security-and-privacy>
 21. Ikram, M., Asghar, H.J., Kaafar, M.A., Mahanti, A., Krishnamurthy, B.: Towards Seamless Tracking-Free Web: Improved Detection of Trackers via One-class Learning. *Proceedings on Privacy Enhancing Technologies* **2017**(1), 79–99 (2016). <https://doi.org/10.1515/popets-2017-0006>
 22. Ikram, M., Kaafar, M.A.: A first look at mobile Ad-Blocking apps. 2017 IEEE 16th International Symposium on Network Computing and Applications, NCA 2017 pp. 1–8 (2017). <https://doi.org/10.1109/NCA.2017.8171376>
 23. Ikram, M., Masood, R., Tyson, G., Kaafar, M.A., Loizon, N., Ensafi, R.: The chain of implicit trust: An analysis of the web third-party resources loading (2019). <https://doi.org/10.1145/3308558.3313521>, <https://doi.org/10.1145/3308558.3313521>
 24. Ikram, M., Vallina-Rodriguez, N., Seneviratne, S., Kaafar, M.A., Paxson, V.: An analysis of the privacy and security risks of android vpn permission-enabled apps p. 349–364 (2016). <https://doi.org/10.1145/2987443.2987471>
 25. Inc., D.: What is the drift cookie security and privacy policy? (2019), [https://gethelp.drift.com/hc/en-us/articles/360019665133-What-is-the-Drift-Cookie-Security-and-Privacy-Policy-](https://gethelp.drift.com/hc/en-us/articles/360019665133-What-is-the-Drift-Cookie-Security-and-Privacy-Policy-360019665133-What-is-the-Drift-Cookie-Security-and-Privacy-Policy-)
 26. Inc., L.: Privacy policy (2021), <https://www.livechat.com/legal/privacy-policy/>
 27. Inc., T.: Privacy policy (2021), <https://www.tidio.com/privacy-policy/>
 28. Ivanov, S., Webster, C.: Adoption of Robots, Artificial Intelligence and Service Automation by Travel, Tourism and Hospitality Companies - A cost-benefit Analysis. International Scientific Conference “Contemporary tourism – traditions and innovations”, 19– 21 October 2017, Sofia University. pp. 1–9 (2017)
 29. Lal, I.: 12 best chatbots to transform your conversation landscape in 2021 (2021), <https://surveysparrow.com/blog/best-chatbot-platforms/#section2>
 30. Leah: The 9 best chatbots of 2021 (2021), <https://www.userlike.com/en/blog/best-chatbots>
 31. Masood, R., Vatsalan, D., Ikram, M., Kaafar, M.A.: Incognito: A method for obfuscating web data pp. 267–276 (2018)
 32. Michiels, E.: Modelling chatbots with a cognitive system allows for a differentiating user experience. In: PoEM Doctoral Consortium (2017)
 33. Shawar, B.A., Atwell, E.: Chatbots: Are they really useful? LDV Forum **22**, 29–49 (2007)
 34. Starov, O., Gill, P., Nikiforakis, N.: Are You Sure You Want to Contact Us? Quantifying the Leakage of PII via Website Contact Forms. *PETS* **2016**(1), 20–33 (2015)
 35. Venkatadri, G., Lucherini, E., Sapiezynski, P., Mislove, A.: Proceedings on Privacy Enhancing Technologies (2019). <https://doi.org/10.2478/popets-2019-0013>
 36. Werner Geyser: Best ai chatbot platforms for 2021 (2021), <https://influencermarketinghub.com/ai-chatbot-platforms/>
 37. Ye, W., Li, Q.: Chatbot Security and Privacy in the Age of Personal Assistants. Proceedings - 2020 IEEE/ACM Symposium on Edge Computing, SEC 2020 pp. 388–393 (2020). <https://doi.org/10.1109/SEC50012.2020.00057>