

# “AI is from the devil.” Behaviors and Concerns Toward Personal Data Sharing with LLM-based Conversational Agents

Noé Zufferey  
ETH Zurich  
Switzerland  
noe.zufferey@gess.ethz.ch

Karola Marky  
Ruhr University of Bochum  
Germany  
Karola.Marky@ruhr-uni-bochum.de

Sarah Abdelwahab Gaballah  
Ruhr University of Bochum  
Germany  
Sarah.Gaballah@ruhr-uni-bochum.de

Verena Zimmermann  
ETH Zurich  
Switzerland  
verena.zimmermann@gess.ethz.ch

## Abstract

With the increased performance of large language models (LLMs), conversational agents (CA), such as ChatGPT, are nowadays available to any individual requiring little technical knowledge and skills. Initial studies that have investigated related privacy risks primarily focused on either technical aspects and misuse of these tools, or captured overall perceptions of CA users in small-scale qualitative evaluations. Complementing and extending previous work, we used a quantitative user-centered approach to analyze and compare the behaviors and concerns of users and non-users. We conducted a survey study ( $N = 422$ ) with (1) service users, i.e., users of CA services, (2) local users, i.e., users of a local instance of CA (partially local users, or fully local users), and (3) non-users. We collected self-reported usage patterns and personal data-sharing behavior as well as privacy concerns related to different types of personal data (e.g., health data, demographics, or opinions). Furthermore, we analyze individuals’ intention to use CA services in multiple scenarios. Our findings show that users of CA services generally have fewer privacy concerns than non-users. While users rarely share data related to personal identifiers and account credentials, they tend to often share data related to lifestyle, health, standard of living, and opinions. Surprisingly, partially local users tend to share more data with CA services as they also generally use CA services more often and for more diverse purposes. Also, while the majority of CA services users declared not being willing to prioritize CA services as an information source in the described scenarios such as seeking legal advice, between about one-quarter and one-third of partially local users would use CA services for all scenarios. Furthermore, half of the users were willing to stop using CA for privacy reasons (e.g., in case of data leaks), whereas a large majority of non-users reported not using CAs simply because they do not have the need or the opportunity. Our work highlights the high privacy risks for CA services users as CA services largely expand the amount of any type of personal information that can be collected by companies.

## Keywords

LLM, Conversational Agent, Privacy, Chabot, HCI

## 1 Introduction

The performance of artificial intelligence (AI), in particular, Large Language Models (LLMs) has rapidly increased over the last few years. Recently, efficient solutions for text and image generation have been put into practice. For instance, many conversational agents (CA) based on LLMs, such as ChatGPT<sup>1</sup>, are nowadays available to any individual requiring little technical knowledge and skills. Hence, anyone can use them in daily life. Such CAs are used in a variety of domains, including healthcare [51, 53], finance [17], and education [2].

To provide their services, LLM-based CAs need to be trained on huge datasets, consisting of publicly available information as well as data provided by the users, such as prompts, conversations, and uploaded content. Hence, sensitive user data may be stored and used to further refine a CA’s capabilities, as in the case of ChatGPT [39]. This practice, in turn, introduces new risks sensitive data shared by users could be memorized by the CA and unintentionally exposed later on, or even extracted – and possibly misused – on purpose [8, 9]. As a consequence, sensitive details from user conversations could be misused, and/or accessed by unauthorized third parties, or inadvertently embedded in the model. A notable example is the recent Samsung incident<sup>2</sup>, where employees unintentionally leaked confidential source code while using a CA for work-related tasks. The privacy risks are worsened by the human-like nature of LLM-based CA, which could encourage users to disclose more personal information [54].

Several studies have recently investigated privacy risks in the context of LLM-based CAs primarily focused on either technical efforts aimed at measuring [9, 29], mitigating risks during training or related to inferences [15, 24] or captured overall perceptions of CA users in small-scale qualitative evaluations [54]. Whereas these efforts are crucial for providing security and privacy by design, it is also essential to take into account the types of personal data users share and how frequently they share it in order to assess the associated risks. Additionally, understanding people’s privacy

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license visit <https://creativecommons.org/licenses/by/4.0/> or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.



*Proceedings on Privacy Enhancing Technologies* 2025(3), 5–28  
© 2025 Copyright held by the owner/author(s).  
<https://doi.org/10.56553/popets-2025-0086>

<sup>1</sup><https://openai.com/index/chatgpt/>

<sup>2</sup><https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/> (last-accessed 25-Nov-2024)

perceptions in-depth also helps researchers develop better LLM-based tools that protect the privacy of individuals while maintaining a reasonable level of utility (i.e., how beneficial is the use of a given tool) and user adoption [18].

In this article, we present a large-scale online survey study to analyze how, why, and in what context individuals would or would not interact with LLM-based CA. We specifically quantify the perceptions of CA users and also non-users to capture differences among them. We also consider individuals using their own local instance of LLM-based CA. Indeed, with some affordable (e.g., 2-core CPU, 2GB RAM, and 5GB storage) one can run a decent CA using locally installed software such as *AnythingLLM*<sup>3</sup> and some open-source models such as *LLama3.1 8B*. In this work, we address the following research questions:

- **RQ1:** What are the differences between the privacy concerns of users and non-users in the context of LLM-based CAs, as well as between CA service-only users and users who also use a local instance of CA?
- **RQ2:** What types of personal data do or would individuals reveal to LLM-based CAs, and what are the differences regarding their usage habits?
- **RQ3:** What is the impact of the granularity level of data on individuals' willingness to share personal data with LLM-based CAs?
- **RQ4:** What type of information source (including CAs) do individuals prioritize to solve a given problem in their daily life and how does it relate to general privacy concerns and data-sharing behaviors?

To investigate these research questions, we conducted an online survey with N=422 participants. The survey differentiated between (1) individuals who use CA services, (2) individuals who, at least partially, use local instances of CAs (i.e., locally processed model and no cloud storage), and (3) non-users of CAs.

We identified and quantified key differences between these three groups and assessed how privacy concerns might influence the adoption of CA services. To do so, we collected self-reported personal data-sharing behavior related to different types of personal data (e.g., health data, personal identifiers, demographics, or opinions). Our results show that users rarely share data related to 📄 Identifiers and General Information, as, for example, 📄 Personal Identifiers and 🔑 Account Credentials. However, they tend to often share other types of data, including multiple pieces of information related to ❤️ Lifestyle and Health, 👤 Personal Characteristics and Emotions, and 🏠 Standard of Living and Opinions. Furthermore, for most data types, there is a significant difference in data-sharing behavior between service users and partially local users, i.e., individuals who use a local instance of a CA in addition to their usage of CA services. Surprisingly, the former tend to share more data with CA services. These results are consistent with our findings about usage patterns, which show that partially local users also generally use CA services more often and for more diverse purposes.

We also analyze individuals' intention to use (or not use) CA services in multiple scenarios involving the respective types of personal data. Our results, on the one hand, reveal a general trend regarding the relationships of individuals with LLM-based CAs and,

on the other hand, quantify the main concerns that could deter someone from using these tools. In particular, most CA service users declared not willing to prioritize the use of CA services as an information or advice source in the described scenarios. Nevertheless, between about one-quarter and one-third of partially local users reported that they would actually do so for all scenarios.

In complement to these quantitative results, open-text questions about reasons for using or not using CA services revealed key insights. Half of the users are willing to stop using CA for privacy reasons (e.g., in case of data leaks), whereas a large majority of non-users declared they simply do not use CAs because they do not have the need or the opportunity. However, our findings about how privacy of CA services having direct access to given data types show that users of CA services generally have fewer privacy concerns than non-users.

Whereas a few of the previous studies partially addressed similar questions, none of them explored these aspects using a quantitative approach. In this article, we report a study with quantitative results about chatbot users' data-sharing behavior and concerns to fill this gap. Moreover, we also compare the behaviors and concerns of users with those of non-users and users of local instances, highlighting the differences between these groups. This comparison provides valuable insights into the relation between privacy concerns, usage purposes, and usage behaviors.

## 2 Related Work

Many studies have investigated the users' data-sharing behavior and privacy concerns in the context of different emerging technologies, e.g., overall investigations of smart cities [49], smart homes [55, 56], messaging apps [27], or social media [30]. Other studies particularly focused on specific data types, such as browsing history [10], geolocations [43], as well as health and fitness data in multiple contexts including social media [31], pandemics [7], wearable activity trackers [58], and donation apps [20].

### 2.1 Security & Privacy in LLM-Based CAs

Since LLM-based CAs are still relatively new, research on the security and privacy risks these CAs may pose remains limited compared to other technologies. This subsection summarizes research on the potential privacy threats related to the use of CA services.

Carlini et al. [9] explored how training data can be extracted from LLMs with a case study on GPT-2. They showed that pieces of data from the training set can be extracted directly from the model and that such data may contain multiple types of personally identifiable information, conversations, and even universally unique identifiers. Additionally, Michelle et al. [37] explored how ChatGPT could be used to identify individuals' gender based on their names and country of residence, demonstrating the potential of LLM-based tools to infer user's personal data. Moreover, Staab et al. [48] highlighted that LLMs can be used to deduce personal information, such as gender, location, or age from written text. Even more recently, Peters and Matz [42] studied how LLMs, such as GPT-3.5 and GPT-4, could be used to infer personality traits like extraversion, neuroticism, and openness from Facebook status.

<sup>3</sup><https://anythingllm.com/> (last-accessed 17-Feb-2025)

Chen and Esmaeilzadeh [11] discussed the security and privacy issues that arise when using LLM-based technologies in a health-related context. They classified multiple types of medical data and AI application domains and presented the potential risks for each of them. Noticeably, they highlighted risks linked to data breaches, such as patient re-identification in the context of LLM-based medical diagnosis tools and virtual health assistants.

As for them, Matz et al. [34] studied how LLMs could be used to generate personalized content to influence individuals. They also showed how personal data, such as personality traits, can be leveraged by LLM-based CAs to make these tools more convincing and improve their chances of persuading the user (e.g., for targeted marketing).

Wu et al. [52] explored the potential risks of integrating ChatGPT into daily life, focusing on security, privacy, and ethical concerns. They found that ChatGPT poses new security threats, such as aiding in generating attack codes and phishing websites. They also highlighted privacy violations due to data collection and identified concerns around transparency.

## 2.2 User Studies on LLM-Based CAs

There are multiple user studies that have been conducted on LLM-Based CAs. For example, Miyazaki et al. [38] analyzed tweets (now X posts) about CAs finding an overall positive perception of such tools by individuals. Further, Skjuve et al. [47] investigated the main characteristics of interactions between individuals and ChatGPT and mostly qualitatively focused on overall user (un)satisfaction. Another user study on user satisfaction was done by Deng et al. [14]. They did a sentiment analysis of conversations between ChatGPT and users to evaluate the users' satisfaction with their usage of the CA. Most of the user's messages were evaluated as positive or neutral, despite a non-negligible amount of negative comments.

Trust in LLM-based CAs has been examined in many studies. Jung et al. [26] conducted a focus group and interview study to understand to what extent ChatGPT users tend to trust LLM-based CAs in the context of information seeking. Most of their respondents tended to express more trust for search engines or crowd-sourced encyclopedias (i.e., Wikipedia) than for CAs. Additionally, Dekkal et al. [13] studied the relation between the perceived trust in CAs and their adoption. Their findings show that the feeling of creepiness [36] when using CAs has a negative impact on trust and thus adoption. The results also indicate that privacy concerns have a negative impact, although smaller.

Shahsavari and Choudhury [46] conducted a survey study to understand the intention to use ChatGPT in the context of health-related information gathering. Their findings show that most of their respondents were inclined to use CA services for self-diagnosis. In their article, they also highlighted the importance of multidisciplinary collaboration to develop safer LLM-based tools.

Few studies have focused on privacy. In their literature review, Leschanowsky et al. [28] found a lack of research about privacy concerns towards CAs while stressing the importance of such studies. They argue to better understand CA users regarding privacy concerns to gather valuable insight to develop trustworthy tools. Also, Zhang et al. [54] conducted an exploratory analysis of conversations with CAs as well as a qualitative study to better understand the

privacy-utility trade-off faced by CA users analyzing their concerns and mental models with semi-structured interviews (N=19). Despite multiple concerns, many users shared personal data, probably due to a misunderstanding of the functioning of CAs.

**Summary.** Related work highlights significant gaps in understanding privacy concerns and data-sharing behavior in the context of LLM-based CAs, with only a few studies addressing these issues. While prior research has explored some aspects of privacy concerns, there is limited knowledge about the differences in these concerns and behaviors between users and non-users. Furthermore, no research has investigated distinctions between users of online-only CA services and those who also use local instances of CAs, or examined how individuals prioritize CAs as an information source in relation to privacy concerns. To address these gaps, we examine differences in data-sharing behavior and privacy concerns across user types, the influence of usage habits on sharing, the impact of data granularity on willingness to share, and the prioritization of CAs as an information source. Our study provides a large-scale quantitative analysis, in contrast to relevant studies that primarily rely on qualitative evaluations, such as interviews and focus groups. We believe that these quantitative results will bring valuable information to guide future research about CA privacy and security. Especially, they provide interesting insights into the most sensitive types of data shared with CAs, the associated risks of what individuals are already sharing, as well as essential information for developing privacy-enhancing technologies (PETs).

## 3 Methodology

To answer our research questions, we primarily collected quantitative data on data-sharing attitudes, practices, and privacy concerns in the context of CA services, along with qualitative data to help us interpret the quantitative results in an online survey ( $N = 422$ ).

For our survey, we define three main types of respondents: (1) service users, i.e., those who only use LLM-based CA services, such as ChatGPT, (2) local users, i.e., those who use a local instance of LLM-based CA (they include partially local users, who use both CA services and their local instance, but also fully local users), and (3) non-users, i.e., those who never tried interacting with LLM-based CAs, or those who tried but declared not using them.

In the following, we detail the surveyed data-sharing scenarios and data types, the study procedure, the recruitment process, and sample, as well as ethical considerations.

### 3.1 Personal Data Types and Scenarios

The survey included questions about various types of personal data and different scenarios describing situations in which individuals might be inclined to use CA services.

**Personal Data Types** were defined during a literature-driven iterative coding process based on multiple research articles and GDPR<sup>4</sup>. The first author of this article initially selected recent related research articles [4, 5, 12, 21, 45, 50] as well as GDPR examples of personal data. From these sources, they compiled a list of all cited data types. After that, they proceeded to a first inductive coding round by creating main categories to classify and merge related data

<sup>4</sup>General Data Protection Regulation - <https://gdpr-info.eu/>

types [35]. Next, the second author of this paper revised the initial classification. A merged version was subsequently reviewed by the last author of this paper. Finally, the first and second authors of this paper discussed the relevance of this classification and made minor adjustments. At the end of this process, we eventually defined four main data type categories: 🗂️ *Identifiers and General Information*, ❤️ *Lifestyle and Health*, 👤 *Personal Characteristics and Emotions*, and 🏠 *Standard of Living and Opinions*. Each of these categories consists of five subcategories. All categories and subcategories are listed in Table 4 in Appendix A.

**Data-sharing Scenarios** were defined during a focus group session involving the first and last author as well as four members of their institution who were not involved in this study. Participants of this session were asked to think of a situation in which they needed help and/or external information to solve a personal situation. They wrote down the situation as well as where/from whom they would seek help first. Finally, they all had a look at another person’s problem and added potential information/help sources. The first author collected all ideas and compiled them into multiple scenarios with four categories of information sources: CA services (defined as “chatbot” in the questionnaire), search engines (e.g., Google Search), acquaintances (e.g., friends), and experts (e.g., physicians). In total, six different scenarios were displayed in the questionnaire (at least one for each main data type category). The scenarios were the following: (1) *looking for a diagnosis based on symptoms*, (2) *checking whether some action is legal*, (3) *advice for a date*, (4) *assessment of personality traits*, (5) *looking for advice about money investment*, and (6) *looking for advice about professional orientation*.

### 3.2 Procedure

In the following, we explain each section of the survey in detail. Before taking part in the main survey, participants underwent a screening through a screener survey detailed in Section 3.3. The whole main questionnaire and the corresponding answers are available in [Supplementary Material](#).<sup>5</sup>

**Sec. A: Introduction.** The respondents were first presented with an informed consent sheet. If they agreed to participate in the study, they were asked to answer the questions of the screener survey again to account for potential changes, such as trying CAs after the initial screening. The screener survey questions included whether they use LLM-based CAs, and if so, whether they use CA services, a local instance, or both. Next, we asked CA service users questions about CA service usage: how frequently they use CA services, how much time they spend on average, and their primary purposes of CA usage. We also asked local users why they (sometimes) prefer to use a local instance rather than CA services. As for non-users, they were asked why they do not use CAs (open-ended question). For the following questions apart from those specifically related to local usage of CA, we asked local users to only take into consideration online CA services, and not their local instance (as it does not raise any privacy issue).

**Sec. B: Data Sharing.** This part of the survey was only displayed to CA users. It consisted of four sub-parts each focusing on one of the

four categories of personal data with five different types related to each category (see Section 3.1 for details). For each type, we asked respondents to rate how frequently they share related data on a five-point scale (ranging from “I never share” to “I usually share”). We asked to take into consideration the data-sharing frequency related to the number of times when it would be relevant to share. For example, we asked them to select “I usually share” if they usually share a given type of data when it is relevant to their CA usage. We proceeded like this as the sharing behavior of a given data type is related to a specific context, i.e., individuals will not just randomly share data, but do it for a purpose.

**Sec. C: Concerns.** This part was divided into four sub-parts similar to the previous part. For each data type presented in each sub-part, we asked respondents to answer, on a five-point scale, how worried they would be if a CA service had access to this data (from “Not worried at all” to “Extremely worried”).

**Sec. D: Granularity.** We then presented five different examples of pieces of personal data (e.g., homeplace, consumption habits). For each data type, we presented multiple sharing options varying by granularity (e.g., for homeplace: exact address, town, county/region, country, continent) and asked for each of them, on a five-level scale how comfortable they would be if they had to share it (from “Not at all comfortable” to “Extremely Comfortable”). We ensured that for each (see Section 3.1), there was at least one corresponding example of a related data type.

**Sec. E: Scenarios.** After that, we investigated six different scenarios that covered each of the four main data categories at least once (see Section 3.1). All these scenarios described an everyday situation for which they would need to ask for help or seek information to resolve it and/or make a decision. For each of these scenarios, we presented multiple information/help sources (i.e., CA service, search engine, someone close, and professional assistance). The respondents were asked to sort the sources from “most likely” to “least likely” to be used by them to solve their problem.

**Sec. F: Mind Changing.** We ended the CA-related parts by asking two open-ended questions: First, users and partially local users were asked what would make them stop using CA services, and non-users what would make them start. Second, all respondents were asked if they had any additional comments on CAs and privacy.

**Sec. G: IUIPC and Demographics.** To measure differences in general privacy concerns between users and non-users, we used the 8-item IUIPC [23]. Finally, we collected multiple demographics.

In all previously described survey parts, each time we presented questions related to data types, the order of the data types was randomized to avoid any bias in the aggregated results due to the survey structure.

### 3.3 Recruitment & Screening

We ran an a priori power analysis to set the number of respondents. For a medium effect size and  $1 - \beta = 0.99$  probability of type II errors, *G\*Power* [19] estimated significant results ( $p < 0.05$ ) with groups of a size of  $N = 127$ , resulting in a total of 381 required participants. This number is in the same range as previous survey studies on similar topics (cf., Liao [31]:  $N = 553$ , Velykoivanenko et al. [50]:  $N = 227$ ).

<sup>5</sup>URL: <https://osf.io/mqwjr/>; Note that, as some questions are randomized or conditionally displayed, they were not displayed to respondents in the same order as presented in this material.

We recruited our survey respondents via Prolific, which is considered a reliable crowdsourcing platform for scientific research [41]. Before deploying our main survey, we conducted a screening survey to ensure a balanced representation of CA usage. While Prolific’s default screening options include CA usage, our additional screening helped gather the most accurate responses. That is because it allowed us to exclude participants who may have changed their CA usage (e.g., started or stopped using CAs) since completing Prolific’s screening question, as well as those uncertain about their usage, to avoid potentially confusing answers.

To ensure a joint understanding, we first introduced a definition of CA services before asking three screening questions: (1) whether they had used CA services (2) how many days per week, on average, they interact with CA services, and (3) apart from their (non)-usage of CA services, whether they use any local CA instance.

We collected the data of  $N = 1993$  respondents. This enabled us not only to select eligible respondents but also to compute general statistics on the general usage of CAs.

**3.3.1 Screener Survey Sample.** We first deployed our questionnaire worldwide to individuals speaking English fluently using Prolific’s pre-screening options and collected answers 1993 people. Among these individuals, 86% of them declared using LLM-based CA services at least a few times a year, 6% declared to exclusively use their own local instance of CAs, 7% declared to not use CAs (either they never tried, or tried but do not usually use), and 1% of the respondents did not know if they use CAs or not. Surprisingly, among CA services users, 36% declared also to use their own local instance. Regarding their living place, 32% of the screener participants were from European countries other than the UK, 23% were from the UK, 20% were from South Africa, 10% were from the USA, and 5% are from Canada. The remaining 6% were from diverse countries around the world. After analyzing the usage habits of CAs regarding countries, we found that respondents from South Africa highly contributed to the high proportion of local users, as 60% of them declared to use their own local instance of CAs in addition to CA services, and 15% of them declared to use their own local instance exclusively.

After analyzing these records, we decided to deploy the screener survey again in order to collect data from more non-users. To achieve this, we added to Prolific’s pre-screening options to only deploy the survey for people who did not declare using CA services. This resulted in 200 new answers, through which we identified 85 potential additional non-users for the main survey.

We then deployed the main survey to the selected screener respondents, trying to keep a balance between service users, (partially) local users, and non-users.

**3.3.2 Main Survey.** For the main survey, we collected answers from 488 respondents and after data cleaning, we ultimately kept data from 422 respondents. In the data cleaning process, we excluded 66 respondents who completed the survey in less than three minutes, and whose answers to the open-ended questions did not make sense [33], e.g., who provided random words or characters, or who failed one of the two attention-check questions.

The final sample answered the survey in 15 min and 22 sec on average with a standard deviation of 8 min and 25 sec. Among these individuals,  $N = 138$  of them reported using LLM-based CA

services only (i.e., no local instance) at least a few times a year,  $N = 122$  reported using their own local instance of CAs in addition to CA services,  $N = 41$  reported using only their own local instance,  $N = 121$  reported not using CAs (either they never tried, or tried but do not use).

In terms of living countries, 32% of the respondents were from the UK, 24% were from other European countries, 23% were from South Africa, 8% were from the USA, and 6% were from Canada. The remaining 7% were from diverse countries around the world or did not report their country.

Regarding their gender, 228 respondents (54%) declared being women, 183 of them (43%) declared being men, 6 respondents declared being non-binary or both a man and a woman, and one preferred not to disclose their gender.

### 3.4 Survey Refinement and Data Reliability

Before launching the survey, we performed three in-person cognitive pretests to identify any issues with its design or instructions with researchers from the first author’s institution. Each time, the test subject was instructed to rephrase the questions, in their own words, to describe what they thought was asked, then to answer it. Between each test, we slightly improved the phrasing of the questions for which this was necessary. This process showed that the survey instructions and questions were overall clear with only a few minor improvements necessary.

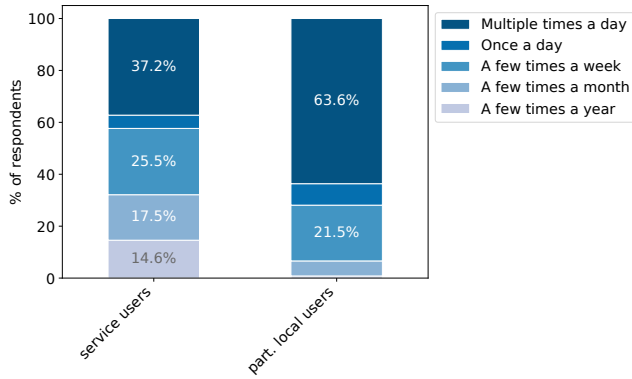
### 3.5 Ethical Considerations

We followed the ethical guidelines established by our institutional ethics committees. These guidelines require that user studies limit the collection of personal data to ensure the privacy of participants. Before answering the survey, respondents need to be informed about and agree to participate in a study through a consent form outlining the conditions of participation, details regarding the data being collected, the procedure for withdrawing from the study, and information about financial incentives. The institutional review board (IRB) at our institutions reviewed and approved the consent form and the study design itself. Following Prolific’s recommendations, we paid the screener respondents £ 0.15, and the respondents of the main survey £ 3 who answered correctly to at least one of the checks in line with Prolific guidelines (even if we did not keep any of their data).

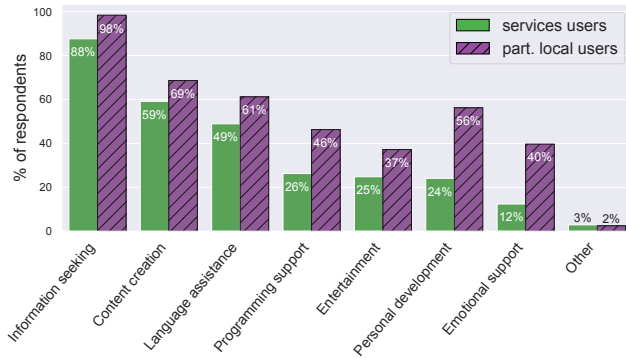
## 4 Results

This section presents the survey results, starting with the frequency and main usage purpose of CAs followed by data-sharing behavior with CA services and related privacy concerns. Finally, we detail results related to preferences regarding data-sharing and information sources. Given the well-defined research questions and simple, straight-forward responses in the open-ended questions, the first author coded the answers based on a codebook developed through open coding [44]. Then, the second author reviewed and provided feedback in line with the recommendations provided by Orloff et al. [40].





**Figure 1: Self-reported usage frequency of CA services.** The figure compares the data-sharing behavior of service users ( $N = 138$ ), and partially local users ( $N = 122$ ).

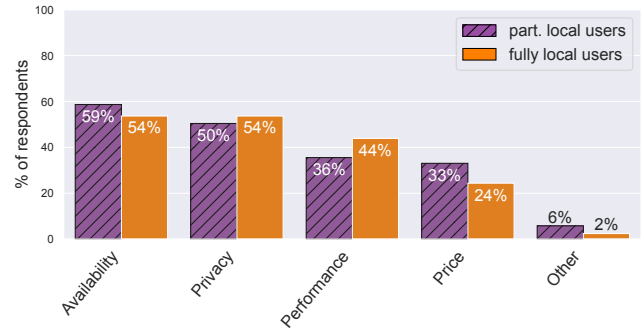


**Figure 2: Self-reported usage purpose of CA services.** The figure compares the data-sharing behavior of service users ( $N = 138$ ), and partially local users ( $N = 122$ ).

#### 4.1 Usage Frequency and Purposes

To analyze the self-reported usage frequency and purposes with LLM-based CA services, we differentiated service users ( $N = 138$ ), i.e., individuals who only use CA services, such as ChatGPT, and partially local users ( $N = 122$ ), i.e., individuals who *also* use a local instance of CA (this use is complementary to their use of CA services). The latter were explicitly asked to report only their usage frequency of *CA services only* and not to include their usage behavior toward local instances in the report. *ChatGPT* was by far the most used CA service with 91.9% of service users and partially local users declaring using it whereas *Copilot* was used by 46.9% of them, *Gemini* by 46.5% of them, and *Claude* by 13.1% of them. Other CA services such as *GigaChat*, *Grok*, *Le Chat*, and *Ninja* have been cited a few times.

**4.1.1 Partially local users use CA services more frequently.** As depicted in Figure 1, partially local users, perhaps surprisingly, reported using CA services substantially more frequently than service users. Specifically, more than 70% of partially local users declared using CA services at least once a day, whereas only slightly more

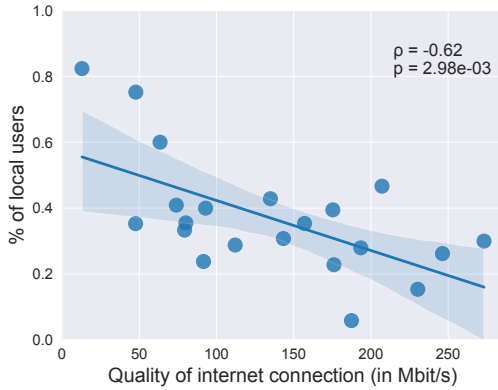


**Figure 3: Comparison of self-reported usage purpose of CA local instance, i.e., reasons why they (sometimes) use local instance instead of CA services.** The figure compares the data-sharing behavior of fully local users ( $N = 41$ ) and partially local users ( $N = 122$ ). There is no statistical difference in both groups.

than 42% of service users reported a similar frequency. Furthermore, almost 15% of the service users stated using CA services only a few times a year, whereas almost none of the partially local users (0.8%) reported such a low usage frequency. This difference in usage frequency of CA services might be due to local users being part of the most enthusiastic individuals about CAs, and, thus, are willing to test more diverse use cases and CA options. However, this interestingly also shows that using a local instance does not necessarily detain users from using CA services. This hypothesis is supported by Figure 2, which shows that partially local users have more diverse usage purposes for CA services. Therefore, our findings suggest that individuals who use local instances of LLM-based CAs, do not necessarily do it *instead* of using CA services, but more likely *in addition* to them. Noticeably, more than half of the partially local users use CA services for personal development (e.g., life coaching, goals setting) and 40% of them for emotional support (e.g., talking, advice seeking).

**4.1.2 Local users mainly aim for availability.** To better understand the local users' motivations, we asked them about the purpose of using CA local instances, more specifically, why they (sometimes) use local instances instead of CA services. The main reason for this is availability, as shown in Figure 3. We calculated a correlation (see Figure 4) based on the 21 most represented countries in our first-screener batch (which only considered language as a pre-screening criterion). The correlation suggests that the lower the internet connection quality in a country, the more likely citizens are to use their own local CA instance. Hence, the lack of internet connection in some countries might also lead some users to rely on local tools rather than on cloud/online-based ones. We also noticed that privacy is one of the most reported reasons for using local CAs. Perhaps surprisingly, half of the partially local users mentioned privacy as a reason for using a CA local instance, even though results regarding data sharing behavior with CA services (see Section 4.2) show that partially local users tend to share more personal data

<sup>6</sup><https://www.speedtest.net/global-index>, accessed on Nov. 14, 2024



**Figure 4: Correlation between the ratio of citizens who use a local instance of CA and the average quality internet connection in the country<sup>6</sup> for the 21 most represented countries in our first screener (no pre-screen except for English-speaking) dataset.**

**Table 1: Reasons for non-users not to use CA services ( $N = 103$ ).**

Reason for not using CA services	% of answers
No need or opportunity	65% ( $n = 67$ )
Lack of trust	10% ( $n = 11$ )
Lack of knowledge	8% ( $n = 8$ )
Prefer human contact	7% ( $n = 7$ )
Ethical concerns	6% ( $n = 6$ )
Afraid of using	4% ( $n = 4$ )

with CA services compared to service users. On the one hand, this could be due to the well-known privacy paradox [3, 22]. On the other hand, this may be because the questionnaire included predefined multiple-choice options (i.e., checkbox) for this question. An open-text question might have revealed fewer privacy-related reasons. Overall, regarding the purpose of using a CA local instance, we did not find any significant differences between partially local users and fully local users. However, this may be due to the small sample size for the latter group.

**4.1.3 Non-users mainly have no opportunity of use.** In addition, we also asked non-users why they prefer not to use CA services. Table 1 shows the categorization of the different answers to this question. Our results suggest that a large majority of non-users (65%) do not use CA services because they do not need to or have never had the opportunity. [W, 40-49 y.o., Europe]: “I do not need it as I try to do the tasks myself.” Only 10% of the respondents reported not using CA services because they do not trust them, either related to their personal data usage or to the reliability of their answers. [W, 18-29 y.o., Europe]: “I don’t trust them with my data, or to provide accurate information.” Other respondents (6%) did not directly express a lack of trust but reported not using CA services for ethical reasons. [M, 21-29 y.o., Europe]: “I do not support the idea (morally and environmentally) of this type of AI, plus I prefer to write myself.”<sup>7</sup>

<sup>7</sup>a few typos in this sentence have been corrected by this article’s authors.

Also, 8% of the respondents simply do not know how to use CAs, and 7% reported preferring human contact. Finally, 4% declared being scared of using such tools. [M, 18-29 y.o., USA]: “AI is from the devil.”

## 4.2 Data-Sharing with CA Services

Our results on data-sharing behavior with CA services confirm the trend we observed regarding the difference in CA usage between partially local users ( $N = 122$ ) and service users ( $N = 138$ ). Partially local users were specifically instructed to report only their data-sharing behavior with CA services, excluding their interactions with their local instance.

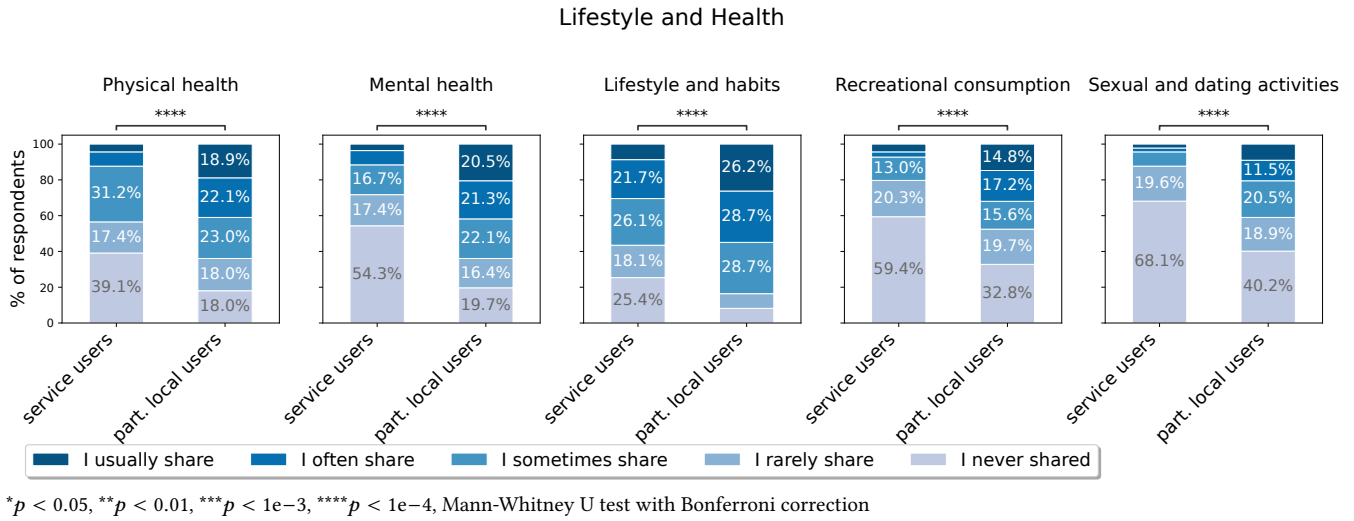
The findings, as illustrated in Figure 5 and 6, show that local users share significantly more data with CA services about their Lifestyle and Health, and also about their Standard of Living and Opinions ( $p < 1e-4$ , Mann-Whitney U test with Bonferroni correction). For nearly all types of personal data, partially local users reported sharing more data with CA services than service users, with statistical significance. The largest differences were observed in mental health and religion-related data. Specifically, 54.3% of service users reported never sharing data related to Mental Health, compared to 19.7% of partially local users. Similarly, 62.3% of service users reported never sharing data related to Religion, compared to 29.5% of partially local users. Generally, the results suggest that most of the individuals who use local instances of CA, do not seem to do that to avoid sharing some pieces of personal information with CA services.

Noticeably, Identifiers and General Information is the category of data types generally shared less often. For example, 92% of service users and 83% of partially local users reported having never shared data related to Account Credentials, and 94.6% of service users and 77% of partially local users reported having never shared data related to Criminal Records.

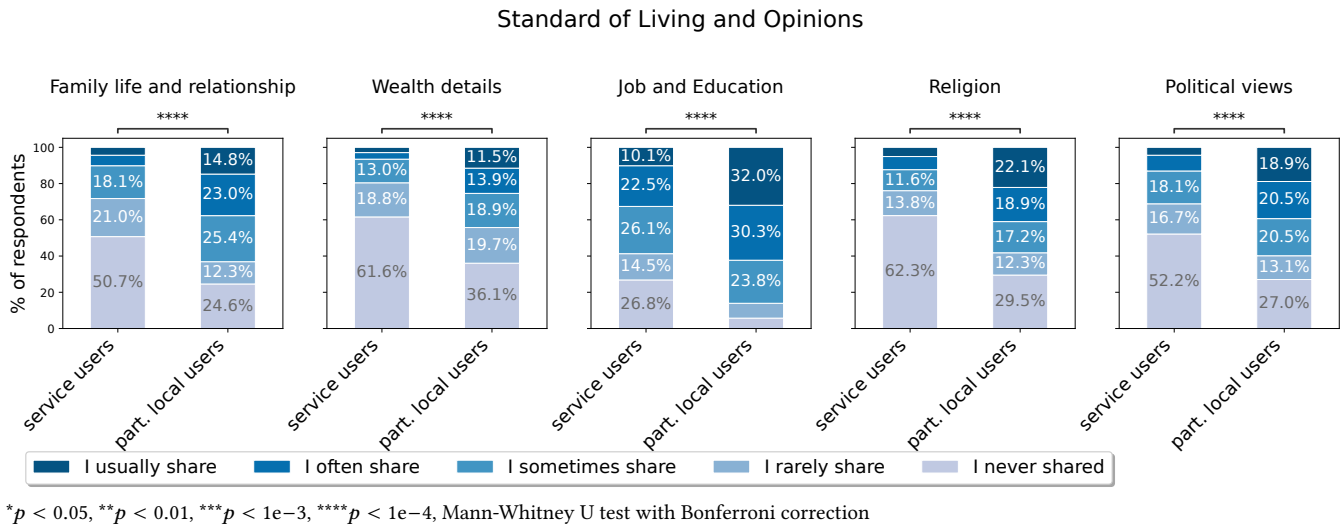
Apart from the difference between partially local users and service users, we can also notice that some data types are widely shared with CA services in general. For example, almost three-quarters of service users and more than 85% of partially local users reported having already shared data related to their Gender, Lifestyle and Habits, and Job and Education.

It is important to note that participants reporting certain data types as never shared does not necessarily indicate that they perceive this data as highly sensitive or are unwilling to share it. Instead, this could be because participants use CA services primarily for tasks that do not involve sharing such data types. For instance, data about religion was frequently reported as not shared (see Figure 6); however, most participants indicated they were not worried at all about sharing this type of data (see Figure 9).

**4.2.1 Differences between Countries and Regions.** The results about data sharing are in general similar across all the studied regions. This is for example the case about the tendency of partially local users to share more data than service users. However, when looking at differences between the four most represented regions in our dataset, i.e., Europe (Apart from the UK), UK, South Africa, as well as USA and Canada (both together), we can note that, in general, individuals from Europe tend to share fewer data than the others, especially compared to individuals from South Africa. For example,



**Figure 5: Self-reported data-sharing frequency of information related to lifestyle and health. The figure compares the data-sharing behavior of service users ( $N = 138$ ), and partially local users ( $N = 122$ ).**



**Figure 6: Self-reported data-sharing frequency of information related to standard of living and opinions. The figure compares the data-sharing behavior of service users ( $N = 138$ ), and partially local users ( $N = 122$ ).**

Figure 7 compares the data-sharing behavior related to lifestyle and health of all user types (services users as well as partially local users) and compares the results of individuals living in UK ( $N = 63$ ), Europe ( $N = 77$ ), South Africa ( $N = 68$ ), and USA or Canada ( $N = 30$ ).

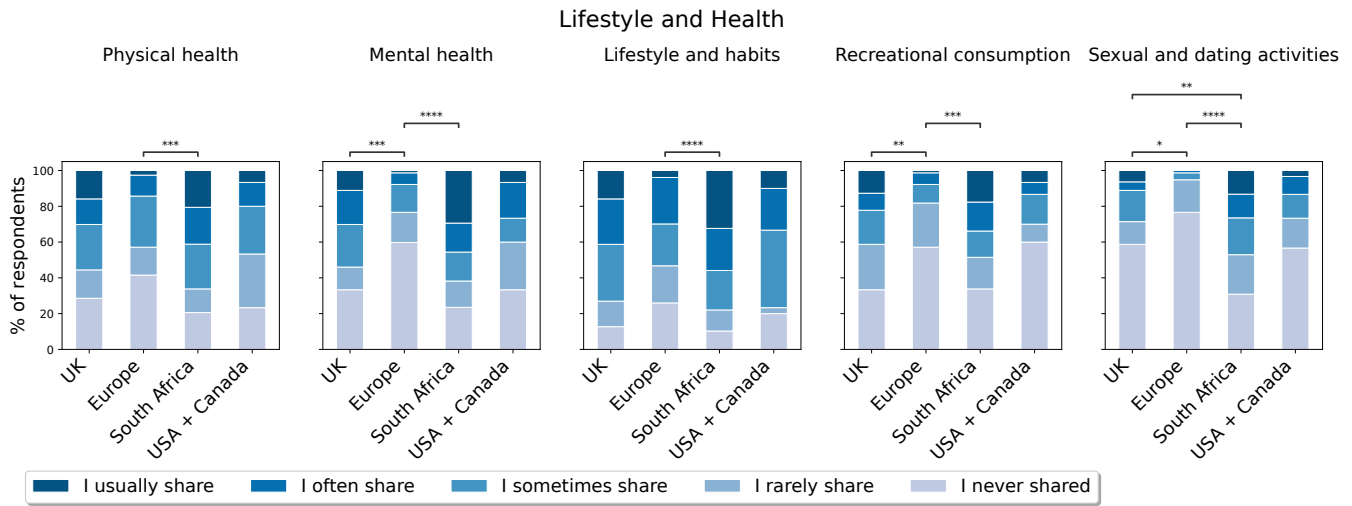
### 4.3 Privacy Concerns

In this section, we describe our findings related to privacy concerns of all investigated user types. Local users were explicitly asked to report only their privacy concerns toward *CA services only*.

**4.3.1 Non-users express more concerns than users.** Our findings demonstrate highly pronounced differences in privacy concerns

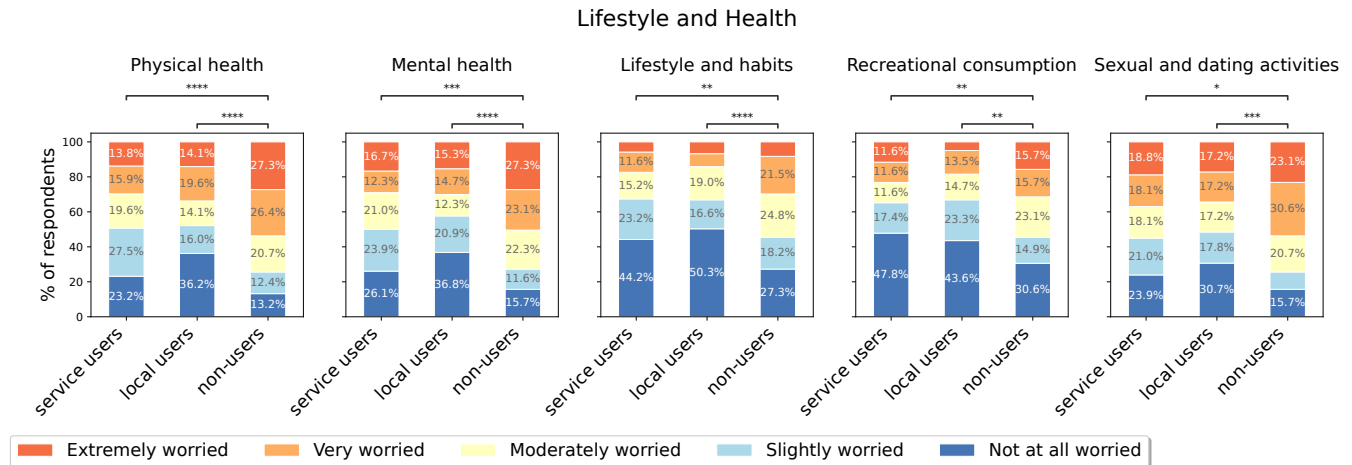
about CA services having access to personal data. Three-quarters of the non-users are at least moderately worried about the fact that CA services could have access to data related to Physical Health, whereas it is only the case for less than half of service users and local users (see Figure 8). This difference between users (service users and local users) and non-users is specifically strong regarding data related to lifestyle and health where there is a significant difference ( $p < 0.05$ , Mann-Whitney U test with Bonferroni correction) for all related personal data types. This is not the case for the other categories (see Figures 9,17, 18). Perhaps surprisingly, there is no significant difference in concerns between service users and local users for any of the different data types. As previously described,






\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 1\text{e-}3$ , \*\*\*\* $p < 1\text{e-}4$ , Mann-Whitney U test with Bonferroni correction




**Figure 7: Self-reported data-sharing frequency of information related to  lifestyle and health. The figure compares the data-sharing behavior of both types of individuals who use CA services (service users as well as partially local users) and compares the results of individuals living in UK ( $N = 63$ ), Europe ( $N = 77$ ), South Africa ( $N = 68$ ), and USA or Canada ( $N = 30$ )**




\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 1e-3$ , \*\*\*\* $p < 1e-4$ , Mann-Whitney U test with Bonferroni correction

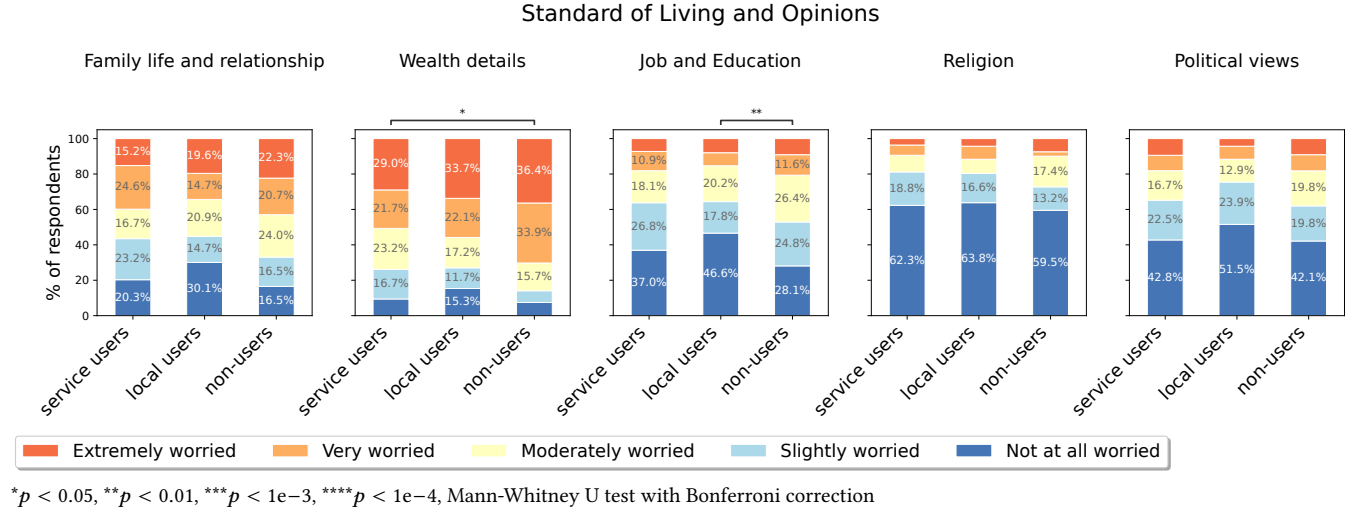
Figure 8: Self-reported privacy concerns about CA services having access to personal data related to  lifestyle and health. The figure compares the privacy concerns of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ).

partially local users tend to share more data. However, this lack of difference in concerns may be due to the fact that service users use CA services less often, and for fewer purposes, so they need to share less personal data to achieve their goals. Hence, our findings suggest that privacy concerns and the data-sharing behavior of CA users are, in general, well-aligned.

Apart from  Personal Identifiers, and  Account Credentials, which may cause direct obvious security threats, the types of data individuals reported to be the most concerned about being accessible by CA services are  Location and Mobility with 74.6% of service users, 71.2% of local users, and 85.9% of non-users declaring

being at least moderately concerned, as well as  Wealth Details with 73.9% of service users, 73.0% of local users, and 86.0% of non-users declaring being at least moderately concerned. These findings are in line with prior work about privacy concerns as these data types were also shown as considered the most sensitive ones by individuals [12]. Figure 17 and 18 in Appendix C show complementary results about privacy concerns for other types of personal data.

4.3.2 *Non-users tend to have slightly higher IUIPC-8 scores.* In addition to concerns about data types, we measured the general privacy



**Figure 9: Self-declared privacy concerns about CA services having access to personal data related to 🏠 Standard of Living and Opinions.** The figure compares the privacy concerns of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ).

**Table 2: IUIPC-8 scores for all three factors, i.e., collection, control, and awareness.** For each factor, values with a  $\dagger$  differ with weak evidence  $p < 0.1$  and the ones with a \* symbol differ significantly  $p < 0.05$  (t-test).

types of user	Collection		Control		Awareness	
	mean	std	mean	std	mean	std
service users	4.55 $\dagger$	1.76	4.75 $\dagger$	1.02	5.38	0.82
local users	4.66	1.25	4.79	1.08	5.17*	0.77
non-users	4.75 $\dagger$	1.31	4.93 $\dagger$	1.01	5.48*	0.82

concerns of service users, local users, and non-users, with the IUIPC-8 scale [23]. Table 2 shows scores of our respondents for all three IUIPC factors, i.e., collection, control, and awareness. Our results show weak evidence ( $p < 0.1$ ) (t-test) of difference between service users and non-users for the collection and control factors, whereas they show a significant difference ( $p < 0.05$ ) (t-test) between local users and non-users for awareness, which may suggest a (slight) correlation between (high) privacy concerns and the (non)-use of CA services.

#### 4.3.3 Individuals express more concerns about specific examples.

All types of individuals showed more concern when we asked them to evaluate how comfortable they would be to share more specific types of personal data (i.e., age, consumption habits, home place, lifestyle and hobbies, and political views). This might be because they had to evaluate their concerns in the context of more concrete examples. For example, whereas 65.3% of service users, 75.4% of local users, and 61.9% of non-users reported being at not all worried or slightly worried with a CA service having access to data related to their 🗳️ Political Views (see Figure 9). Figure 10 shows that only 19.6% of service users, 34.3% of local users, and 19.8% of non-users reported being very or extremely comfortable with the idea of sharing the list of votes and/or political support to CA service.

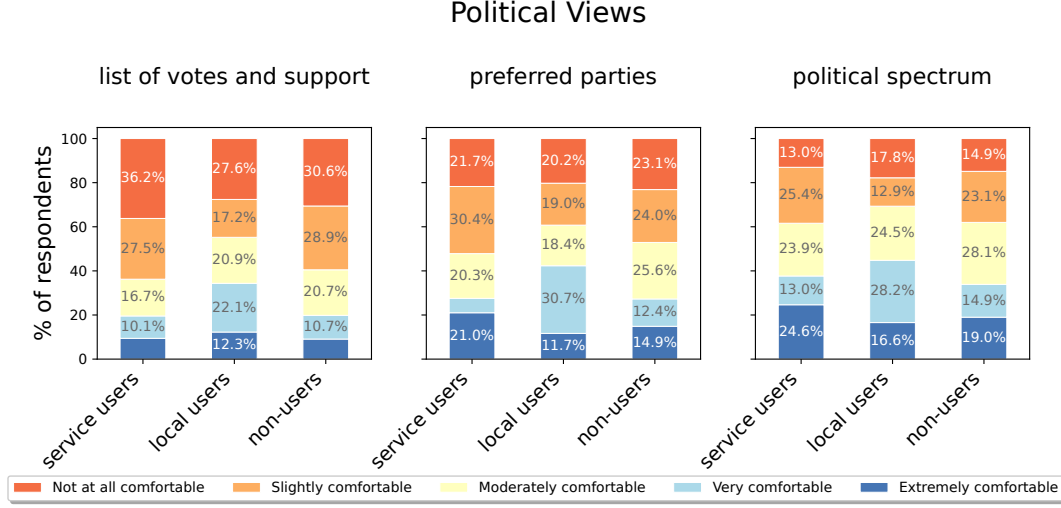
**Table 3: Reasons for users (services users as well as partially local users) to stop using CA services ( $N = 258$ ).**

Reasons to stop using CA services	% of answers
Privacy risks	53% ( $n = 136$ )
Untrustful information	17% ( $n = 43$ )
Price increase	15% ( $n = 38$ )
Unsatisfactory performance	15% $n = 12$
Ethic	2% ( $n = 6$ )
Other	8% ( $n = 20$ )
Nothing	14% ( $n = 36$ )

Furthermore, we found that the lower the granularity of the shared data, the more comfortable they were to share (see Figure 10 and Figures in Appendix D), which is in line with prior works that have shown that individuals are more inclined to share personal data with a lower granularity level in several different contexts [21, 32] and that related techniques as generalization and aggregation have been proven effective to preserve users' privacy [16, 57]. Regarding lower granularity, Figure 10 shows that only 37.6% of service users, 44.8% of local users, and 33.9% of non-users reported being very or extremely comfortable with the idea of sharing their position on the political spectrum to CA service. Such results, compared to the respondents' reported concerns about data related to 🗳️ Political Views in general, suggest that individuals (users as well as non-users) may have difficulties picturing privacy risks when they are not provided with specific concrete examples. It has indeed already been shown that the more individuals are exposed to concrete examples of privacy risks, the more they tend to be concerned about it [6]. Finally, the privacy concerns are, in general, similar across all the studied regions.

#### 4.3.4 Users are willing to stop using CA services if their privacy is at risk.

In addition to privacy concerns in general, we asked users



**Figure 10: Self-reported privacy concerns about sharing data related to political views with different granularity. The figure compares the answers of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ).**

(service users as well as partially local users), as an open-text question, what would make them stop using CA services. We collected  $N = 258$  answers that were coded into multiple (non-exclusive) categories. Table 3 summarizes the related findings. Among all these respondents, 53% declared that they would stop using CA services for reasons related to privacy. Some of them reported privacy reasons in general. [M, 40-49 y.o., Asia]: “I think the security/privacy issues could be the main reason not to interact or to stop using chatbots in general.” Others directly referred to data leaks and sharing to third-party [M, 18-29, Europe]: “Knowing that the information I share is not private and they can get leaked/sold to others.” or to CA services directly asking for too much personal information. [M, 18-29, Europe]: “When they will be requiring too much information to be somewhat useful.” Others explicitly described experiences from daily CA service usage that may make them stop using. [W, 50-59 y.o., South Africa]: “If they start revealing things to me about stuff I’ve done in my past.” These findings highlight the fact that users, as they share a large amount of personal data with CA services, either tend to trust the companies to keep their data safe or are not fully aware of how their data is processed, previous work has indeed shown that users tend to be less willing to share data when they are presented more transparent explanations of how their data is collected and processed [1].

Apart from privacy reasons, 17% of the answers were related to untrustful information. The large majority of individuals use CA services for information seeking (see Figure 2). Therefore, some of them declared being willing to stop using CA services if the quality or reliability of information declines. [M, 18-29, Europe]: “I’d probably stop using chatbots if they often provided me with false information.” This is interesting as prior work showed that multiple users reported having already experienced CA services providing them with false or biased information [47]. Finally, 15% declared that they would stop using CA services due to the service no longer being free or an increase in subscription costs, 8% if the performance

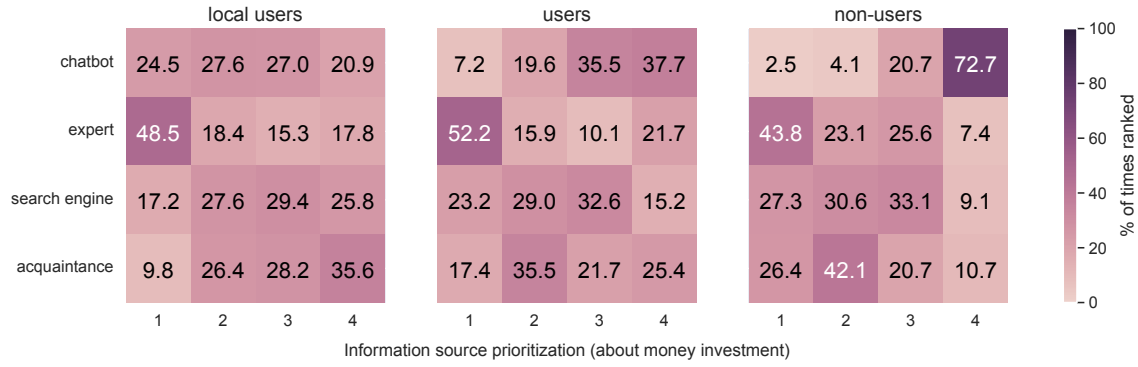
of the tool decreases (e.g., if response time becomes slow), 2% for reasons related to other ethical concerns, 8% for other reasons like the deployment of new, more powerful, tools, and 14% explicitly stated that nothing would make them stop using CA services.

#### 4.4 Data Sources Prioritization

In addition to data-sharing behavior and privacy concerns, we evaluated how individuals would prioritize different information/advice sources compared to CA services. For that purpose, we described six different daily-life scenarios (refer to Section 3.1). Then, we asked the participants to rank information/advice sources from the most likely to the least likely for them to use between (1) CA services, (2) experts in the corresponding field, (3) search engines, and (4) acquaintances.

For scenarios about money investment and physical health diagnosis, human experts were the preferred information source for all types of individuals. This partially contrasts the findings by Shahsavari and Choudhury [46] who found that most users were inclined to use CA services for self-diagnosis. However, CA services were the second-most choice for local users and more local users also tend to substantially choose CA services as a second option compared to service users. Between 40% and slightly more than half of the respondents (depending on whether they are service users, local users, or non-users) would prioritize an investment expert as an information/advice source about money investment compared to other sources (see Figure 11). However, almost one-quarter of local users would first seek advice from CA services, compared to less than 10% of service users. This also confirms our previously described trend of local users being more engaged with LLM-based CAs, even when it comes to CA online services.

For scenarios involving law, professional orientation, and dating, the rankings of CA services are similar. However, both service users and non-users are more likely to prioritize other sources, such as search engines or acquaintances, over experts. For instance,



**Figure 11: Heatmap describing how respondents would prioritize different types of information/advice sources in the context of money investment. The figure compares the answers of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ). Each row shows, for one given information/advice source, the percentage of times it has been chosen at the first, second, third, and fourth position. Each row column shows, for each position, how many times (in percentage) each information/advice source has been chosen at this specific position.**

regarding professional orientation, 27.6% of the local users would prioritize CA services, while it is the case for only 8% of services users and 1.7% of non-users (see Figure 27 in Appendix C). Notably, a large majority of participants would prioritize acquaintances when seeking dating advice.

Finally, personality assessment was the scenario where CA services were chosen the most frequently with 33.5% of local users, 17.4% of service users, and 4.1% of non-users reporting that they would prioritize them (see Figure 26 in Appendix C).

## 5 Discussion

Our survey study reveals that partially local users (who use both a local version of CA and CA online services) mainly use local instances for availability reasons (e.g., a lack of internet connection), and privacy (RQ1). However, partially local users reported to generally share far more data with CA services than service users (RQ2). This is in line with the fact that, whereas users use CA services for many different reasons, ranging from text generation to emotional support, local users use CA services for more diverse purposes and share more data than exclusive online service users. This may be due to the intensive use of CAs in general, leading to sharing more data despite their privacy concerns. Also, it may be due to the well-known privacy paradox where individuals tend to claim that they care deeply about their privacy but continue to use tools that may be harmful to it [3, 22].

Users of CA services expressed fewer privacy concerns than non-users regarding the idea of CA services having access to their personal data (RQ1). However, when we gave the participants specific examples such as lists of votes, birth dates, or the town where they live, this difference appeared to be smaller, especially between service users and non-users (RQ3). This suggests that, when taking privacy into consideration, users may have difficulties picturing concrete examples, which may lead them to underestimate the privacy risks. This finding partially aligns with Zhang et al. [54] who found that despite several concerns, users shared data with CAs potentially based on a misunderstanding of CAs. Our research

indicates that this misunderstanding may partly result from a lack of concreteness, making it difficult to grasp what happens to which data and what the implications could be.

Furthermore, most of the respondents declared that they would not prioritize CA services as an information/advice source in all the presented scenarios. Yet, local users were more likely to do it as compared to the other groups (RQ4).

The overall low level of privacy concerns could also be attributed to a lack of knowledge about LLM-based CA services, how these services work, and how they might use the data that users share with them. Indeed, although more than half of users reported that they would stop using CA services for privacy-related reasons, they continued using these services despite factors that align with their stated reasons for quitting being given. For example, many users indicated that they would stop using CA services if these services accessed or stored their personal data, which can occur when users share their personal data with these services. Additionally, many users expressed that they would quit using CA services if their data could be leaked to others, even though this is a concern with many LLM-based technologies [9]. This may also be explained with the privacy paradox.

Non-users expressed more privacy concerns about CA services having access to their personal data than users. However, when it comes to reasons for non-users to not use CA services, the large majority of them expressed a lack of need, and opportunity, and only 10% of them referred to trust and/or privacy issues. This difference is probably related to the hypothetical entity that can access personal data. Indeed, when users expressed reasons to stop related to privacy, they often refer to breaches or the sale of their data, hence they are more concerned about their data being available to other entities than CA services.

Our results about privacy concerns and data-sharing behavior are generally in line with previous user studies in other domains such as online services [12], IoT [21, 32], online social networks [1], or privacy concerns in general [6]. Nonetheless, we believe that the interactive nature of CAs as well as the more diverse possible

purposes for use may lead users to disclose more personal information in the long term than in other contexts. Indeed, whereas some more specialized services or devices often collect specific types of data using complex sensors and information processing, the textual interface of CA services allows users to easily input any type of data into multiple formats, such as text, files, or pictures. This therefore potentially expands the amount of personal information that can be collected by companies holding CA services to almost any type of personal data, and to any granularity level.

### 5.1 Limitations and Future Work

Our work has a few limitations. First of all, we deployed our study worldwide. On the one hand, that may help gather knowledge about a more diverse group of individuals and analyze a wider range of concerns and behaviors; on the other hand, it may also introduce cultural and regional biases. Studies on more specific populations (e.g., focused on one particular country) may bring additional interesting findings. Furthermore, our data related to data-sharing behaviors are self-reported, which may not directly reflect the users' exact behaviors. Similarly, our results about data-sharing behavior do not take into account the actual purpose of the user, and they might share data without realizing it. Also, whereas we have information on the types of used CA services, we did not collect anything about local instances (e.g., which model, software). Finally, whereas some of our questions are randomized to minimize biases, questions presenting multiple options of answers as well as questions asked at the end of the survey (i.e., reason to stop using, IUIPC-8) may be slightly biased by multiple previous questions related to data sharing and privacy concerns.

As for future works, we suggest additional research on privacy in LLM-based technologies that consider the user perspective. First, it would be important to explore how CA service users balance privacy and utility by conducting studies focusing on the complexity of different usage purposes and the personal data types that would be required to achieve these purposes. In addition to survey-based approaches, in-the-wild studies capturing actual user interactions and data-sharing behaviors of CA service users would shed light on hypothetical as compared to actual privacy risks and their implications for individual users. This could be done for example by collecting logs of interaction with CA services, or simply by asking users to ask the CA what kind of information it has about them. Other studies about the actual implications of sharing data types with CA services are needed. For example, further research to investigate the extent to which personal information could be inferred from users' interaction with CA services, or how CA services might influence individuals during their interactions with the services to share more personal data. Beyond individual sharing behavior and its implications for the individual, research on the level of groups or society at large would be highly relevant to receive a more holistic picture of the impact of privacy risks, e.g., related to the (mis-)use of personal data for spreading misinformation or influencing users in other ways. Based on that, a study on how privacy risk awareness-raising interventions or the use of transparency-enhancing technologies (TETs) [25] impact data-sharing behaviors would bring valuable insight. Such research would highlight potential risks and help developing safer tools. Moreover, it would also be highly useful in

the context of communicating the risks to a wider public to be sure that all users can make well-informed decisions about sharing or not sharing a piece of personal data. To implement those decisions, the development of privacy-enhancing technologies (PETs) such as a private/incognito mode that, if activated, processes data only locally, would support user. This would obviously require more storage space (i.e., for the model) and computational power from the used device, but our results already showed that many already use their own local instance of CA. Another idea could be to develop a more user-friendly turnkey local solution that could be adopted by a large public, for example by advertising the privacy aspects of such options. Apart from privacy reasons, many users declared that getting incorrect information from CA services may lead them to stop using. According to prior work, multiple users have experienced such a situation [47]. In this context, it may be interesting to explore how the actual experience affected CA service use and to develop new more specialized tools based on LLMs, as fact-checkers that could evaluate the trustworthiness of information reported by CAs.

## 6 Conclusion

In this article, we report on the findings related to our survey study with N=422 users and non-users of LLM-based CAs. Our results reveal valuable insights related to the data-sharing behaviors and privacy concerns toward these CA services and discuss how users and non-users perceive these emerging tools in the context of privacy. Our findings show that there is a significant difference between service users (individuals who only use CA services as ChatGPT) and partially local users (individuals who would use a local instance of CA in addition to services) in terms of data-sharing behavior with CA services. Indeed, partially local users seem to be more hardcore users and tend to not only use CA services more frequently and for more various purposes than service users, but they also generally share more personal data than service users. Additionally, our results reveal a significant difference in privacy concerns between non-users and users (all types). Non-users avoid using CA services mostly because of the lack of occasion or purpose, followed by lack of trust only as a secondary reason. Current service users reported being willing to abandon these services for privacy-related reasons, although these intentions seem to not always be executed despite given reasons. These findings highlight the importance of privacy risk assessment of such tools, as well as the need for privacy awareness-raising campaigns and the implementation of PETs and TETs in emerging technologies, as, for example, local data-processing options and fact-checking oriented LLMs.

### Data Availability Statement

The data that support the findings of this article and the used survey are openly available in the following URL: <https://osf.io/mqwjr/>.

### Acknowledgments

We thank Anastasija Collen, Neele Roch, and Qianjung Zheng for participating in the cognitive pre-tests for the main survey. Grammarly and ChatGPT were used across sections to detect typos and grammar mistakes and slightly improve complex sentences.



























# References

- [1] Patricia Arias-Cabarcos, Saina Khalili, and Thorsten Strufe. 2023. 'Surprised, Shocked, Worried': User Reactions to Facebook Data Collection from Third Parties. *Proc. on Privacy Enhancing Technologies (PoPETs)* 2023, 1 (Jan. 2023), 384–399. <https://doi.org/10.56553/popets-2023-0023>
- [2] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI* (2023). <https://doi.org/10.2139/ssrn.4337484>
- [3] Susanne Barth and Menno D. T. de Jong. 2017. The privacy paradox – Investigating discrepancies between expressed privacy concerns and actual online behavior – A systematic literature review. *Telematics and Informatics* 34, 7 (Nov. 2017), 1038–1058. <https://doi.org/10.1016/j.tele.2017.04.013>
- [4] Rahime Belen-Saglam, Jason R. C. Nurse, and Duncan Hodges. 2022. An Investigation Into the Sensitivity of Personal Information and Implications for Disclosure: A UK Perspective. *Frontiers in Computer Science* 4 (June 2022). <https://doi.org/10.3389/fcomp.2022.908245>
- [5] Richard Brown, Elizabeth Sillence, Lynne Coventry, Emma Simpson, Jo Gibbs, Shema Tariq, Abigail C. Durrant, and Karen Lloyd. 2022. Understanding the attitudes and experiences of people living with potentially stigmatised long-term health conditions with respect to collecting and sharing health and lifestyle data. *Digital Health* 8 (Jan. 2022), 20552076221089798. <https://doi.org/10.1177/20552076221089798>
- [6] Raphaëlle Butori and Caroline Lancelot Milgten. 2023. A construal level theory approach to privacy protection: The conjoint impact of benefits and risks of information disclosure. *Journal of Business Research* 168 (Nov. 2023), 114205. <https://doi.org/10.1016/j.jbusres.2023.114205>
- [7] Laura Calloway, Hilda Hadan, Shakhidhar Gopavaram, Shirang Mare, and L. Jean Camp. 2020. Privacy in Crisis: Participants' Privacy Preferences for Health and Marketing Data during a Pandemic. In *Proc. of the Workshop on Privacy in the Electronic Society (WPES)*. Association for Computing Machinery, New York, NY, USA, 181–189. <https://doi.org/10.1145/3411497.3420223>
- [8] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying Memorization Across Neural Language Models. In *ICLR*. <https://arxiv.org/pdf/2202.07646>
- [9] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. *USENIX Security Symposium* 2021 (2021).
- [10] Juan Pablo Carrascal, Christopher Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira. 2013. Your browsing behavior for a big mac: economics of personal information online. In *Proc. of the international conference on World Wide Web (WWW)*. ACM, 189–200. <https://doi.org/10.1145/2488388.2488406>
- [11] Yan Chen and Pouyan Esmailzadeh. 2024. Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges. *Journal of Medical Internet Research* 26, 1 (March 2024), e53008. <https://doi.org/10.2196/53008>
- [12] Hui Na Chua, Jie Sheng Ooi, and Anthony Herbrand. 2021. The effects of different personal data categories on information privacy concern and disclosure. *Computers & Security* 110 (Nov. 2021), 102453. <https://doi.org/10.1016/j.cose.2021.102453>
- [13] Massilva Dekkal, Manon Arcand, Sandrine Prom Tep, Lova Rajaobelina, and Line Ricard. 2024. Factors affecting user trust and intention in adopting chatbots: the moderating role of technology anxiety in insurtech. *Journal of Financial Services Marketing* 29, 3 (Sept. 2024), 699–728. <https://doi.org/10.1057/s41264-023-00230-y>
- [14] Yuyang Deng, Zhao Ni, and Huang Xin. 2023. Early ChatGPT User Portrait through the Lens of Data. In *Conf. on Big Data*. IEEE. <https://www.computer.org/csdl/proceedings-article/bigdata/2023/10386415/1TUOYsrRIFC>
- [15] Christophe Dupuy, Radhika Arava, Rahul Gupta, and Anna Rumshisky. 2022. An Efficient DP-SGD Mechanism for Large Scale NLU Models. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4118–4122. <https://doi.org/10.1109/ICASSP43922.2022.9746975>
- [16] Gunther Eibl and Dominik Engel. 2015. Influence of Data Granularity on Smart Meter Privacy. *IEEE Transactions on Smart Grid* 6, 2 (March 2015), 930–939. <https://doi.org/10.1109/TSG.2014.2376613>
- [17] Aldasoro et al. 2024. Intelligent financial system: how AI is transforming finance. <https://www.bis.org/publ/work1194.pdf>
- [18] Li et al. 2024. Human-Centered Privacy Research in the Age of Large Language Models. In *Extended Abstracts of the Conf. on Human Factors in Computing Systems (CHI AE)*. ACM. <https://dl.acm.org/doi/full/10.1145/3613905.3643983>
- [19] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 4 (Nov. 2009), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- [20] Sarah Abdelwahab Gaballah, Lamya Abdullah, Ephraim Zimmer, Sascha Fahl, Max Mühlhäuser, and Karola Marky. 2025. "It's Not My Data Anymore": Exploring Non-Users' Privacy Perceptions of Medical Data Donation Apps. *Proc. on Privacy Enhancing Technologies (PoPETs)* Proc. on Privacy Enhancing Technologies (PoPETs) (2025).
- [21] Sandra Gabriele and Sonia Chiasson. 2020. Understanding Fitness Tracker Users' Security and Privacy Knowledge, Attitudes and Behaviours. In *Proc. of the Conf. on Human Factors in Computing Systems (CHI)*. ACM, 1–12. <https://doi.org/10.1145/3313831.3376651>
- [22] Nina Gerber, Paul Gerber, and Melanie Volkamer. 2018. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & Security* 77 (Aug. 2018), 226–261. <https://doi.org/10.1016/j.cose.2018.04.002>
- [23] Thomas Groß. 2021. Validity and Reliability of the Scale Internet Users' Information Privacy Concerns (IUIPC). *Proc. on Privacy Enhancing Technologies (PoPETs)* 2021, 2 (April 2021), 235–258. <https://doi.org/10.2478/popets-2021-0026>
- [24] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon So. 2023. Knowledge Unlearning for Mitigating Privacy Risks in Language Models. In *Proc. of the Meeting of the Association for Computational Linguistics*.
- [25] Milena Janic, Jan Pieter Wijbenga, and Thijs Veugen. 2013. Transparency Enhancing Tools (TETs): An Overview. In *Workshop on Socio-Technical Aspects in Security and Trust*. 18–25. <https://doi.org/10.1109/STAST.2013.11>
- [26] Yongnam Jung, Cheng Chen, Eunhae Jang, and S. Shyam Sundar. 2024. Do We Trust ChatGPT as much as Google Search and Wikipedia?. In *Extended Abstracts of the Conf. on Human Factors in Computing Systems (CHI AE)*. ACM, Honolulu HI USA, 1–9. <https://doi.org/10.1145/3613905.3650862>
- [27] Angeliki Kalapodi and Nicolas Sklavos. 2021. The Concerns of Personal Data Privacy, on Calling and Messaging, Networking Applications. In *Security in Computing and Communications*, Sabu M. Thampi, Guojun Wang, Danda B. Rawat, Ryan Ko, and Chun-I Fan (Eds.). Springer, Singapore, 275–289. [https://doi.org/10.1007/978-981-16-0422-5\\_20](https://doi.org/10.1007/978-981-16-0422-5_20)
- [28] Anna Leschanowsky, Slias Anna, Birgit Popp, and Tom Bäckström. 2024. Evaluating privacy, security, and trust perceptions in conversational AI: A systematic review. (2024). <https://www.sciencedirect.com/science/article/pii/S0747563224002127>
- [29] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. 2023. Multi-step Jailbreaking Privacy Attacks on ChatGPT. In *EMNLP*.
- [30] Yao Li, Eugenia Ha Rim Rho, and Alfred Kobas. 2022. Cultural differences in the effects of contextual factors and privacy concerns on users' privacy decision on social networking sites. *Behaviour & Information Technology* 41, 3 (Feb. 2022), 655–677. <https://doi.org/10.1080/0144929X.2020.1831608>
- [31] Yuting Liao. 2019. Sharing Personal Health Information on Social Media: Balancing Self-presentation and Privacy. In *Proc. of the Int'l Conf. on Social Media and Society (SMSociety)*. Association for Computing Machinery, New York, NY, USA, 194–204. <https://doi.org/10.1145/3328529.3328560>
- [32] Byron Lowens, Vivian Genaro Motti, and Kelly Caine. 2017. Wearable Privacy: Skeletons in The Data Closet. In *International Conference on Healthcare Informatics (ICHI)*. IEEE, Park City, UT, USA, 295–304. <https://doi.org/10.1109/ICHI.2017.29>
- [33] Tenga Matsuura, Ayako A. Hasegawa, Mitsuaki Akiyama, and Tatsuya Mori. 2021. Careless Participants Are Essential for Our Phishing Study: Understanding the Impact of Screening Methods. In *European Symp. on Usable Security*. ACM, Karlsruhe Germany, 36–47. <https://doi.org/10.1145/3481357.3481515>
- [34] S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. 2024. The potential of generative AI for personalized persuasion at scale. *Scientific Reports* 14, 1 (Feb. 2024), 4692. <https://doi.org/10.1038/s41598-024-53755-0>
- [35] Philipp Mayring. 2015. Qualitative Content Analysis: Theoretical Background and Procedures. In *Approaches to Qualitative Research in Mathematics Education: Examples of Methodology and Methods*, Angelika Bikner-Ahsbals, Christine Knipping, and Norma Presmeg (Eds.). Springer Netherlands, Dordrecht, 365–380. [https://doi.org/10.1007/978-94-017-9181-6\\_13](https://doi.org/10.1007/978-94-017-9181-6_13)
- [36] Francis T. McAndrew and Sara S. Koehnke. 2016. On the nature of creepiness. *New Ideas in Psychology* 43 (Dec. 2016), 10–15. <https://doi.org/10.1016/j.newideapsych.2016.03.003>
- [37] Alexopoulos Michelle, Lyons Kelly, Mahetaji Kaushar, Barnes Marcus Emmanuel, and Gutwillinger Rogan. 2023. Gender Inference: Can ChatGPT Outperform Common Commercial Tools?. In *Proc. of the Conf. on Computer Science and Software Engineering (CASCON '23)*. IBM, USA, 161–166.
- [38] Kunihiro Miyazaki, Taichi Murayama, Takayuki Uchiba, Jisun An, and Haewoon Kwak. 2024. Public perception of generative AI on Twitter: an empirical study based on occupation and usage. *EPJ Data Science* (2024). <https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-023-00445-y>
- [39] OpenAI. 2023. Data Usage for Consumer Services FAQ. <https://help.openai.com/en/articles/7039943-data-usage-for-consumer-services-faq>
- [40] Anna-Marie Orloff, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. 2023. Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of an Interview and Survey Study on Security Advice. In *Proc. of the Conference on Human Factors in Computing Systems (CHI) (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. <https://doi.org/10.1145/3544548.3580766>

- [41] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (March 2018), 22–27. <https://doi.org/10.1016/j.jbef.2017.12.004>
- [42] Heinrich Peters and Sandra C Matz. 2024. Large language models can infer psychological dispositions of social media users. *PNAS Nexus* 3, 6 (May 2024), pgae231. <https://doi.org/10.1093/pnasnexus/pgae231>
- [43] Christopher J. Riederer, Sebastian Zimmeck, Coralie Phanord, Augustin Chain-treau, and Steven M. Bellovin. 2015. "I don't have a photograph, but you can have my footprints": Revealing the Demographics of Location Data. In *Proc. of the Conf. Online Social Networks*. ACM, Palo Alto California USA, 185–195. <https://doi.org/10.1145/2817946.2817968>
- [44] Johnny Saldana. 2021. *The Coding Manual for Qualitative Researchers*.
- [45] Eva-Maria Schomakers, Chantal Lidynia, and Martina Ziefle. 2020. All of me? Users' preferences for privacy-preserving data markets and the importance of anonymity. *Electronic Markets* 30, 3 (Sept. 2020), 649–665. <https://doi.org/10.1007/s12525-020-00404-9>
- [46] Yeganeh Shahsavari and Avishek Choudhury. 2023. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. *JMIR Human Factors* 10 (May 2023), e47564. <https://doi.org/10.2196/47564>
- [47] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. In *Conf. on Conversational User Interfaces (CUI)*. ACM. <https://dl.acm.org/doi/abs/10.1145/3571884.3597144>
- [48] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Beyond Memorization: Violating Privacy via Inference with Large Language Models. *ICLR* (2024).
- [49] Liesbet van Zoonen. 2016. Privacy concerns in smart cities. *Government Information Quarterly* 33, 3 (July 2016), 472–480. <https://doi.org/10.1016/j.giq.2016.06.004>
- [50] Lev Velykoivanenko, Kavous Salehzadeh Niksirat, Noé Zufferey, Mathias Humbert, Kévin Huguenin, and Mauro Cherubini. 2021. Are Those Steps Worth Your Privacy?: Fitness-Tracker Users' Perceptions of Privacy and Utility. In *Proc. of the Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, Vol. 5. ACM, 1–41. <https://doi.org/10.1145/3494960>
- [51] W. Patrick Walters and Mark Murcko. 2020. Assessing the impact of generative AI on medicinal chemistry. *Nature Biotechnology* 38, 2 (2020), 143–145. <https://doi.org/10.1038/s41587-020-0418-2>
- [52] Xiaodong Wu, Ran Duan, and Jianbing Ni. 2023. Unveiling security, privacy, and ethical concerns of ChatGPT. *Journal of Information and Intelligence* (Oct. 2023). <https://doi.org/10.1016/j.jiixd.2023.10.007>
- [53] Jinsung Yoon, Lydia N. Drumright, and Mihaela van der Schaar. 2020. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics* 24, 8 (Aug. 2020), 2378–2388. <https://doi.org/10.1109/JBHI.2020.2980262>
- [54] Zhiping Zhang, Michelle Jia, Hao-Ping (Hank) Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. 2024. "It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. In *Proc. of the Conference on Human Factors in Computing Systems (CHI) (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–26. <https://doi.org/10.1145/3613904.3642385>
- [55] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. 2018. User Perceptions of Smart Home IoT Privacy. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 200:1–200:20. <https://doi.org/10.1145/3274469>
- [56] Verena Zimmermann, Paul Gerber, Karola Marky, Leon Böck, and Florian Kirchbuchner. 2019. Assessing Users' Privacy and Security Concerns of Smart Home Technologies. *i-com* 18, 3 (Dec. 2019), 197–216. <https://doi.org/10.1515/icom-2019-0015>
- [57] Noé Zufferey, Mathias Humbert, Romain Tavenard, and Kévin Huguenin. 2023. Watch your Watch: Inferring Personality Traits from Wearable Activity Trackers. In *USENIX Security Symposium*. 193–210. <https://www.usenix.org/conference/usenixsecurity23/presentation/zufferey>
- [58] Noé Zufferey, Kavous Salehzadeh Niksirat, Mathias Humbert, and Kévin Huguenin. 2023. "Revoked just now!" Users' Behaviors Toward Fitness-Data Sharing with Third-Party Applications. *Proc. on Privacy Enhancing Technologies (PoPETs)* 2023, 1 (2023), 47–67. <https://doi.org/10.56553/popets-2023-0004>

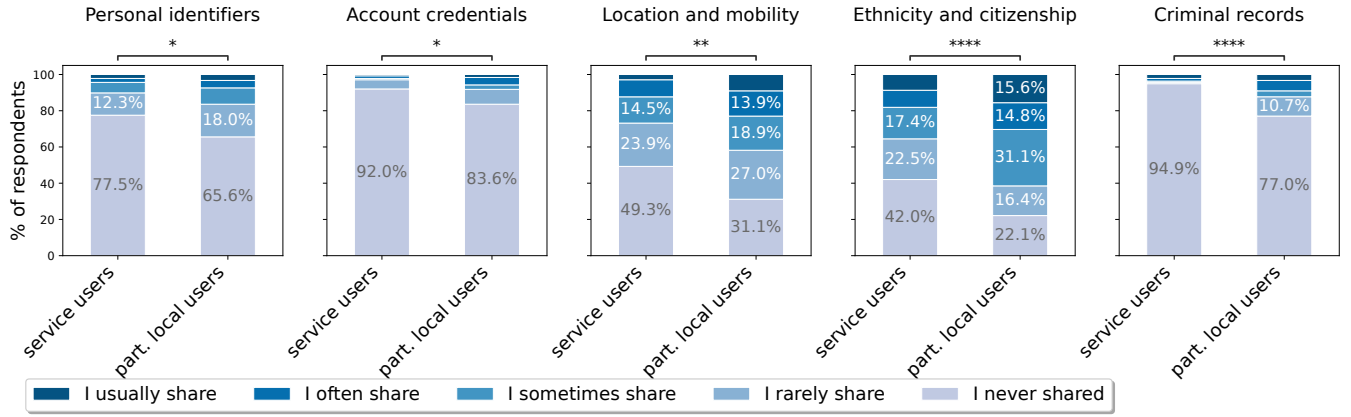
## A Personal Data Types

**Table 4: Personal Data Categories**


Categories	Subcategories	Examples
 Identifiers and General Information	 Personal Identifiers	Name, ID Card Number, Email Address, Phone Number
	 Account Credentials	User Name, Password
	 Location and Mobility	Homeplace, Workplace, Current Location
	 Ethnicity and Citizenship	
	 Criminal Records	
 Lifestyle and Health	 Mental Health	
	 Physical Health	Diagnosis, Medication, Symptoms, Menstrual Cycle Details
	 Lifestyle and Habits	Sport Activities, Hobbies, Sleeping Time, Eating Habits
	 Recreational Consumption	Alcohol, Smoking Habits, Recreational Drugs
	 Sexual and Dating Activities	
 Personal Characteristics and Emotions	 Sexual orientation	
	 Mental State and Personality	Mood, Emotions, Mindest, Personality Traits
	 Gender	
	 Age	
	 Physical Traits	Size, Skin/Hair/Eye Color
 Standard of Living and Opinions	 Family Life and Relationship	Civil Status, Number of Children
	 Wealth Details	Salary, Savings Amount, Land Ownership
	 Job and Education	Employment Status, Type of Job, Study/Industry Field, Company
	 Religion	
	 Political Views	

## B Data-sharing Behavior

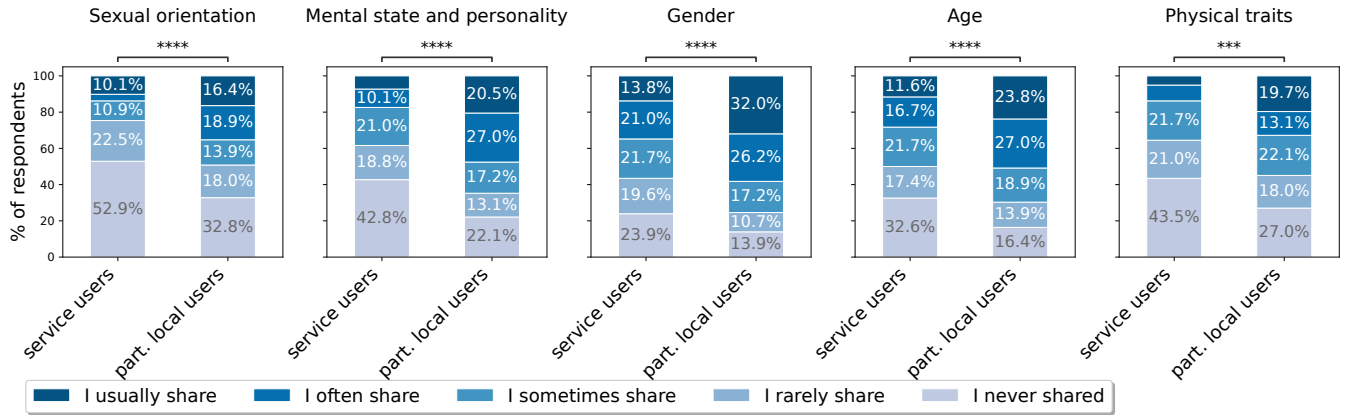
### Identifiers and General Information




\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 1e-3$ , \*\*\*\* $p < 1e-4$ , Mann-Whitney U test with Bonferroni correction

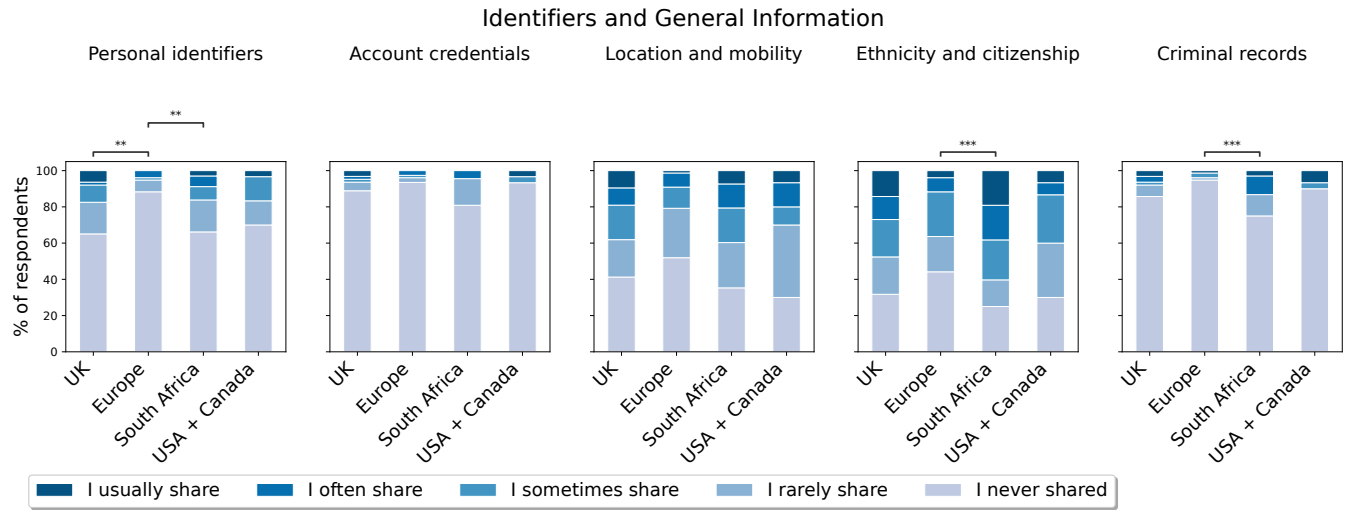
**Figure 12: Self-declared data-sharing frequency of information related to  identifiers and general information. The figure compares the data-sharing behavior of service users ( $N = 138$ ), and partially local users ( $N = 122$ ).**


### Personal Characteristics and Emotions

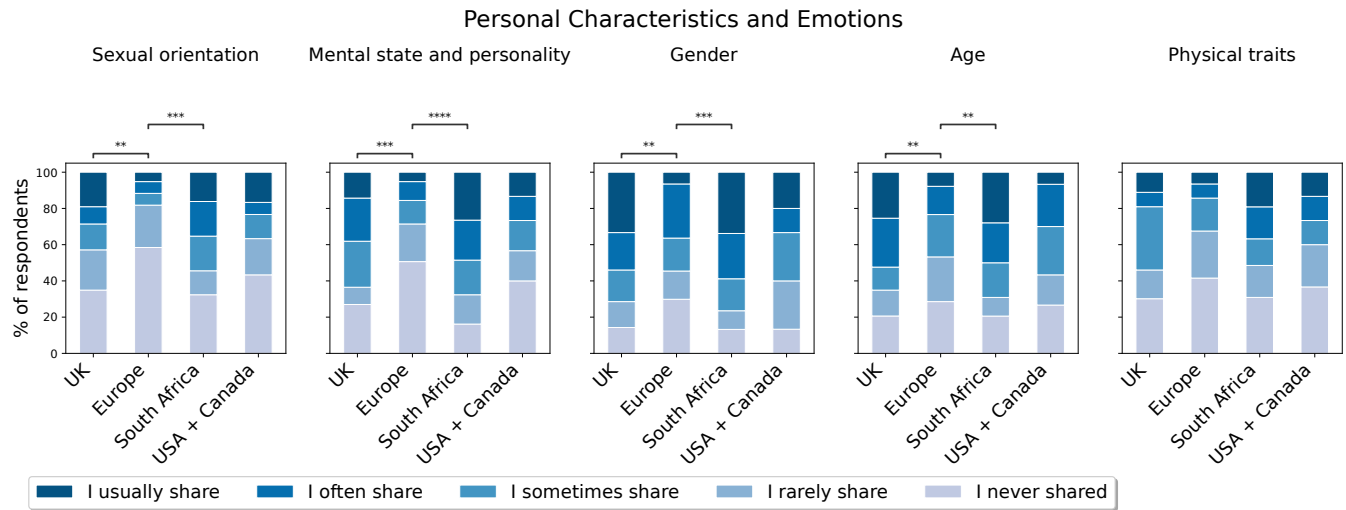



\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 1e-3$ , \*\*\*\* $p < 1e-4$ , Mann-Whitney U test with Bonferroni correction

**Figure 13: Self-declared data-sharing frequency of information related to  personal characteristics and emotions. The figure compares the data-sharing behavior of service users ( $N = 138$ ), and partially local users ( $N = 122$ ).**

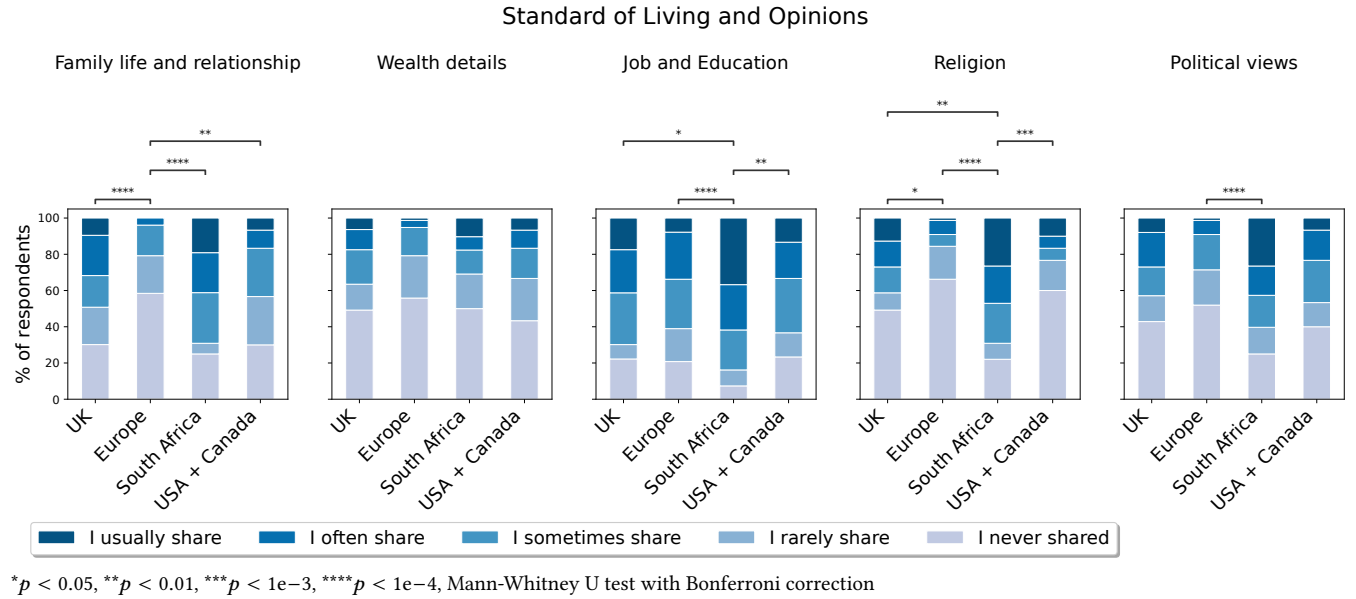


**Figure 14: Self-reported data-sharing frequency of information related to  identifiers and general information.** The figure compares the data-sharing behavior of both types of individuals who use CA services (service users as well as partially local users) and compares the results of individuals living in UK ( $N = 63$ ), Europe ( $N = 77$ ), South Africa ( $N = 68$ ), and USA or Canada ( $N = 30$ )



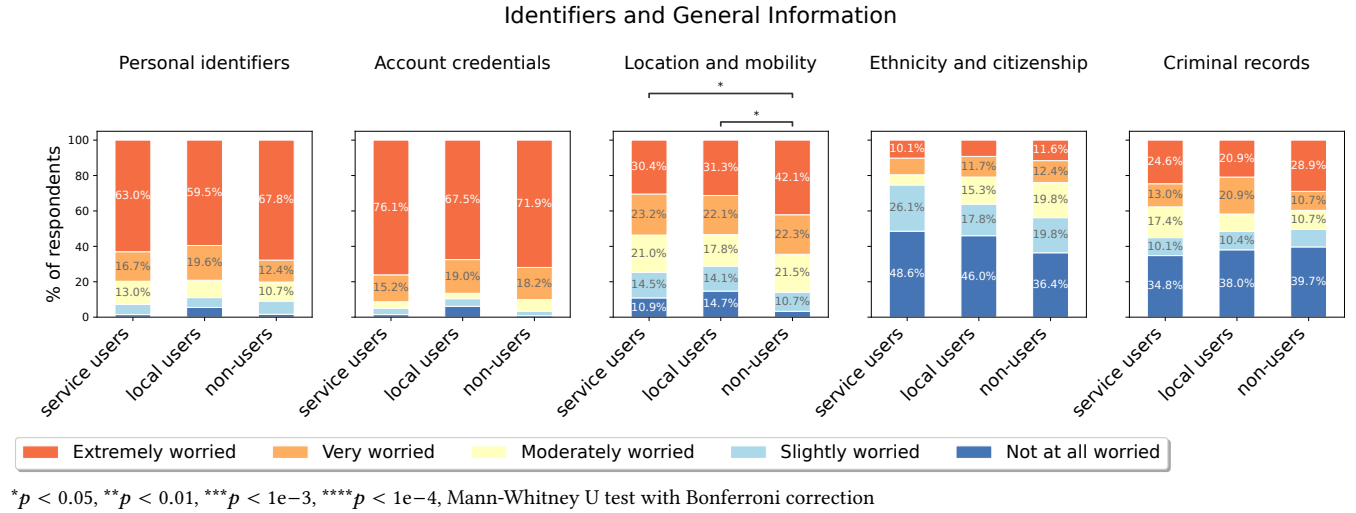
**Figure 15: Self-reported data-sharing frequency of information related to  personal characteristics and emotions.** The figure compares the data-sharing behavior of both types of individuals who use CA services (service users as well as partially local users) and compares the results of individuals living in UK ( $N = 63$ ), Europe ( $N = 77$ ), South Africa ( $N = 68$ ), and USA or Canada ( $N = 30$ )



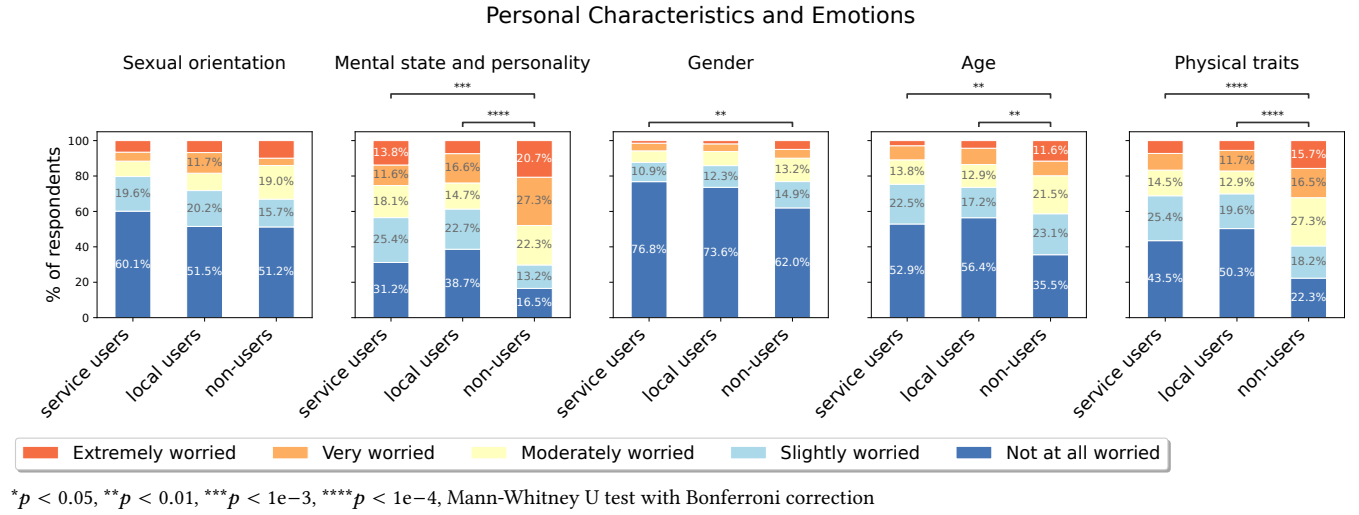


**Figure 16: Self-reported data-sharing frequency of information related to 🏠 Standard of Living and Opinions.** The figure compares the data-sharing behavior of both types of individuals who use CA services (service users as well as partially local users) and compares the results of individuals living in UK ( $N = 63$ ), Europe ( $N = 77$ ), South Africa ( $N = 68$ ), and USA or Canada ( $N = 30$ )

## C Privacy Concerns

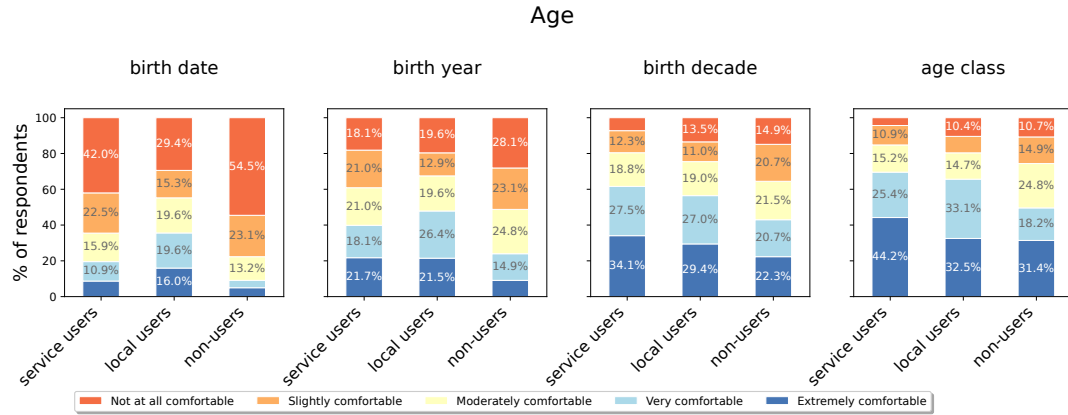


**Figure 17: Self-declared privacy concerns about CA services having access to personal data related to 📄 identifiers and general information.** The figure compares the privacy concerns of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ).

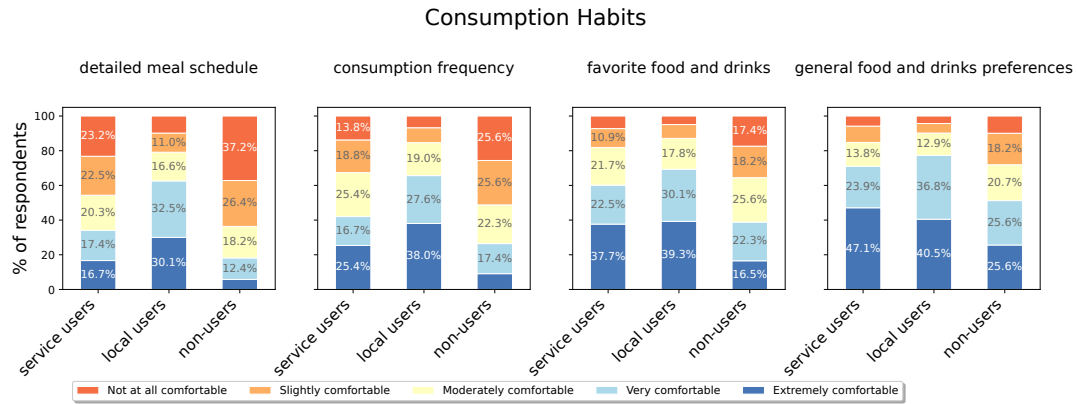


**Figure 18: Self-reported privacy concerns about CA services having access to personal data related to personal characteristics and emotions.** The figure compares the privacy concerns of service users (( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ).

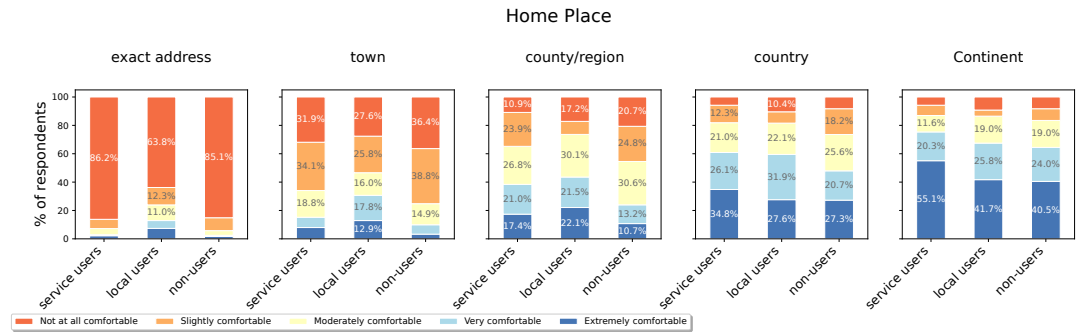
## D Granularity Preferences of Data Sharing



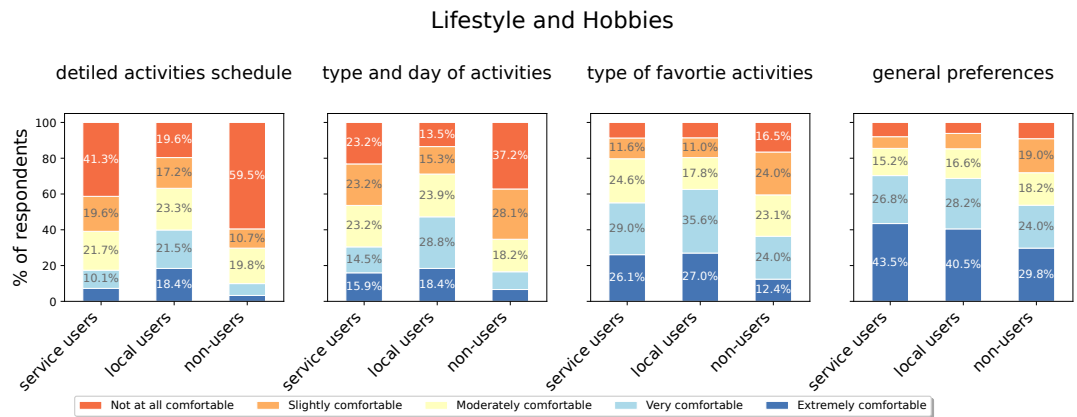
**Figure 19: Self-reported privacy concerns about sharing data related to age with different granularity.** The figure compares the privacy concerns of service users (( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ).



**Figure 20: Self-reported privacy concerns about sharing data related to consumption habits with different granularity. The figure compares the privacy concerns of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ).**



**Figure 21: Self-reported privacy concerns about sharing data related to home place with different granularity. The figure compares the privacy concerns of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ).**



**Figure 22: Self-reported privacy concerns about sharing data related to lifestyle and hobbies with different granularity. The figure compares the privacy concerns of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ).**

E Information Source Ranking

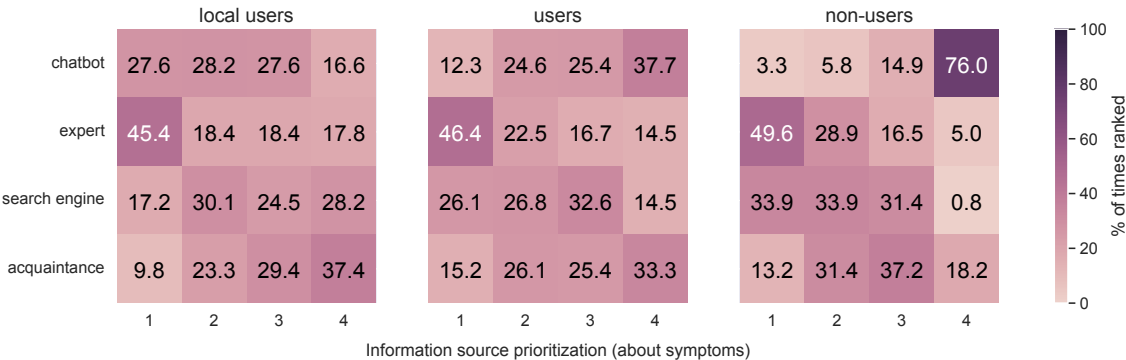


Figure 23: Heatmap describing how respondents would prioritize different types of information/advice sources in the context of health diagnosis. The figure compares the privacy concerns of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ). Each row shows, for one given information/advice source, the percentage of times it has been chosen at the first, second, third, and fourth position. Each row column shows, for each position, how many times (in percentage) each information/advice source has been chosen at this specific position.

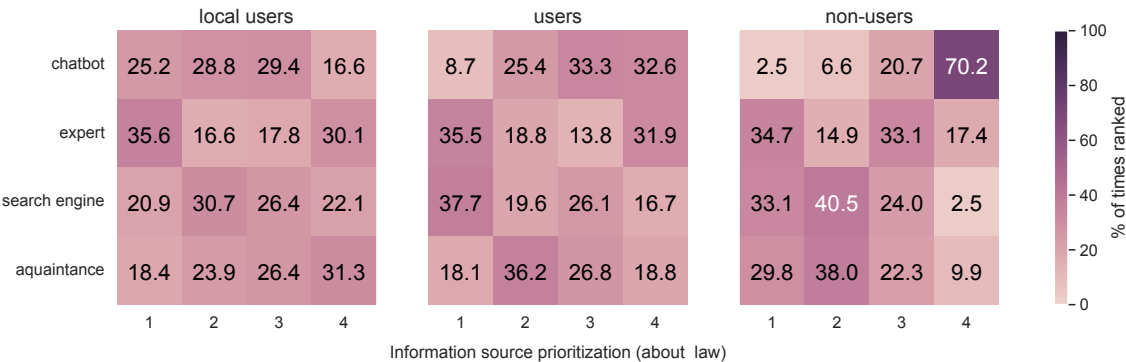


Figure 24: Heatmap describing how respondents would prioritize different types of information/advice sources in the context of law. The figure compares the privacy concerns of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ). Each row shows, for one given information/advice source, the percentage of times it has been chosen at the first, second, third, and fourth position. Each row column shows, for each position, how many times (in percentage) each information/advice source has been chosen at this specific position.

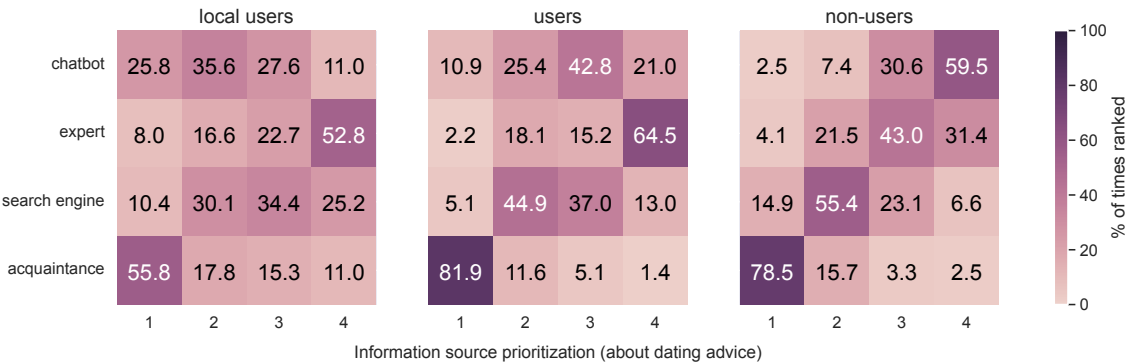


Figure 25: Heatmap describing how respondents would prioritize different types of information/advice sources in the context of dating. The figure compares the privacy concerns of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ). Each row shows, for one given information/advice source, the percentage of times it has been chosen at the first, second, third, and fourth position. Each row column shows, for each position, how many times (in percentage) each information/advice source has been chosen at this specific position.

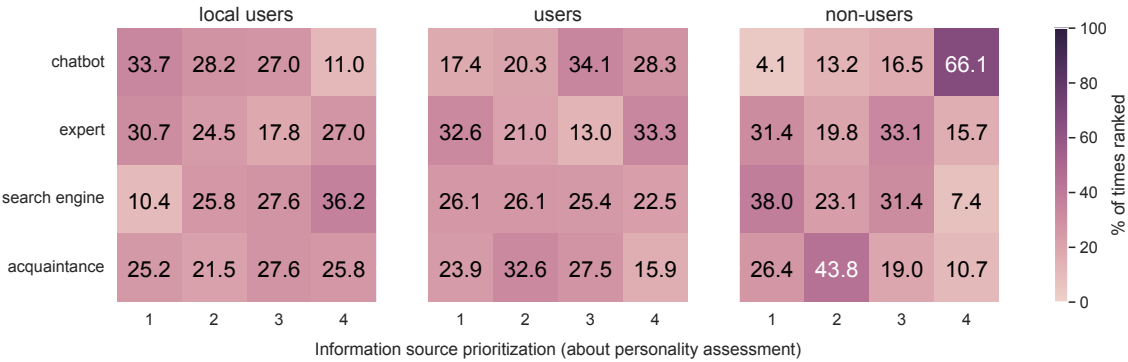
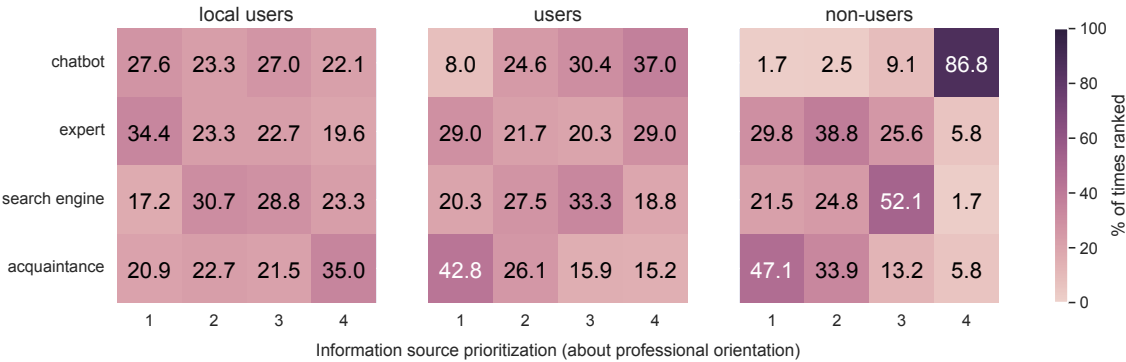


Figure 26: Heatmap describing how respondents would prioritize different types of information/advice sources in the context of personality assessment. The figure compares the privacy concerns of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ). Each row shows, for one given information/advice source, the percentage of times it has been chosen at the first, second, third, and fourth position. Each row column shows, for each position, how many times (in percentage) each information/advice source has been chosen at this specific position.





**Figure 27: Heatmap describing how respondents would prioritize different types of information/advice sources in the context of professional orientation. The figure compares the privacy concerns of service users ( $N = 138$ ), local users ( $N = 163$ ), and non-users ( $N = 121$ ). Each row shows, for one given information/advice source, the percentage of times it has been chosen at the first, second, third, and fourth position. Each row column shows, for each position, how many times (in percentage) each information/advice source has been chosen at this specific position.**